REVIEW

CrossMark

Improving Methodological Standards in Behavioral Interventions for Cognitive Enhancement

C. Shawn Green[1] · Daphne Bavelier[2] · Arthur F. Kramer[3,4] · Sophia Vinogradov[5,6,7] · Ulrich Ansorge[8] · Karlene K. Ball[9] · Ulrike Bingel[10] · Jason M. Chein[11] · Lorenza S. Colzato[12,13] · Jerri D. Edwards[14] · Andrea Facoetti[15] · Adam Gazzaley[16] · Susan E. Gathercole[17] · Paolo Ghisletta[18,19,20] · Simone Gori[21] · Isabela Granic[22] · Charles H. Hillman[23] · Bernhard Hommel[24] · Susanne M. Jaeggi[25] · Philipp Kanske[26] · Julia Karbach[27,28] · Alan Kingstone[29] · Matthias Kliegel[30] · Torkel Klingberg[31] · Simone Kühn[32] · Dennis M. Levi[33] · Richard E. Mayer[34] · Anne Collins McLaughlin[35] · Danielle S. McNamara[36] · Martha Clare Morris[37] · Mor Nahum[38] · Nora S. Newcombe[39] · Rogerio Panizzutti[40,41] · Ruchika Shaurya Prakash[42] · Albert Rizzo[43] · Torsten Schubert[44] · Aaron R. Seitz[45] · Sarah J. Short[46] · Ilina Singh[47] · James D. Slotta[48] · Tilo Strobach[49] · Michael S. C. Thomas[50] · Elizabeth Tipton[51,52] · Xin Tong[53] · Haley A. Vlach[54] · Julie Loebach Wetherell[55,56] · Anna Wexler[57] · Claudia M. Witt[58]

Abstract
There is substantial interest in the possibility that cognitive skills can be improved by dedicated behavioral training. Yet despite the large amount of work being conducted in this domain, there is not an explicit and widely agreed upon consensus around the best methodological practices. This document seeks to fill this gap. We start from the perspective that there are many types of studies that are important in this domain—e.g., feasibility, mechanistic, efficacy, and effectiveness. These studies have fundamentally different goals, and, as such, the best-practice methods to meet those goals will also differ. We thus make suggestions in topics ranging from the design and implementation of control groups, to reporting of results, to dissemination and communication, taking the perspective that the best practices are not necessarily uniform across all study types. We also explicitly recognize and discuss the fact that there are methodological issues around which we currently lack the theoretical and/or empirical foundation to determine best practices (e.g., as pertains to assessing participant expectations). For these, we suggest important routes forward, including greater interdisciplinary collaboration with individuals from domains that face related concerns. Our hope is that these recommendations will greatly increase the rate at which science in this domain advances.

Keywords Cognitive enhancement · Behavioral intervention methodology

Introduction

The past two decades have brought a great deal of attention to the possibility that certain core cognitive abilities, including those related to processing speed, working memory, perceptual abilities, attention, and general intelligence, can be improved by dedicated behavioral training (Au et al. 2015; Ball et al. 2002; Bediou et al. 2018; Deveau et al. 2014a; Karbach and Unger 2014; Kramer et al. 1995; Schmiedek et al. 2010;

Strobach and Karbach 2016; Valdes et al. 2017). Such a prospect has clear theoretical scientific relevance, as related to our understanding of those cognitive sub-systems and their malleability (Merzenich et al. 2013). It also has obvious practical relevance. Many populations, such as children diagnosed with specific clinical disorders or learning disabilities (Franceschini et al. 2013; Klingberg et al. 2005), individuals with schizophrenia (Biagianti and Vinogradov 2013), traumatic brain injury (Hallock et al. 2016), and older adults (Anguera et al. 2013; Nahum et al. 2013; Whitlock et al. 2012), may show deficits in these core cognitive abilities, and thus could reap significant benefits from effective interventions.

There are also a host of other circumstances outside of rehabilitation where individuals could potentially benefit from enhancements in cognitive skills. These include, for instance,

✉ C. Shawn Green
cshawn.green@wisc.edu

Extended author information available on the last page of the article
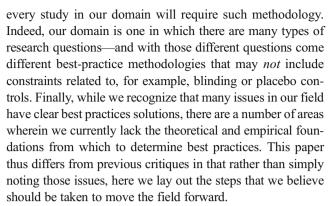
Springer

improving job-related performance in individuals whose occupations place heavy demands on cognitive abilities, such as military and law enforcement personnel, pilots, high-level athletes, and surgeons (Deveau et al. 2014b; Schlickum et al. 2009). Finally, achievement in a variety of academic domains, including performance in science, technology, engineering, and mathematics (STEM) fields, in scientific reasoning, and in reading ability, have also been repeatedly linked to certain core cognitive capacities. These correlational relations have in turn then sparked interest in the potential for cognitive training to produce enhanced performance in the various academic areas (Rohde and Thompson 2007; Stieff and Uttal 2015; Wright et al. 2008).

However, while there are numerous published empirical results suggesting that there is reason for optimism that some or all of these goals are within our reach, the field has also been subject to significant controversy, concerns, and criticisms recommending that such enthusiasm be appreciably dampened (Boot et al. 2013; Melby-Lervag and Hulme 2013; Shipstead et al. 2012; Simons et al. 2016). Our goal here is not to adjudicate between these various positions or to rehash prior debates. Instead, the current paper is forward looking. We argue that many of the disagreements that have arisen in our field to-date can be avoided in the future by a more coherent and widely agreed-upon set of methodological standards in the field. Indeed, despite the substantial amount of research that has been conducted in this domain, as well as the many published critiques, there is not currently an explicitly delineated scientific consensus outlining the best methodological practices to be utilized when studying behavioral interventions meant to improve cognitive skills.

The lack of consensus has been a significant barrier to progress at every stage of the scientific process, from basic research to translation. For example, on the basic research side, the absence of clear methodological standards has made it difficult-to-impossible to easily and directly compare results across studies (either via side-by-side contrasts or in broader meta-analyses). This limits the field's ability to determine what techniques or approaches have shown positive outcomes, as well as to delineate the exact nature of any positive effects—e.g., training effects, transfer effects, retention of learning, etc. On the translational side, without such standards, it is unclear what constitutes scientifically acceptable evidence of efficacy or effectiveness. This is a serious problem both for researchers attempting to demonstrate efficacy and for policy makers attempting to determine whether efficacy has, in fact, been demonstrated.

Below, we lay out a set of broad methodological standards that we feel should be adopted within the domain focused on behavioral interventions for cognitive enhancement. As will become clear, we strongly maintain that a "gold standard methodology," as exists in clinical or pharmaceutical trials, is not only a goal that our field can strive toward, but is indeed one that can be fully met. We also appreciate though that not every study in our domain will require such methodology. Indeed, our domain is one in which there are many types of research questions—and with those different questions come different best-practice methodologies that may *not* include constraints related to, for example, blinding or placebo controls. Finally, while we recognize that many issues in our field have clear best practices solutions, there are a number of areas wherein we currently lack the theoretical and empirical foundations from which to determine best practices. This paper thus differs from previous critiques in that rather than simply noting those issues, here we lay out the steps that we believe should be taken to move the field forward.

We end by noting that although this piece is written from the specific perspective of cognitive training, the vast majority of the issues that are covered are more broadly relevant to any research domain that employs behavioral interventions to change human behavior. Some of these interventions do not fall neatly within the domain of "cognitive training," but they are nonetheless conducted with the explicit goal of improving cognitive function. These include interventions involving physical exercise and/or aerobic activity (Hillman et al. 2008; Voss et al. 2010), mindfulness meditation (Prakash et al. 2014; Tang et al. 2007), video games (Colzato et al. 2014; Green and Bavelier 2012; Strobach et al. 2012), or musical interventions (Schellenberg 2004). The ideas are further clearly applicable in domains that, while perhaps not falling neatly under the banner of "cognitive enhancement," clearly share a number of touch points. These include, for instance, many types of interventions deployed in educational contexts (Hawes et al. 2017; Mayer 2014). Finally, the issues and ideas also apply in many domains that lie well outside of cognition. These range from behavioral interventions designed to treat various clinical disorders such as post-traumatic stress disorder (PTSD) or major depressive disorder (Rothbaum et al. 2014), to those designed to decrease anti-social behaviors or increase pro-social behaviors (Greitemeyer et al. 2010). The core arguments and approaches that are developed here, as well as the description of areas in need of additional work, are thus similarly shared across these domains. Our hope is thus that this document will accelerate the rate of knowledge acquisition in all domains that study the impact of behavioral interventions. And as the science grows, so will our knowledge of how to deploy such paradigms for practical good.

Behavioral Interventions for Cognitive Enhancement Can Differ Substantially in Content and Target(s) and Thus A Common Moniker like "Brain Training" Can Be Misleading

As the literature exploring behavioral interventions for cognitive enhancement has grown, so too has the number of unique approaches adopted in this endeavor. For example, some

research groups have used reasonably unaltered standard psychology tasks as training paradigms (Schmiedek et al. 2010; Willis et al. 2006), while others have employed "gamified" versions of such tasks (Baniqued et al. 2015; Jaeggi et al. 2011; Owen et al. 2010). Some groups have used off-the-shelf commercial video games that were designed with only entertainment-based goals in mind (Basak et al. 2008; Green et al. 2010), while others have utilized video games designed to mimic the look and feel of such commercial games, but with the explicit intent of placing load on certain cognitive systems (Anguera et al. 2013). Some groups have used a single task for the duration of training (Jaeggi et al. 2008), while others have utilized training consisting of many individual tasks practiced either sequentially or concurrently (Baniqued et al. 2015; Smith et al. 2009). Some groups have used tasks that were formulated based upon principles derived from neuroscience (Nahum et al. 2013), while others have used tasks inspired by Eastern meditation practices (Tang et al. 2007). In all, the range of approaches is now simply enormous, both in terms of the number of unique dimensions of variation, as well as the huge variability within those dimensions.

Unfortunately, despite vast differences in approach, there continues to exist the tendency of lumping all such interventions together under the moniker of "brain training," not only in the popular media (Howard 2016), but also in the scientific community (Bavelier and Davidson 2013; Owen et al. 2010; Simons et al. 2016). We argue that such a superordinate category label is not a useful level of description or analysis. Each individual type of behavioral intervention for cognitive enhancement (by definition) differs from all others in some way, and thus will generate different patterns of effects on various cognitive outcome measures. There is certainly room for debate about whether it is necessary to *only* consider the impact of each unique type of intervention, or whether there exist categories into which unique groups of interventions can be combined. However, we urge caution here, as even seemingly reasonable sub-categories, such as "working memory training," may still be problematic (Au et al. 2015; Melby-Lervag and Hulme 2013). For instance, the term "working memory training" can easily promote confusion regarding whether working memory was a targeted outcome or a means of training. Regardless, it is clear that "brain training" is simply too broad a category to have descriptive value.

Furthermore, it is notable that in those cases where the term "brain training" is used, it is often in the context of the question "Does brain training *work*?" (Howard 2016; Walton 2016). However, in the same way that the term "brain training" implies a common mechanism of action that is inconsistent with the wide number of paradigms in the field, the term "work" suggests a singular target that is inconsistent with the wide number of training targets in the field. The cognitive processes targeted by a paradigm intended to improve functioning in individuals diagnosed with schizophrenia may be quite different from those meant to improve functioning in a healthy older adult or a child diagnosed with ADHD. Similarly, whether a training paradigm serves to recover lost function (e.g., improving the cognitive skills of a 65-year old who has experienced age-related decline), ameliorate abnormal function (e.g., enhancing cognitive skills in an individual with developmental cognitive deficits), or improve normal function (e.g., improving speed of processing in a healthy 21-year old) might all fall under the description of whether cognitive training "works"—but are absolutely not identical and in many cases may have very little in common.

In many ways then, the question "Does brain training work?" is akin to the question "Do drugs work?" As is true of the term "brain training," the term "drugs" is a superordinate category label that encompasses an incredible variety of chemicals—from those that were custom-designed for a particular purpose to those that arose "in the wild," but now are being put to practical ends. They can be delivered in many different ways, at different doses, on different schedules, and in endless combinations. The question of whether drugs "work" is inherently defined with respect to the stated target condition(s). And finally, drugs with the same real-world target (e.g., depression), may act on completely different systems (e.g., serotonin versus dopamine versus norepinephrine).

It is undoubtedly the case, at least in the scientific community, that such broad and imprecise terms are used as a matter of expository convenience (e.g., as is needed in publication titles), rather than to actually reflect the belief that all behavioral interventions intended to improve cognition are alike in mechanisms, design, and goals (Redick et al. 2015; Simons et al. 2016). Nonetheless, imprecise terminology can easily lead to imprecise understanding and open the possibility for criticism of the field. Thus, our first recommendation is for the field to use well-defined and precise terminology, both to describe interventions and to describe an intervention's goals and outcomes. In the case of titles, abstracts, and other text length-limited parts of manuscripts, this would include using the most precise term that these limitations allow for (e.g., "working memory training" would be more precise than "brain training," and "dual-N-back training" would be even more precise). However, if the text length limitations do not allow for the necessary level of precision, more neutral terminology (e.g., as in our title—"behavioral interventions for cognitive enhancement") is preferable with a transition to more precise terminology in the manuscript body.

Different Types of Cognitive Enhancement Studies Have Fundamentally Different Goals

One clear benefit to the use of more precise and better-defined terms is that research studies can be appropriately and clearly delineated given their design and goals. Given the potential

real-world benefits that behavioral interventions for cognitive enhancement could offer, a great deal of focus in the domain to date has been placed on studies that could potentially demonstrate real-world impact. However, as is also true in medical research, demonstration of real-world impact is not the goal of every study.

For the purposes of this document, we differentiate between four broad, but distinct, types of research study:

(i)   Feasibility studies
(ii) Mechanistic studies
(iii) Efficacy studies
(iv) Effectiveness studies

Each type of study is defined by fundamentally different research questions. They will thus differ in their overall methodological approach and, because of these differences, in the conclusions one may draw from the study results. Critically though, if properly executed, each study type provides valuable information for the field going forward. Here, we note that this document focuses exclusively on intervention studies. There are many other study types that can and do provide important information to the field (e.g., the huge range of types of basic science studies—correlational, cross-sectional, longitudinal, etc.). However, these other study types are outside the scope of the current paper and they are not specific for the case of training studies (see the "Discussion" section for a brief overview of other study types).

Below, we examine the goals of each type of study listed above—feasibility, mechanistic, efficacy, and effectiveness studies—and discuss the best methodological practices to achieve those goals. We recommend that researchers state clearly at the beginning of proposals or manuscripts the type of study that is under consideration, so that reviewers can assess the methodology relative to the research goals. And although we make a number of suggestions regarding broadly defined best methodological practices within a study type, it will always be the case that a host of individual-level design choices will need to be made and justified on the basis of specific well-articulated theoretical models.

Feasibility, Mechanistic, Efficacy, and Effectiveness Studies—Definitions and Broad Goals

Feasibility Studies The goal of a feasibility study is to test the viability of a given paradigm or project—almost always as a precursor to one of the study designs to follow. Specific goals may include identifying potential practical or economic problems that might occur if a mechanistic, efficacy, or effectiveness study is pursued (Eldridge et al. 2016; Tickle-Degnen 2013). For instance, it may be important to know if participants can successfully complete the training task(s) as designed (particularly in the case of populations with deficits). Is the training task too difficult or too easy? Are there side effects that might induce attrition (e.g., eye strain, motion sickness, etc.)? Is training compliance sufficient? Do the dependent variables capture performance with the appropriate characteristics (e.g., as related to reliability, inter-participant variability, data distribution, performance not being at ceiling or floor, etc.)? What is the approximate effect size of a given intervention and what sample size would be appropriate to draw clear conclusions?

Many labs might consider such data collection to be simple "piloting" that is never meant to be published. However, there may be value in re-conceptualizing many "pilot studies" as feasibility studies where dissemination of results is explicitly planned (although note that other groups have drawn different distinctions between feasibility and pilot studies, see for instance, Eldridge et al. 2016; Whitehead et al. 2014). This is especially true in circumstances in which aspects of feasibility are broadly applicable, rather than being specific to a single paradigm. For instance, a feasibility study assessing whether children diagnosed with ADHD show sufficient levels of compliance in completing an at-home multiple day behavioral training paradigm unmonitored by their parents could provide valuable data to other groups planning on working with similar populations.

Implicit in this recommendation then is the notion that the value of feasibility studies depends on the extent to which aspects of feasibility are in doubt (e.g., a feasibility study showing that college-aged individuals can complete ten 1-h in-lab behavioral training sessions would be of limited value as there are scores of existing studies showing this is true). We thus suggest that researchers planning feasibility studies (or pilot studies that can be re-conceptualized as feasibility studies) consider whether reasonably minor methodological tweaks could not only demonstrate feasibility of their own particular paradigm, but also speak toward broader issues of feasibility in the field. While this might not be possible in all cases or for all questions (as some issues of feasibility are inherently and inseparably linked to a very specific aspect of a given intervention), we argue that there are, in fact, a number of areas wherein the field currently lacks basic feasibility information. Published data in these areas could provide a great deal of value to the design of mechanistic, efficacy, or effectiveness studies across many different types of behavioral interventions for cognitive enhancement. These include issues ranging from compliance with instructions (e.g., in online training studies), to motivation and attrition, to task-performance limitations (e.g., use of keyboard, mouse, joysticks, touch screens, etc.). As such, we believe there is value in researchers considering whether, with a slight shift in approach and emphasis, the piloting process that is often thought of happening "before the science has really begun" could contribute more directly to the base of knowledge in the field.

Finally, it is worth noting that a last question that can potentially be addressed by a study of this type is whether there is enough evidence in favor of a hypothesis to make a full-fledged study of mechanism, efficacy, or effectiveness potentially feasible and worth undertaking. For instance, showing the potential for efficacy in underserved or difficult-to-study populations could provide inspiration to other groups to examine related approaches in that population.

The critical to-be-gained knowledge here includes an estimate of the expected effect size, and in turn, a power estimate of the sample size that would be required to demonstrate statistically significant intervention effects (or convincing null effects). It would also provide information about whether an effect is likely to be clinically significant (which could require a higher effect size than what is necessary to reach statistical significance). While feasibility studies will not be conclusive (and all scientific discourse of such studies should emphasize this fact), they can provide both information and encouragement that can add to scientific discourse and lead to innovation. By recognizing that feasibility studies have value, for instance, in indicating whether larger-scale efficacy trials are worth pursuing, it will further serve to reduce the tendency for authors to overinterpret results (i.e., if researchers know that their study will be accepted as valuable for providing information that efficacy studies could be warranted, authors may be less prone to directly and inappropriately suggesting that their study speaks to efficacy questions).

Mechanistic Studies The goal of a mechanistic study is to identify the mechanism(s) of action of a behavioral intervention for cognitive enhancement. In other words, the question is not whether, but how. More specifically, mechanistic studies test an explicit hypothesis, generated by a clear theoretical framework, about a mechanism of action of a particular cognitive enhancement approach. As such, mechanistic studies are more varied in their methodological approach than the other study types. They are within the scope of fundamental or basic research, but they do often provide the inspiration for applied efficacy and effectiveness studies. Thus, given their pivotal role as hypothesis testing grounds for applied studies, it may be helpful for authors to distinguish when the results of mechanistic studies indicate that the hypothesis is sufficiently mature for practical translation (i.e., experimental tests of the hypothesis are reproducible and are likely to produce practically relevant outcomes) or is instead in need of further confirmation. Importantly, we note that the greater the level of pressure to translate research from the lab to the real world, the more likely it will be that paradigms and/or hypotheses will make this transition prematurely or that the degree of real-world applicability will be overstated (of which there are many examples). We thus recommend appropriate nuance if authors of mechanistic studies choose to discuss potential real-world implications of the work. In particular, the discussion should be used to explicitly comment on whether the data indicate readiness for translation to efficacy or effectiveness studies, rather than giving the typical full-fledged nods to possible direct real-world applications (which are not among the goals of a mechanistic study).

Efficacy Studies The goal of efficacy studies is to validate a given intervention as the cause of cognitive improvements above and beyond any placebo or expectation-related effects (Fritz and Cleland 2003; Marchand et al. 2011; Singal et al. 2014). The focus is not on establishing the underlying mechanism of action of an intervention, but on establishing that the intervention (when delivered in its planned "dose") produces the intended outcome when compared to a placebo control or to another intervention previously proven to be efficacious. Although efficacy studies are often presented as asking "Does the paradigm produce the intended outcome?" they would be more accurately described as asking, "Does the paradigm produce the anticipated outcome in the exact and carefully controlled population of interest when the paradigm is used precisely as intended by the researchers?" Indeed, given that the goal is to establish whether a given intervention, as designed and intended, is efficacious, reducing unexplained variability or unintended behavior is key (e.g., as related to poor compliance, trainees failing to understand what is required of them, etc.).

Effectiveness Studies As with efficacy studies, the goal of effectiveness studies is to assess whether a given intervention produces positive impact of the type desired and predicted, most commonly involving real-world impact. However, unlike efficacy studies—which focus on results obtained under a set of carefully controlled circumstances—effectiveness studies examine whether significant real-world impact is observed when the intervention is used in less than ideally controlled settings (e.g., in the "real-world"; Fritz and Cleland 2003; Marchand et al. 2011; Singal et al. 2014). For example, in the pharmaceutical industry, an efficacy study may require that participants take a given drug every day at an exact time of day for 30 straight days (i.e., the exact schedule is clearly defined and closely monitored). An effectiveness study, in contrast, would examine whether the drug produces benefits when it is used within real-world clinical settings, which might very well include poor compliance with instructions (e.g., taking the drug at different times, missing doses, taking multiple doses to "catch up," etc.). Similarly, an efficacy study in the pharmaceutical domain might narrowly select participants (e.g., in a study of a drug for chronic pain, participants with other comorbid conditions, such as major depressive disorder, might be excluded), whereas an effectiveness trial would consider all individuals likely to be prescribed the drug, including those with comorbidities.

Effectiveness studies of behavioral interventions have historically been quite rare as compared to efficacy studies, which is a major concern for real-world practitioners (although there are some fields within the broader domain of psychology where such studies have been more common—e.g., human factors, engineering psychology, industrial organization, education, etc.). Indeed, researchers seeking to use behavioral interventions for cognitive enhancement in the real-world (e.g., to augment learning in a school setting) are unlikely to encounter the homogenous and fully compliant individuals who comprise the participant pool in efficacy studies. This in turn may result in effectiveness study outcomes that are not consistent with the precursor efficacy studies, a point we return to when considering future directions.

Critically, although we describe four well-delineated categories in the text above, in practice, studies will tend to vary along the broad and multidimensional space of study types. This is unlikely to change, as variability in approach is the source of much knowledge. However, we nonetheless recommend that investigators should be as clear as possible about the type of studies they undertake starting with an explicit description of the study goals (which in turn constrains the space of acceptable methods).

Methodological Considerations as a Function of Study Type

Below, we review major design decisions including participant sampling, control group selection, assignment to groups, and participant and researcher blinding, and discuss how they may be influenced by study type.

Participant Sampling across Study Types

One major set of differences across study types lies in the participant sampling procedures—including the population(s) from which participants are drawn and the appropriate sample size. Below, we consider how the goals of the different study types can, in turn, induce differences in the sampling procedures.

Feasibility Studies In the case of feasibility studies, the targeted population will depend largely on the subsequent planned study or studies (typically either a mechanistic study or an efficacy study). More specifically, the participant sample for a feasibility study will ideally be drawn from a population that will be maximally informative for subsequent planned studies. Note that this will most often be the exact same population as will be utilized in the subsequent planned studies. For example, consider a set of researchers who are planning an efficacy study in older adults who live in assisted living communities. In this hypothetical example, before embarking on

the efficacy study, the researchers first want to assess feasibility of the protocol in terms of: (1) long-term compliance and (2) participants' ability to use a computer-controller to make responses. In this case, they might want to recruit participants for the feasibility study from the same basic population as they will recruit from in the efficacy study.

This does not necessarily have to be the case though. For instance, if the eventual population of interest is a small (and difficult to recruit) population of individuals with specific severe deficits, one may first want to show feasibility in a larger and easier to recruit population (at least before testing feasibility in the true population of interest). Finally, the sample size in feasibility studies will often be relatively small as compared to the other study types, as the outcome data simply need to demonstrate feasibility.

Mechanistic and Efficacy Studies At the broadest level, the participant sampling for mechanistic and efficacy studies will be relatively similar. Both types of studies will tend to sample participants from populations intended to reduce unmeasured, difficult-to-model, or otherwise potentially confounding variability. Notably, this does not necessarily mean the populations will be homogenous (especially given that individual differences can be important in such studies). It simply means that the populations will be chosen to reduce unmeasured differences. This approach may require excluding individuals with various types of previous experience. For example, a mindfulness-based intervention might want to exclude individuals who have had any previous meditation experience, as such familiarity could reduce the extent to which the experimental paradigm would produce changes in behavior. This might also require excluding individuals with various other individual difference factors. For example, a study designed to test the efficacy of an intervention paradigm meant to improve attention in typically developing individuals might exclude individuals diagnosed with ADHD.

The sample size of efficacy studies must be based upon the results of a power analysis and ideally will draw upon anticipated effect sizes observed from previous feasibility and/or mechanistic studies. However, efficacy studies are often associated with greater variability as compared with mechanistic and feasibility studies. Hence, one consideration is whether the overall sample in efficacy studies should be even larger still. Both mechanistic and efficacy studies could certainly benefit from substantially larger samples than previously used in the literature and from considering power issues to a much greater extent.

Effectiveness Studies In effectiveness studies, the population of interest is the population that will engage with the intervention as deployed in the real-world and thus will be recruited via similar means as would be the case in the real-world. Because recruitment of an unconstrained participant sample

will introduce substantial inter-individual variability in a number of potential confounding variables, sample sizes will have to be correspondingly considerably larger for effectiveness studies as compared to efficacy studies. In fact, multiple efficacy studies using different populations may be necessary to identify potential sources of variation and thus expected power in the full population.

Control Group Selection across Study Types

A second substantial difference in methodology across study types is related to the selection of control groups. Below, we consider how the goals of the different study types can, in turn, induce differences in control group selection.

Feasibility Studies In the case of feasibility studies, a control group is not necessarily required (although one might perform a feasibility study to assess the potential informativeness of a given control or placebo intervention). The goal of a feasibility study is not to demonstrate mechanism, efficacy, or effectiveness, but is instead only to demonstrate viability, tolerability, or safety. As such, a control group is less relevant because the objective is not to account for confounding variables. If a feasibility study is being used to estimate power, a control group (even a passive control group) could be useful, particularly if gains unrelated to the intervention of interest are expected (e.g., if the tasks of interest induce test-retest effects, if there is some natural recovery of function unattributable to the training task, etc.).

Mechanistic Studies To discuss the value and selection of various types of control groups for mechanistic studies (as well as for efficacy, and effectiveness studies), it is worth briefly describing the most common design for such studies: the pre/post design (Green et al. 2014). In this design, participants first undergo a set of pre-test (baseline) assessments that measure performance along the dimensions of interest. The participants are then either randomly or pseudo-randomly assigned to a treatment group. For instance, in the most basic design, the two treatment groups would be an active intervention and a control intervention. The participants then complete the training associated with their assigned group. In the case of behavioral interventions for cognitive enhancement, this will often involve performing either a single task or set of tasks for several hours spaced over many days or weeks. Finally, after the intervention is completed, participants perform the same tasks they completed at pre-test as part of a post-test (ideally using parallel-test versions rather than truly identical versions). The critical measures are usually comparisons of pre-test to post-test changes in the treatment group as compared to a control group. For example, did participants in the intervention group show a greater improvement in performance from pre-test to post-test as compared to the participants in the

control group? The purpose of the control group is thus clear—to subtract out any confounding effects from the intervention group data (including simple test-retest effects), leaving only the changes of interest. This follows from the assumption that everything is, in fact, the same in the two groups with the exception of the experimental manipulation of interest.

In a mechanistic study, the proper control group may appear to be theoretically simple to determine—given some theory or model of the mechanism through which a given intervention acts, the ideal control intervention is one that isolates the posited mechanism(s). In other words, if the goal is to test a particular mechanism of action, then the proper control will contain all of the same "ingredients" as the experimental intervention other than the proposed mechanism(s) of action. Unfortunately, while this is simple in principle, in practice it is often quite difficult because it is not possible to know with certainty all of the "ingredients" inherent to either the experimental intervention or a given control.

For example, in early studies examining the impact of what have come to be known as "action video games" (one genre of video games), the effect of training on action video games was contrasted with training on the video game Tetris as the control (Green and Bavelier 2003). Tetris was chosen to control for a host of mechanisms inherent in video games (including producing sustained arousal, task engagement, etc.), while not containing what were felt to be the critical components inherent to action video games specifically (e.g., certain types of load placed on the perceptual, cognitive, and motor systems). However, subsequent research has suggested that Tetris may indeed place load on some of these processes (Terlecki et al. 2008). Had the early studies produced null results—i.e., if the action video game trained group showed no benefits as compared to the Tetris trained group—it would have been easy to incorrectly infer that the mechanistic model was incorrect, as opposed to correctly inferring that both tasks in fact contained the mechanism of interest.

Because of this possibility, we suggest that there is significant value for mechanistic studies to consider adding a second control group—what we would call a "business as usual" control—to aid in the interpretation of null results. Such a control group (sometimes also referred to as a "test-retest" control group or passive control group) undergoes no intervention whatsoever, but completes the pre- and post-testing within the same timeframe as the trained groups. If *neither* the intervention group nor the active control group shows benefits relative to this second control group, this is strong evidence against either the mechanistic account itself, against the ability of the intervention to activate the proposed mechanism, or against the ability of the outcome measures to detect effects of the intervention (Roberts et al. 2016). Conversely, if *both* the intervention and the active control show a benefit relative to the business-as-usual control group, a range of other

possibilities are suggested. For instance, it could be the case that both the intervention and active control group have properties that stimulate the proposed mechanism. It could also be the case that there is a different mechanism of action inherent in the intervention training, control training, or both, that produces the same behavioral outcome. Finally, it could be that the dependent measures are not sensitive enough to detect real differences between the outcomes. Such properties might include differential expectancies that lead to the same outcome including the simple adage that sometimes doing almost anything is better than nothing, that the act of being observed tends to induce enhancements, or any of a host of other possibilities.

Efficacy Studies For efficacy studies, the goal of a control group is to subtract out the influence of a handful of mechanisms of "no interest"—including natural progression and participant expectations. In the case of behavioral interventions for cognitive enhancement, natural progression will include, for instance, mechanisms: (1) related to time and/or development, such as children showing a natural increase in attentional skills as they mature independent of any interventions; and (2) those related to testing, such as the fact that individuals undergoing a task for a second time will often have improved performance relative to the first time they underwent the task. Participant expectations, meanwhile, would encompass mechanisms classified as "placebo effects." Within the medical world, these effects are typically controlled via a combination of an inert placebo control condition (e.g., sugar pill or saline drip) and participant and experimenter blinding (i.e., neither the participant nor the experimenter being informed as to whether the participant is in the active intervention condition or the placebo control condition). In the case of behavioral interventions for cognitive enhancement, it is worth noting, just as was true of mechanistic studies, that there is not always a straightforward link between a particular placebo control intervention and the mechanisms that placebo is meant to control for. It is always possible that a given placebo control intervention, that is meant to be "inert," could nonetheless inadvertently involve mechanisms that are of theoretical interest.

Given this, in addition to a placebo control (which we discuss in its own section further below), we suggest here that efficacy studies also include a business-as-usual control group. This will help in cases where the supposed "inert placebo" control turns out to be not inert with respect to the outcomes of interest. For instance, as we will see below, researchers may wish to design an "inert" control that retains some plausibility as an active intervention for participants, so as to control for participant expectations. However, in doing so, they may inadvertently include "active" ingredients. Notably, careful and properly powered individual difference studies examining the control condition conducted prior to the efficacy study will reduce this possibility.

More critically perhaps, in the case of an efficacy study, such business-as-usual controls have additional value in demonstrating that there is no harm produced by the intervention. Indeed, it is always theoretically possible that both the active and the control intervention may inhibit improvements that would occur due to either natural progression, development, maturation, or in comparison with how individuals would otherwise spend their time. This is particularly crucial in the case of any intervention that replaces activities known to have benefits. This would be the case, for instance, of a study examining potential for STEM benefits where classroom time is replaced by an intervention, or where a physically active behavior is replaced by a completely sedentary behavior.

Effectiveness Studies For effectiveness studies, because the question of interest is related to benefits that arise when the intervention is used in real-world settings, the proper standard against which the intervention should be judged is business-as-usual—or in cases where there is an existing proven treatment or intervention, the contrast may be against normal standard of care (this latter option is currently extremely rare in our domain, if it exists at all). In other words, the question becomes: "Is this use of time and effort in the real world better for cognitive outcomes than how the individual would otherwise be spending that time?" Or, if being compared to a current standard of care, considerations might also include differential financial costs, side effects, accessibility concerns, etc.

We conclude by noting that the recommendation that many mechanistic and all efficacy studies include a business-as-usual control has an additional benefit beyond aiding in the interpretation of the single study at hand. Namely, such a broadly adopted convention will produce a common control group against which all interventions are contrasted (although the outcome measures will likely still differ). This in turn will greatly aid in the ability to determine effect sizes and compare outcomes across interventions. Indeed, in cases where the critical measure is a difference of differences (e.g., (post-performance$_{\text{intervention}}$ − pre-performance$_{\text{intervention}}$) − (post-performance$_{\text{control}}$ − pre-performance$_{\text{control}}$), there is no coherent way to contrast the size of the overall effects when there are different controls across studies. Having a standard business-as-usual control group allows researchers to observe which interventions tend to produce bigger or smaller effects and take that information into account when designing new interventions. There are of course caveats, as business-as-usual and standard-of-care can differ across groups. For example, high SES children may spend their time in different ways than low-SES children, rendering it necessary to confirm that apples-to-apples comparisons are being made.

Assignment to Groups

While the section above focused on the problem of choosing appropriate control interventions, it is also important to

consider how individuals are assigned to groups. Here, we will consider all types of studies together (although this is only a concern for feasibility studies in cases where the feasibility study includes multiple groups). Given a sufficiently large number of participants, true random assignment can be utilized. However, it has long been recognized that truly random assignment procedures can create highly imbalanced group membership, a problem that becomes increasingly relevant as group sizes become smaller. For instance, if group sizes are small, it would not be impossible (or potentially even unlikely) for random assignment to produce groups that are made up of almost all males or almost all females or include almost all younger individuals or almost all older individuals (depending on the population from which the sample is drawn). This in turn can create sizeable difficulties for data interpretation (e.g., it would be difficult to examine sex as a biological variable if sex was confounded with condition).

Beyond imbalance in demographic characteristics (e.g., age, sex, SES, etc.), true random assignment can also create imbalance in initial measured performance; in other words—pre-test (or baseline) differences. Pre-test differences in turn create severe difficulties in interpreting changes in the typical pre-test → training → post-test design. As just one example, consider a situation where the experimental group's performance is worse at pre-test than the control group's performance. If, at post-test, a significant improvement is seen in the experimental group, but not in the control group, a host of interpretations are possible. Such a result could be due to: (1) a positive effect of the intervention, (2) it could be regression to the mean due to unreliable measurements, or (3) it could be that people who start poorly have more room to show simple test-retest effects, etc. Similar issues with interpretation arise when the opposite pattern occurs (i.e., when the control group starts worse than the intervention group).

Given the potential severity of these issues, there has long been interest in the development of methods for group assignment that retain many of the aspects and benefits of true randomization while allowing for some degree of control over group balance (in particular in clinical and educational domains; Chen and Lee 2011; Saghaei 2011; Taves 1974; Zhao et al. 2012). A detailed examination of this literature is outside of the scope of the current paper. However, such promising methods have begun to be considered and/or used in the realm of cognitive training (Green et al. 2014; Jaeggi et al. 2011; Redick et al. 2013). As such, we urge authors to consider various alternative group assignment approaches that have been developed (e.g., creating matched pairs, creating homogeneous sub-groups or blocks, attempting to minimize group differences on the fly, etc.) as the best approach will depend on the study's sample characteristics, the goals of the study, and various practical concerns (e.g., whether the study enrolls participants on the fly, in batches, all at once, etc.). For instance, in studies employing extremely large task batteries, it may not be feasible to create groups that are matched for pre-test performance on all measures. The researchers would then need to decide which variables are most critical to match (or if the study was designed to assess a smaller set of latent constructs that underlie performance on the larger set of various measures, it may be possible to match based upon the constructs). In all, our goal here is simply to indicate that not only can alternative methods of group assignment be consistent with the goal of rigorous and reproducible science, but in many cases, such methods will produce more valid and interpretable data than fully random group assignment.

Can Behavioral Interventions Achieve the Double-Blind Standard?

One issue that has been raised in the domain of behavioral interventions is whether it is possible to truly blind participants to condition in the same "gold standard" manner as in the pharmaceutical field. After all, whereas it is possible to produce two pills that look identical, one an active treatment and one an inert placebo, it is not possible to produce two behavioral interventions, one active and one inert, that are outwardly perfectly identical (although under some circumstances, it may be possible to create two interventions where the manipulation is subtle enough to be perceptually indistinguishable to a naive participant). Indeed, the extent to which a behavioral intervention is "active" depends entirely on what the stimuli are and what the participant is asked to do with those stimuli. Thus, because it is impossible to produce a mechanistically active behavioral intervention and an inert control condition that look and feel identical to participants, participants may often be able to infer their group assignment.

To this concern, we first note that even in pharmaceutical studies, participants can develop beliefs about the condition to which they have been assigned. For instance, active interventions often produce some side effects, while truly inert placebos (like sugar pills or a saline drip) do not. Interestingly, there is evidence to suggest: (1) that even in "double-blind" experiments, participant blinding may sometimes be broken (i.e., via the presence or absence of side effects; Fergusson et al. 2004; Kolahi et al. 2009; Schulz et al. 2002) and (2) the ability to infer group membership (active versus placebo) may impact the magnitude of placebo effects (Rutherford et al. 2009, although see Fassler et al. 2015).

Thus, we would argue that—at least until we know more about how to reliably measure participant expectations and how such expectations impact on our dependent variables—efficacy studies should make every attempt to adopt the same standard as the medical domain. Namely, researchers should employ an active control condition that has some degree of face validity as an "active" intervention from the participants' perspective, combined with additional attempts to induce participant blinding (noting

further that attempts to assess the success of such attempts is perhaps surprisingly rare in the medical domain—Fergusson et al. 2004; Hrobjartsson et al. 2007).

Critically, this will often start with participant recruitment—in particular using recruitment methods that either minimize the extent to which expectations are generated or serve to produce equivalent expectations in participants, regardless of whether they are assigned to the active or control intervention (Schubert and Strobach 2012). For instance, this may be best achieved by introducing the overarching study goals as examining which of two active interventions is most effective, rather than contrasting an experimental intervention with a control condition. This process will likely also benefit retention as participants are more likely to stay in studies that they believe might be beneficial.

Ideally, study designs should also, as much as is possible, include experimenter blinding, even though it is once again more difficult in the case of a behavioral intervention than in the case of a pill. In the case of two identical pills, it is completely possible to blind the experimental team to condition in the same manner as the participant (i.e., if the active drug and placebo pill are perceptually indistinguishable, the experimenter will not be able to ascertain condition from the pill alone—although there are perhaps other ways that experimenters can nonetheless become unblinded; Kolahi et al. 2009). In the case of a behavioral intervention, those experimenter(s) who engage with the participants during training will, in many cases, be able to infer the condition (particularly given that those experimenters are nearly always lab personnel who, even if not aware of the exact tasks or hypotheses, are reasonably well versed in the broader literature). However, while blinding those experimenters who interact with participants during training is potentially difficult, it is quite possible and indeed desirable to ensure that the experimenter(s) who run the pre- and post-testing sessions are blind to condition and perhaps more broadly to the hypotheses of the research (but see the "Suggestions for Funding Agencies" section below, as such practices involve substantial extra costs).

Outcome Assessments across Study Types

Feasibility Studies The assessments used in behavioral interventions for cognitive enhancement arise naturally from the goals. For feasibility studies, the outcome variables of interest are those that will speak to the potential success or failure of a subsequent mechanistic, efficacy, or effectiveness studies. These may include the actual measures of interest in those subsequent studies, particularly if one purpose of the feasibility study is to estimate possible effect sizes and necessary power for those subsequent studies. They may also include a host of measures that would not be primary outcome variables in subsequent studies. For instance, compliance may be a primary outcome variable in a feasibility study, but not in a

subsequent efficacy study (where compliance may only be measured in order to exclude participants with poor compliance).

Mechanistic Studies For mechanistic studies, the outcomes that are assessed should be guided entirely by the theory or model under study. These will typically make use of in-lab tasks that are either thought or known to measure clearly defined mechanisms or constructs. Critically, for mechanistic studies focused on true learning effects (i.e., enduring behavioral changes), the assessments should always take place after potential transient effects associated with the training itself have dissipated. For instance, some video games are known to be physiologically arousing. Because physiological arousal is itself linked with increased performance on some cognitive tasks, it is desireable that testing takes place after a delay (e.g., 24 h or longer depending on the goal), thus ensuring that short-lived effects are no longer in play (the same holds true for efficacy and effectiveness studies).

Furthermore, there is currently a strong emphasis in the field toward examining mechanisms that will produce what is commonly referred to as *far transfer* as compared to just producing *near transfer*. First, it is important to note that the distinction between what is "far" and what is "near" is commonly a qualitative, rather than quantitative one (Barnett and Ceci 2002). Near transfer is typically used to describe cases where training on one task produces benefits on tasks meant to tap the same core construct as the trained task using slightly different stimuli or setups. For example, those in the field would likely consider transfer from one "complex working memory task" (e.g., the O-Span) to another "complex working memory task" (e.g., Spatial Span) to be an example of near transfer. Far transfer is then used to describe situations where the training and transfer tasks are not believed to tap the exact same core construct. In most cases, this means partial, but not complete, overlap between the training and transfer tasks (e.g., working memory is believed to be one of many processes that predict performance on fluid intelligence measures, so training on a working memory task that improves performance on a fluid intelligence task would be an instance of far transfer).

Second, and perhaps more critically, the inclusion of measures to assess such "far transfer" in a mechanistic study are only important to the extent that such outcomes are indeed a key prediction of the mechanistic model. To some extent, there has been a tendency in the field to treat a finding of "only near transfer" as a pejorative description of experimental results. However, there are a range of mechanistic models where only near transfer to tasks with similar processing demands would be expected. As such, finding near transfer can be both theoretically and practically important. Indeed, some translational applications of training may only require near transfer (although true real-world application will essentially always require some degree of transfer across content).

Thus, in general, we would encourage authors to describe the similarities and differences between trained tasks and outcome measures in concrete, quantifiable terms whenever possible (whether these descriptions are in terms of task characteristics—e.g., similarities of stimuli, stimulus modality, task rules, etc.—or in terms of cognitive constructs or latent variables).

We further suggest that assessment methods in mechanistic studies would be greatly strengthened by including, and clearly specifying, tasks that are *not* assumed to be susceptible to changes in the proposed mechanism under study. If an experimenter demonstrates that training on Task A, which is thought to tap a specific mechanism of action, produces predictable improvements in some new Task B, which is also thought to tap that same specific mechanism, then this supports the underlying model or hypothesis. Notably, however, the case would be greatly strengthened if the same training did not also change performance on some other Task C, which does not tap the underlying specific mechanism of action. In other words, only showing that Task A produces improvements on Task B leaves a host of other possible mechanisms alive (many of which may not be of interest to those in cognitive psychology). Showing that Task A produces improvements on Task B, but not on Task C, may rule out other possible contributing mechanisms. A demonstration of a double dissociation between training protocols and pre-post assessment measures would be better still, although this may not always be possible with all control tasks. If this suggested convention of including tasks not expected to be altered by training is widely adopted, it will be critical for those conducting future meta-analyses to avoid improperly aggregating across outcome measures (i.e., it would be a mistake, in the example above, for a meta-analysis to directly combine Task B and Task C to assess the impact of training on Task A).

Efficacy Studies The assessments that should be employed in efficacy studies lie somewhere between the highly controlled, titrated, and precisely defined lab-based tasks that will be used most commonly in mechanistic studies, and the functionally meaningful real-world outcome measurements that are employed in effectiveness studies. The broadest goal of efficacy studies is, of course, to examine the potential for real-world impact. Yet, the important sub-goal of maintaining experimental control means that researchers will often use lab-based tasks that are thought (or better yet, known) to be associated with real-world outcomes. We recognize that this link is often tenuous in the peer-reviewed literature and in need of further well-considered study. There are some limited areas in the literature where real-world outcome measures have been examined in the context of cognitive training interventions. Examples include the study of retention of driving skills (in older adults; Ross et al. 2016) or academic achievement (in children; Wexler et al. 2016) that have been measured in both

experimental and control groups. Another example is psychiatric disorders (e.g., schizophrenia; Subramaniam et al. 2014), where real-world functional outcomes are often the key dependent variable.

In many cases though, the links are purely correlational. Here, we caution that such an association does not ensure that a given intervention with a known effect on lab-based measures will improve real-world outcomes. For instance, two measures of cardiac health—lower heart-rate and lower blood pressure—are both correlated with reductions in the probability of cardiac-related deaths. However, it is possible for drugs to produce reductions in heart-rate and/or blood pressure without necessarily producing a corresponding decrease in the probability of death (Diao et al. 2012). Therefore, the closer that controlled lab-based efficacy studies can get to the measurement of real-world outcomes, the better. We note that the emergence of high-fidelity simulations (e.g., as implemented in virtual reality) may help bridge the gap between well-controlled laboratory studies and a desire to observe real-world behaviors (as well as enable us to examine real-world tasks that are associated with safety concerns - such as driving). However, caution is warranted as this domain remains quite new and the extent to which virtual reality accurately models or predicts various real-world behaviors of interest is at present unknown.

Effectiveness Studies In effectiveness studies, the assessments also spring directly from the goals. Because impact in the real-world is key, the assessments should predominantly reflect real-world functional changes. We note that "efficiency," which involves a consideration of both the size of the effect promoted by the intervention *and* the cost of the intervention is sometimes utilized as a critical metric in assessing both efficacy and effectiveness studies (larger effects and/or smaller costs mean greater efficiency; Andrews 1999; Stierlin et al. 2014). By contrast, we are focusing here primarily on methodology associated with accurately describing the size of the effect promoted by the intervention in question (although we do point out places where this methodology can be costly). In medical research, the outcomes of interest are often described as patient relevant outcomes (PROs): outcome variables of particular importance to the target population. This presents a challenge for the field, though, as there are currently a limited number of patient-relevant "real-world measures" available to researchers, and these are not always applicable to all populations.

One issue that is often neglected in this domain is the fact that improvements in real-world behaviors will not always occur immediately after a cognitive intervention. Instead, benefits may only emerge during longer follow-up periods, as individuals consolidate enhanced cognitive skills into more adaptive real-world behaviors. As an analogy from the visual domain, a person with nystagmus (constant, repetitive, and
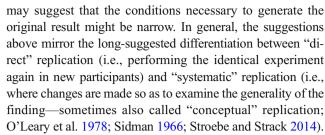
uncontrolled eye-movements) may find it difficult to learn to read because the visual input is so severely disrupted. Fixing the nystagmus would provide the system with a stronger opportunity to learn to read, yet would not give rise to reading in and of itself. The benefits to these outcomes would instead only be observable many months or years after the correction.

The same basic idea is true of what have been called "sleeper" or "protective" effects. Such effects also describe situations where an effect is observed at some point in the future, regardless of whether or not an immediate effect was observed. Specifically, sleeper or protective benefits manifest in the form of a reduction in the magnitude of a natural decline in cognitive function (Jones et al. 2013; Rebok et al. 2014). These may be particularly prevalent in populations that are at risk for a severe decline in cognitive performance. Furthermore, there would be great value in multiple long-term follow-up assessments even in the absence of sleeper effects to assess the long-term stability or persistence of any findings. Again, like many of our other recommendations, the presence of multiple assessments increases the costs of a study (particularly as attrition rates will likely rise through time).

Replication—Value and Pitfalls

There have been an increasing number of calls over the past few years for more replication in psychology (Open Science 2012; Pashler and Harris 2012; Zwaan et al. 2017). This issue has been written about extensively, so here we focus on several specific aspects as they relate to behavioral interventions for cognitive enhancement. First, questions have been raised as to how extensive a change can be made from the original and still be called a "replication." We maintain that if changes are made from the original study design (e.g., if outcome measures are added or subtracted; if different control training tasks are used; if different populations are sampled; if a different training schedule is used), then this ceases to be a replication and becomes a test of a new hypothesis. As such, only studies that make no such changes could be considered replications. Here, we emphasize that because there are a host of cultural and/or other individual difference factors that can differ substantially across geographic locations (e.g., educational and/or socioeconomic backgrounds, religious practices, etc.) that could potentially affect intervention outcomes, "true replication" is actually quite difficult. We also note that when changes are made to a previous study's design, it is often because the researchers are making the explicit supposition that such changes yield a better test of the broadest level experimental hypothesis. Authors in these situations should thus be careful to indicate this fact, without making the claim that they are conducting a replication of the initial study. Instead, they can indicate that a positive result, if found using different methods, serves to demonstrate the validity of the intervention across those forms of variation. A negative result meanwhile

may suggest that the conditions necessary to generate the original result might be narrow. In general, the suggestions above mirror the long-suggested differentiation between "direct" replication (i.e., performing the identical experiment again in new participants) and "systematic" replication (i.e., where changes are made so as to examine the generality of the finding—sometimes also called "conceptual" replication; O'Leary et al. 1978; Sidman 1966; Stroebe and Strack 2014).

We are also aware that there is a balance, especially in a world with ever smaller pools of funding, between replicating existing studies and attempting to develop new ideas. Thus, we argue that the value of replication will depend strongly on the type of study considered. For instance, within the class of mechanistic studies, it is rarely (and perhaps never) the case that a single design is the only way to test a given mechanism.

As a pertinent example from a different domain, consider the "facial feedback" hypothesis. In brief, this hypothesis holds that individuals use their own facial expressions as a cue to their current emotional state. One classic investigation of this hypothesis involved asking participants to hold a pen either in their teeth (forcing many facial muscles into positions consistent with a smile) or their lips (prohibiting many facial muscles from taking positions consistent with a smile). An initial study using this approach produced results consistent with the facial feedback hypothesis (greater positive affect when the pen was held in the teeth; Strack et al. 1988). Yet, multiple attempted replications largely failed to find the same results (Acosta et al. 2016).

First, this is an interesting exemplar given the distinction between "direct" and "conceptual" replication above. While one could argue that the attempted replications were framed as being direct replications, changes were made to the original procedures that may have substantively altered the results (e.g., the presence of video recording equipment; Noah et al. 2018). Yet, even if the failed replications were truly direct, would null results falsify the "facial feedback" hypothesis? We would suggest that such results would not falsify the broader hypothesis. Indeed, the pen procedure is just one of many possible ways to test a particular mechanism of action. For example, recent work in the field has strongly indicated that facial expressions should be treated as trajectories rather than end points (i.e., it is not just the final facial expression that matters, the full set of movements that gave rise to the final expression matter). If pen procedure does not effectively mimic the full muscle trajectory of a smile, it might not be the best possible test of the broader theory.

Therefore, such a large-scale replication of the pen procedure—one of many possible intervention designs—provides limited evidence for or against the posited mechanism. It instead largely provides information about the given intervention (which can easily steer the field in the wrong direction; Rotello et al. 2015). It is unequivocally the case that our understanding of the links between tasks and mechanisms

is often weaker than we would like. Given this, we suggest that, in the case of mechanistic studies, there will often be more value in studies that are "extensions," which can provide converging or diverging evidence regarding the mechanism of action, rather than in direct replications.

Conversely, the value of replication in the case of efficacy and effectiveness studies is high. In these types of studies, the critical questions are strongly linked to a single well-defined intervention. There is thus considerable value in garnering additional evidence about that very intervention.

Best Practices when Publishing

In many cases, the best practices for publishing in the domain of behavioral interventions for cognitive enhancement mirror those that have been the focus of myriad recent commentaries within the broader field of psychology (e.g., a better demarcation between analyses that are planned and those that were exploratory). Here, we primarily speak to issues that are either unique to our domain or where best practices may differ by study type.

In general, there are two mechanisms for bias in publishing that must be discussed. The first is publication bias (also known as the "file drawer problem"; Coburn and Vevea 2015). This encompasses, among other things, the tendency for authors to be more likely to submit for publication studies with positive results (most often positive results that confirm their hypotheses, but also including positive, but unexpected results), while failing to submit studies with negative or more equivocal results. It also includes the related tendency for reviewers and/or journal editors to be less likely to accept studies that show non-significant or null outcomes. The other bias is p-hacking (Head et al. 2015). This is when a study collects many outcomes and only the statistically significant outcomes are reported. Obviously, if only positive outcomes are published, it will result in a severely biased picture of the state of the field.

Importantly, the increasing recognition of the problems associated with publication bias has apparently increased the receptiveness of journals, editors, and reviewers toward accepting properly powered and methodologically sound null results. One solution to these publication bias and p-hacking problems is to rely less on p values when reporting findings in publications (Barry et al. 2016; Sullivan and Feinn 2012). Effect size measures provide information on the size of the effect in standardized form that can be compared across studies. In randomized experiments with continuous outcomes, Hedges' g is typically reported (a version of Cohen's d that is unbiased even with small samples); this focuses on changes in standard deviation units. This focus is particularly important in the case of feasibility studies and often also mechanistic studies, which often lack statistical power (see Pek and Flora 2017 for more discussion related to reporting effect sizes). Best practice in these studies is to report the effect sizes and

p values for all comparisons made, not just those that are significant or that make the strongest argument. We also note that this practice of full reporting applies also to alternative methods to quantify statistical evidence, such as the recently proposed Bayes factors (Morey et al. 2016; Rouder et al. 2009). It would further be of value in cases where the dependent variables of interest were aggregates (e.g., via dimensionality reduction) to provide at least descriptive statistics for all variables and not just the aggregates.

An additional suggestion to combat the negative impact of selective reporting is preregistration of studies (Nosek et al. 2017). Here, researchers disclose, prior to the study's start, the full study design that will be conducted. Critically, this includes pre-specifying the confirmatory and exploratory outcomes and/or analyses. The authors are then obligated, at the study's conclusion, to report the full set of results (be those results positive, negative, or null). We believe there is strong value for preregistration both of study design and analyses in the case of efficacy and effectiveness studies where claims of real-world impact would be made. This includes full reporting of all outcome variables (as such studies often include sizable task batteries resulting in elevated concerns regarding the potential for Type I errors). In this, there would also potentially be value in having a third party curate the findings for different interventions and populations and provide overviews of important issues (e.g., as is the case for the Cochrane reviews of medical findings).

The final suggestion is an echo of our previous recommendations to use more precise language when describing interventions and results. In particular, here, we note the need to avoid making overstatements regarding real-world outcomes (particularly in the case of feasibility and mechanistic studies). We also note the need to take responsibility for dissuading hyperbole when speaking to journalists or funders about research results. Although obviously scientists cannot perfectly control how research is presented in the popular media, it is possible to encourage better practices. Describing the intent and results of research, as well as the scope of interpretation, with clarity, precision, and restraint will serve to inspire greater confidence in the field.

Need for Future Research

While the best practices with regard to many methodological issues seem clear, there remain a host of areas where there is simply insufficient knowledge to render recommendations.

The Many Uncertainties Surrounding Expectation Effects

We believe our field should strive to meet the standard currently set by the medical community with regard to blinding

and placebo control (particularly as there is no evidence indicating that expectation effects are larger in cognitive training than in the study of any other intervention). Even if it is impossible to create interventions and control conditions that are perceptually identical (as can be accomplished in the case of an active pill and an inert placebo pill), it is possible to create control conditions that participants find plausible as an intervention. However, we also believe that we, as a field, have the potential to go beyond standards set by the medical community. This may arise, for instance, via additional research on potential benefits from placebo effects. Indeed, the practice of entirely avoiding expectation-based effects may remove an incredibly powerful intervention component from our arsenal (Kaptchuk and Miller 2015).

At present, empirical data addressing whether purely expectation-based effects are capable of driving gains within behavioral interventions for cognitive enhancement is extremely limited and what data does exist is mixed. While one short-term study indicated that expectation effects may have an impact on measured cognitive performance (Foroughi et al. 2016), other studies have failed to find an impact of either expectations or deliberately induced motivational factors (Katz et al. 2018; Tsai et al. 2018).

Although a number of critiques have indicated the need for the field to better measure and/or control for expectation effects, these critiques have not always recognized the difficulties and uncertainties associated with doing so. More importantly, indirect evidence suggests that such effects could serve as important potential mechanisms for inducing cognitive enhancement if they were purposefully harnessed. For instance, there is work suggesting a link between a variety of psychological states that could be susceptible to influence via expectation (e.g., beliefs about self-efficacy) and positive cognitive outcomes (Dweck 2006). Furthermore, there is a long literature in psychology delineating and describing various "participant reactivity effects" or "demand characteristics," which are changes in participant behavior that occur due to the participants' beliefs about or awareness of the experimental conditions (Nichols and Maner 2008; Orne 1962). Critically, many sub-types of participant reactivity result in enhanced performance (e.g., the Pygmalion effect, wherein participants increase performance so as to match high expectations; Rosenthal and Jacobson 1968).

There is thus great need for experimental work examining the key questions of how to manipulate expectations about cognitive abilities effectively and whether such manipulations produce significant and sustainable changes in these abilities (e.g., if effects of expectations are found, it will be critical to dissociate expectation effects that lead to better test-taking from those that lead to brain plasticity). In this endeavor, we can take lessons from other domains where placebo effects have not only been explored, but have begun to be purposefully harnessed, as in the literature on pain (see also the

literature on psychotherapy; Kirsch 2005). Critically, studies in this vein have drawn an important distinction between two mechanisms that underlie expectation and/or placebo effects. One mechanism is through direct, verbal information given to participants; the other one is learned by participants, via conditioning, and appears even more powerful in its impact on behavior (Colloca and Benedetti 2006; Colloca et al. 2013).

Consider, for instance, a study examining the effectiveness of a placebo cream in reducing pain experienced when a high-temperature probe is applied to the skin (Voudouris et al. 1985). In this study, participants first rate their level of pain when the high-temperature probe is applied at a setting of 80 out of 100—sufficient to produce moderate-to-elevated levels of pain. The placebo cream is then applied to the skin. The participant is given the explicit verbal description of the cream as an analgesic that should diminish the level of pain that is experienced. The high-temperature probe is then reapplied to the skin at a setting of 80 out of 100 (as before) and the participant again rates the level of pain. If the rated level of pain is reduced after the application of the placebo cream, this would be taken as evidence for a verbally expectation-induced placebo effect.

In order to induce the conditioning version of the placebo effect, an additional step is inserted in the process above. Namely, after the application of the cream and the description of the cream as an analgesic, participants are told that they will be given the same temperature stimulus as previously, but in fact are given a lower temperature (e.g., a 60 out of 100—one that will naturally produce noticeably less pain). This, theoretically, should provide evidence to the participant of the cream's effectiveness. In other words, it creates an association between the cream and reduced pain. The final step then proceeds as above (with the probe applied at the same 80 out of 100 temperature). If the experienced pain in this version of the study is less than in the verbal instruction version, it is taken as evidence for an additional benefit of conditioning-induced expectations.

In practice, when these two methods of inducing expectations have been contrasted, conditioning-based expectations have typically been found to produce significantly larger and more reliable placebo effects. Similar studies, examining not just the impact of verbal instruction in inducing beliefs, but that of conditioning, would be of significant value in our domain (Colloca et al. 2013).

Finally, although much work needs to be done to determine if expectation-based mechanisms can provide enduring benefits for cognitive skills, it is worth discussing one major concern about the use of placebos in the real-world—namely the potentially negative impact of broken blinding. In other words, if expectation-based mechanisms are used in behavioral interventions for cognitive enhancement, will the benefits that arise from such mechanisms disappear immediately if participants are made aware that some aspect of the training

was meant to induce "placebo like" effects? While this is an open question in need of investigation in our domain, there is again reason for optimism based upon the persistence of benefits observed after unblinding in some medical research. Indeed, there are research studies in the domain of pain perception (as well as various clinical disorders) in which participants are unblinded at study onset—expressly told that they are taking a placebo—and yet the benefits typically associated with placebos are nonetheless still observed, particularly when a prior conditioning mechanism is already in place (Carvalho et al. 2016; Kaptchuk et al. 2010; Kelley et al. 2012; Sandler and Bodfish 2008).

Beyond the need for more research on how to potentially harness expectation effects, there would also be real value in more research on how to best measure whether such expectations arise in existing studies (Rubin 2016). Given such measurements, it would then be possible to determine whether expectations, if present, impact the observed results. Indeed, while placebo effects (if they exist in our domain) could potentially be useful tools for efficacy or effectiveness studies to harness, they are often unwanted potential confounds in mechanistic studies (i.e., in any case where the mechanism of interest is not related explicitly to placebo effects).

Unfortunately, there simply is not, at the moment, a gold standard for making measurements of expectations in the realm of behavioral interventions for cognitive enhancement. Instead, there are a host of critical open questions:

For instance, *will participants be truthful when they are asked about their expectations? If not, how can they be encouraged to be more truthful?* This is a concern because there are an enormous number of instances in the broader psychological literature where participants have been seen to give less than truthful responses about their beliefs or expectations about a hypothesis.

Do participants have the capacity to explain their expectations in such a way that the expectations can be coded and used to potentially explain variance in outcome? The types of expectations that participants may be able to explicitly verbalize may be too vague or too unlinked to the true hypothesis space (e.g., that they will "get better" or "be smarter") to assess their potential impact.

When is the proper time to elicit participant expectations? If expectations are elicited prior to the commencement of training, it is possible that this act will serve to directly produce expectations that might otherwise have not existed. If expectations are elicited after the conclusion of training, it is possible that the expectations will not reflect the beliefs that were held during the training itself (with the added possibility that the expectations changed fluidly throughout training).

Given the huge number of unknowns regarding how to best accomplish the measurement of expectations, more research is clearly needed. We would thus suggest that researchers conducting all types of studies measure expectations, perhaps in concert with partners from domains of psychology more well-versed in such assessments.

Future Issues to Consider with Regard to Assessments

One key future consideration is the relative lack of measures that are known to be reliable indicators of real-world functional performance. One possibility going forward is that the explosion of new forms of wearable technology could potentially provide a host of reliable, valid, and scalable dependent variables of real-world cognition that can also be tracked over long time periods (e.g., ecological momentary assessments; Shiffman et al. 2008). A second issue is the general lack of knowledge regarding the reliability of many key measures currently used in the field. While there have been wide scale efforts to develop and disseminate measures with known and high reliability (Beaumont et al. 2013; Hendershot et al. 2015; Hodes, Insel, Landis,, and Research 2013; Weintraub et al. 2013), it is nonetheless quite common for different research groups to make use of adapted versions of established tasks (where the consequences of the given manipulations in terms of measurement reliability are often not known or described) or else completely different tasks. Furthermore, there are a host of cases where the reliability of certain tasks is known to be rather poor, but these tasks are nonetheless commonly utilized (which is a particular concern in the case of null results, as a null result using an unreliable measure has an increased probability of being a Type II error).

Another set of open questions is related to the composition of pre/post-test batteries. Many of these questions arise due to tension between the desire to utilize multiple tasks with at least partially overlapping processing demands and concern regarding the impact of ever-larger task batteries. The former is required to evaluate performance at the level of latent variables, which has a number of benefits (Engle et al. 1999). Indeed, it has been repeatedly noted in psychology that even our simplest tasks are unlikely to be truly "process pure." Thus, if our theories and associated hypotheses are at the level of cognitive constructs, it would be sensible for our measurements to similarly be made at that level. This necessitates a sufficiently large and varied battery of tasks to allow these constructs to be extracted. Furthermore, individual tasks can be more or less similar to trained tasks (meaning that transfer to those tasks could be construed as "nearer" or "farther" even if the tasks are posited to tap the same basic skills). These issues are reduced (though perhaps not eliminated) by utilizing multiple partially overlapping measures of the construct of interest. In this though, it is important to consider the risk associated with defining constructs that are too general, thus concealing more specific effects. For example, lumping together verbal and visuo-spatial working memory tasks in one "working memory factor" might prevent finding an effect that is specific for visuo-spatial working memory tasks. The

same is true of a "creativity factor," as convergent creativity tasks, such as the Remote Associates Test (where participants must find the word that links three given words—e.g., cottage, Swiss, cake), are sometimes affected by different factors than divergent creativity tasks, such as the Alternative Uses Test (where participants are asked to come up with as many uses as possible for a single item—e.g., a brick; Hommel et al. 2011). This would suggest that tasks or measures should be combined in such a way as to measure the smallest possible part or process of a construct.
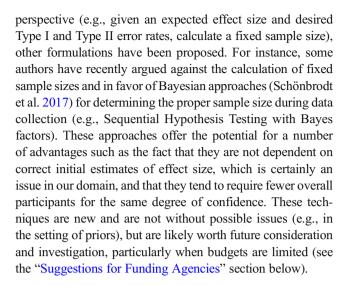
However, while the benefits of larger task batteries are myriad, there are potential downsides as well. For instance, a possible confound of larger task batteries is the very real potential for participant burden and motivational or cognitive fatigue (Holtzer et al. 2011). Indeed, some tasks are known to be mentally tiring. If participants are forced to perform many of these tasks in sequence, their performance on each may be substantially lower than it would have been if the tasks were performed alone. This is a particularly large concern given that fatigue effects may be disproportionately large in many populations that would be of most interest to researchers in cognitive training (e.g., children, individuals with deficits, etc.; Bryant and Deluca 2004). Furthermore, the possibility of fatigue effects may create another confounding mechanism through which certain types of training may create benefits (i.e., that benefits may arise due to a reduction in experienced cognitive fatigue rather than "true" enhancements).

Other issues with ever-larger task batteries include a host of possible temporal sequential dependencies between tasks. These include "carry-over" effects (e.g., in a case where Task A encourages participants to adopt the strategy to respond quicker sacrificing accuracy and this carries over to Task B, where that strategy is sub-optimal). They also include learning effects. Indeed, if a task battery contains many tests of a single construct, it is possible that the testing will act as de facto training of the construct. Whether such effects indeed occur and, if so, how they should (or should not) influence our understanding of the effects of an intervention, will be a matter for future work to decide.

Finally, a last potential issue with ever-larger task batteries is the multiple comparisons problem that they often introduce. Larger numbers of dependent variables mean more potential for Type I errors and thus care must be taken in this regard (e.g., either by reducing the variable space, by making clear what analyses were associated with a priori predictions and which were exploratory in nature, by correcting for multiple comparisons, etc.)

Future Issues to Consider with Regard to Sample Size and Statistical Power

With regard to statistical power, while we have focused on the need for properly powered designs from a more traditional perspective (e.g., given an expected effect size and desired Type I and Type II error rates, calculate a fixed sample size), other formulations have been proposed. For instance, some authors have recently argued against the calculation of fixed sample sizes and in favor of Bayesian approaches (Schönbrodt et al. 2017) for determining the proper sample size during data collection (e.g., Sequential Hypothesis Testing with Bayes factors). These approaches offer the potential for a number of advantages such as the fact that they are not dependent on correct initial estimates of effect size, which is certainly an issue in our domain, and that they tend to require fewer overall participants for the same degree of confidence. These techniques are new and are not without possible issues (e.g., in the setting of priors), but are likely worth future consideration and investigation, particularly when budgets are limited (see the "Suggestions for Funding Agencies" section below).

Discussion

Suggestions for Funding Agencies

As noted numerous times above, many of the proposed best practices are expensive—considerably more so than currently common methodologies. For example, double-blinding, in the most rigorous scenario, requires separate personnel to administer pre/post-tests and to administer training (and perhaps even separate personnel to administer intervention training and control training so as to further diminish the possibility of experimenter bias). Budgets for studies that utilize these best practices will grow accordingly. The same is true for studies that recruit larger sample sizes. For instance, our suggestion to include a "business as usual" control for all mechanistic and efficacy studies (in addition to any other active control groups) will significantly increase sample sizes. Finally, the assessment of possible sleeper or protective or other longitudinal effects will require not just larger budgets, but more stable funding. Indeed, there are few current funding mechanisms that can handle the measurement of effects that may only be seen several years post-intervention. Given the scope and scale of the possible positive impacts that would be realized via convincingly demonstrated behavioral interventions for cognitive enhancement (e.g., within the USA, this would contribute to the missions of the NIA, NCI, NEI, NIMH, NIAAA, NINCD, NIDA, and NINDS, in that all of these agencies work with populations who can gain direct benefits from successful approaches to mental fitness), as well as the opportunity costs that might be paid if less robust (and less expensive) methods are employed, we feel strongly that an increase in funding to accommodate best practice studies is of the utmost importance.

One major issue with regard to funding agencies within the USA regards the sweeping new definitions of "clinical trials" now in use at the National Institutes of Health (Health 2014).

In essence, under these new definitions, all four types of studies would likely fall under the definition of a clinical trial. Indeed, under this broad new definition, an "intervention" means "a manipulation of the subject or subject's environment for the purpose of modifying one or more health-related biomedical or behavioral processes and/or endpoints." This is clearly the case in all four types of studies discussed above, as well as a host of "basic science" studies.

While we enthusiastically share the NIH goals of increasing data transparency and accountability, we believe it is critical for studies to be treated in a manner consistent with their goals and methods. Treating all studies as an intervention, irrespective of the study's goals and methodology, runs directly counter to this position. The negative consequences of such a broad classification include confusion within the scientific population and within the general public. For example, such a classification scheme would require that a naive individual from the general public be capable of reading a full study and understanding that even though the study is identified by the NIH as a "successful clinical trial," it is in fact a feasibility study examining whether older adults can learn to use a joystick in a given video game environment has limited-to-no "clinical" value. Additionally, there are numerous and obvious negative consequences to the progress of research that are associated with this additional administrative burden (Wolfe and Kanwisher 2018).

Suggestions for Regulatory Bodies

The huge surge in publicly available products sold with the promise of increasing cognitive health has resulted in an increasingly urgent discussion as to whether or which governmental agencies should regulate this industry and what standards should be utilized. In the USA, there are two main federal regulatory bodies of interest—the Food and Drug Administration (FDA) and the Federal Trade Commission (FTC).

The purview of the FDA (at least as it relates to the current topic) largely involves evaluating the safety and efficacy of drugs and devices that make claims regarding diagnosing, curing, or treating disease. As such, a cognitive training paradigm marketed for treating a specific medical condition, such as ADHD, would fall under the oversight of the FDA. However, if the exact same paradigm were marked as "enhancing" normal cognition, it likely would not fall under the purview of the FDA. Indeed, in a recent FDA guidance document (which contains nonbinding recommendations), the Agency stated that it did not intend to regulate low-risk products (such as mobile applications) marketed for general wellness purposes (e.g., improving mental acuity or concentration). This, as well as a similar document from 2015 regarding mobile medical health applications, seems to shift the regulatory burden for low-risk enhancement products to the FTC, which is tasked with protecting consumers from unfair or deceptive business practices, such as misleading advertising. Indeed, in recent years, the FTC has acted against a number of companies marketing products that claim to enhance cognition ("FEDERAL TRADE COMMISSION, Plaintiff, v. LUMOS LABS, INC., a corporation d/b/a LUMOSITY and KUNAL SARKAR, and MICHAEL SCANLON, Individually and as officers of LUMOS LABS, INC. Defendants. ," 2016).

Our core suggestion here is that the public will be best served by closer communication between scientists, funding agencies, and regulators so as to hone regulatory guidelines and ensure that they are based upon the evidence and true overall current state of the field. For example, as this paper documents, there remain a host of open questions related to placebo effects and how to best control for them in the domain of behavioral interventions for cognitive enhancement. Therefore, caution and careful consideration is needed before directly utilizing, or otherwise adapting, standards derived from other domains to the domain of behavioral interventions for cognitive enhancement. This is particularly true when those outside domains have vastly different histories, open questions, and methodological affordances (e.g., as is true of the standards developed in the context of clinical trials for pharmaceuticals). Our hope is thus that this manuscript will not only be directly considered by individuals positioned to make decisions in the relevant funding and regulatory agencies, but can also be used as a bridge for scientists when approaching such agencies.

A Call for Greater Interdisciplinary Collaboration

Many of the issues we have considered are not unique to our field and will best be addressed through an expansion of interdisciplinary research networks. Some research domains have already faced key challenges we identify here and thus can offer guidance based upon their best practices. Other research domains are currently facing some of the same key challenges and thus could offer partnerships to address those issues broadly. For instance, there are already pockets of expertise regarding how to best induce and optimally harness expectation effects within the medical community (e.g., those who treat pain, anxiety, or depression) as well as within social psychology (e.g., those who study motivation, achievement, and identity contingencies). Below, we highlight those areas where a common body of knowledge and set of best practices across disciplines would benefit all, as well as indicating various unique challenges.
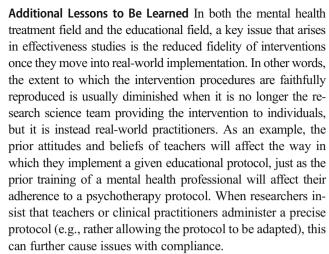
Greater Precision of Description The tendency for a wide range of quite disparate procedures to be lumped under a common moniker is certainly not unique to the cognitive training field. For example, there are a wide variety of practices that have traditionally been subsumed under the term "mindfulness training"; yet, recent evidence provides empirical support for differential behavioral and neural prophylaxis for

various types of meditative practices (Davidson and Dahl 2017; Lutz et al. 2015). The same basic issue is equally pertinent when considering the treatment of mental disorders (e.g., a huge range of techniques and practices fall under the label of "behavioral therapy" or even "medication management"), when discussing the impact of exercise on cognitive health (e.g., aerobic exercise is not the same as resistance exercise such as weightlifting and there are even important distinctions within those sub-categories), and within the realm of educational interventions (e.g., as pertains to the question of whether there are benefits to "video games" in the classroom). In each case, it is important to use a more nuanced terminology that clearly describes the distinct, specific mental and physical processes being harnessed by the given intervention (for an example from contemplative science, see Davidson and Dahl 2018; Van Dam et al. 2018). Partnerships with these fields could elucidate more general principles for the correct units of aggregation to test broader hypotheses.

Control Group Selection and Assignment Many of the issues discussed above relate to the appropriate selection of control conditions. For instance, how might one create different conditions (e.g., one active and one control) while still blinding participants to experimenter intent? Here, lessons might be learned by fields that have grappled with this same issue. For instance, in studies on the effects of aerobic exercise on cognitive health, common control conditions involve plausible potential benefits, but do not harness the systemic changes hypothesized to be critical to the effects of aerobic exercise (e.g., toning and stretching which are largely non-aerobic; Erickson et al. 2011). Other fields may broadly find themselves in the same general position as our own. For instance, in education, new interventions have most commonly been contrasted against status quo instruction. Yet, simple novelty (e.g., doing something other than the status quo curriculum) is nearly always associated with a significant increase in student achievement (Bayraktar 2001). These issues have sometimes been addressed in education (as well as many other domains) via crossover designs (where the groups receive both the active and control experiences, but in different orders; Hills and Armitage 2004). This design type though is far less common in the cognitive training domain (noting though that this design also has its own set of drawbacks).

Issues related to participant assignment are also shared broadly across many research domains. For instance, truly random assignment is often impossible in educational and clinical contexts. Children are not randomly assigned to classrooms or school districts. Patients are not randomly assigned to practitioners or health-care facilities. Instead, a host of demographic and individual-difference level characteristics are nearly always correlated with such factors. How to effectively deal with these issues is another area where interdisciplinary partnerships would be of clear value.

Additional Lessons to Be Learned In both the mental health treatment field and the educational field, a key issue that arises in effectiveness studies is the reduced fidelity of interventions once they move into real-world implementation. In other words, the extent to which the intervention procedures are faithfully reproduced is usually diminished when it is no longer the research science team providing the intervention to individuals, but it is instead real-world practitioners. As an example, the prior attitudes and beliefs of teachers will affect the way in which they implement a given educational protocol, just as the prior training of a mental health professional will affect their adherence to a psychotherapy protocol. When researchers insist that teachers or clinical practitioners administer a precise protocol (e.g., rather allowing the protocol to be adapted), this can further cause issues with compliance.

A common paradigm used by educational scientists to deal with low fidelity is referred to as "designed-based" research (Sandoval and Bell 2004). In this paradigm, researchers first start with a study that identifies a learning pathway. The goal of this study is to reveal an active ingredient in learning; that is, to identify the causal agent that produces enhanced learning. This study is then followed up by iterative application studies, in which teachers apply the curriculum several times, adapting it to their particular classroom. Researchers study these iterations to examine how the protocol evolves with individual teacher use. Finally, after the curriculum has been implemented several times, researchers return to a broad study of effectiveness across classrooms, determining whether the benefit of the intervention persists. Indeed, design-based paradigms allow researchers to know how implementation of curriculum evolves, and whether the efficacy of the intervention changes with this evolution.

Another issue that arises when behavioral interventions are transitioned from the research lab into real-world implementation is how to choose outcomes that are meaningful to the individual participant, rather than outcomes that are based on research measures which may or may not fit in with each individual person's real-world goals. Goal-attainment scaling (GAS) is a study design method from mental health services research that allows the researcher to evaluate individually meaningful effectiveness across different kinds of interventions. GAS is used to scale each individual's goals for an intervention, permitting for a valid assessment of individual or group outcomes.

In all, given that true effectiveness studies have been quite rare in the domain of behavioral interventions for cognitive health, greater partnerships with domains that have faced the challenges associated with this type of study would likely be beneficial.

Additional Points for Consideration

There Is Not Necessarily a Linear Progression from Feasibility to Effectiveness Although we have discussed study types in

order—from feasibility, to mechanistic, to efficacy, to effectiveness—we by no means suggest that there is a linear, or even ordered, trajectory of studies that must always be followed. It is much more useful to think of these study types (as well as a host of others) as an interconnected web, with each type of study potentially informing every other study type. For instance, interesting and unexpected participant responses made to a specific component of a full-scale efficacy study might indicate new untested mechanisms of action that could be explored via a mechanistic study. A product that is already in wide public use might spur researchers to begin with an effectiveness study, which, if successful, would then spawn mechanistic studies to understand the mechanisms through which the intervention acts. Or an effectiveness study conducted in younger adults might indicate the value of a feasibility study assessing whether the same intervention could be used in older adults.

There Are Many Other Study Types and Designs beyond those Discussed Here The four study types we have considered in detail do not remotely represent the entire set of possible study types that will prove informative for our field. Other study types may focus, for instance, on narrow questions related to underlying neuroplastic processes or seek to delineate how individual differences (from differences in experience to differences in genetics) impact outcomes. As such, they may use quite different methods (e.g., purely correlational within-group designs, observational longitudinal studies, latent change modeling, etc.). Furthermore, some fruitful approaches may combine one of the four main study types discussed here with some of these alternative methods. For instance, the medical domain has increasingly moved toward more individualized interventional approaches (e.g., cancer therapies that are based upon an individual's genetic makeup, rather than one-size fits all brute force approaches). Assessing the extent to which inter-individual differences predictably mediate the outcome of behavioral interventions for cognitive enhancement, and thus whether more individualized interventions can be developed, might require a mechanistic, efficacy, or effectiveness study that is combined with an individual-differences correlational design.

Additional authors, particularly in the translational clinical field, have discussed the critical need for studies that might occur after efficacy and effectiveness trials. For instance, the stage model proposed by the NIH discusses the need for studies that target implementation and dissemination (i.e., determining what steps and/or procedures are most useful in ensuring that scientifically validated interventions come into common use among real-world practitioners; Onken et al. 2014). As our field progresses, we will need to determine methods for training professional staff to appropriately apply and evaluate cognitive enhancement interventions, especially in clinical populations.

Limitations of the RCT Model Randomized controlled trials have long been the gold standard for evaluating the efficacy of interventions and thus they have been the focus of this paper. However, the RCT model contains several inherent limitations (Kaptchuk 2001), particularly as it relates to our field. As a prime example, current digital technologies which are used in current cognitive training research are rapidly evolving over very short timeframes. In the midst of this rapidly shifting technological landscape, traditional RCT research designs may be less practical, useful and feasible since they require that an intervention remain stable across the evaluation period (Mohr et al. 2015). The need for long-term stability may be crippling for interventions that rely on rapidly evolving technologies, their usage patterns, and their accessibility. Indeed, if digital tools are forced to remain calcified for the 3–5 years that are usually required to conduct a high-quality RCT, potentially critical opportunities to improve outcomes could be missed. The most appropriate solution to this dilemma is far from clear. Instead, how best to deal with this reality is something that the field will need to address. For example, this may take the form of moving toward a model that evaluates an intervention based on its mechanistic principles, target definition and target engagement, and clinical usage outcomes—rather than surface features of the graphics or design or the specific technology or platform on which it is delivered.

Finally, the field of cognitive enhancement may very well be one in which the impact of many (or even most) interventions are modest and might reflect multiple interacting factors in addition to the intervention itself (diet, sleep, exercise, specific learning techniques, specific ways to present materials, specific mindsets, specific motivational techniques, two-generation interventions, etc.). Thus, studies that evaluate each specific intervention in isolation via an RCT may not yield robust results. However, studies that involve combinations of interventions plus "enhancing factors" may be impractical and unwieldy, as well as make it impossible to separate those factors that produce real benefits from those that do not. In addition, it is probable that in some instances, the greatest real-world generalization will occur when cognitive enhancing interventions are sequenced with other psychosocial or vocational treatments, or when lower-level cognitive training is followed by higher-level metacognitive interventions. Sequential multi-arm randomized trials are one approach to these questions. In particular, the possibility that large-scale improvements are only achievable via the cumulative contribution of many small effects is something that the field will need to consider when assessing otherwise "null" or "not clinically relevant" effects.

Conclusions

While many of the methodological recommendations we have made in previous sections can be found scattered in the

literature, we explicitly focus here on their application as an ensemble of best practices for the study of behavioral interventions for cognitive enhancement. We also commit, as a group of 48 scientists, to be cognizant of these recommendations in our research studies to the best of our abilities and resources. While many of the suggestions will require only minor adjustments from researchers, we believe that they will have serious implications for the field's ability to move forward.

Box 1. Key Recommendations for Researchers

Communication/dissemination/publishing
- Avoid unhelpfully broad category labels such as "brain training"
- Make the study goals, and thus the type of study (e.g., feasibility, mechanistic, efficacy, effectiveness), explicit from the outset
- Focus on (and report) effect sizes, not just p-values
- Fully report statistics, not just the results consistent with hypothesis
- Avoid hyperbole in language
  - If a study was not designed to address translational goals, discuss readiness to assess translation rather than give full-fledged nods to real-world application

Control-group selection and assignment
- Proper control group(s) depend on study goals/type
- Mechanistic and efficacy studies would benefit from consistent inclusion of a passive control in addition to relevant active control groups
- In many cases, alternative methods of group assignment are likely preferable to truly random methods

Assessments
- A greater emphasis on real-world measures would be useful in efficacy and effectiveness studies
- Assessment batteries would benefit from the inclusion of tasks explicitly not expected to be altered by the intervention(s), rather than only tasks that are (potentially) expected to be altered
- Long-term follow-up assessments would have definite value (e.g., in detecting sleeper effects not seen immediately after training).

Blinding
- The field has the capability not only to meet, but to exceed the double-blind standard in the medical field
- Participant blinding: Participants cannot be blinded to the content of their training, but can be blinded to intention/hypotheses
  - Every effort should be taken to make 'inert' control conditions plausible as an active intervention (i.e., blind participants to intention)
- Experimenter blinding: Separate personnel, blinded to participant condition, can be used to administer pre-test and post-test assessments

Replication
- Direct replication has value in the case of efficacy & effectiveness studies
- Systematic replication (i.e., "extensions") may have greater value for mechanistic studies

---

Box 2. Key Recommendations for Journal Editors, Funding Agencies, and Regulatory Bodies

Journal Editors

- Recognize that all types of studies (e.g., feasibility, mechanistic, efficacy, and effectiveness), if well-executed, have value in our field
- Increased receptiveness toward accepting properly powered and methodologically sound null results

Funding Agencies

- Many of the recommendations here are more expensive to implement than currently common methods (e.g., double-blinding requires separate personnel for administering pre-test/post-test assessments and for administering training – perhaps even different personnel for administering different training conditions)
- Many of the recommendations require more stable funding sources than are currently typical (e.g., the ability to perform follow-up assessments to examine sleeper effects several years after the end of training)
- Studies should be treated in a manner commensurate with their goals and methods
  - Treating all studies in this domain as interventions will likely have a number of negative consequences

Regulatory Bodies

- The public will be best served by closer communication between scientists, funding agencies, and regulators so as to hone regulatory guidelines and ensure that they are based upon the evidence and true overall current state of the field

---

Box 3. Future Directions

Uncertainties Around Expectation Effects
- Research is needed to examine whether expectation effects can in fact influence intervention outcomes in this domain
  - If so, research is needed on how to best harness these for good
- Research is needed to how best to assess expectations in this domain
  - How should questions be worded? How will truthfulness be promoted? When should expectations be assessed?

Assessments
- Creation/validation of reliable measures of real-world impact are deeply needed
- Research needed on how/whether the assessment battery composition affects outcomes
  - This will allow the development of best-practices for the various types of studies

Partnerships with related domains
- Many domains share all or most of the methodological challenges in our field
- These include control task selection, assignment to groups, assessment batteries, etc.
- Cues can be taken from related fields that have already moved forward in addressing some issues
  - Several areas have sought to purposefully harness participant expectations (pain, certain mental disorders)
  - Several areas have considered how to create plausible, but inert control groups (e.g., aerobic exercise + stretching/toning or sub-aerobic exercise)

Compliance with Ethical Standards

Conflict of Interest The following authors have declared conflict(s) of interest. Bavelier is a founding partner and on the scientific advisory board of Akili Interactive, Boston; Vinogradov is a consultant for Posit Science Corp, Alkermes, Inc., and Mindstrong, Inc.; Ball owns stock in the Visual Awareness Research Group (formerly Visual Awareness, Inc.) and Posit Science, Inc., the companies that market the Useful Field of View Test and speed of processing training software (now the Double Decision exercise in BrainHQ), and is a member of the Posit Science Scientific Advisory Board; Gazzaley is a co-founder, scientific advisor, and BOD member for Akili Interactive Lab and has several patents filed at UCSF for video game enhancement technologies; Jaeggi has an indirect financial interest in the MIND Research Institute, Irvine, CA, whose interests are related to this work; Levi is a member of the Scientific Advisory Board of NovaSight; Morris is on the scientific advisory boards of Neurotrack and of the AARP Global Council on Brain Health; Nahum

is a paid consultant for Posit Science; Panizzutti is the founder of NeuroForma LTDA, a company with a financial interest in computerized cognitive training; Seitz is a founder and stakeholder in Carrot Neurotechnology, a company that sells a vision brain game called ULTIMEYES. The other authors declare that they have no conflict of interest.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Acosta, A., Adams, R.B., Jr., Albohn, D.N., Allard, E.S., Beek, T., Benning, S. D., … Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11(6), 917–928. doi:https://doi.org/10.1177/1745691616674458.

Andrews, G. (1999). Efficacy, effectiveness and efficiency in mental health service delivery. *The Australian and New Zealand Journal*

of Psychiatry, 33(3), 316–322. https://doi.org/10.1046/j.1440-1614.1999.00581.x.

Anguera, J.A., Boccanfuso, J., Rintoul, J.L., Al-Hashimi, O., Faraji, F., Janowich, J., … Gazzaley, A. (2013). Video game training enhances cognitive control in older adults. *Nature, 501*(7465), 97–101. doi:https://doi.org/10.1038/nature12486.

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: a meta-analysis. *Psychonomic Bulletin & Review, 22*(2), 366–377. https://doi.org/10.3758/s13423-014-0699-x.

Ball, K., Berch, D. B., Helmers, K. F., Jobe, J. B., Leveck, M. D., Marsiske, M., … Group, A. S. (2002). Effects of cognitive training interventions with older adults: a randomized controlled trial. JAMA, 288(18), 2271–2281.

Baniqued, P., Allen, C.M., Kranz, M.B., Johnson, K., Sipolins, A., Dickens, C., …, Kramer, A.F. (2015). Working memory, reasoning, and task switching training: transfer effects, limitations, and great expectations?. PLoS One. doi:https://doi.org/10.1371/journal.pone.0142169.

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.

Barry, A. E., Szucs, L. E., Reyes, J. V., Ji, Q., Wilson, K. L., & Thompson, B. (2016). Failure to report effect sizes: the handling of quantitative results in published health education and behavior research. *Health Education & Behavior, 43*(5), 518–527. https://doi.org/10.1177/1090198116669521.

Basak, C., Boot, W. R., Voss, M. W., & Kramer, A. F. (2008). Can training in a real-time strategy video game attenuate cognitive decline in older adults. *Psychology and Aging, 23*(4), 765–777.

Bavelier, D., & Davidson, R. J. (2013). Brain training: games to do you good. *Nature, 494*(7438), 425–426. https://doi.org/10.1038/494425a.

Bayraktar, S. (2001). A meta-analysis of the effectiveness of computer-assisted instruction in science education. *Journal of Research on Technology in Education, 34*(2), 173–188.

Beaumont, J. L., Havlik, R., Cook, K. F., Hays, R. D., Wallner-Allen, K., Korper, S. P., …, Gershon, R. (2013). Norming plans for the NIH toolbox. *Neurology, 80*(11 Suppl 3), S87–92. doi:https://doi.org/10.1212/WNL.0b013e3182872e70.

Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Green, C. S., & Bavelier, D. (2018). Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. *Psychological Bulletin, 144*(1), 77–110. https://doi.org/10.1037/bul0000130.

Biagianti, B., & Vinogradov, S. (2013). Computerized cognitive training targeting brain plasticity in schizophrenia. *Progress in Brain Research, 207*, 301–326. https://doi.org/10.1016/B978-0-444-63327-9.00011-4.

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. https://doi.org/10.1177/1745691613491271.

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

Carvalho, C., Caetano, J. M., Cunha, L., Rebouta, P., Kaptchuk, T. J., & Kirsch, I. (2016). Open-label placebo treatment in chronic low back pain: a randomized controlled trial. *Pain, 157*(12), 2766–2772. https://doi.org/10.1097/j.pain.0000000000000700.

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

Coburn, K. M., & Vevea, J. L. (2015). Publication bias as a function of study characteristics. *Psychological Methods, 20*(3), 310–330. https://doi.org/10.1037/met0000046.

Colloca, L., & Benedetti, F. (2006). How prior experience shapes placebo analgesia. *Pain, 124*(1–2), 126–133. https://doi.org/10.1016/j.pain.2006.04.005.

Colloca, L., Klinger, R., Flor, H., & Bingel, U. (2013). Placebo analgesia: psychological and neurobiological mechanisms. *Pain, 154*(4), 511–514. https://doi.org/10.1016/j.pain.2013.02.002.

Colzato, L. S., van den Wildenberg, W. P., & Hommel, B. (2014). Cognitive control and the COMT Val(1)(5)(8)Met polymorphism: genetic modulation of videogame training and transfer to task-switching efficiency. *Psychological Research, 78*(5), 670–678. https://doi.org/10.1007/s00426-013-0514-8.

Davidson, R. J., & Dahl, C. J. (2017). Varieties of contemplative practice. *JAMA Psychiatry, 74*(2), 121–123. https://doi.org/10.1001/jamapsychiatry.2016.3469.

Davidson, R. J., & Dahl, C. J. (2018). Outstanding challenges in scientific research on mindfulness and meditation. *Perspectives on Psychological Science, 13*(1), 62–65. https://doi.org/10.1177/1745691617718358.

Deveau, J., Jaeggi, S. M., Zordan, V., Phung, C., & Seitz, A. R. (2014a). How to build better memory training games. *Frontiers in Systems Neuroscience, 8*, 243. https://doi.org/10.3389/fnsys.2014.00243.

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. https://doi.org/10.1016/j.cub.2014.01.004.

Diao, D., Wright, J. M., Cundiff, D. K., & Gueyffier, F. (2012). Pharmacotherapy for mild hypertension. *Cochrane Database of Systematic Reviews*(8). doi:https://doi.org/10.1002/14651858.CD006742.pub2.

Dweck, C. (2006). *Mindset: the new psychology of success*. New York: Random House.

Eldridge, S. M., Lancaster, G. A., Campbell, M. J., Thabane, L., Hopewell, S., Coleman, C. L., & Bond, C. M. (2016). Defining feasibility and pilot studies in preparation for randomised controlled trials: development of a conceptual framework. *PLoS One, 11*(3), e0150205. https://doi.org/10.1371/journal.pone.0150205.

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

Erickson, K. I., Voss, M. W., Prakash, R. S., Basak, C., Szabo, A., Chaddock, L., et al. (2011). Exercise training increases size of hippocampus and improves memory. *Proceedings of the National Academy of Sciences of the United States of America, 108*(7), 3017–3022. https://doi.org/10.1073/pnas.1015950108.

Fassler, M., Meissner, K., Kleijnen, J., Hrobjartsson, A., & Linde, K. (2015). A systematic review found no consistent difference in effect between more and less intensive placebo interventions. *Journal of Clinical Epidemiology, 68*(4), 442–451. https://doi.org/10.1016/j.jclinepi.2014.11.018.

FEDERAL TRADE COMMISSION, (2016) Plaintiff, V.. Lumos Labs, Inc., a corporation d/b/a Lumosity and Kunal Sarkar, and Michael Scanlon, Individually and as officers of LumoS Labs, Inc. Defendants. , No. Case No. 3:16-cv-00001-sk (United StateS DistricT Court for the Northern District of California, San Francisco Division.

Fergusson, D., Glass, K. C., Waring, D., & Shapiro, S. (2004). Turning a blind eye: the success of blinding reported in a random sample of randomised placebo controlled trials. *BMJ, 328*(7437), 432. https://doi.org/10.1136/bmj.37952.631667.EE.

Foroughi, C. K., Monfort, S. S., Paczynski, M., McKnight, P. E., & Greenwood, P. M. (2016). Placebo effects in cognitive training. *PNAS, 113*, 7470–7474. https://doi.org/10.1073/pnas.1601243113.

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. https://doi.org/10.1016/j.cub.2013.01.044.

Fritz, J. M., & Cleland, J. (2003). Effectiveness versus efficacy: more than a debate over language. *The Journal of Orthopaedic and Sports Physical Therapy, 33*(4), 163–165. https://doi.org/10.2519/jospt.2003.33.4.163.

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

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.

Green, C. S., Strobach, T., & Schubert, T. (2014). On methodological standards in training and transfer experiments. *Psychological Research, 78*(6), 756–772. https://doi.org/10.1007/s00426-013-0535-3.

Greitemeyer, T., Osswald, S., & Brauer, M. (2010). Playing prosocial video games increases empathy and decreases schadenfreude. *Emotion, 10*(6), 796–802. https://doi.org/10.1037/a0020194.

Hallock, H., Collins, D., Lampit, A., Deol, K., Fleming, J., & Valenzuela, M. (2016). Cognitive training for post-acute traumatic brain injury: a systematic review and meta-analysis. *Frontiers in Human Neuroscience, 10*, 537. https://doi.org/10.3389/fnhum.2016.00537.

Hawes, Z., Moss, J., Caswell, B., Naqvi, S., & MacKinnon, S. (2017). Enhancing children's spatial and numerical skills through a dynamic spatial approach to early geometry instruction: effects of a 32-week intervention. *Cognition and Instruction, 35*(3), 236–264.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology, 13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106.

Health, N. I. o. (2014). Notice of Revised NIH Definition of "Clinical Trial". (NOT-OD-15-015). Retrieved from https://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-015.html. Accessed 1 Nov 2018.

Hendershot, T., Pan, H., Haines, J., Harlan, W. R., Marazita, M. L., McCarty, C. A., …, Hamilton, C. M.. (2015). Using the PhenX toolkit to add standard measures to a study. Curr Protoc Hum Genet, 86, 1 21 21–17. https://doi.org/10.1002/0471142905.hg0121s86.

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.

Hills, M., & Armitage, P. (2004). The two-period cross-over clinical trial. 1979. *British Journal of Clinical Pharmacology, 58*(7), S703–S716; discussion S717-709. https://doi.org/10.1111/j.1365-2125.2004.02275.x.

Hodes, R. J., Insel, T. R., Landis, S. C., & Research, N. I. H. B. f. N. (2013). The NIH toolbox: setting a standard for biomedical research. *Neurology, 80*(11 Suppl 3), S1. https://doi.org/10.1212/WNL.0b013e3182872e90.

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. https://doi.org/10.1080/13825585.2010.517826.

Hommel, B., Colzato, L. S., Fischer, R., & Christoffels, I. K. (2011). Bilingualism and creativity: benefits in convergent thinking come with losses in divergent thinking. *Frontiers in Psychology, 2*, 273. https://doi.org/10.3389/fpsyg.2011.00273.

Howard, J. (2016). Do brain-training exercises really work? *CNN*. Retrieved from https://www.cnn.com/2016/10/20/health/brain-training-exercises/index.html. Accessed 1 Nov 2018.

Hrobjartsson, A., Forfang, E., Haahr, M. T., Als-Nielsen, B., & Brorson, S. (2007). Blinded trials taken to the test: an analysis of randomized clinical trials that report tests for the success of blinding. *International Journal of Epidemiology, 36*(3), 654–663. https://doi.org/10.1093/ije/dym020.

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 of the United States of America, 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 of the United States of America, 108*, 10081–10086.

Jones, R. N., Marsiske, M., Ball, K., Rebok, G., Willis, S. L., Morris, J. N., & Tennstedt, S. L. (2013). The ACTIVE cognitive training interventions and trajectories of performance among older adults. *Journal of Aging and Health, 25*(8 Suppl), 186S–208S. https://doi.org/10.1177/0898264312461938.

Kaptchuk, T. J. (2001). The double-blind, randomized, placebo-controlled trial: gold standard or golden calf? *Journal of Clinical Epidemiology, 54*(6), 541–549.

Kaptchuk, T. J., & Miller, F. G. (2015). Placebo effects in medicine. *The New England Journal of Medicine, 373*(1), 8–9. https://doi.org/10.1056/NEJMp1504023.

Kaptchuk, T. J., Friedlander, E., Kelley, J. M., Sanchez, M. N., Kokkotou, E., Singer, J. P., et al. (2010). Placebos without deception: a randomized controlled trial in irritable bowel syndrome. *PLoS One, 5*(12), e15591. https://doi.org/10.1371/journal.pone.0015591.

Karbach, J., & Unger, K. (2014). Executive control training from middle childhood to adolescence. *Frontiers in Psychology, 5*, 390. https://doi.org/10.3389/fpsyg.2014.00390.

Katz, B., Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (2018). The effect of monetary compensation on cognitive training outcomes. Learning and Motivation, 63(77–90).

Kelley, J. M., Kaptchuk, T. J., Cusin, C., Lipkin, S., & Fava, M. (2012). Open-label placebo for major depressive disorder: a pilot randomized controlled trial. *Psychotherapy and Psychosomatics, 81*(5), 312–314. https://doi.org/10.1159/000337053.

Kirsch, I. (2005). Placebo psychotherapy: synonym or oxymoron? *Journal of Clinical Psychology, 61*(7), 791–803. https://doi.org/10.1002/jclp.20126.

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. *J Am Acad Child Adolesc Psychiatry, 44*(2), 177–186.

Kolahi, J., Bang, H., & Park, J. (2009). Towards a proposal for assessment of blinding success in clinical trials: up-to-date review. *Community Dentistry and Oral Epidemiology, 37*(6), 477–484. https://doi.org/10.1111/j.1600-0528.2009.00494.x.

Kramer, A. F., Larish, J., & Strayer, D. L. (1995). Training for attentional control in dual-task settings: a comparison of young and old adults. *Journal of Experimental Psychology: Applied, 1*, 50–76.

Lutz, A., Jha, A. P., Dunne, J. D., & Saron, C. D. (2015). Investigating the phenomenological matrix of mindfulness-related practices from a neurocognitive perspective. *The American Psychologist, 70*(7), 632–658. https://doi.org/10.1037/a0039585.

Marchand, E., Stice, E., Rohde, P., & Becker, C. B. (2011). Moving from efficacy to effectiveness trials in prevention research. *Behaviour Research and Therapy, 49*(1), 32–41. https://doi.org/10.1016/j.brat.2010.10.008.

Mayer, R. E. (Ed.). (2014). *Computer games for learning: an evidence-based approach*. Cambridge: MIT Press.

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

Merzenich, M. M., Nahum, M., & Van Vleet, T. M. (2013). Neuroplasticity: introduction. *Progress in Brain Research, 207*, xxi–xxvi. https://doi.org/10.1016/B978-0-444-63327-9.10000-1.

Mohr, D. C., Schueller, S. M., Riley, W. T., Brown, C. H., Cuijpers, P., Duan, N., et al. (2015). Trials of intervention principles: evaluation methods for evolving behavioral intervention technologies. *Journal of Medical Internet Research, 17*(7), e166. https://doi.org/10.2196/jmir.4391.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology, 72*, 6–18.

Nahum, M., Lee, H., & Merzenich, M. M. (2013). Principles of neuroplasticity-based rehabilitation. *Progress in Brain Research, 207*, 141–171. https://doi.org/10.1016/B978-0-444-63327-9.00009-6.

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

Noah, T., Yaacov, S., & Mayo, R. (2018). When both the original study and its failed replication are correct: feeling observed eliminates the facial-feedback effect. *Journal of Personality and Social Psychology, 114*(5), 657–664.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2017). The preregistration revolution. OSF Preprints. doi:https://doi.org/10.17605/OSF.IO/2DXU5.

O'Leary, K. D., Rosenbaum, A., & Hughes, P. C. (1978). Direct and systematic replication: a rejoinder. *Journal of Abnormal Child Psychology, 6*(3), 295–297.

Onken, L. S., Carroll, K. M., Shoham, V., Cuthbert, B. N., & Riddle, M. (2014). Reenvisioning clinical science: unifying the discipline to improve the public health. *Clinical Psychological Science: A Journal of the Association for Psychological Science, 2*(1), 22–34. https://doi.org/10.1177/2167702613497932.

Open Science, C. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*(6), 657–660. https://doi.org/10.1177/1745691612462588.

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.

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.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*(6), 531–536. https://doi.org/10.1177/1745691612463401.

Pek, J., & Flora, D. B. (2017). Reporting effect sizes in original psychological research: a discussion and tutorial. Psychological Methods. doi:https://doi.org/10.1037/met0000126.

Prakash, R. S., De Leon, A. A., Patterson, B., Schirda, B. L., & Janssen, A. L. (2014). Mindfulness and the aging brain: a proposed paradigm shift. *Frontiers in Aging Neuroscience, 6*, 120. https://doi.org/10.3389/fnagi.2014.00120.

Rebok, G. W., Ball, K., Guey, L. T., Jones, R. N., Kim, H. Y., King, J. W., … Group, A. S. (2014). Ten-year effects of the advanced cognitive training for independent and vital elderly cognitive training trial on cognition and everyday functioning in older adults. Journal of the American Geriatrics Society, 62(1), 16–24. doi:https://doi.org/10.1111/jgs.12607.

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.

Redick, T. S., Shipstead, Z., Wiemers, E. A., Melby-Lervag, M., & Hulme, C. (2015). What's working in working memory training? An educational perspective. *Educational Psychology Review, 27*(4), 617–633. https://doi.org/10.1007/s10648-015-9314-6.

Roberts, G., Quach, J., Spencer-Smith, M., Anderson, P. J., Gathercole, S., Gold, L., … Wake, M. (2016). Academic outcomes 2 years after working memory training for children with low working memory: a randomized clinical trial. JAMA Pediatrics, 170(5), e154568. doi: https://doi.org/10.1001/jamapediatrics.2015.4568.

Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence, 35*(1), 83–92.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: teacher expectation and pupils' intellectual development*. New York: Holt, Rinehart and Winston.

Ross, L. A., Edwards, J. D., O'Connor, M. L., Ball, K. K., Wadley, V. G., & Vance, D. E. (2016). The transfer of cognitive speed of processing training to older adults' driving mobility across 5 years. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences, 71*(1), 87–97. https://doi.org/10.1093/geronb/gbv022.

Rotello, C. M., Heit, E., & Dube, C. (2015). When more data steer us wrong: replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review, 22*(4), 944–954. https://doi.org/10.3758/s13423-014-0759-2.

Rothbaum, B. O., Price, M., Jovanovic, T., Norrholm, S. D., Gerardi, M., Dunlop, B., … Ressler, K. J. (2014). A randomized, double-blind evaluation of D-cycloserine or alprazolam combined with virtual reality exposure therapy for posttraumatic stress disorder in Iraq and Afghanistan War veterans. The American Journal of Psychiatry, 171(6), 640–648. doi:https://doi.org/10.1176/appi.ajp.2014.13121625.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225.

Rubin, M. (2016). The Perceived Awareness of the Research Hypothesis Scale: assessing the influence of demand characteristics. In.

Rutherford, B. R., Sneed, J. R., & Roose, S. P. (2009). Does study design influence outcome?. The effects of placebo control and treatment duration in antidepressant trials. *Psychotherapy and Psychosomatics, 78*(3), 172–181. https://doi.org/10.1159/000209348.

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

Sandler, A. D., & Bodfish, J. W. (2008). Open-label use of placebos in the treatment of ADHD: a pilot study. *Child: Care, Health and Development, 34*(1), 104–110. https://doi.org/10.1111/j.1365-2214.2007.00797.x.

Sandoval, W. A., & Bell, P. (2004). Design-based research methods for studying learning in context: introduction. *Educational Psychologist, 39*(4), 199–201.

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.

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

Schönbrodt, F. D., Wagenmakers, E. J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: efficiently testing mean differences. *Psychological Methods, 22*(2), 322–339.

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.

Schulz, K. F., Chalmers, I., & Altman, D. G. (2002). The landscape and lexicon of blinding in randomized trials. *Annals of Internal Medicine, 136*(3), 254–259.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology, 4*, 1–32.

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

Sidman, M. (1966). *Tactics of scientific research: evaluating experimental data in psychology.* Oxford: Basic Books.

Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. (2016). Do "Brain-Training" programs work? *Psychological Science in the Public Interest, 17*(3), 103–186. https://doi.org/10.1177/1529100616661983.

Singal, A. G., Higgins, P. D., & Waljee, A. K. (2014). A primer on effectiveness and efficacy trials. *Clinical Translational Gastroenterology, 5*, e45. https://doi.org/10.1038/ctg.2013.13.

Smith, G. E., Housen, P., Yaffe, K., Ruff, R., Kennison, R. F., Mahncke, H. W., & Zelinski, E. M. (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.

Stieff, M., & Uttal, D. (2015). How much can spatial training improve STEM achievement. *Educational Psychology Review, 27*(4), 607–615.

Stierlin, A. S., Herder, K., Helmbrecht, M. J., Prinz, S., Walendzik, J., Holzmann, M., … Kilian, R. (2014). Effectiveness and efficiency of integrated mental health care programmes in Germany: study protocol of an observational controlled trial. BMC Psychiatry, 14, 163. doi:https://doi.org/10.1186/1471-244X-14-163.

Strack, F., Martin, L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology, 54*(5), 768–777.

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

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

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*(1), 59–71. https://doi.org/10.1177/1745691613514450.

Subramaniam, K., Luks, T. L., Garrett, C., Chung, C., Fisher, M., Nagarajan, S., & Vinogradov, S. (2014). Intensive cognitive training in schizophrenia enhances working memory and associated prefrontal cortical efficiency in a manner that drives long-term functional gains. *NeuroImage, 99*, 281–292. https://doi.org/10.1016/j.neuroimage.2014.05.057.

Sullivan, G. M., & Feinn, R. (2012). Using effect size-or why the P value is not enough. *Journal of Graduate Medical Education, 4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1.

Tang, Y. Y., Ma, Y., Wang, J., Fan, Y., Feng, S., Lu, Q., et al. (2007). Short-term meditation training improves attention and self-regulation. *Proceedings of the National Academy of Sciences of the United States of America, 104*(43), 17152–17156. https://doi.org/10.1073/pnas.0707678104.

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

Terlecki, M. S., Newcombe, N. S., & Little, M. (2008). Durable and generalized effects of spatial experience on mental rotation: gender differences in growth patterns. *Applied Cognitive Psychology, 22*, 996–1013.

Tickle-Degnen, L. (2013). Nuts and bolts of conducting feasibility studies. *The American Journal of Occupational Therapy, 67*(2), 171–176. https://doi.org/10.5014/ajot.2013.006270.

Tsai, N., Buschkuehl, M., Kamarsu, S., Shah, P., Jonides, J., & Jaeggi, S. M. (2018). (Un)Great expectations: the role of placebo effects in cognitive training. *Journal of Applied Research in Memory and Cognition*, 7(4), 564–573. https://doi.org/10.1016/j.jarmac.2018.06.001.

Valdes, E.G., Andel, R., Lister, J.J., Gamaldo, A., & Edwards, J.D. (2017). Can cognitive speed of processing training improve everyday functioning among older adults with psychometrically defined mild cognitive impairment? Journal of Aging and Health, 898264317738828. https://doi.org/10.1177/0898264317738828.

Van Dam, N. T., van Vugt, M. K., Vago, D. R., Schmalzl, L., Saron, C. D., Olendzki, A., et al. (2018). Mind the hype: a critical evaluation and prescriptive agenda for research on mindfulness and meditation. *Perspectives on Psychological Science, 13*(1), 36–61. https://doi.org/10.1177/1745691617709589.

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. https://doi.org/10.3389/fnagi.2010.00032.

Voudouris, N. J., Peck, C. L., & Coleman, G. (1985). Conditioned placebo responses. *Journal of Personality and Social Psychology, 48*(1), 47–53.

Walton, A.G. (2016). Do brain training games work, or is it the placebo effect? *Forbes.com*. Retrieved from https://www.forbes.com/sites/alicegwalton/2016/06/21/does-brain-training-work-or-is-it-all-placebo/#3b654dc67497. Accessed 1 Nov 2018.

Weintraub, S., Dikmen, S.S., Heaton, R.K., Tulsky, D.S., Zelazo, P.D., Bauer, P.J., … Gershon, R.C. (2013). Cognition assessment using the NIH toolbox. Neurology, 80(11 Suppl 3), S54–64. https://doi.org/10.1212/WNL.0b013e3182872ded.

Wexler, B.E., Iseli, M., Leon, S., Zaggle, W., Rush, C., Goodman, A., … Bo, E. (2016). Cognitive priming and cognitive training: immediate and far transfer to academic skills in children. Scientific Reports, 6, 32859. doi:https://doi.org/10.1038/srep32859.

Whitehead, A. L., Sully, B. G., & Campbell, M. J. (2014). Pilot and feasibility studies: is there a difference from each other and from a randomised controlled trial? *Contemporary Clinical Trials, 38*(1), 130–133. https://doi.org/10.1016/j.cct.2014.04.001.

Whitlock, L. A., McLaughlin, A. C., & Allaire, J. C. (2012). Individual differences in response to cognitive training: using a multi-modal, attentionally demanding game-based intervention for older adults. *Computers in Human Behavior, 28*(4), 1091–1096.

Willis, S.L., Tennstedt, S.L., Marsiske, M., Ball, K., Elias, J., Koepke, K.M., … Group, A.S. (2006). Long-term effects of cognitive training on everyday functional outcomes in older adults. JAMA, 296(23), 2805–2814.

Wolfe, J. M., & Kanwisher, N. G. (2018). Not your parent's NIH clinical trial. *Nature Human Behaviour, 2*, 107–109.

Wright, R., Thompson, W. L., Ganis, G., Newcombe, N. S., & Kosslyn, S. M. (2008). Training generalized spatial skills. *Psychonomic Bulletin and Review, 15*(4), 763–771.

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 https://doi.org/10.1177/0962280212436447.

Zwaan, R.A., Etz, A., Lucas, R.E., & Donnellan, M. B. (2017). Making replication mainstream. The Behavioral and Brain Sciences, 1–50. https://doi.org/10.1017/S0140525X17001972.

Affiliations

C. Shawn Green[1] · Daphne Bavelier[2] · Arthur F. Kramer[3,4] · Sophia Vinogradov[5,6,7] · Ulrich Ansorge[8] · Karlene K. Ball[9] · Ulrike Bingel[10] · Jason M. Chein[11] · Lorenza S. Colzato[12,13] · Jerri D. Edwards[14] · Andrea Facoetti[15] · Adam Gazzaley[,16] · Susan E. Gathercole[17] · Paolo Ghisletta[18,19,20] · Simone Gori[21] · Isabela Granic[22] · Charles H. Hillman[23] · Bernhard Hommel[24] · Susanne M. Jaeggi[25] · Philipp Kanske[26] · Julia Karbach[27,28] · Alan Kingstone[29] · Matthias Kliegel[30] · Torkel Klingberg[31] · Simone Kühn[32] · Dennis M. Levi[33] · Richard E. Mayer[34] · Anne Collins McLaughlin[35] · Danielle S. McNamara[36] · Martha Clare Morris[37] · Mor Nahum[38] · Nora S. Newcombe[39] · Rogerio Panizzutti[40,41] · Ruchika Shaurya Prakash[42] · Albert Rizzo[43] · Torsten Schubert[44] · Aaron R. Seitz[45] · Sarah J. Short[46] · Ilina Singh[47] · James D. Slotta[48] · Tilo Strobach[49] · Michael S. C. Thomas[50] · Elizabeth Tipton[51,52] · Xin Tong[53] · Haley A. Vlach[54] · Julie Loebach Wetherell[55,56] · Anna Wexler[57] · Claudia M. Witt[58]

[1] Department of Psychology, University of Wisconsin, Madison, WI, USA

[2] Psychology and Education Sciences (FPSE) & Campus Biotech, University of Geneva, Geneva, Switzerland

[3] Department of Psychology, Northeastern University, Boston, MA, USA

[4] Beckman Institute, University of Illinois at Urbana-Champaign, Champaign, IL, USA

[5] Department of Psychiatry, University of Minnesota, Minneapolis, MN, USA

[6] University of Minnesota Medical School, Minneapolis, MN, USA

[7] University of Minnesota Physicians, Minneapolis, MN, USA

[8] Faculty of Psychology, University of Vienna, Vienna, Austria

[9] Center for Research on Applied Gerontology & Department of Psychology, University of Alabama at Birmingham, Birmingham, AL, USA

[10] Department of Neurology, University Hospital Essen, University Duisburg-Essen, Duisburg, Germany

[11] Department of Psychology, Temple University, Philadelphia, PA, USA

[12] Cognitive Psychology Unit & Leiden Institute for Brain and Cognition, Leiden University, Leiden, Netherlands

[13] Department of Cognitive Psychology, Institute of Cognitive Neuroscience, Faculty of Psychology, Ruhr University Bochum, Bochum, Germany

[14] Department of Psychiatry and Behavioral Neurosciences, University of South Florida, Tampa, FL, USA

[15] Developmental and Cognitive Neuroscience Laboratory, Department of General Psychology, University of Padua, Padua, Italy

[16] Departments of Neurology, Physiology & Psychiatry, University of California, San Francisco, CA, USA

[17] MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, England

[18] Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland

[19] Swiss Distance Learning University, Brig, Switzerland

[20] Swiss National Centre of Competence in Research LIVES – Overcoming vulnerability: Life course perspectives, University of Lausanne, Laussane, Switzerland; University of Geneva, Geneva, Switzerland

[21] Department of Human and Social Sciences, University of Bergamo, Bergamo, Italy

[22] Behavioural Science Institute, Radboud University, Nijmegen, Netherlands

[23] Department of Psychology & Department of Physical Therapy, Movement & Rehabilitation Sciences, Northeastern University, Boston, MA, USA

[24] Institute of Psychology, Leiden University, Leiden, Netherlands

[25] School of Education and School of Social Sciences, Department of Cognitive Sciences, University of California, Irvine, CA, USA

26 Faculty of Psychology, Clinical Psychology and Behavioral Neuroscience, Technische Universität Dresden, Dresden, Germany

27 Department of Psychology, University of Koblenz-Landau, Landau, Germany

28 Center for Research on Individual Development and Adaptive Education of Children at Risk (IDeA), Frankfurt, Germany

29 Department of Psychology, University of British Columbia, Vancouver, Canada

30 Department of Psychology & Center for the Interdisciplinary Study of Gerontology and Vulnerability, University of Geneva, Geneva, Switzerland

31 Department of Neuroscience, Karolinska Institutet, Solna, Sweden

32 Clinic for Psychiatry and Psychotherapy, University Clinic Hamburg-Eppendorf, Hamburg, Germany

33 School of Optometry & Graduate Group in Vision Science, University of California, Berkeley, CA, USA

34 Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA, USA

35 Department of Psychology, North Carolina State University, Raleigh, NC, USA

36 Department of Psychology, Arizona State University, Tempe, AZ, USA

37 Rush Institute for Healthy Aging, Rush University, Chicago, IL, USA

38 School of Occupational Therapy, Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel

39 Department of Psychology, Temple University, Philadelphia, PA, USA

40 Instituto de Ciencias Biomedicas & Instituto de Psiquiatria, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

41 Global Brain Health Institute, Trinity College Dublin, Dublin, Ireland

42 Department of Psychology, The Ohio State University, Columbus, OH, USA

43 Institute for Creative Technologies, University of Southern California, Los Angeles, CA, USA

44 Institute of Psychology, Martin-Luther-University Halle-Wittenberg, Halle, Germany

45 Department of Psychology & Brain Game Center, University of California, Riverside, CA, USA

46 Department of Educational Psychology & Center for Healthy Minds, University of Wisconsin, Madison, WI, USA

47 Department of Psychiatry and Wellcome Centre for Ethics and the Humanities, University of Oxford, Oxford, UK

48 Department of Curriculum, Teaching and Learning, Boston College, Chestnut Hill, USA

49 Department of Psychology, Medical School Hamburg, Hamburg, Germany

50 Centre for Educational Neuroscience, Department of Psychological Sciences, Birkbeck, University of London, London, UK

51 Teachers College, Columbia University, New York, NY, USA

52 Department of Statistics and Institute for Policy Research, Northwestern University, Evanston, IL, USA

53 Department of Psychology, University of Virginia, Charlottesville, VA, USA

54 Department of Educational Psychology, University of Wisconsin, Madison, WI, USA

55 VA San Diego Healthcare System, San Diego, CA, USA

56 University of California-San Diego, San Diego, CA, USA

57 Department of Medical Ethics & Health Policy, University of Pennsylvania, Philadelphia, PA, USA

58 Institute for Complementary and Integrative Medicine, University Hospital and University of Zurich, Zürich, Switzerland