HHS Public Access

Author manuscript

Biol Psychiatry Cogn Neurosci Neuroimaging. Author manuscript; available in PMC 2019 August 11.

Published in final edited form as:

Biol Psychiatry Cogn Neurosci Neuroimaging. 2018 November ; 3(11): 903–904. doi:10.1016/j.bpsc.2018.09.010.

How to Resolve Controversies in Cognitive Training: Let the Data Speak!

Robert M. Bilder

Departments of Psychiatry and Biobehavioral Sciences and Psychology, UCLA David Geffen School of Medicine and College of Letters and Science, and UCLA Jane & Terry Semel Institute for Neuroscience and Human Behavior, Los Angeles, CA

Keywords

cognition; cognitive remediation; cognitive training; psychometrics; schizophrenia; aging

Controversies in Computerized Testing (1) marks a maturational milestone in our appreciation of computerized cognitive testing (CCT) and cognitive remediation (CR). We appear to have progressed through the natural phases of the “hype cycle” from the “peak of inflated expectations”, through the “trough of disillusionment” into the “plateau of productivity” (2). As with most novel technologies, progress seldom emerges without controversy. Harvey and colleagues’ perspective should help overcome polarized opinions about the efficacy of CCT and CR that have generated competing “position” or “consensus” statements (3, 4), and advance us to a position where we can let the data do the talking.

While this article advocates a positive view of CCT and CR based on a careful consideration of the existing evidence, it makes a major contribution by pinpointing the sources of the controversy and proposing that these center on differing definitions of the interventions and the kinds of efficacy. Sharpening these definitions may offer the most direct path to further progress.

Regarding definitions of intervention, the authors highlight that CCT is usually implemented in the context of cognitive remediation (CR), which incorporates other ingredients like skills training. Some reviews exclude studies of CCT if it is done in a CR context. While more studies are needed to tease apart the active ingredients of these treatments, a comprehensive overview supports the efficacy of CCT whether or not it is embedded in CR. This suggests that relevant data sources should be considered as broadly as possible to enhance generalization of results across a variety of real-world settings. In this case, letting the data speak means assuring that all the data are heard, not selected subsets of the data.

With respect to defining efficacy, Harvey and colleagues point out that the critical 2016 review by Simon et al (5) did not clearly distinguish near, far and environmental transfer effects. Harvey and colleagues argue that CCT alone probably yields near transfer effects,

while CR that incorporates CCT probably yields both near and far transfer effects. But exactly how near is near and how far is far? This focus on definitions again brings us more directly to the underlying data. Consider Simon's and colleagues' concept of transfer, specifically:

"The tasks that did show differential improvements following cognitive training were distinct from the tasks that were used during training, but they tapped some of *the same underlying constructs* [italics added] ... Improvements were limited to these trained domains, suggesting relatively narrow and focused training benefits rather than broad improvements to cognition more generally." (p 135)

First, we must appreciate the core concern that a different measurement of the same or a similar construct still comprises "transfer" according to most definitions, including that of Harvey and colleagues. A second more subtle issue is: Why do Simon and colleagues believe the other measures tap the same underlying construct? This example shows how the use of "domain" and "construct" labels may be misleading; depending on the boundaries chosen by an investigator, transfer effects may be seen as "near," "far," or "null." Given that we possess statistical tools to specify how much overlap there is among measures of different psychological constructs, why should definitions of near and far transfer depend on opinion or consensus rather than empirical data?

The excellent and influential meta-analysis of Wykes and colleagues provides potentially informative examples (6). They estimate a global "primary endpoint" for each study by averaging all the cognitive tests administered, and "cognitive domains" by averaging scores on tests assigned to domains following consensus definitions. They show reasonable assignments of tests to domains, but as the authors point out, "cognitive outcomes reflect processing in several cognitive domains," and others might make different decisions. For example, "continuous performance tests" (CPTs) were considered measures of "attention/vigilance" (following consensus reached in the MATRICS initiative), but other research uses the AX-CPT as a measure of working memory (7) following consensus reached in the RDoC initiative. These comments are not intended to second-guess the assignments made by Wykes and colleagues, but to suggest that ideally, decisions about lumping and splitting should be based on data rather than consensus. The observation of greater heterogeneity across studies for domains that "lump" measures (5 of 7 domains showed significant heterogeneity) compared to individual tests (only 1 of 4 specific tests showed heterogeneity) may reflect that the domains are not ideally specified, and support the value of using individual tests if we lack clear domain-specification.

These psychometric concerns are not just academic side-issues, but are at the heart of determining if there are genuine differences in the outcomes of CCT and CR, and to the extent that there may be differences between tests or well-defined domains, these data provide the key to defining transfer effects. Wykes and colleagues reported effect sizes (Cohen's *d*) for CR effects on "social cognition" ($d=.65$, 95% CI = .33 to .97) and "reasoning and problem solving" ($d=.57$, 95% CI = .22 to .92) that are nominally higher than the effects on "visual learning and memory" ($d=.15$, 95% CI = -.08 to .38), "attention/vigilance" ($d=.25$, 95% CI = .08 to .42), and "speed of processing" ($d=.26$, 95% CI = .07 to .45). Do these data inform our understanding of near/far transfer? Might the more complex

CR contexts yield greater impact on more complex cognitive processes? Is it harder to change more “basic” cognitive functions, or might there be psychometric limitations on these tests that prevent detection of change? If we can better specify exactly what we are measuring we will ultimately get better answers to these questions.

The article by Harvey and colleagues takes a step in the right direction by attempting to clarify and operationalize what we mean by “near” and “far” transfer effects. Their definition of near transfer is relatively clear: improvement on any non-trained task other than trained tasks would qualify. There could still be room for argument, however, if training and testing overlap substantively in process or content. For example, if training involved digit span procedures and testing involved letter span procedures, would that qualify? It might be clearer to have operational criteria that specify the content, process and psychometric commonalities and differences between training and testing paradigms. The proposed definition of far transfer is more complicated: “improvement on cognitively demanding functional tasks.” The authors suggest two kinds of task that would satisfy this criterion. First are measures of “functional capacity,” or tests of the ability to perform functional skills. This is appealing on the surface, but functional capacity tests typically behave like other cognitive tests and there is little psychometric evidence that these are “closer” to real-world outcomes than other cognitive tests. Second are measures of real-world functioning. It is hard to argue that this would not always qualify as far transfer, but is that a bar too high? Might this not lead us to abandon many potentially promising treatments?

Progress in refining definitions of near and far transfer may benefit from a sharper psychometric focus. There are multiple ways forward:

1. Increase use of common data elements: If more studies used shared outcome measures mega- and meta-analyses would be clearer, sources of heterogeneity easier to identify, and potential confounds easier to eliminate.
2. Establish psychometric criteria for domain definitions and boundaries: If investigators wish to declare “domains,” these constructs should be specified in advance and satisfy psychometric standards for reliability, precision and sensitivity across the range of the relevant trait. We could also specify statistical boundaries to quantify what is “near” and what is “far”, based on shared variance or other psychometric distance parameters.
3. Share data openly: burgeoning opportunities for open science contributions, and increased scope of NIMH Data Archives that accommodate cognitive measures, provide the basis for data-pooling and shared analysis.
4. Apply modern psychometric and factor analytic strategies: To determine the commonalities among tests and enable more robust pooling of data across studies, we should use modern psychometric approaches to test-linking, harmonization and alignment, that allow estimation of individual differences on shared traits even though different tests may be used (8).

Advances like these can promote “bottom-up,” data-driven analyses that will inform our understanding of the structure of cognitive dimensions impacted by CCT and CR, define

empirical, quantitative standards for transfer effects, and determine more precisely what are the active ingredients of novel treatments that maximize clinical benefits.

Acknowledgements:

This work was supported by grants from the National Institute of Mental Health (R01MH101478 and U01MH105578).

Disclosures: The author has grant funding from the National Institutes of Health, has received consulting fees from Think Now, Inc., and is a shareholder of JKBC, LLC, a developer of internet-based cognitive assessments.

References

1. Harvey PD, McGurk SR, Mahrcke H, Wykes TJBPCN, Neuroimaging (2018): Controversies in Computerized Cognitive Training
2. Fenn J, LeHong HJG, 7 (2011): Hype cycle for emerging technologies, 2011.
3. Allaire J, Bäckman L, Balota D, Bavelier D, Bjork R, Bower GJMpfhd, et al. (2014): A consensus on the brain training industry from the scientific community
4. <https://www.cognitivetrainingdata.org/the-controversy-does-brain-training-work/response-letter/cognitive-training-data-signatories/> (2016).
5. Simons DJ, Boot WR, Charness N, Gathercole SE, Chabris CF, Hambrick DZ, et al. (2016): Do “Brain-Training” Programs Work? Psychol Sci Public Interest 17:103–186. [PubMed: 27697851]
6. Wykes T, Huddy V, Cellard C, McGurk S, Czobor P (2011): A meta-analysis of cognitive remediation for schizophrenia: methodology and effect sizes. The American journal of psychiatry 168:472–485. [PubMed: 21406461]
7. MacDonald AW (2008): Building a clinically relevant cognitive task: case study of the AX paradigm. Schizophr Bull 34:619. [PubMed: 18487225]
8. Bilder RM, Reise SP (2018, in press): Neuropsychological Tests of the Future: How Do We Get There from Here? The Clinical Neuropsychologist1–26.