Psychon Bull Rev
DOI 10.3758/s13423-016-1196-1

THEORETICAL REVIEW

Effects of cognitive training on the structure of intelligence

John Protzko[1]

Abstract Targeted cognitive training, such as *n*-back or speed of processing training, in the hopes of raising intelligence is of great theoretical and practical importance. The most important theoretical contribution, however, is not about the malleability of intelligence. Instead, I argue the most important and novel theoretical contribution is understanding the causal structure of intelligence. The structure of intelligence, most often taken as a hierarchical factor structure, necessarily prohibits transfer from subfactors back up to intelligence. If this is the true structure, targeted cognitive training interventions will fail to increase intelligence not because intelligence is immutable, but simply because there is no causal connection between, say, working memory and intelligence. Seeing the structure of intelligence for what it is, a causal measurement model, allows us to focus testing on the presence and absence of causal links. If we can increase subfactors without transfer to other facets, we may be confirming the correct causal structure more than testing malleability. Such a blending into experimental psychometrics is a strong theoretical pursuit.

Keywords Psychometrics/testing · Cognitive training

A lot of attention has been directed toward raising intelligence through targeted cognitive training interventions. These interventions involve training a specific component of intelligence, most often processing speed (PS) or working memory (WM). The goal of much of this research in nonclinical populations is to enhance intellectual life. Here, I argue, there is a second equally important outcome of this research: it allows us to *test* the structure of human cognitive ability.

This article is organized as follows: First, I present a sampling of models regarding the structure of intelligence. Each of these models represent top-down processes (as is standard in much intelligence research). Then, I describe the logic of causality found within structural equation models, with a small diversion to some additional assumptions of latent variable modeling. Following from these two points, we immediately see that a top-down causal structure makes upward causation from subfactors (like working memory) to general intelligence impossible. I then introduce targeted cognitive training, interventions that target distinct cognitive processes for improvement. I argue that these targeted interventions allow one to test the causal assumptions made in hierarchical and nonhierarchical models of the structure of human cognitive abilities. Failures to find transfer may be indicative of the correct causal direction between intelligence and its subprocesses.

Intelligence and its structure

Intelligence, broadly defined, may be

> a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings—"catching on," "making sense" of things, or "figuring out" what to do. (Gottfredson, 1997, p. 13).

✉ John Protzko
protzko@gmail.com

[1] Psychological and Brain Sciences, University of California Santa Barbara, Santa Barbara, CA 93106, USA

Published online: 14 November 2016

This conceptual definition carries with it a measurement model relating observed and latent variables (e.g., Bollen, 1989). Specifically, it defines the measurement of intelligence as a single latent variable that is the reason disparate cognitive tests and/or latent subfactors are correlated with one another. Whether a single latent variable model or a bifactor model or a hierarchical model, the reason tests of cognitive ability correlate with one another is because they are measuring the same underlying thing: intelligence (Jensen, 1998).

Different models, however, alter how we understand the *structure* of intelligence. A single-factor model of intelligence (see Fig. 1, top left) poses that all cognitive tests, from vocabulary to reaction time to math skills, correlate with one another because they all are measures of intelligence and little else.

A hierarchical factor model (see Fig. 1b) acknowledges that, for example, vocabulary tests and tests of general knowledge and verbal reasoning tasks correlate with one another because they are all measuring verbal ability. This could be the same with other latent variables such as processing speed, inductive ability, and memory. Those latent variables, however, correlate with one another. The proposed reason they correlate with one another is that they are measuring intelligence. A bifactor model (see Fig. 1c) allows for the latent subfactors, but proposes that the reason the factors correlate with one another is because the cognitive tests are still measuring intelligence *on top of the latent factor* (e.g., verbal ability).[1]

The number of subfactors and specific relations do not matter for the measurement of intelligence, however. The correlation of a hierarchical *g* factor across different batteries and different tests range from .83 to 1—effective unity (Jensen, 1998; Johnson, Bouchard, Krueger, McGue, & Gottesman, 2004; Thorndike, 1987). The different structures are all measuring the same thing.

Investigations into the psychometric structure of intelligence generally rely on testing different models and comparing model fit statistics. As a simple concrete example, one study compared different structures of intelligence on the same batteries of tests to see which fit best. Some of the structures included a Gf/Gc nonhierarchical model (see Fig. 2a), a hierarchical verbal/perceptual model (Fig. 2b), a three-strata hierarchical multifactor model (Fig. 2c), and a hierarchical multifactorial four-strata model (Fig. 2d). It was the final model that provided the best fit to the data (Johnson & Bouchard, 2005).

Based off this analysis, the authors concluded that the structure of intelligence is verbal, perceptual, and image rotation ability (VPR).[2]

[1] There are, of course, other nonhierarchical models of test score patterns that do not have a single intelligence factor. These models often involved correlated latent traits. While in this article I only discuss intelligence models, the causal arguments throughout apply to all latent variable models, not just hierarchical ones.

[2] All hierarchical models share the same basic assumptions. I present only the VPR model for simplicity, as I do not wish to engage in the debate about "which" model may best fit the data. For another, more expansive hierarchical model, see the CHC model (McGrew, 2009).

Model fit is one way to test competing models, but it does not ensure that we identify the correct model. What is more important to the discussion is given the same number of variables and parameters, the direction of the arrows has little to no effect on model fit. Our models of intelligence can be entirely misspecified because, for example, verbal ability is causally related upwards to intelligence. This would not likely show up in model fit statistics, but may instead be discovered through experimental research.

Causality in latent-variable models

These modeling approaches toward understanding what the structure of intelligence is involve structural equation models with directed arrows. These models are, by their very nature, causal models that reflect what we take to be traits that exist in the world independent of their measures (e.g. Borsboom, Mellenbergh, & Van Heerden, 2003; Pearl, 2009). So every arrow in a given model of the structure of intelligence is a strong causal claim about the effects of, say, a latent variable *causing* responses on another factor or test.

Equally important is all of the causal arrows that are absent. Under no model are there arrows going upward in the hierarchy (e.g., no arrows from verbal ability to *g*), and no arrows going directly or indirectly from one latent variable to another (e.g., processing speed and executive functioning are causally unrelated and independent). These are all strong causal claims that are inferred from covariance patterns and model fit statistics (e.g., Hausman & Woodward, 1999).

Consider another causal graph. The probability of rain and readings on a barometer are correlated—the higher the barometric reading, the higher the probability of rain. This is because they both share a common cause—atmospheric pressure. I can make this causal connection explicit in Fig. 3.

Figure 3 represents how atmospheric pressure causes both rain and changes in barometer readings (indicated by the arrows). If we were to artificially change the reading on a barometer (putting it in a pressure-cooker, for example), it would not alter the probability of rain. This fact is noted by the absence of any arrows from *barometer reading* to *rain*. It is important to note that there is no difference between the causal claims being made in Fig. 3 versus those being made in Figs. 1 and 2.

This creates a problem given hierarchical models of intelligence. If the models in Figs. 1 and 2 are indeed correct in that they represent the true causal structure of intelligence and its subfactors, then every intervention that attempts to raise a cognitive subtrait, be it working memory, processing speed, verbal ability, will *necessarily* show no transfer whatsoever to intelligence. This is because there is no causality from any test or subfactor going back to *g*.

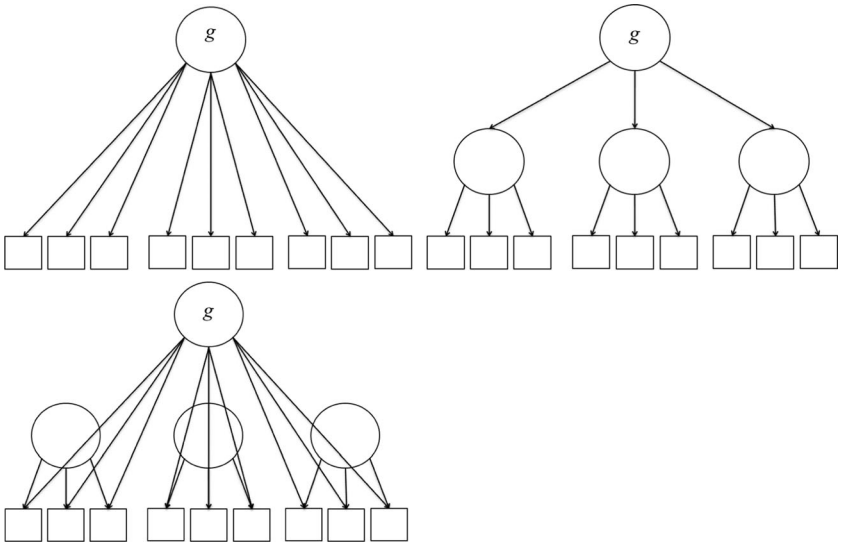

Fig. 1 Three different models of the structure of intelligence: unified (**a**), hierarchical (**b**) and bifactor (**c**)

Take a simplified version of the VPR model in Fig. 4. This model of intelligence makes the strong causal claim that if one were to increase verbal ability, we would see no increases in *g*, in any other latent variable, or in any other test that is not causally related to verbal ability. The absence of causal arrows in this structure are claims to those facts. Increases in verbal ability will show resulting increases only in those tests that are causally connected to verbal ability, *and nothing else*. I argue that it is exactly the presence and absence of causal connections that allow for experimental testing, as well as the strongest theoretical contribution of targeted cognitive training.

Between- and within-subjects modeling

Before delving into targeted cognitive training, I must also briefly introduce another technical assumption of latent variable modeling. Investigations into the structure of intelligence typically rely on between-subjects modeling; it tests the patterns of intercorrelations between variables across persons in cross-sectional data. It may be that this also reflects within-subjects' processes, meaning the causal relations are retained within an individual. Between- and within-subjects models, however, are not *necessarily* the same (e.g., Borsboom et al., 2003).

The critical difference of between- and within-subjects models regards the interpretation of the probabilistic nature of item responses (Borsboom et al., 2003; Holland, 1990). For example, to answer why an individual answers a given problem correctly or incorrectly, in the between-subjects interpretation, out of all people with a given ability *X*, choosing one randomly is the probability that we will choose one who will get the problem right. So for a simple math problem, *the probability that a smart person will get it right is high* means that if we get a group of people with high ability (*X*) together, choosing one randomly to answer the simple math problem means we have a high probability of choosing someone who will get it right. In the within-subjects interpretation, it means that, for any given person, it is not guaranteed that they will get a problem right or wrong. A high ability (*X*) person who knows the answer to the simple math problem can still get it wrong (they misspeak, misread the problem, perform a simple calculation error, etc.), although the high ability makes it unlikely. To give a hierarchical model built on between-subjects differences a within-subjects interpretation (John has a high probability of being right because he has a high *g*) is called the local homogeneity assumption (see the M-IRT model in Kovacs & Conway, 2016, for an explicit model of intelligence granting this assumption).

An extreme version of this difference between the two interpretaitons would posit that "intelligence" does not exist within the individual, it is simply a mathematical extraction of the correlations between the mental processes that *do* happen between people. A person has different abilities to understand verbal argument, to hold multiple pieces of information in his or her working memory, and to rotate objects, for example. These abilities are correlated with each other within the person. When trying to solve a difficult cognitive problem, however, there is no "intelligence" that exists in the person's brain. The appearance of a strong *g* factor appears as a function of these intercorrelations across people. The within-people mental processes are all that are used to solve a problem. Such an argument can point out how between- and within-subjects approaches can give different results.

This difference of between- and within-subjects interpretations begins to fall apart if, as measurement models have it, we interpret the relations in a measurement model as causal. Via extrapolation from Simpson's paradox, there cannot be a

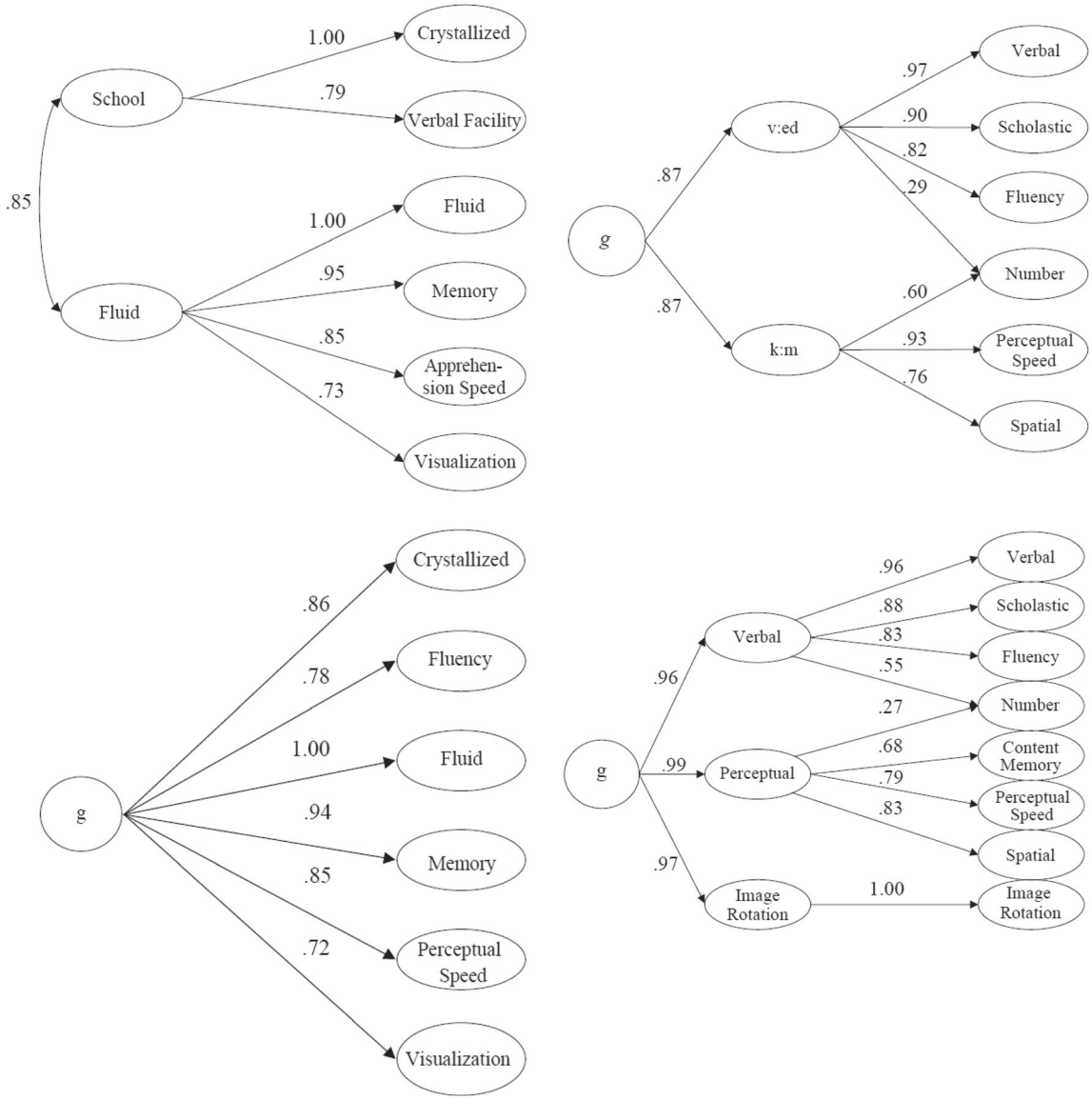

Fig. 2 Different patterns of intelligence structure; taken from Johnson & Bouchard, 2005

cause within a population that is also not a cause within a subpopulation; the reason, such a statistical result may appear, is due to an improper formulation of covariation as causation, along with appropriate effect modifiers (Pearl, 2009). If we grant that individuals can count as subpopulations, it then follows that there can be no causal effect in a population that is also not a causal effect within at least one of its members (see Weinberger, 2015 for the elaboration of this argument).[3]

For a concrete example, suppose we run an intervention where we randomly assign people to control or to working memory training groups. Members of the control group act as counterfactual estimates of those in the experimental group, allowing us to estimate the causal impact (Rubin, 2005; Pearl, 2009). In this experiment there is no variation within the individual—they are either in the experimental or control group. The results, however, are direct estimates of within-subject

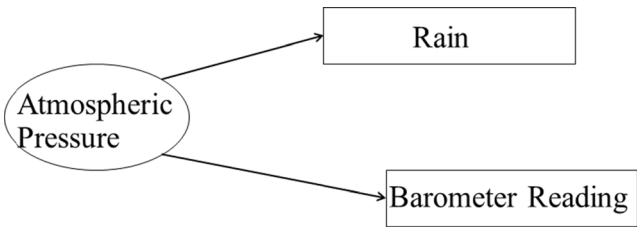

Fig. 3 A causal model about *atmospheric pressure*, *rain,* and *barometer readings*

[3] This argument is formulated in terms of causality as responsiveness in an SEM framework; however, it should be pointed out that such frameworks are mathematically equivalent to counterfactual or potential-outcomes frameworks more commonly used in randomized controlled trials in their underlying assumptions and implications (Galles & Pearl, 1998).

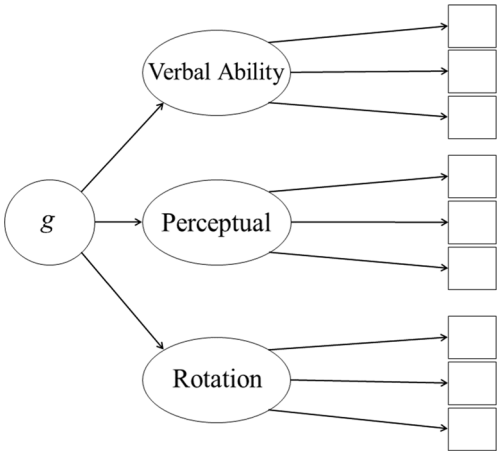

Fig. 4 Simplified VPR model of intelligence

treatment effects, despite the fact that they are between-subjects' estimates (Weinberger, 2015).

It is, however, possible to formulate interventions that change a person's *g* score without changing anything about that person. This occurs because intelligence scores, whether *g* or IQ, are relational variables, only existing in relation to other people (for modern arguments of the existence of *g* within individuals, see Colom, Chuderski, & Santarnecchi, 2016; Deary, Cox, & Ritchie, 2016; Gottfredson, 2016; and citations within). Thus, I can measure intelligence of an individual from a sample of high school seniors and find that they have a *g*-score 1 standard deviation above the mean, which I will call $g = 1$. Now, if I take that same individual and put them in a sample of college seniors, this same person now has $g = 0$, because they are even with the rest of the new sample (college seniors are more intelligent than high school seniors; see Borsboom, 2015, for this argument).

The problem with this argument is that it fails to cache itself out in observable differences. This becomes clear if we take a simpler example—working memory. Working memory can be considered the ability to hold and process and manipulate information in one's mind (Diamond, 2013). Suppose we construct a latent working memory from three observed tests: *n*-back, operations span, and digits backwards (see Fig. 5).

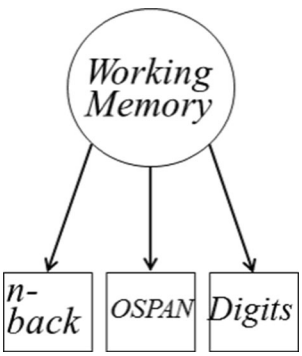

Fig. 5 Latent variable of *working memory* and its causal effect on three tests

Digits backwards is a test where participants hear a string of numbers and then repeat the numbers back in reverse order (e.g., 3-7-1-6 is repeated back 6-1-7-3). If we move a person from high-school population (wm = 1) to college population (wm = 0), given the links from latent working memory to the digits test are indeed causal (e.g., Borsboom et al., 2003; Pearl, 2009), we should expect that simply moving the person into this new population would cause him to answer larger spans of digits correctly. Such a result would unsurprisingly not occur, which leaves us with the following: (1) we must accept that the causal claims from latent working memory to digit span backwards performance are indeed false; (2) there is no within-subjects working memory, and it exists entirely as a between-subjects variable; (3) moving the person to a new sample intervention does not represent a counterfactual causal intervention. Exploration of these points is beyond the scope of this article, but it is important to understand the implications that causal structure has on measurement models. Having introduced the difficulties of between-/within-subjects modeling, we can now move on to discussing raising intelligence and targeted cognitive training in isolation.

Raising intelligence through targeted cognitive training

There have been hundreds of attempts to raise intelligence. Some of these attempts succeed, some of them fail. Because attempts to raise intelligence using randomized controlled trials allow us to strongly explore causality, there is also an implication for understanding the causal connections in the structure of intelligence.

In many attempts to raise intelligence it is difficult to pin down precise mechanisms because they are broad. They can involve learning to play a musical instrument (e.g., Schellenberg, 2004), iodine supplementation (e.g., Shrestha, 1994), completely altering the early environment of children for the first few years of their life (e.g., Ramey et al., 1992). The problem with such interventions for understanding the structure of intelligence is that they are too broad. It is not possible to understand if program effects show causality between latent variables or falsify causal connections between latent factors and subordinate measures.

Targeted cognitive training, however, does not suffer from the same breadth. A brief, and in no way complete, review of the cognitive training literature, specifically with regard to transfer across processes, follows. The nature of training and extent of transfer allows us to investigate the causal connections presented in any model of the structure of intelligence.

In targeted cognitive training, a single specific test or process is trained. Researchers select a given aspect of cognition and train the underlying process. Such training is adaptive, growing with the performance of the individuals. It is separate from

retest effects in that the individual is not simply practicing the items that will be the eventual posttest. Instead, the underlying cognitive parameters are targeted for improvement.

Within the structure of intelligence, a subfactor or broad ability is targeted for improvement. The most common targets have been memory, speed of processing, and executive functioning. Because every cognitive ability correlates with every other ability to some extent (the positive manifold; see Jensen, 1998), the necessary ingredient for transfer is guaranteed to be present. Then, either a single or a battery of tests that load heavily onto the targeted process are chosen for training. The underlying process (e.g., complex reaction time) is then trained by increasing difficulty with the aim of improving the underlying process. Provided the underlying process is accurately targeted in isolation, transfer to other abilities or processes can be seen as evidence of a causal connection between them. We highlight some examples from the literature.

Memory training has been used in the elderly to improve their basic ability to store and recall information. Interventions often involve practice in recalling learned words and sentences as well as already-memorized information (e.g., labels on prescription drugs; Ball et al., 2002). This type of training increases verbal episodic memory, but has no effect on speed of processing or reasoning ability (Ball et al., 2002). Thus, we might conclude that verbal episodic memory is causally unrelated to speed of processing and reasoning.

Speed of processing is the ability to discriminate stimuli, react, and solve simple problems in a short time. Training often involves practicing reaction time to either auditory (e.g., Anderson, White-Schwoch, Parbery-Clark, & Kraus, 2013) or visual stimuli (e.g., Edwards et al., 2005). Auditory training has resulted in increases not only in speed of processing but also in short-term memory tasks (Anderson et al., 2013). This suggests that there may be a causal effect between auditory speed of processing training and auditory short-term memory. That, or the training targets both processes.

Visual speed of processing training has received much more attention, but has also found less heartening results. Training is often on the Useful Field of Vision test, where participants respond as quickly as possible to a target that shows up in some area of peripheral vision. While training has often found improvement in this task, most studies have found no transfer to other measures of processing speed (see Edwards et al., 2002, 2005; Vance et al., 2007; Wadley et al., 2006, for example). However, one study found transfer to one type of reaction time measure, but not another (choice, not simple; Roenker, Cissell, Ball, Wadley, & Edwards, 2003; see Takeuchi et al., 2011, for inspection time training to congruent Stroop performance). This suggests that UFOV practice may not be training the underlying processing speed. There has also been a lack of transfer from speed of processing training to executive function measures as well as measures of intelligence (Edwards et al., 2002; Takeuchi et al., 2011; Wadley et al., 2006). There has been some evidence of UFOV training to visual attention (Vance et al., 2007). This may again reflect the demands of the procedure and not the causal connections of the underlying traits.

Executive function training is the testing of such processes as response inhibition or working memory. It has amassed a large body of research, with furious debate on the nature and existence and interpretation of effects (see Jaeggi, Buschkuehl, Jonides, & Perrig, 2008; Au, Buschkuehl, Duncan, & Jaeggi, 2015; Au, Sheehan, Tsai, et al., 2015; Melby-Lervåg & Hulme, 2015; Simons et al., 2016 for example). In short, whether *n*-back and similar types of training can improve tests of intelligence is under debate, with effects being often small with wide confidence intervals.

The nature of transfer to other constructs has received considerably less attention than transfer to intelligence tests. Reports conflict as to whether executive function training can increase other abilities, such as verbal abilities (for no effects, see Chooi & Thomspon, 2012; Redick et al., 2013; Thompson et al., 2013) or processing speed (for effects, see Seidler et al., 2010; Heinzel et al., 2014; no effects see: Chooi & Thompson, 2012; Redick et al., 2013; Thompson et al., 2013).

Overall, memory training has received the least attention and may also show the least potential for transfer. This may suggest that short- and long-term memory are correctly specified as being causally unrelated to other aspects of intelligence. Speed of processing has shown scattered transfer, but may not be robustly trained. Failure to transfer to other measures of speed of processing suggests that interventions such as UFOV training may not be improving the underlying construct. Until that is done, we cannot be sure of the place in the causal structure of intelligence. Executive function training has the most potential to investigate the causal structure of intelligence, but it is also the most controversial. Investigations into *n*-back training have shown the training to be not solely to working memory, and transfer with auditory training is largely absent. Research has started to test the theoretical causal connections between working memory and other traits, such as analogical reasoning. Increasing working memory without a concomitant increase in it's supposed connections represents a positive step in the direction of testing, and falsifying, individual differences links with targeted training (see Richey, Phillips, Schunn, & Schneider, 2014).

In addition, the first look into latent transfer has failed to uphold *n*-back training as a way to improve latent working memory (Colom et al., 2013). After training 56 participants in either adaptive *n*-back or a passive control group for 3 months, changes were observed in performance on certain working memory (dot matrix and reading span) and attention measures (Simon task). These improvements, however, did not appear in the latent construct analyses. It should be pointed out,

however, that there were a number of results in the "trending" direction, including for latent fluid intelligence. Whether such evidence falsifies the training effect is far from equivocal. With a focus of future experiments testing latent changes, we can better understand the repercussions on our models.

Let us consider a hypothetical experiment that trains reaction time. The investigators see significant effects on latent processing speed, suggesting that the intervention has successfully increased the latent factor. Now, if they investigate transfer to executive functioning, or upwards to intelligence, and they find no transfer, what can we conclude from this?

An immediate interpretation, especially among those who believe intelligence is immutable, would be "see, you can't change intelligence!" But such an interpretation may not be warranted. If our hierarchical structure of intelligence is indeed correct, with causal connections pointing in the right direction, then we should never expect targeted cognitive training of processing speed to increase intelligence—simply because intelligence causes, but is not caused by, processing speed. Therefore, the failure of transfer does not reflect the immutability of intelligence but instead the correct causal structure between intelligence and its subfactors.[4]

Nonhierarchical models

Although we have focused on hierarchical models, they are not the only theories of the relationships between cognitive variables. One such class of models involve feedback loops between cognitive processes (e.g., Kovacs & Conway, 2016; Van der Maas et al., 2006). Although these models have their own limitations (see, e.g., Gignac, 2014; Oberauer, 2016; Protzko, 2015), they come with the same magnitude of causal assumptions that may be individually tested. Isolating these causal assumptions can lead to successful testing of nonhierarchical models as well.

As an example, the underlying assumption of one such nonhierarchical model is that *g* emerges from different overlapping cognitive processes (Kovacs & Conway, 2016). The extent of this emergence is dictated by the weakest executive process—the weaker the process, the lower the emergent *g* (really, positive manifold). Therefore, a successful test of such a model would be to isolate the weakest links in each individuals' executive processing and conduct a training study targeting those processes. If an increase in latent *g* is observed, we may have evidence for such a nonhierarchical model of intelligence. The arguments in this article are not restricted to hierarchical models.

The future direction of targeted cognitive training

Aside from the noble pursuit of attempting to increase our intellectual lives, I argue that targeted cognitive training has a deeper theoretical implication: being able to test the causal links in the proposed structure of intelligence, telling us what intelligence *is*.

Much of the research, however, has not been able to cleanly answer these questions. The main problem has been a focus on manifest instead of latent variables. In almost every case of targeted cognitive training (see Colom et al., 2013, for the exception), interpretations are based on the results seen on test scores, not on the constructs under investigation. Thus, changes in processing speed as a construct is interpreted solely from changes in the scores on the Digit Symbol Substitution Test, for example. The problem is that performance on the Digit Symbol Substitution Test is likely multiply determined. There may be *g* effects, working memory effects, short-term memory effects, and other, unmodeled effects (see Schneider, 2013; Wechsler, 2003). While such manifest improvements are necessary for evidence that an increase in the construct has occurred, they are not sufficient, as the source of the performance increases cannot be detected from changes in total scores alone. The observed test not only suffers from measurement error but also from causal influences from intelligence and possibly other subfactors, depending on the true structure of intelligence. Thus, the interpretation may be unwarranted that one has increased processing speed from increases on a test. Such thinking hints at an underlying opertionalistic approach toward measurement, which has long been discredited in the sciences (e.g., Maul, Irribarra, T., & Wilson, 2016).

To remedy this, investigators should focus on small batteries of tests comprising the constructs under investigation. Typically, this requires at least three to four tests per latent variable (Anderson & Rubin, 1956). If we want, at a minimum, 10 participants per group per variable (e.g., Velicer & Fava, 1998), the absolute minimum number of participants needed to investigate, say, whether *n*-back training increases latent working memory or latent executive functioning, would be 60 participants. This says nothing of power, which further compounds the search for transfer to more distant constructs. Using structural equation modeling, researchers can investigate whether cognitive training alters construct-level

[4] One may be tempted to turn to Bayesian confirmation theory for help in exploring the null (e.g., Dienes, 2014). We find the invocation of Bayes factors in interpreting null hypotheses to be a largely unnecessary step, for the following reason. In a study with $n = 20$ per group, a result leading to a value of $p = .1$ would result in a Bayes factor of 1, indicating complete indifference between the null and alternate. Thus, with $n = 20$, all ps $> .1$, will always lead to the data being in favor of the null. The question no longer becomes one of whether, but to what degree, do the data support no causal connection between subfactors or with intelligence. With larger sample sizes, nonsignificance becomes more strongly in favor of the null. Thus, Bayes factors can be used to determine to what extent lack of transfer supports a lack of causality, but provided any $p > .1$ at minimally adequate sample sizes, the Bayes factors will always support such lack of causality. If we are willing to accept this state of affairs, then all nonsignificant failure of transfer supports the top-down hierarchical structure of intelligence.

variables, which can then inform the causal links to other constructs. The most likely reform to make this common practice would be creating a simple macro for commonly used, user-friendly statistics programs (e.g., Hayes & Preacher, 2014).

Furthermore, studies of targeted cognitive training have subjected participants to multiple posttests in the pursuit of transfer. Including multiple dependent variables, however, comes with a cost. It is necessary to either model the data concurrently (as in, say, a structural equations model, taking into account correlations between dependent variables) or incur a penalty to critical p values for each new dependent variable taken (see Bird & Hadzi-Pavlovic, 2014, for some methods of protection). One cannot subject participants to a dozen posttests, find one or two effects at $p < .05$, and be confident that a true underlying change has been made.

It is also important to ensure that the training only targets the processes under consideration. If an intervention trains both underlying processing speed and working memory, for example, we cannot be sure if transfer represents common training or true causal connections between processes. Some interventions (e.g., Smith et al., 2009; van Ravenzwaaij, Boekel, Forstmann, Ratcliff, & Wagenmakers, 2014) are too broad to categorize as targeted cognitive training; so the results cannot be readily applied to understanding the structure of human abilities. One notable example tested visual n-back (the default of most executive function training) versus auditory n-back. If the training was only increasing underlying working memory, then this visual versus auditory difference should not matter for transfer. The authors found different patterns of results on matrix reasoning subtests, with the visual n-back groups showing improvement, whereas no improvement was seen in auditory n-back (Stephenson & Halpern, 2013). This suggests we need to take a closer look at what the training is actually doing.[5]

There are other concerns, which I shall not repeat here, such as over reliance on pre × post interactions (see Huck & McLean, 1975, for a classic criticism) or active versus passive control groups. In addition, some studies use truncated tests or tests only using half of the items (e.g., Lawlor-Savage & Goghari, 2016), which can lower the reliability of the tests, harming interpretability. Though this may also affect the ability to investigate transfer at the latent level, we are unsure of any explicit tests of this assertion. Such concerns as these apply to all experimental research, not just targeted cognitive training.

So what is the way forward for experimental psychometrics in understanding the causal structure of human cognitive abilities? Only a sketch is provided here, with future research being the real arbiter of progress. The question becomes how do we test the direction of causality and the presence of causal links in our understanding of the structure of human cognitive abilities.

First, a cognitive process must be identified for improvement. Then, a method for improvement must be found that, for theoretical reasons, would only involve the process under investigation. As discussed with auditory versus verbal n-back, this can often be difficult. Having identified a process and isolating a method of improvement, the first step would be to show that improvements on the training task are indeed improvements to the underlying latent process. A battery of posttests, not including the training task, analyzed in a randomized controlled trial with adequate sample sizes, using latent variable modeling, would be the first test (e.g., Lubke, Dolan, & Kelderman, 2001). After establishing that a method of training can improve the underlying process, follow-up experiments could be run using batteries of tests isolating distant latent processes. The success or failure of transfer (see note 3 as well) "upward" to latent g or "sideways" to other latent subfactors can help us understand the causal connections underlying the correlation of human cognitive abilities. This is only a rough sketch of the required steps, but taking causality seriously requires serious changes to the way we investigate the questions surrounding targeted cognitive training.

Conclusion

Attempting to enhance our cognitive lives through targeted training is a most noble pursuit, and one we fully endorse. Interventions that train specific cognitive abilities, searching for transfer to intelligence, may be doomed to fail not because intelligence is immutable, but because the causal relations between intelligence and its subfactors are unidirectional.

All of our models are incomplete; they do not capture every cause and effect within a given system (e.g., MacCallum, 2003; Meehl, 1990). While this incompleteness from 100 % representation is usually deemed acceptable, falsity, the wrong specifications of our models (faulty causal paths, reference to nonexistent variables, etc.) is not. The goal, however, is to increase how much our models correspond to the world. Although a popular method among individual differences researchers is to compare model fit statistics of competing models, we have the ability to directly test the causal assumptions underlying the structure of intelligence. We can test the falsity of the models. We can provide greater strides in fully understanding human cognitive ability.

The causal assumptions of our models of the structure of intelligence are not different because they exist largely in the latent realm. They do not require special studies or data different from any other exploration of causality. What is important, however, is that we recognize these causal claims and take them seriously. With targeted training and the appropriate analysis, we can explore the validity of these causal connections.

[5] At the time of writing this article, the debate around n-back training is currently vogue, but the arguments within apply to all future approaches to targeted cognitive training, whatever the process.

Much as we cannot move the hand on the barometer in the hopes it will change the weather, the structure of intelligence may make it impermeable to changing subfactors in the hopes of upward effects. Seeing factor analytic structures (such as Fig. 1) as models of strong causal claims and not just graphs to represent data, we can understand one of the true benefits of targeted cognitive training as allowing us to explore what intelligence *is*.

References

Anderson, S., White-Schwoch, T., Parbery-Clark, A., & Kraus, N. (2013). Reversal of age-related neural timing delays with training. *Proceedings of the National Academy of Sciences, 110*(11), 4357–4362.

Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 5.*

Au, J., Buschkuehl, M., Duncan, G. J., & Jaeggi, S. M. (2015). There is no convincing evidence that working memory training is NOT effective: A reply to Melby-Lervåg and Hulme (2015). *Psychonomic Bulletin & Review*, 1–7. doi:10.3758/s13423-015-0967-4.

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.

Ball, K., Berch, D. B., Helmers, K. F., Jobe, J. B., Leveck, M. D., Marsiske, M., & ACTIVE Study Group. (2002). Effects of cognitive training interventions with older adults: A randomized controlled trial. *The Journal of the American Medical Association, 288*(18), 2271–2281.

Bird, K. D., & Hadzi-Pavlovic, D. (2014). Controlling the maximum familywise Type I error rate in analyses of multivariate experiments. *Psychological Methods, 19*(2), 265–280.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Borsboom, D. (2015). What is causal about individual differences?: A comment on Weinberger. *Theory & Psychology, 25*(3), 362–368.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203–219.

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

Colom, R., Román, F. J., Abad, F. J., Shih, P. C., Privado, J., Froufe, M., & Jaeggi, S. M. (2013). Adaptive n-back training does not improve fluid intelligence at the construct level: Gains on individual tests suggest that training may enhance visuospatial processing. *Intelligence, 41*(5), 712–727.

Colom, R., Chuderski, A., & Santarnecchi, E. (2016). Bridge over troubled water: Commenting on Kovacs and Conway's process overlap theory. *Psychological Inquiry, 27*(3), 181–189.

Deary, I. J., Cox, S. R., & Ritchie, S. J. (2016). Getting Spearman off the skyhook: One more in a century (since Thomson, 1916) of attempts to vanquish g. *Psychological Inquiry, 27*(3), 192–199.

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

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*(781), 1–17.

Edwards, J. D., Wadley, V. G., Myers, R. S., Roenker, D. L., Cissell, G. M., & Ball, K. K. (2002). Transfer of a speed of processing intervention to near and far cognitive functions. *Gerontology, 48*(5), 329–340.

Edwards, J. D., Wadley, V. G., Vance, D. E., Wood, K., Roenker, D. L., & Ball, K. K. (2005). The impact of speed of processing training on cognitive and everyday performance. *Aging & Mental Health, 9*(3), 262–271.

Galles, D., & Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science, 3*(1), 151–182.

Gignac, G. E. (2014). Dynamic mutualism versus g factor theory: An empirical test. *Intelligence, 42,* 89–97.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24*(1), 13–23.

Gottfredson, L. S. (2016). A g theorist on why Kovacs and Conway's process overlap theory amplifies, not opposes, g theory. *Psychological Inquiry, 27*(3), 210–217.

Hausman, D., & Woodward, J. (1999). Independence, invariance, and the causal Markov condition. *British Journal for the Philosophy of Science, 50,* 1–63.

Hayes, A. F., & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology, 67*(3), 451–470.

Heinzel, S., Schulte, S., Onken, J., Duong, Q. L., Riemer, T. G., Heinz, A., & Rapp, M. A. (2014). Working memory training improvements and gains in non-trained cognitive tasks in young and older adults. *Aging, Neuropsychology, and Cognition, 21*(2), 146-173.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55,* 577–601.

Huck, S. W., & McLean, R. A. (1975). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin, 82*(4), 511–518.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences, 105*(19), 6829–6833.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.

Johnson, W., & Bouchard, T. J. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence, 33*(4), 393–416.

Johnson, W., Bouchard, T. J., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one g: Consistent results from three test batteries. *Intelligence, 32*(1), 95–107.

Kovacs, K., & Conway, A. R. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry, 27*(3), 151–177.

Lawlor-Savage, L., & Goghari, V. M. (2016). Dual n-back working memory training in healthy adults: A randomized comparison to processing speed training. *PLOS ONE, 11*(4), e0151817.

Lubke, G. H., Dolan, C. V., & Kelderman, H. (2001). Investigating group differences on cognitive tests using Spearman's hypothesis: An evaluation of Jensen's method. *Multivariate Behavioral Research, 36*(3), 299–324.

MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research, 38*(1), 113–139.

Maul, A., Irribarra, D. T., & Wilson, M. (2016). On the philosophical foundations of psychological measurement. *Measurement, 79,* 311–320.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*(1), 1–10.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1,* 108–141.

Melby-Lervåg, M., & Hulme, C. (2015). There is no convincing evidence that working memory training is effective: A reply to Au et al.

(2014) and Karbach and Verhaeghen (2014). *Psychonomic Bulletin & Review*, 1–7. doi:10.3758/s13423-015-0862-z

Oberauer, K. (2016). Parameters, not processes, explain general intelligence. *Psychological Inquiry, 27*(3), 231–235.

Pearl, J. (2009). *Causality*. New York, NY: Cambridge University Press.

Protzko, J. (2015). The environment in raising early intelligence: A meta-analysis of the fadeout effect. *Intelligence, 53,* 202–210.

Ramey, C. T., Bryant, D. M., Wasik, B. H., Sparling, J. J., Fendt, K. H., & La Vange, L. M. (1992). Infant Health and Development Program for low birth weight, premature infants: Program elements, family participation, and child intelligence. *Pediatrics, 89*(3), 454–465.

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.

Richey, J. E., Phillips, J. S., Schunn, C. D., & Schneider, W. (2014). Is the link form working memory to analogy causal? No analogy improvement following working memory training gains. *PLOS ONE, 9*(9), e106616.

Roenker, D. L., Cissell, G. M., Ball, K. K., Wadley, V. G., & Edwards, J. D. (2003). Speed-of-processing and driving simulator training result in improved driving performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 45*(2), 218–233.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association, 100*(469), 322–331.

Schneider, W. J. (2013). What if we took our models seriously? Estimating latent scores in individuals. *Journal of Psychoeducational Assessment, 31,* 186–201.

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

Seidler, R. D., Bernard, J. A., Buschkuehl, M., Jaeggi, S., Jonides, J., & Humfleet, J. (2010). *Cognitive training as an intervention to improve driving ability in the older adult*, (No. M-CASTL 2010-01).

Shrestha, R. (1994). *Effect of iodine and iron supplementation on physical, psychomotor and mental development in primary school children in Malawi*. Wageningen, The Netherlands: Landbouwuniversiteit te Wageningen.

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.

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 the Improvement in Memory with Plasticity-based Adaptive Cognitive Training (IMPACT) study. *Journal of the American Geriatrics Society, 57*(4), 594–603.

Stephenson, C. L., & Halpern, D. F. (2013). Improved matrix reasoning is limited to training on tasks with a visuospatial component. *Intelligence, 41*(5), 341–357.

Takeuchi, H., Taki, Y., Hashizume, H., Sassa, Y., Nagase, T., Nouchi, R., & Kawashima, R. (2011). Effects of training of processing speed on neural systems. *The Journal of Neuroscience, 31*(34), 12139–12148.

Thompson, T. W., Waskom, M. L., Garel, K. L. A., Cardenas-Iniguez, C., Reynolds, G. O., Winter, R., & Gabrieli, J. D. (2013). Failure of working memory training to enhance cognition or intelligence. *PLOS ONE, 8*(5), e63614.

Thorndike, R. L. (1987). Stability of factor loadings. *Personality and Individual Differences, 8*(4), 585–586.

Van Der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review, 113*(4), 842–861.

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

Vance, D., Dawson, J., Wadley, V., Edwards, J., Roenker, D., Rizzo, M., & Ball, K. (2007). The accelerate study: The longitudinal effect of speed of processing training on cognitive performance of older adults. *Rehabilitation Psychology, 52*(1), 89–96.

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3,* 231–251.

Wadley, V. G., Benz, R. L., Ball, K. K., Roenker, D. L., Edwards, J. D., & Vance, D. E. (2006). Development and evaluation of home-based speed-of-processing training for older adults. *Archives of Physical Medicine and Rehabilitation, 87*(6), 757–763.

Wechsler, D. (2003). *Wechsler intelligence scale for children* (4th ed.). San Antonio, TX: Harcourt Assessment.

Weinberger, N. (2015). If intelligence is a cause, it is a within-subjects cause. *Theory & Psychology, 25*(3), 346–361.