Intelligence 48 (2015) 1–14

Contents lists available at ScienceDirect

Intelligence

Do we really become smarter when our fluid-intelligence test scores improve?

Taylor R. Hayes, Alexander A. Petrov *, Per B. Sederberg

Department of Psychology, Ohio State University, United States

ARTICLE INFO

Article history:
Received 14 January 2014
Received in revised form 24 September 2014
Accepted 10 October 2014
Available online 8 November 2014

Keywords:
Fluid intelligence
Cognitive training
Skill acquisition
Strategy
Eye movements

ABSTRACT

Recent reports of training-induced gains on fluid intelligence tests have fueled an explosion of interest in cognitive training—now a billion-dollar industry. The interpretation of these results is questionable because score *gains* can be dominated by factors that play marginal roles in the scores themselves, and because intelligence gain is not the only possible explanation for the observed control-adjusted far transfer across tasks. Here we present novel evidence that the test score gains used to measure the efficacy of cognitive training may reflect strategy refinement instead of intelligence gains. A novel scanpath analysis of eye movement data from 35 participants solving Raven's Advanced Progressive Matrices on two separate sessions indicated that one-third of the variance of score gains could be attributed to test-taking strategy alone, as revealed by characteristic changes in eye-fixation patterns. When the strategic contaminant was partialled out, the residual score gains were no longer significant. These results are compatible with established theories of skill acquisition suggesting that procedural knowledge tacitly acquired during training can later be utilized at posttest. Our novel method and result both underline a reason to be wary of purported intelligence gains, but also provide a way forward for testing for them in the future.

1. Introduction

Can intelligence be improved with training? For the most part, the numerous training methods attempted through the years have yielded disappointing results for healthy adults (e.g., Detterman & Sternberg, 1982). Nonetheless, if an effective training method could be designed, it would have immense practical implications. Therefore, when Jaeggi, Buschkuehl, Jonides, and Perrig (2008) recently published some encouraging experimental results, they were greeted with remarkable enthusiasm. Cognitive enhancement is now a billion-dollar industry Brain sells (2013). Millions of customers buy "brain building" games and subscribe to "mental gyms" on-line where they perform various "cognitive workouts" in the hope of raising their IQ (Hurley, 2012). Hundreds of millions of dollars are being invested in educational (e.g., Cogmed, http://www.cogmed.com), military, and commercial programs (e.g., Lumosity, http://www.lumosity.com) on the assumption that intelligence can be improved through training. But can it really? Given the massive societal resources that are at stake and the checkered track record of similar initiatives in the past (e.g., Detterman & Sternberg, 1982; Melby-Lervåg & Hulme, 2013; Owen et al., 2010), this claim must be evaluated very carefully. Here we present novel evidence that suggests reasons for skepticism. The evidence is not definitive and the question remains open. It leads directly to three other questions: (i) What is intelligence? (ii) How can we measure intelligence? and (iii) How can we measure *gains* of intelligence? The first two of those have been debated and researched for over a century (see, e.g., Neisser et al., 1996, for an authoritative review). The last question, however, has not received the attention it deserves. One goal of

* Corresponding author at: Department of Psychology, 200B Lazenby Hall, Ohio State University, Columbus, OH 43210, United States.
E-mail address: apetrov@alexpetrov.com (A.A. Petrov).

http://dx.doi.org/10.1016/j.intell.2014.10.005

this article is to point out how methodologically challenging it is to measure the *change* of a latent variable.

With respect to the first two questions, we adopt the popular (though not universally accepted) psychometric approach that both defines and measures fluid intelligence as the latent variable explaining the intercorrelations in performance on tasks such as analogy making, reasoning, and problem solving. This approach is grounded in the fact that individual differences in performance across a wide variety of cognitive tasks are positively correlated (Spearman, 1927). Through factor analysis, the matrix of intercorrelations can be explained in terms of a hierarchical arrangement with a general intelligence factor *G* at the apex and various more specialized abilities arrayed below it (Carroll, 1993; Jensen, 1998). The second tier in the hierarchy includes the distinction between crystallized (*Gc*) and fluid (*Gf*) intelligence (Carroll, 1993; Cattell, 1963). *Gc* refers to overlearned skills and static knowledge such as vocabulary, which undoubtedly accumulate with experience. In contrast, *Gf* refers to the ability to detect patterns and relations, solve problems, and "figure things out" in novel environments. Empirically, fluid intelligence predicts many forms of achievement, especially school achievement (Gottfredson, 1997). There is strong evidence that *Gf* is highly heritable—between 50% and 75% of the variance of intelligence test scores in healthy adults is linked to genetic variation (Neisser et al., 1996). Although heritability does not entail immutability (Dickens & Flynn, 2001), most psychometricians conceptualize *Gf* as a stable trait that is relatively immune to interventions in adulthood (Carroll, 1993; Jensen, 1998).

This is why a recent study by Jaeggi et al. (2008) triggered such excitement and controversy. The study used a pretest-train-posttest design with an untrained control group. A titrated, adaptive dual n-back task was practiced for up to 18 sessions in the experimental group ($N = 34$) but not in the control group ($N = 35$). All participants were pre- and post-tested on two parallel short-form versions of a matrix-based *Gf* test—either Raven's Advanced Progressive Matrices (Raven, Raven, & Court, 1998) or BOMAT (Hossiep, Turck, & Hasella, 1999). Whereas the results showed statistically significant score gains in both groups, the average gain in the trained group was significantly higher than that in the control ($p < 0.05$, $\eta_p^2 = 0.07$, Jaeggi et al., 2008). The latter finding—a significant *control-adjusted gain*—was interpreted as an improvement in *Gf* and fueled the current boom in the cognitive enhancement industry, as well as a big controversy in the scientific literature. Of particular relevance to the controversy is that the original study (Jaeggi et al., 2008) had various methodological shortcomings (Moody, 2009) and subsequent attempts to replicate the putative improvement in *Gf* have produced mixed results (e.g., Chooi & Thompson, 2012; Harrison et al., 2013; Jaeggi, Buschkuehl, Jonides, & Shah, 2011; Jaeggi et al., 2010; Redick et al., 2012; Thompson et al., 2013). This rapidly growing field is characterized by large variations in reported effect sizes (see Melby-Lervåg & Hulme, 2013, for a meta-analysis of 23 studies), polarization of opinion, and contradictory reviews (e.g., Buschkuehl & Jaeggi, 2010; Morrison & Chein, 2011, on the optimistic side; Melby-Lervåg & Hulme, 2013; Shipstead, Redick, & Engle, 2012, on the skeptical side).

The neurobiological interpretation of *Gf* (M. Anderson, 2005; Duncan et al., 2000) emphasizes its linkage to factors such as processing speed (Jensen, 2006; Sheppard & Vernon, 2008) and working memory capacity (Fry & Hale, 2000; Gray & Thompson, 2004; Halford, Cowan, & Andrews, 2007; Kane & Engle, 2002). The interest in the latter linkage surged after Jaeggi et al.'s (2008) publication because their participants trained on a WM task. The hypothesis that fuels the current enthusiasm is that WM training increases WM capacity (near transfer), which in turn improves *Gf* (far transfer). There is a strong analogy with athletics, where swimming workouts, for example, increase cardiovascular capacity, which in turn improves the general athletic ability. Thus, Jaeggi et al. (2011) characterize WM as "taking the place of the cardiovascular system."

This hypothesis is simple and elegant but the methodology for testing it empirically is fraught with difficulties because an objective method for measuring *Gf gains* is required. The commonly used test–retest method is seriously flawed. The overwhelming majority of studies use test–retest score gains to measure *Gf* gains. This practice is based on the misleading intuition that if a test such as Raven's APM is a valid measure of *Gf*, then a *gain* in the score on this test is a valid measure of *Gf gain*. This is not necessarily true because, in addition to *Gf*, the scores reflect non-*Gf* factors such as visuospatial ability, motivation, and test-taking strategy. The latter factors—and hence the test scores—can improve while *Gf* itself remains stable. Indeed, Raven's APM scores increase significantly on repeated testing without any targeted training (e.g., Bors & Forrin, 1995; Bors & Vigneau, 2003; Denney & Heidrich, 1990). Worse, a large meta-analysis of 64 test–retest studies (te Nijenhuis, van Vianen, & van der Flier, 2007) indicates a strong *negative* correlation between score gains and the *G* loadings of test items. To control for such "mere retest" effects, the common practice in the field is to compare the score gains in the treatment group to those in an untreated control group. Cognitive enhancement advocates (e.g., Jaeggi et al., 2008) acknowledge the interpretive problems of unadjusted score gains but assume that control-adjusted gains necessarily measure real gains in *Gf*. As we argue below, however, this assumption is incorrect because the adjustment does not guarantee validity either.

These methodological difficulties can be illustrated by analogy with athletics. In a classic study of motor skill learning (Hatze, 1976), an athlete practiced kicking a target as rapidly as possible. His performance improved at first and then plateaued. However, after seeing a film about kicking technique, the athlete immediately improved his time considerably and with additional practice was able to reach a much higher asymptote. For our purposes, this illustrates the relationships between the following three variables. The first is kicking time, which was the only objective measurement. The second variable is general athletic ability, which includes factors such as cardiovascular capacity, agility, muscle strength, and so forth. The third is kicking technique—the optimal way to execute a kick so as to minimize kicking time, all else being equal. Importantly, because the kicking time reflects a mixture of athletic ability and technique, gains in kicking time can occur without any change in athletic ability. Indeed, watching a movie could not have changed the strength or agility of the participant in Hatze's (1976) experiment. Analogously, gains in test scores can occur without any change in "brainpower" factors such as WM capacity or processing speed.

This brings us to the central topic of transfer across tasks. The most widely used inference pattern in the cognitive