Studies in History and Philosophy of Biological and Biomedical Sciences xxx (2017) 1–6

Contents lists available at ScienceDirect

Studies in History and Philosophy of Biological and Biomedical Sciences

journal homepage: www.elsevier.com/locate/shpsc

Essay Review

Flynn, the Age-Table Method, and a metatheory of intelligence

Lucas J. Matthews*, Eric Turkheimer

Department of Psychology, University of Virginia, Charlottesville, VA 22904-4400, United States

Does your family make you smarter: Nature, nurture, and human autonomy, James Flynn. Cambridge University Press, Cambridge (2016). 258 pp., Softcover, ISBN-10: 1316604462

1. Flynn 1.0: the Flynn effect

In the 1980s, working by himself, James Flynn, an American expatriate philosopher in Otago, New Zealand, noticed something extraordinary: IQ scores had been rising for the past fifty years.[1] And not by just a little: scores had increased by as much as a standard deviation (SD), or 15 points on the familiar IQ scale with a mean of 100. These increases, known as the "Flynn Effect," changed the course of IQ research for good.

The Flynn effect owed its great importance to the tenor of mainstream science about IQ at the time. 75 years of psychometric research had shown that intelligence was a measurable and stable trait that predicted important life outcomes. Twin and adoption studies had shown that measured IQ was substantially heritable. Arthur Jensen had completed most of an imposing body of work tying these two phenomena together and exploring their implications for social policy. Most of these implications seemed to lean significantly rightward. The Bell Curve (Herrnstein & Murray, 1994) would be published a few years later.

The larger problem was not the heritability of intelligence *per se* but the apparent absence of environmental effects. Classical behavior genetics partitions the variability in a trait like IQ into three components: the additive effects of genes (A), environmental effects that make siblings *similar* [C], and environmental effects that make them *different* (E). Although at one time the idea that something like intelligence might be related to genetics was a bitterly divisive idea, by Flynn and Jensen's time the notion had become more widely accepted, as it had become clear that all human differences, behavioral no different than physical, bore some relationship to genotype. The absence of shared environmental effects was more persistently troubling, however. Poverty, for starters, is a shared environmental effect, because siblings raised together share their family's socioeconomic status. Twin and adoption studies suggested the presence of modest shared environmental effects on IQ in early childhood, but the effects declined rapidly as children grew up, disappearing entirely by early adulthood. It was the persistent absence of shared environmental effects in twin studies, paired with the difficulties of demonstrating reliable effects for programs designed to boost IQ, that motivated Jensen (1969) conclusions about the innateness of observed IQ differences among the races and Herrnstein and Murray (1994) similar conclusions about the origins of social class. If genes make a difference for IQ and poverty doesn't, then the outlook for a socially progressive world seems grim.

What causes the Flynn Effect? Something so recent is unlikely genetic in origin. Flynn and William Dickens proposed a feedback loop explanation for historical IQ gains. Consider basketball. Twin studies would likely reveal that basketball ability – running, jumping, dribbling, etc. – is highly heritable. Yet the average ability of basketball players today greatly exceeds that of players in the 1950s. One way to reconcile such large changes over time with a genetic trait is to appeal to the notion of 'reciprocal effects.' As basketball in the 1950s became televised and more popular, greater numbers of talented players competed, which further increased popularity and, in turn, drew more talented players. Dickens and Flynn showed that such 'multiplier effects' could produce large increases over a short period of time, genetic influence notwithstanding. The extension to IQ is obvious, if hard to prove. The modern cognitive climate is very different from the still largely agrarian world before World War II, and equally so from the agrarian and/or economically impoverished environments of the modern undeveloped world.

Notably, Flynn did not discover his eponymous effect using experimental studies, but in the appendices of IQ testing manuals. Test administrators produce their familiar IQ scores using standardization tables that translate raw totals of correct answers to scaled scores on each subtest, normed to a bell-shaped population with a mean of 10 and a SD of three for each age group. The scaled subtest scores are then summed and re-normed to produce IQ scores with a mean of 100 and a SD of 15. Intelligence tests were renormed as they were updated and revised over the years, leaving behind a trail of standardization tables that evince the Flynn Effect.

Although standardization tables were perfectly suited to document the Flynn Effect, it's important to note how little information

* Corresponding author.
E-mail addresses: lucasjohnmatthews@virginia.edu (L.J. Matthews), ent3c@virginia.edu (E. Turkheimer).

[1] See (Flynn, 1984a; 1984b; 1986; 1987).

http://dx.doi.org/10.1016/j.shpsc.2017.07.003

they actually contain. The tables for seven-year-olds and nine-year-olds are not tracking the abilities of children as they grow up – they record the performance of completely different individuals. There are no data on the families of the individuals in the tables, nothing on their parents or the environments they provide, nothing on their siblings, never mind their twins. They are a way to convert raw number-correct scores to normally distributed scaled scores, and nothing more.

2. Flynn 2.0: the age-table method

The original Flynn Effect (call it ‘Flynn 1.0’) was an anti-determinist, environmentalist salve on the hereditarian wounds of the late 20th Century, but it did not heal them entirely. Flynn 1.0 is an effect of culture and generational time, not families. Family effects are what environmentalists really want. Can it possibly be true that being raised by impoverished or uninterested or absent parents, socialized in dangerous neighborhoods and educated in under-resourced schools, has no effect on IQ scores? That is what much of the behavior genetic evidence seems to suggest. Once genetic effects are controlled siblings raised in the same family, sharing the same parents, streets, and schools, are scarcely more similar in IQ than two people chosen at random from the population.[2]

Questions of nature and nurture are difficult to study in humans because most children are raised by their biological parents. If accomplished and attentive parents have successful children, the reasons are as likely to be genetic as they are environmental. Social scientists cannot randomly assign children to rearing conditions and see what happens, but there is a long tradition of attempts to wring knowledge of environmental effects from the “natural experiments” of child-rearing, particularly adoption. Most of the time children are adopted from relatively poorer to relatively better-off homes, and one can ask whether such adoptions produce increases in IQ relative to biological mothers or nonadopted siblings. The results of such studies suggest that fairly reliable improvements in IQ are associated with adoption, but like many conclusions based on the complex natural arrangements of human beings, those conclusions are implacably controversial, batted back and forth by environmentalists and hereditarians until the objective mind glazes over. Why were the children adopted in the first place? Why were the siblings to whom they are compared *not* adopted? Were the IQ scores administered to the biological mothers accurate? Non-experimental (that is, without random assignment to rearing conditions) science about controversial topics like the malleability of IQ never reaches consensus. The noise is stronger than the signal, leaving ideological predilection in unfettered control.

Yet in *Does your family make you smarter?* Flynn makes an extraordinary pronouncement: “I will *settle* the great debate about whether family effects are still significant at the age of 17” (p. 30, italics added). We cannot emphasize enough what a daring claim this is. If he is right, it would conclude what is arguably the most long-standing controversy in human individual differences. Even more amazing is how he proposes to do it, definitively inferring the environmental effect of families from standardization tables that include no information about either families or the environment. He refers to his method as the “Age-Table Method.”

Reviewing the Age-Table Method presents several difficult problems. Flynn works outside the traditional constraints of quantitative genetics and social science statistics, and for that matter ordinary scientific reporting – there are no equations, no path diagrams, no ACE models, just back-of-the-envelope calculations that are at once extremely elaborate and poorly documented. Flynn’s illustrious track record, sketchy exposition, and sublime intellectual confidence combine to leave one with the distinct feeling that something important is missing. Moreover, Flynn’s conclusion – that there are real environmental effects of families that fade as children pass through adolescence – is almost certainly true, based on evidence other than what is presented in the book.[3] Finally, Flynn’s big picture views on the other controversial aspects of the theory of intelligence, like the existence of *g*, which we will explain in more detail later, and genetically-based differences among human groups – align closely with our own. Nevertheless, we have to insist that the evidentiary foundation of the Age-Table Method is entirely invalid.

What follows is not going to be simple, so have a seat. Different intelligence tests are standardized slightly differently, and much of Flynn’s procedure is taken up with accounting for differences among the tests and their revisions over time. Although many of the improvised techniques for these purposes are clever, here we’ll focus on the foundation upon which they are built. For our example, we will use the Stanford-Binet vocabulary scale, which has the advantage of being standardized continuously for individuals between two and 80 years of age; otherwise, one has to follow Flynn through a series of clever but extremely complex operations to link up performance on separate tests for children and adults.

Flynn uses standardization tables in reverse to calculate the number of vocabulary words a test-taker has to know in order to fall at a certain percentile of the scaled score distribution. In order to obtain an average score for her age (a scaled score of 10), a seven-year-old must get 27 items correct. What interests Flynn is how many new words this person would have to learn in order to remain average later in life, particularly at the age (55) at which vocabulary scores reach their maximum. For 55-year-olds, 44.5 (some incidental averaging produces the decimal) correct responses are required to be at the mean, so vocabulary must grow by 17.5 words in order to remain average. Now, compare all this to what is required to remain at an unusually proficient level, two SDs above the mean, or at a scaled score of 16. For a seven-year-old, that would require 37 correct responses; for a 55-year-old, it would require 58.5. So, an average seven-year-old must learn 17.5 new words to remain average, but a very proficient seven-year old must learn 21.5 words to maintain the same level. The opposite is true for those underperforming at the other end of the spectrum: at two SDs below the mean, a seven-year-old would score 16 while a 55-year-old would score 28.5. Thus those underperforming must only learn 12.5 words in order to remain two SDs below the mean from age seven to 55. The pattern that emerges suggests high achievers must learn more words than low achievers in order to remain in their respective percentiles. Why does the improvement in performance necessary to sustain performance at a given level differ between high, median, and low scorers?

Flynn reasons (correctly, as far as it goes) that children come from varying levels of “cognitive quality” (Flynn’s unfortunate and old-fashioned phrase) of family environment. Specifically, high-performing children tend to come from family environments that are of a lower cognitive quality than themselves, and low-performing children tend to come from family environments that are of a higher cognitive quality than themselves. In other words, students who score in the 99th percentile on SATs likely come from a family environment that is lower than the 99th percentile, while students who score in the 1st percentile on SATs are likely to come

[2] See Plomin and Daniels (1987) or David C. Rowe (1993).

[3] See Turkheimer, 1991 and Kendler, Turkheimer, Ohlsson, Sundquist, & Sundquist, 2015.

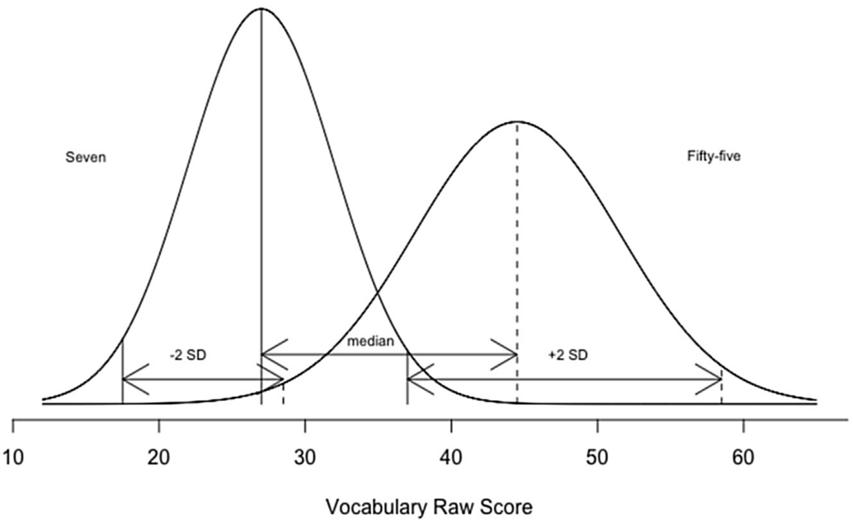

Fig. 1. Two normal distributions, one for seven-year-olds with a mean of 27 and SD of five, and another for 55-year-olds with a mean of 58.5 and SD of seven.

from family environments higher than the 1st percentile. The idea is that lower cognitive quality environments hold the high-performing children back, creating relatively larger gaps between the children's performance and that of equivalently high performing adults who are no longer held back by their lower-quality families, because they no longer live in the family home. For low-performing children, it's the opposite: their performance is enhanced by their relatively high quality families, compared to low-performing adults who no longer get any environmental boost from their families. Children at the median are matched to their parents' abilities, and thus unaffected. Like we said, it's a complicated theory, which Flynn presents as though it were perfectly obvious. But before proceeding further into the weeds, it is important to note that Flynn's interpretation is, at best, pure speculation. Standardization tables, to repeat, provide no actual information about family environments. Flynn can only *guess* the percentile of cognitive quality of family environment from which a child comes. So, the best Flynn can do is to argue that the pattern he observes across age at the upper and lower quantiles of the distribution looks like what one might expect if the environmental effect of families were responsible for the differences. Unfortunately for Flynn's theory, there is a far simpler possibility that renders further speculation superfluous.

An interesting fact uncovered by Flynn's method is that the SD of raw scores increases throughout development. That is to say, the distribution of raw scores spreads out. In seven-year-olds, the difference between a performance two SDs below and above the mean is 21 correct words; by 17 it is 27 words, and by 55 it is 30 words. The SD of raw scores is about five at age seven, and about seven by age 55. Suppose we forget all about the actual IQ scores of actual people as documented in the standardization tables, and just draw two schematic normal distributions based on what we know: one for seven-year-olds with a mean of 27 and a SD of five, and one for 55-year-olds with a mean of 58.5 and a SD of seven. We get Fig. 1:

The three two-headed arrows in the figure are the Flynn 2.0 quantities: the differences in raw score between the seven and 55-year-old distributions at two SDs above and below the mean, and at the median. What do you notice? The Flynn 2.0 effect is clearly there – a larger gap between seven-year-olds and 55-year-olds above the median than at the median, and a larger gap at the median than two SDs below. In Flynn-like descriptive language, because the 55-year-old distribution is both to the right of the seven-year-old distribution *and* more spread out, the upper tail of the 55-year-olds stretches further away from the seven-year-olds. If you redraw the figure without the increase in SD, Flynn 2.0 completely disappears. So, it turns out all that is required to produce Flynn 2.0 is an increase in the SD of raw scores that parallels the increase in the mean. Flynn 2.0 amounts to the observation that the SD of raw scores increases with age. Flynn has to be willing to argue that this increase is itself somehow a consequence of family environments, but it is difficult to see how this could be true.

But it isn't worth the time trying to justify how environmental effects of families could produce age-related increases in variance, because there is a much simpler explanation: scores on vocabulary tests grow. Growth is a particular variety of age-related change. Some quantities that change fluctuate over time: one might be more depressed today, less so tomorrow, then back up again next week. Growth processes, physical growth being the obvious example, only go up. In a growth process, there is variability in terms of where people start out, and then more variability in each successive time period in terms of how much they add on. The dynamics of growth processes are complex, but in most cases growth produces increasing variability with age.[4] Except in the most unusual of circumstances, raw vocabulary only increases as children age.

In other words, we think that Flynn 2.0 is a side effect of this unsurprising statistical property of growth. It is as true of height as it is of vocabulary score. See, for example, Fig. 2 adapted from Tanner and Whitehouse (1976), which depicts the same side effect of height growth over time. The fanning-out of the trace lines of the upper and lower quantiles of height in boys between age one and 19 represents the growth-related increase in the SD of height. If one uses the graph to approximate the 3rd, 50th and 97th percentiles at ages four and 18, one obtains 92, 100 and 109 cm at age four, and 163, 175, and 188 at age 18. The differences of 71, 75, and 79 cm produce Flynn 2.0. Presumably, the fact that tall four-year-olds have to grow more to maintain their relative standing at 18 is not an environmental consequence of being raised around relatively shorter family members.

[4] Elliot Tucker-Drob (personal communication) points out that growth processes can produce decreasing variances when there is a strong negative correlation between where people start and how fast they grow.

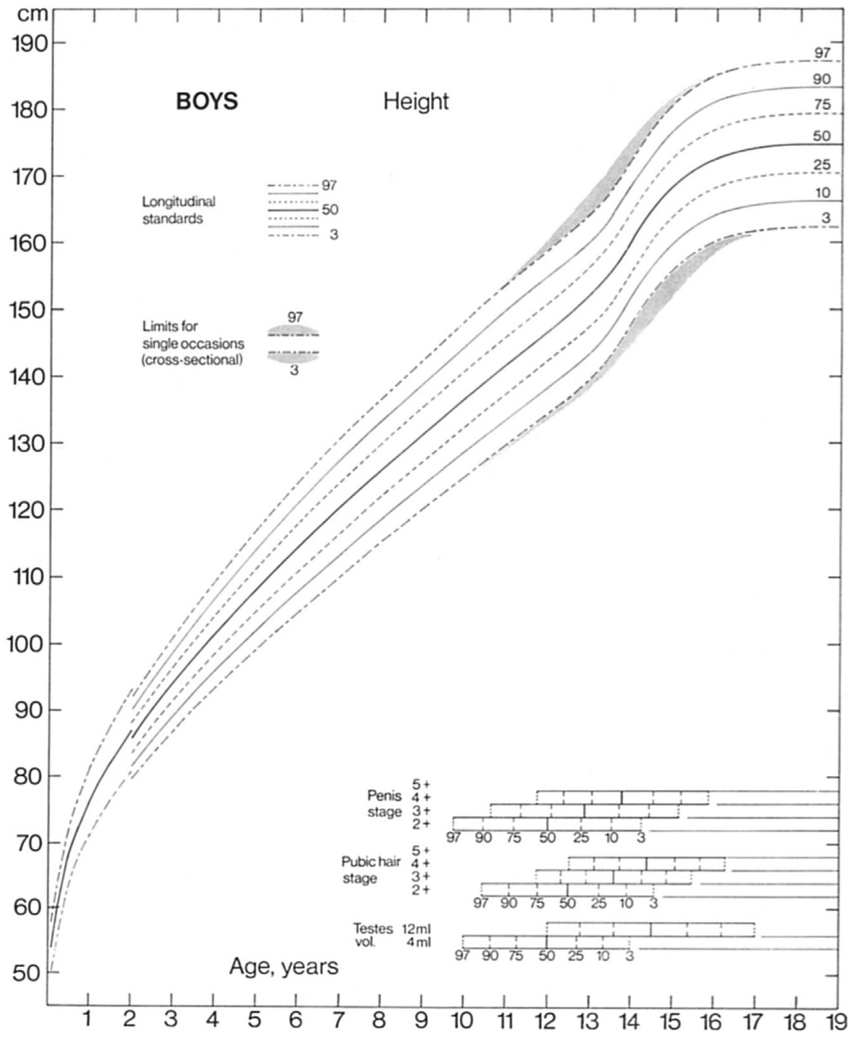

Fig. 2. Adapted from Fig. 1 of Tanner and Whitehouse (1976), here we see that height exhibits the same side-effect highlighted by Flynn's Age-Table Method applied to IQ subtests — mean and variance increase over time.

3. Part II: a meta-theory of intelligence

The Age-Table Method sets the table for Flynn's 'Meta-Theory of Intelligence', which comes in Part II of the book. In typically casual expository fashion, a lot goes on in Part II: an application of the Age-Table Method to another IQ test (Raven's Progressive Matrices), a rehash of the history and philosophy of astronomy, a defense of Flynn 1.0 from Jensen and his followers, and a review of new and old intelligence research. Here we'll focus on what we take to be most pertinent to philosophical perspectives of Flynn's latest contribution to the intelligence literature: an argument for what looks like pluralism in the guise of a "meta-theory" of intelligence.

The greatest challenge in reviewing Flynn's philosophy of intelligence, unfortunately, is understanding exactly where it lies. Like his numerical method discussed above, the metatheory of intelligence is nearly incoherent at times. Most notably, Flynn completely redefines standard terminology, such as 'metatheory' and 'heuristic', and then uses those terms in different ways, which often neither match standard definitions nor his own vague pronouncements. A metatheory isn't a theory about theories, it's a set of heuristics. Moreover, 'heuristics' aren't procedures, they're advice for theory builders. Type-one and Type-two errors aren't false positives and negatives, they're relations between scientists and Flynn's own ill-defined concepts (metatheories, heuristics, and errors). Then heuristics breed sub-heuristics, and Type-one and Type-two errors are extended to Type-three. Although once trained as a philosopher, Flynn makes a mistake frustratingly common among scientists-doing-philosophy: he writes as though he's trailblazing through an uncharted history and philosophy of science. Like his numerical methods, Flynn's inchoate ideas completely sidestep robust and coherent literature on the very concepts about which he writes.[5]

On a positive note, however, we think under the quasi-indecipherable and conceptually erratic prose is a point of real interest to philosophers of science. When all is said and done, the Meta-Theory of Intelligence sounds a lot like a call for philosophical pluralism about intelligence research. In developing his criticism of Jensen, for example, Flynn argues that "a heuristic should not be specific but broad enough to allow competing theories to fit under its umbrella and compete in terms of the evidence" (p. 113).

[5] See, for example, Wallis (2010) and Chow (2015) for analyses of 'metatheory' and 'heuristic' respectively. Flynn's versions of these concepts make no connection.

Elsewhere, "it is a mistake to use a precise theory-embedded concept to guide or unify intelligence research… We should make do with a heuristic and sub-heuristics broad enough to allow theories to freely compete at a 'lower' level" (p. 128). To our best approximation, Flynn's position is that research demands attention to distinct aspects of intelligence (differences within a cohort, trends over time, and brain physiology), each of which benefit from a free competition of theories and methodologies from multiple disciplines.

The *raison d'être* for Flynn's metatheory appears to be a rejection of Jensen's monism about intelligence research. The disagreement stems from an apparent incompatibility between Flynn 1.0 and the universally accepted empirical fact that all ability subtests are positively correlated. Those who score well on one subtest (e.g., vocabulary) also tend to score well on other subtests (e.g., block design). Spearman (1904) devised the statistical method of factor analysis to analyze this most basic empirical finding about intelligence tests, sometimes called the 'positive manifold.' Spearman demonstrated that to a considerable (but decidedly imperfect) degree, all of those positive correlations could be explained by the influence of a single unobserved latent construct to which they are all positively related. The nature of this latent construct, which Spearman called '*g*,' has been at the heart of the IQ question ever since.

The question about *g* is whether it is simply a statistical entity – a mathematical description of a matrix of positive correlations – as opposed to being an empirical instantiation of something deeper, more causal and more biological. Is *g* a *description* of the positive manifold, or does *g cause* the positive manifold? Stephen J. Gould (1981) endorsed the former position, arguing that Jensen had reified what is really only a statistical abstraction. In Flynn (2016) terms, Jensen reified Spearman's statistical concept by construing *g* as the 'irreplaceable fuel of intelligence' (p. 112). For Jensenites, reification warrants *g*'s use as a criterion of evidence for hypotheses about intelligence: if a phenomenon does not conform to the mathematics of *g*, it is not really about intelligence at all. The Jensen standard of evidence appeals to a factor analytic quantity called '*g*-loading.' The more positively correlated a subtest is with the general domain, the higher its *g*-loading. To use this as a criterion of evidence, one compares a ranking of IQ subtests by *g*-loading to a ranking of IQ subtests by some purported intelligence effect. Jensen used this technique to delegitimize the Flynn Effect, demonstrating that a ranking of IQ subtests by gains over time were not measures of *real* intelligence, as they fail to align with a ranking of the same IQ subtests by *g*-loading. In Flynn's terms, *g*-loading rendered evidence for the Flynn Effect 'hollow.'

Flynn's metatheory of intelligence comprises two responses. The first argument is historical. Tracing the history of astronomy, Flynn argues that 19th Century Newtonians who refused to consider theories that did not tally with Newton's mechanics were impeding scientific progress by failing to consider ideas that, later, would advance understandings of the world. Flynn's philosophy of science construes this failure as a "Type-three" error in which one elevates a theory-embedded concept to the level of metatheory. Thus Flynn's historical argument is that those who use *g* as a standard of evidence for intelligence research are committing a similar, type-three error. Although we are sympathetic to Flynn's call to refrain from using *g* as a standard of evidence, the historical argument from astronomy is unpersuasive. It simply does not follow that, because celestial mechanics should not have been used as a criterion of evidence in the 19th century, neither should *g* – or *any* theory-embedded concept for that matter - be used as a criterion of evidence today. Flynn's history of astronomy is brief and vague, spanning the course of a mere eight pages. It ignores a wealth of technical history and philosophy of science and astronomy that could have done a better job articulating scientific change (Kuhnian paradigms, for example).

Flynn's second line of criticism is more persuasive, as it highlights recent empirical evidence that conflicts with the notion that *g* could measure something biological in the brain. Flynn, te Nijenhuis, and Metzen (2014), for example, demonstrate that *g*-loadings fail to capture the difference between two sets of IQ tests *known* to be biological – a comparison of tests taken by those with normal brains and those suffering from various neurological disorders, such as prenatal cocaine exposure and traumatic brain injury. If *g* were biological, then the *g*-loading of tests taken by normal brains would match the *g*-loading of tests taken by neurologically dysfunctional brains – but they don't.

The positive element of Flynn's metatheory is a tripartite set of 'subheuristics' for three key areas of intelligence research: (1) individual differences within a cohort, (2) trends over time, and (3) brain physiology. Regarding differences within a cohort, he recommends that IQ instruments (i.e., IQ tests) cater to ability patterns of specific groups (i.e., subcultures) to maximize predictive success. Regarding trends over time, Flynn offers his own explanation for long-term national IQ gains: modernization presents cognitive challenges that has produced real changes in our abilities, but not smoothly across subtests: abstract thinking ability has increased greatly, but ability to do mental arithmetic (no longer an especially useful skill) has not. Regarding physiology, Flynn suggests feats of cognition are like feats of athleticism: performing cognitive tasks calls on resources from different aspects of the brain in the same way that performing various athletic tasks calls for orchestration of different muscle groups. Explaining human intelligence requires multiple theories and methodologies that cannot be contained within a monolithic *g*.

We're also optimistic that Flynn's metatheory of intelligence could advance recent discussions of pluralism by philosophers of science. Although Flynn's metatheory points to a pluralistic approach to intelligence research, he explicitly challenges any hopes for unification. We think this echoes the recent disagreement between *eliminative* and *integrative* pluralists. Does Flynn's account of intelligence research align with Mitchell (2003) and Tabery (2014) *integrative pluralism*, which leaves open the possibility that alternative approaches may eventually co-inform one-another to the point of unification? Or does it support Longino (2013), *ineliminative pluralism*, which maintains that the differences between the methods and assumptions of various approaches to human behavior are so great that there is no room for integration or unification? At the very least, Flynn's perspective sets the table for a philosophical assessment of pluralism in intelligence research.

4. Conclusion

Flynn 1.0 was conjured out of what seemed like an unlikely analysis of old standardization tables, but it worked: his observation of historical increases in the number of raw score items that were required to maintain a percentile standing in the scaled score distribution revolutionized modern thinking about intelligence. The simple observation of Flynn 1.0 — IQ scores are going up — was so compelling that uncertainty about why it occurred didn't place a serious limit on its importance. Flynn 2.0 dips into the same well but comes up empty: tracking age-related variance changes in standardization tables does nothing to identify the environmental effects of families on IQ. And the simple reason that Flynn 2.0 occurs — the SD of raw scores increases with age — isn't particularly interesting in and of itself, and can probably be explained as a statistical artifact anyway.

Nevertheless, Flynn uses his flawed new method to reach conclusions that are plausible on other grounds. We think he's right that family environment affects IQ during childhood, then fades out substantially but not completely. We also agree, *contra* Jensen, that *g* should not be used as a monistic standard of evidence for all intelligence research. (Although it is certainly the case that non-reified versions of *g* and other factor models have an important role to play.) Finally, we're sympathetic to a pluralistic metatheory: understanding intelligence requires orchestrated efforts from multiple fields, addressing multiple questions, and drawing on the resources of various technologies, instruments, and alternative theories.

Does your family make you smarter? Is an odd mix of data analysis and philosophy, but the two enterprises wind up having a lot in common. The style of both is charmingly colloquial and infuriatingly vague. They both, to their great detriment, blithely ignore their respective histories of methodological thinking. In important ways they are both, not to put too fine a point on it, incorrect. But Flynn's intuitions about intelligence are, as ever, scientifically and philosophically grounded and socially progressive. We wind up in the uncomfortable position of disagreeing with almost every detail he calls on to get him where he wants to go, and then ultimately agreeing with his conclusions when he finally gets there.

Acknowledgements

We're very grateful to Sarah Carroll and Evan Giangrande for helping us decipher the age-table method. We'd also like to thank Elliot Tucker-Drob, Paige Harden, and Jim Tabery for providing detailed feedback on an earlier draft of the manuscript. We acknowledge the support of the John Templeton Foundation to the Genetics and Human Agency project at the University of Virginia.

References

Chow, S. J. (2015). Many meanings of 'heuristic'. *The British Journal for the Philosophy of Science, 66*(4), 977–1016.

Flynn, J. R. (1984a). IQ gains and the binet decrements. *Journal of Educational Measurement, 21*(3), 283–290.

Flynn, J. R. (1984b). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*(1), 29–51.

Flynn, J. R. (1986). Sociobiology and IQ trends over time. *Behavioral and Brain Sciences, 9*(1), 192, 192.

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

Flynn, J. R. (2016). *Does your family make you smarter? Nature, nurture, and autonomy*. Oxford: Oxford University Press.

Flynn, J. R., te Nijenhuis, J., & Metzen, D. (2014). The G beyond Spearman's G: Flynn's paradoxes resolved using four exploratory meta-analyses. *Intelligence, 44*(May), 1–10.

Gould, S. J. (1981). *The mismeasure of man. Revised and Expanded*. New York: Norton.

Herrnstein, R. J., & Murray, C. A. (1994). *The bell curve: Intelligence and class structure in american Life*. New York: Simon & Schuster.

Jensen, A. (1969). How much can we boost IQ and scholastic achievement. *Harvard Educational Review, 39*(1), 1–123.

Kendler, K. S., Turkheimer, E., Ohlsson, H., Sundquist, J., & Sundquist, K. (2015). Family environment and the malleability of cognitive ability: A Swedish national home-reared and adopted-away cosibling control study. *Proceedings of the National Academy of Sciences, 112*(15), 4612–4617.

Longino, H. E. (2013). *Studying human behavior: How scientists investigate aggression & sexuality*. Chicago: The University of Chicago Press.

Mitchell, S. D. (2003). *Biological complexity and integrative pluralism*. New York, NY: Cambridge University Press.

Plomin, R., & Daniels, D. (1987). Why are children in the same family so different from one another? *Behavioral and Brain Sciences, 10*(1), 1–16.

Rowe, D. C. (1993). *The limits of family influences: Genes, experience, and behavior*. New York, NY: Guilford.

Spearman, C. (1904). 'General intelligence,' Objectively determined and measured. *The American Journal of Psychology, 15*(2), 201–292.

Tabery, J. G. (2014). *Beyond versus: The struggle to understand the interaction of nature and nurture. Life and Mind: Philosophical Issues in Biology and Psychology*. Cambridge, MA: MIT Press.

Tanner, J. M., & Whitehouse, R. H. (1976). Clinical longitudinal standards for height, weight, height velocity, weight velocity, and stages of puberty. *Archives of Disease in Childhood, 51*(3), 170–179.

Turkheimer, E. (1991). Individual and group differences in adoption studies of IQ. *Psychological Bulletin, 110*(3), 392–405.

Wallis, S. E. (2010). Toward a science of metatheory. *Integral Review, 6*(3), 73–120.

Please cite this article in press as: Matthews, L. J., & Turkheimer, E., Flynn, the Age-Table Method, and a metatheory of intelligence, Studies in History and Philosophy of Biological and Biomedical Sciences (2017), http://dx.doi.org/10.1016/j.shpsc.2017.07.003