Intelligence 82 (2020) 101481

Contents lists available at ScienceDirect

Intelligence

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

The Flynn effect can become embedded in tests: How cross-sectional age norms can corrupt longitudinal research

Patrick O'Keefe*, Joseph Lee Rodgers

Vanderbilt University, Nashville, Tennessee

A B S T R A C T

The Flynn Effect (FE; Flynn, 1984, 1987) is the decades-long increase in measured mean IQ of approximately 1/3 point per year, observed in industrialized nations over the course of at least a century. An obvious and practical implication of the FE is that the FE can cause test norm obsolescence. If norms from 1970 were used today, the average score would be approximately a standard deviation above the original mean. A more subtle effect was suggested by Mingroni (2007): Age-normed tests could have a FE "built-in" through the norming process. His observation can be true in any case where there are cohort differences (between- or within-family); it is almost certain to occur in cases where cross-sectional samples are used to age norm in the presence of cohort effects. We illuminate this process in several ways, because it can significantly impact longitudinal research. If the "built in FE" hypothesis is supported, then the FE potentially affects resulting scores assigned to test-takers from all age-normed cognitive tests exhibiting a FE. A series of graphic simulations demonstrate the logic. Following, analysis of the National Longitudinal Survey of Youth Children data suggest that the Flynn Effect is indeed built into the PIAT-Math scores.

1. Introduction

The Flynn effect (e.g., Flynn, 1984, 1987; Lynn, 1983) is the longitudinal increase in measured IQ over the past century. Many mechanisms/potential theories have attempt to explain the Flynn effect, including improved nutrition (Lynn, 2009), improved test taking strategies (Brand, 1987), changes in life history (Woodley, 2012), improved niche picking (Dickens & Flynn, 2001), and heterosis (Mingroni, 2007). More recently, a peak and subsequent decline in the Flynn effect has been noted in several countries (see Teasdale & Owen, 2008 for early documentation, and Dutton, van der Linden, & Lynn, 2016, for review). A number of authors have noted methodological challenges associated with studying the Flynn Effect (e.g., Rodgers, 1998; Rodgers, 2015). Starting with Flynn's (1984) original paper, studying the Flynn effect is associated with an important set of questions: Is the increase "real", i.e., is it the result of actual increases in cognitive ability? More recently, is the decline noted in some countries an indication of a true turn-around in cognitive ability? Ultimately, can we identify the cause(s), and will that identification have implications for how we evaluate and understand human intelligence?

A subtle and problematic methodological mistake in the behavioral science literature is to interpret a cross-sectional age effect as though it is a longitudinal within-person pattern. This mistake can easily happen when respondents are individuals of different ages obtained and measured in a cross-section. Interpreting cross-sectional age comparisons as though they are age-related within-person changes to be expected across time is an incorrect logical inference. Nevertheless, the psychological (and other) literature is filled with examples of that incorrect logic. The primary reason this logic is fallacious is that at a given point in time, 20- year-olds have aged through an entirely different world than 40-year-olds or 60-year-olds. If the 20-year-olds of 40 years ago are different from today's 20- year-olds, we obviously can't expect tomorrow's 60-year-olds to be the same as today's. Importantly for the present paper, the cross-sectional methods that assume age differences are the same as age changes are highly related to the same methods used to age norm testing instruments.

There are many reasons that persons of different ages are expected to differ, beyond developmental differences. Schaie (1986, 1994) identified a number of processes potentially causing these differences, including changes in childrearing practices, improved health care, changes in the educational system, and shifts in public policy innovations. The Flynn effect can be shown on logical and empirical grounds to cause confounding in cross-sectional studies of aging. Dickinson and Hiscock (2010) compared 20-year-olds to 70-year-olds in terms of verbal and performance-related IQ scores on different versions of the WAIS. Of course, the 20-year-olds performed better; the typical interpretation would attribute that finding to the well-documented decline in IQ (and subscales) that occurs within individuals as they age. But Dickinson and Hiscock posited that the Flynn effect might be at work. Their estimates show that around 85% of the supposed within-person decline in performance was due to the Flynn effect, not within-person change.

* Corresponding author.
E-mail address: Patrick.okeefe@vanderbilt.edu (P. O'Keefe).

https://doi.org/10.1016/j.intell.2020.101481
Received 13 February 2020; Received in revised form 23 July 2020; Accepted 27 July 2020

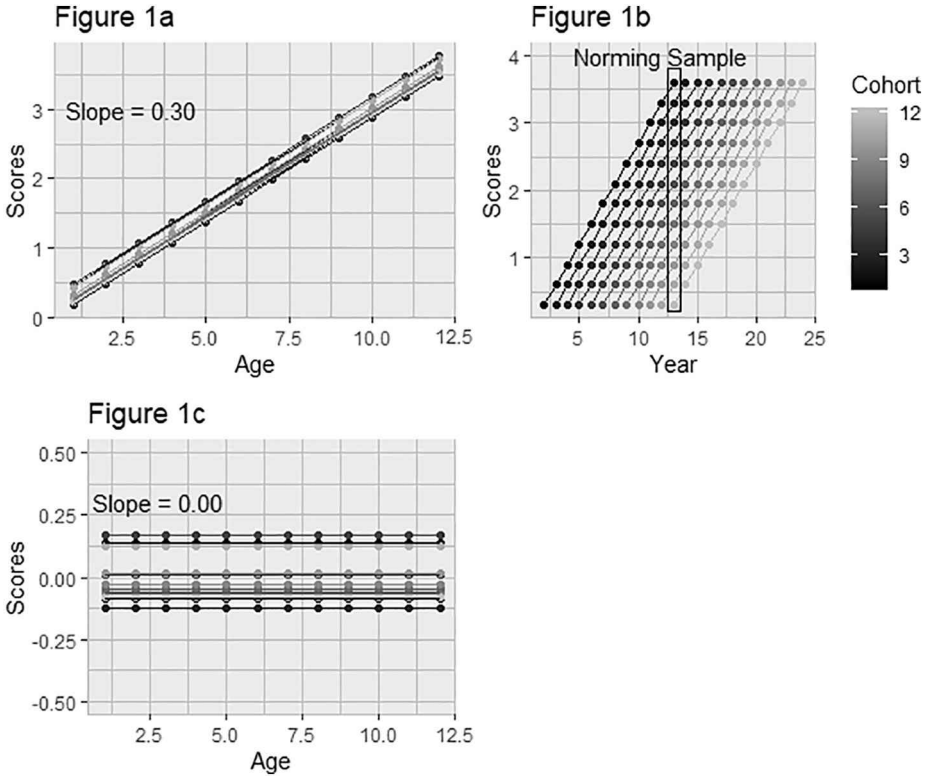

Fig. 1. This figure shows the results of test norming for a test that measures development and has no between cohort differences. Any differences in the plot are due either to age or jittering.

Mingroni (2007) noted how the Flynn effect could result in age differences that are independent of actual developmental differences:

> "…consider a test like the WISC (Wechsler, 1991), which has age appropriate norms for children between the ages of 6 and 16. If these norms were generated in, say, 1970, then the 6-year-old norm would likely have been derived from children in the 1964 birth cohort, but the 16-year-old norm would have been derived from the 1954 birth cohort. A 3-point Flynn effect between the two cohorts would make the 6-year-old norm 3 points more difficult, on average, because it was generated using a later born, and hence higher performing, cohort." (Mingroni, 2007).

In other words, Mingroni's suggestion implies that a cohort difference due to the Flynn effect could be embedded within a test that has been age normed using a cross-sectional sample. As a result, apparent differences across ages, across groups, or due to treatments, could be at least partially reflecting the Flynn effect, and not the processes that were studied. In a similar vein, Rodgers (2014) noted that the well-researched negative birth order relationship with intelligence can also be an artifact caused by the Flynn effect, with cross-sectional research confounding cohort differences with age differences. Our goal in the current paper is to evaluate empirically whether the effect proposed by Mingroni can be observed within the PIAT-Math instrument, taking advantage of the knowledge that the Flynn effect is already identified in this measure.

We know empirically from Dickinson and Hiscock (2010) and Rodgers (2014) that some well-known patterns are at least partially artifacts of the Flynn effect. We know from Mingroni that there are conceptual reasons to believe that the Flynn effect may be embedded within well-known intelligence (and other) instruments. In what instruments? Can it be identified empirically? How does it become embedded within a test? We address these questions in this paper.

2. Background

In this section, we present two motivations for our empirical analyses. First, we discuss conceptually how the Flynn effect can become embedded within a test. Second, we present several graphical results from simulations that show empirically how this embedding process can occur. To show that it can occur does not necessarily mean that it has or routinely will occur (though it implies that it likely has occurred). In the empirical part of our paper, we demonstrate that such embedding can be observed in a well-known dataset.

2.1. Age norms

Across ages, we expect differences in cognitive performance due to development. If an IQ or an achievement test is administered to kindergartners, 6th graders, and 12th-graders, we expect different performance at different ages. Age norms are designed to equate scores across ages. A 6-year-old who scores 100 on an IQ instrument is considered to have scored at approximately an average level; a 12-year-old who scores 100 on the same instrument is also considered to be approximately average. But because of development, a higher performance level is required to be average for a 12-year-old than for a 6-year-old. But by using age norms, researchers can more effectively compare performance across ages.

A "built-in" embedded Flynn Effect comes about because the mean differences between age group performance is assumed to be caused by development – refer to that developmental difference as *d*. This value *d* is represented as a positive mean shift for an older age groups, and is what age norms are supposed to account for. However, there is another factor at play: the negative effect for older cohorts caused by the Flynn Effect. We will call this change *f*. During most of the past century, the Flynn effect has caused scores to move in the opposite direction to development (although recently, the Flynn effect in some locations may have reached an asymptote, and even turned slightly negative; see

Dutton et al., 2016). When norms are created using cross-sectional data, instead of correcting solely for development, *d,* the actual correction will be *d-f*, the observed ability difference between younger and older cohorts. This re-norming process will have the effect of biasing age-based norms downward (in areas with a negative Flynn Effect the impact is reversed). The existence of the *f* component, combined with the age-norming process, is the component to which Mingroni (2007) referred. But few researchers or test users seem to be aware of the difference between d and d-f.

2.2. Graphical Portrayal

We present several figures, based on the simple model above, to visualize the effects previously described. In the first set of patterns, assume a true, linear, within-person effect (development, *d*). There is no systematic between-cohort effect (in graphing, the cohorts are jittered to allow the plots to show each cohort separately). In Fig. 1a individual's scores are plotted against individual ages. Each line represents a single individual, and each individual is from a unique cohort. Differences between cohorts are due to jittering in the plot, and are what cause the separation between individual lines in Fig. 1a. In Fig. 1b we see what each individual's scores look like if we plot over period (year) instead of over age. Each individual has identical within-person patterns of effects, and scores increase linearly with age and time; the increase in scores due to *d*. Assume that individuals are measured in a single year, year 13. If we wished to adjust for age effects in the typical way, we could use this cross-section to establish age norms. Fig. 1c shows the age-normed scores. The third panel (Fig. 1c) shows what we would usually hope to achieve with age-norming, with the within-person slope of 0. This pattern reflects the reality that, apart from development, a person's scores are not changing over time.

The panels in Fig. 2 show a different pattern. This case illustrates a systematic cohort difference of 0.30 points (a Flynn effect) for each year later that a person was born, what we refer to above as *f*. However, there is no within-person change, *d*, as we can see in Fig. 2a (note that, as in Fig. 1, each line represents an individual, and shading distinguishes cohorts). In Fig. 2b we have plotted each person's score by year of observation. In year 13 we have measured an observation from each cohort, which also corresponds to an observation for each age. We can norm our scores using this sample (as we would in a standard cross-sectional age norming), resulting in Fig. 2c. In Fig. 2c a spurious within-person effect has arisen because we have built-in the cohort differences into our age norms. It appears that as individuals age they score better and better on our test. In this case development, *d*, is zero and has no impact on a person (Fig. 2a), and over time an individual has no other change (Fig. 2b), yet if we observed a person over time using age norms we would see apparent change within-person (Fig. 2c) because of *f*.

The third set of plots demonstrates what we expect in the case of a cohort based Flynn effect, *f*, combined with true development, *d*. Within-person there is an effect of age of 0.50, we consider this development *d* (Fig. 3a). There is also an additional effect for cohort, *f*, seen in the separation of cohorts in Fig. 3a. This cohort effect is such that being born one year later results in a 0.3 point increase in scores. This scenario is essentially a combination of the effects from the previous two sets of figures. Again in year 13 there is overlap between all the cohorts and we can "norm" the scores by age (Fig. 3b). As happened in the pure cohort effect case there is still a within-person effect after age norming the tests. This again is the built-in Flynn effect, which we see in Fig. 3c. In fact, the observed within-person change in scores is exactly equal to the between-cohort difference of 0.3. This is the pattern of effects that we believe to be most plausible (i.e., true development mixed with cohort effects).

A reviewer noted that the Flynn effect may not be strictly linear. This is in line with current research (e.g., Pietschnig & Voracek, 2015), although we note that in reported meta-analyses the Flynn effect is still largely monotonically increasing. However, the non-linearity of the Flynn effect has little bearing on the present discussion, as non-linear effects are as easily built into a test as linear ones.

The logical conclusion of the simple model and graphical results presented in this section is that, in the presence of a cohort based secular trend like the Flynn effect, age normed tests that use cross-sectional data will inevitably "build-in" the effect into the age norms. The argument for a built in Flynn effect is based on logic; the purpose of this paper is to develop this logic, and to demonstrate a case using empirical analysis in which this result has actually happened. The methodological basis for our study emerges from the presentations above.

We will search empirically for an embedded Flynn effect by looking at within-person changes in the NLSY PIAT-Math data. An embedded

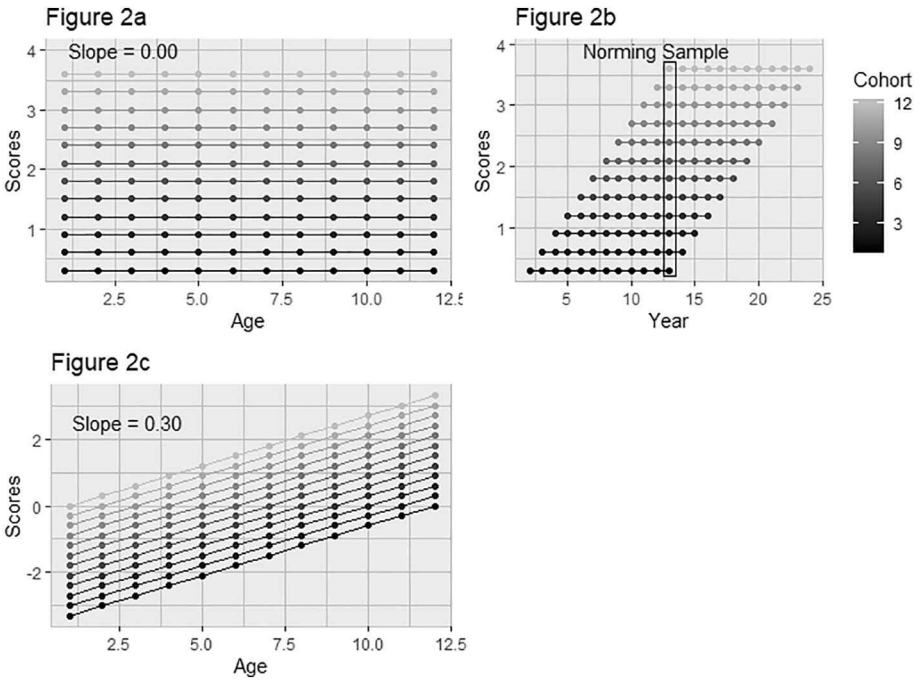

Fig. 2. This figure shows a test with no within-person development, but between-cohort differences. The result of norming is a within-person slope.

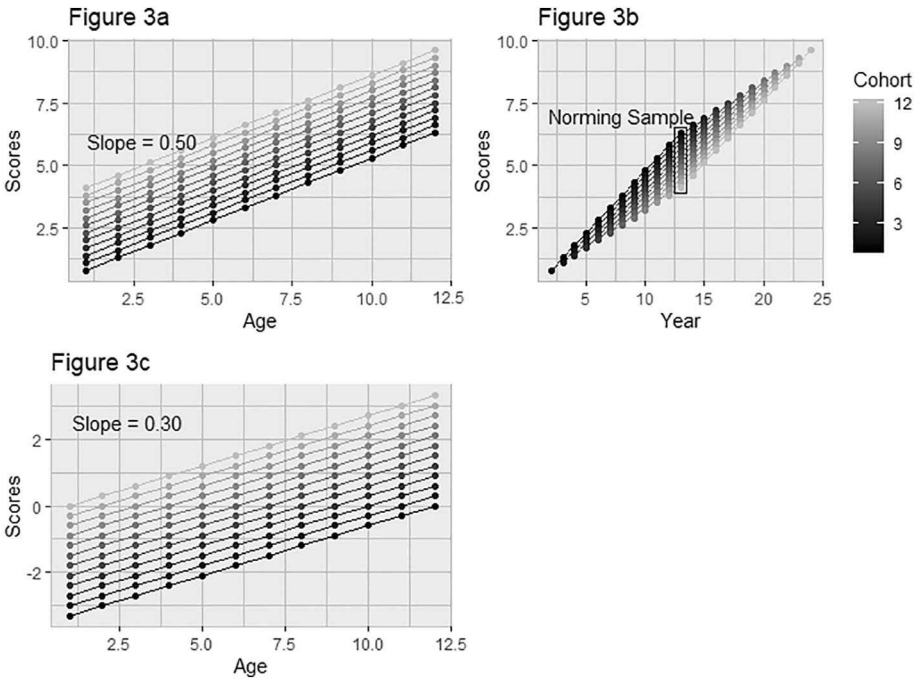

Fig. 3. This figure shows a test with both a developmental effect and a between-cohort effect. After norming a within-person effect is observed due to the cohort differences.

Flynn effect will show up across an individual's scores over time. However, there exist other explanations for a within-person increase, and these explanations should be evaluated as well. First, development could exist, however the age norming is supposed to eliminate the development that occurs in within-individual scores, although if age norms were set incorrectly development may not be fully accounted for. Second, there can be practice effects, caused by "learning a test," if individuals take the test more than once. Thus, in our analysis – defined in the Methods Section below – we will outline two analytic efforts: (1) We will evaluate whether our cognitive ability measure has a within-person trend; and (2) We will rule out two alternative explanations (development and practice) for that within-person trend.

3. Methods

3.1. Participants

Participants were respondents from the National Longitudinal Survey of Youth – Children (NLSYC; Bureau of Labor Statistics, 2014b) survey between the ages of 5 and 13 during the survey years from 1986 through 2014. These children were the children of women from the National Longitudinal Survey of Youth 1979 (NLSY79; Bureau of Labor Statistics, 2014a) sample. The NLSY79 sample was a household probability sample representative of the adolescent population of the United States on December 31, 1978. In addition, the survey design included an oversampling of poor, minority and military respondents in the NLSY79 sample, whose children are represented in the NLSYC sample until 1992 (at which point many of the respondents were dropped for budgetary reasons). The inclusion of the non-representative portion of the sample in our analyses is acceptable because, although there may be mean differences between the groups, age-based differences should be largely unaffected. The primary analyses occur within person and these should not be substantially affected if there are (non-cohort based) group mean differences; only age based differences should affect these within-person results. Informed consent was obtained from all participants at the beginning of the study in accordance with the U.S. Office of Management and Budget and the U.S. Bureau of Labor Statistics. Analysis of the publically available, deidentified, data was declared to be IRB exempt by the Vanderbilt University IRB.

In the present sample there were 9173 children with PIAT-Math scores, with an average of 3.34 observations per child, resulting in 30,664 total observations. In the child sample, 49.0% of the children were female, 51.0% were male. The child's race and ethnicity was based on their mother's report, with three categories: Hispanic, Black, and not Hispanic/not Black. In our sample 20.6% were Hispanic, 30.5% were Black and the remaining 48.9% were non-Hispanic-non-Black. The age of participants was limited between 5 and 13 inclusive, through the design of the NLSYC survey. Data have been collected biennially starting in 1986. For the PIAT-Math instrument, data are available from 1986 through 2014.

3.2. Measures

The primary measure of interest in this study is the Peabody Individual Achievement Test – Math subscale (PIAT-M; Dunn & Markwardt, 1970). The PIAT-Math is an age normed test of mathematical reasoning and ability for children. The original norms were used throughout the whole NLSYC survey (for comparability), despite the introduction of new norms in 1986. The new norms were introduced too late to be incorporated at the beginning of the study and for consistency the old norms were used throughout. In the present study only the PIAT-M scores, child's year of birth, and the year of test administration for each PIAT-M score, are needed. The primary focus regards within-person change; higher level variables are generally unnecessary for answering this specific question.

3.3. Design

This study requires an age-normed measure exhibiting the Flynn effect (in order for there to be a built-in Flynn effect, there must first be a Flynn effect). The PIAT-M satisfies this goal, as it is an age-normed test with the means for every age set equal. Further, the PIAT-M has shown a Flynn Effect in past research (see, Ang, Rodgers, & Wänström, 2010; Rodgers & Wänström, 2007). In addition to the measure, longitudinal data are necessary to test the hypothesis that individual scores increase as children grow older, and pass through age norms. The

National Longitudinal Survey of Youth-Children (NLSYC) provides an ideal dataset of more than 30,000 PIAT-M observations. We outlined three potential causes of a within person increase in PIAT-M scores: a built-in FE, inadequate norming to wash out a developmental effect, and practice effects. Each of these effects may be evaluated within the design of the NLSYC. Prior to testing for additional effects we will ensure that the age norms are set properly. If the age norms were set improperly (e.g., the mean score of 13 year olds was set too high), an apparent within-person increase in scores could occur. We can easily remedy the problem by creating our own sample-based age norms.

3.4. Models

The following models illustrate the multi-level analysis that will be used to answer the questions necessary to determine whether or not the Flynn effect was built-in to the NLYSC PIAT-Math. The explanation of these models is relatively brief; further discussion can be found in the appendix. The first two models (Model 1 & 2) test for an age effect within-individual. This is the most basic requirement. If there is no within-individual age effect at all then it is not relevant to attempt to extract a cause for the change. The two models differ in that the first includes all random effects (slope and intercept) whereas the second only includes a random intercept. The different models were included for practical purposes as in the actual model fitting we encountered estimation difficulties with the original (Model 1) formulation.

Model 1: Multilevel model of the overall age effect, with random and fixed effects. Score is PIAT-Math score, age is age at measurement.

$$score_{ijk} = \beta_{0jk} + \beta_{1jk} * age_{ijk} + e_{ijk}$$

$$\beta_{0jk} = \beta_{00k} + u_{0jk}$$

$$\beta_{00k} = \beta_{000} + u_{00k}$$

$$\beta_{1jk} = \beta_{10k} + u_{1jk}$$

$$\beta_{10k} = \beta_{100} + u_{10k}$$

$$score_{ijk} = \beta_{000} + (\beta_{100} + u_{1jk} + u_{10k}) * age_{ijk} + e_{ijk} + u_{0jk} + u_{00k}$$

$$e_{ijk} \sim N(0, \sigma^2)$$

$$U_2 \sim MVN\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\tau^2_{00} & \\ \tau_{10} & \tau^2_{11}\end{bmatrix}\right)$$

$$U_3 \sim MVN\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\tau^2_{000} & \\ \tau_{100} & \tau^2_{110}\end{bmatrix}\right)$$

Model 2: Multilevel model of the overall age effect, with random intercept and fixed slopes. Score is PIAT-Math score, age is age at measurement.

$$score_{ijk} = \beta_{0jk} + \beta_{100} * age_{ijk} + e_{ijk}$$

$$\beta_{0jk} = \beta_{00k} + u_{0jk}$$

$$\beta_{00k} = \beta_{000} + u_{00k}$$

$$score_{ijk} = \beta_{000} + \beta_{100} * age_{ijk} + e_{ijk} + u_{0jk} + u_{00k}$$

$$e_{ijk} \sim N(0, \sigma^2)$$

$$u_{0jk} \sim N(0, \tau^2_{00})$$

$$u_{00k} \sim N(0, \tau^2_{000})$$

The third model (Model 3) attempts to evaluate whether or not the age effect is due to typical development. If development is the sole cause of the age effect then we would expect that, all else being equal, older children should perform better than younger children and older families should perform better than younger families. This analysis, using the same data, has already been reported in O'Keefe and Rodgers (2017); only the within-person age effect was present. Here we present a condensed version of that model looking only at age effects. Age is group mean centered within each child and then the child mean ages are group mean centered within family; group mean centering is indicated by dots above the relevant variable. To identify that development is the cause of the within-person effect we would expect significant effects for all components of the age variable (means and mean centered components), and we would expect those values to be approximately equal to the magnitude of the overall Flynn effect.

Model 3: Multilevel model of Piat-Math scores by age. Age is group mean centered at the child and family level. This model has a random intercept and fixed slopes.

$$\dot{age}_{ijk} = age_{ijk} - (\dot{age}_{.jk} + \overline{age}_{..k})$$

$$\dot{age}_{.jk} = \overline{age}_{.jk} - \overline{age}_{..k}$$

$$score_{ijk} = \beta_{0jk} + \beta_{100} * \dot{age}_{ijk} + \beta_{200} * \dot{age}_{.jk} + \beta_{300} * \overline{age}_{..k} + e_{ijk}$$

$$\beta_{0jk} = \beta_{00k} + u_{0jk}$$

$$\beta_{00k} = \beta_{000} + u_{00k}$$

$$score_{ijk} = \beta_{000} + \beta_{100} * \dot{age}_{ijk} + \beta_{200} * \dot{age}_{.jk} \beta_{300} * \overline{age}_{..k} + e_{ijk} + u_{0jk} + u_{00k}$$

$$e_{ijk} \sim N(0, \sigma^2)$$

$$u_{0jk} \sim N(0, \tau^2_{00})$$

$$u_{00k} \sim N(0, \tau^2_{000})$$

Models 4 through 7 attempt to disentangle practice effects. If there were a practice effect, then because of the design of the NLSY-C, there are two age groups whose level of practice diverges over time. The five and six year olds have never taken the test before and this is true in every interview period. Conversely, in the first year of interviews the 12 and 13 year olds have also never taken the test, but with each passing year they have subsequently more practice (the 13 year olds in the second round of surveys had taken the test at 11, in the third round of surveys the 13 year olds had taken the test at ages 9 and 11 etc.) This feature of the design means that, if practice is the true cause of the within-person effect, then the five year olds and the 13 year olds should have different slopes. This is simply a preliminary model and so only the 5 and 13 year olds are used as they provide the longest timeframe to examine (Model 4). Models 5 through 7 use an explicit measure of practice effect (the count of previously administered tests), and look to see if any practice effect is reliable across levels. A solely within-person practice effect without higher order effects cannot be a practice effect (but can be explained by a built-in Flynn effect).

Model 4: Model of the PIAT-Math looking for interaction effect between age at observation and year of observation. Random intercept and fixed slopes are modeled.

$$score_{ijk} = \beta_{0k} + \beta_{10} * age_{ik} + \beta_{20} * Year_{ik} + \beta_{30} * age_{ik} * Year_{ik} + e_{ik}$$

$$\beta_{0k} = \beta_{00} + u_{0k}$$

$$score_{ijk} = \beta_{00} + \beta_{10} * age_{ik} + \beta_{20} * Year_{ik} + \beta_{30} * age_{ik} * Year_{ik} + e_{ik} + u_{0k}$$

$$e_{ijk} \sim N(0, \sigma^2)$$

$$u_{0k} \sim N(0, \tau^2_{00})$$

Model 5: Model of PIAT-Math residualized on year of observation. This model directly evaluates practice effects using a count of prior testing administrations. Random intercept and fixed slopes are

modeled.

$$residual\ score_{ijk} = \beta_{0jk} + \beta_{100} * practice_{ijk} + e_{ijk}$$

$$\beta_{0jk} = \beta_{00k} + u_{0jk}$$

$$\beta_{00k} = \beta_{000} + u_{00k}$$

$$residual\ score_{ijk} = \beta_{000} + \beta_{100} * practice_{ijk} + e_{ijk} + u_{0jk} + u_{00k}$$

$$e_{ijk} \sim N(0, \sigma^2)$$

$$u_{0jk} \sim N(0, \tau^2_{00})$$

$$u_{00k} \sim N(0, \tau^2_{000})$$

Model 6: The model is similar to model 5, except that practice effects are group mean centered at the child and family level. Random intercept and fixed slopes are modeled.

$$\dot{practice}_{ijk} = practice_{ijk} - (\ddot{practice}_{.jk} + \overline{practice}_{..k})$$

$$\ddot{practice}_{.jk} = \overline{practice}_{.jk} - \overline{practice}_{..k}$$

$$residual\ score_{ijk} = \beta_{0jk} + \beta_{100} * \dot{practice}_{ijk} + \beta_{200} * \ddot{practice}_{.jk} + \beta_{300} * \overline{practice}_{..k} + e_{ijk}$$

$$\beta_{0jk} = \beta_{00k} + u_{0jk}$$

$$\beta_{00k} = \beta_{000} + u_{00k}$$

$$residual\ score_{ijk} = \beta_{000} + \beta_{100} * \dot{practice}_{ijk} + \beta_{200} * \ddot{practice}_{.jk}\beta_{300} * \overline{practice}_{..k} + e_{ijk} + u_{0jk} + u_{00k}$$

$$e_{ijk} \sim N(0, \sigma^2)$$

$$u_{0jk} \sim N(0, \tau^2_{00})$$

$$u_{00k} \sim N(0, \tau^2_{000})$$

Model 7: Model of the PIAT-Math using centered practice variable and year of observation. Random intercept and fixed slopes are modeled.

$$score_{ijk} = \beta_{0jk} + \beta_{100} * \dot{practice}_{ijk} + \beta_{200} * \ddot{practice}_{.jk} + \beta_{300} * \overline{practice}_{..k} + \beta_{400} * Year + e_{ijk}$$

$$\beta_{0jk} = \beta_{00k} + u_{0jk}$$

$$\beta_{00k} = \beta_{000} + u_{00k}$$

$$score_{ijk} = \beta_{000} + \beta_{100} * \dot{practice}_{ijk} + \beta_{200} * \ddot{practice}_{.jk}\beta_{300} * \overline{practice}_{..k} + \beta_{400} * Year + e_{ijk} + u_{0jk} + u_{00k}$$

$$e_{ijk} \sim N(0, \sigma^2)$$

$$u_{0jk} \sim N(0, \tau^2_{00})$$

$$u_{00k} \sim N(0, \tau^2_{000})$$

4. Results

All analyses were conducted using R and applicable R packages. In particular, the lme4 package was used for multilevel analyses (Bates, Mächler, Bolker, & Walker, 2014). All multilevel models utilized restricted maximum likelihood (reml) for model fitting. Code for the analyses is available in appendix A. The data as structured for these analyses are freely available from the author upon request, and also are freely available on the internet from Center for Human Resource Research, who manages the NLSY datasets. The data can be easily accessed by googling "NLSY, Bureau of Labor Statistics", or by going to nlsinfo.org.

The first analysis in this paper will establish age norms for the PIAT-Math. The PIAT-Math has pre-existing age norms. However, to account for the possibility that the norms may be off slightly, due either to random fluctuations in the original norming sample or to a poorly executed norming procedure, we used the NLSYC to reset the norms. The renorming process used the observations from the 1986 survey. This was the first year of the NLSYC survey and thus avoids any practice effects (and it also avoids the inclusion of the same individual twice or more, which would occur if the whole dataset were used). The means and standard deviations for all age groups between 5 and 13 were used to age standardize all observations in the data set. Thus, all 5 years olds were norm referenced to the age 5 sample mean/standard deviation in 1986; all 6 years olds were norm referenced to the age 6 mean/standard deviation in 1986; etc. There were 1874 observations available as a norming sample with a minimum of 32 observations for the 13-year-old group and a maximum of 462 observations in the six-year-old group. The median number of observations per group was 176. As expected (due if for no other reasons to developmental increase in math ability) the mean raw score increased monotonically across these ages. Although it is possible that this norming could result in unaccounted disparities between age groups in ability, unrelated to the Flynn Effect, there are two reasons to believe that this will not significantly impact our conclusions here. First the original principle is still the same: norms based on cross-sectional observations do not account for ability differences between cohorts in addition to developmental differences, and will result in an apparent within-person effect regardless of any real existence of that effect. Second, the work in O'Keefe and Rodgers (2017) already demonstrated, using the original PIAT-Math norms, that the Flynn Effect operates in such a way that we would expect cohort differences in ability. We use both the original age norms and the renormed version of the PIAT-Math (scaled to have a baseline mean of 100 and baseline standard deviation of 15) in all subsequent analyses as a check to ensure that an improper age norming process is not the primary driver of the relevant results.

After renorming the PIAT-Math, all available data were analyzed using multilevel modeling for children ages 5–13 looking at the possibility of an age effect. Using two models (Model 1 & Model 2, as specified above), one with fixed and random slopes for age and one with only fixed effects, the age effect was statistically significant ($t > 8.5$, $p < .001$). Two models were used because the model with random slopes produced software warnings. Results for fixed effects for model 2 can be found in Table 1. Although we omit the effects for model 1 (because of the issues in computing the model) the effects that were recovered were numerically nearly identical to the effects reported in

Table 1
Model 2, fixed age effect, estimation results.

	Effect		Standard error		t-statistic		*p* value	
	Original norms	NLSY norms	Original norms	NLSY norms	Original norms	NLSY norms	Original norms	NLSY norms
Intercept	98.46	105.83	0.26	0.34	376.80	310.90	< 0.001	< 0.001
Age	0.29	0.28	0.02	0.03	12.92	9.44	< 0.001	< 0.001

Table 2
Model 3, centered age model, estimation results.

	Effect		Standard error		t-statistic		*p* value	
	Original norms	NLSY norms	Original norms	NLSY norms	Original norms	NLSY norms	Original norms	NLSY norms
Intercept	101.01	108.22	0.17	0.21	604.46	508.36	< 0.001	< 0.001
WP age	0.31	0.34	0.02	0.03	12.92	11.12	< 0.001	< 0.001
BP age	0.16	−0.56	0.12	0.16	1.29	−3.54	> 0.05	< 0.001
BF age	−0.47	−1.19	0.16	0.21	−2.88	−5.65	< 0.01	< 0.001

Table 1. In both cases the age effect was approximately 0.3 points per year of age, virtually identical to the commonly cited size of the Flynn Effect.

Further extending the model without random slopes to include child and family mean centered ages (Model 3) did not change the within-person effect for age substantially, although the child and family level variables had negative coefficients with t values less than −3.5 in the NLSY normed sample (Table 2). Overall these results indicate that there is a positive within person effect for age even after the scores have been set to have the same mean and variance for all ages. Importantly, the lack of positive effects at the family and child level suggests that this is not due to incorrect norms or child development. If the within-person result was due to incorrect norms or development we would expect a child and family level effect as well (i.e., older children scoring better on average, and families with older children scoring better).

The second component of this study tests for practice effects using two methods. The first tests for a positive age by year interaction among 5 and 13 year-old participants from 1986 to 1992. Because age 5 is the first interview, a practice effect is implausible for 5 year olds, and any slope corresponds only to the FE. Between the years of 1986–1992 each subsequent cohort of 13 year olds had more test experience than the previous cohort (on average), their slope is a combination of the FE and increased practice effects. To maintain independence between the two groups, the 1994 cohort of 13 year olds is omitted because it consists of the 5 year olds from 1986. This structure also implies that we should use a two, not three, level model (we only need to account for non-independence due to family clustering, not repeated measurement). If practice effects exist they would manifest as an interaction, and the slope for 13 year olds would be steeper than that of 5 year olds (a positive interaction). A multilevel model (Model 4; Table 3) accounting for family clustering found a small negative interaction ($t = -2.26$). This result implies that there is not a detectable practice effect in this analysis.

A second test for a practice effect uses the whole dataset. By counting the number of times a given child is administered the PIAT-Math we can obtain a direct estimate of their practice. The drawback to this method is that children may have uneven spacing between practice rounds if they miss a survey round. The first analysis involved residualizing the normed scores on year using a linear regression model. It makes sense to control for year of testing because the Flynn Effect is known to exist in these scores and because only with subsequent rounds could children have practice. It would be misleading to not control for the passage of time when accounting for practice in light of the Flynn Effect. Because these scores are age normed, it makes little sense to residualize on age. If a count of the number of times a child had taken the PIAT-Math (Model 5; Table 4) is included in a three level model there is a slight negative effect for practice, and if that same variable is group mean centered (Model 6; Table 5) the effect is still negative within person. A final model (Model 7; Table 6) does not residualize PIAT-Math scores on year of testing but instead includes it as a control variable. In this model year of testing has a significant positive effect of about 0.39 ($t = 14.98$, $p < .001$; a Flynn effect) and the within person component of practice is negative, showing no practice effect ($t = -1.17$, $p > .05$), however, this effect was statistically significantly different from what would be predicted if practice was accounting for the within person effect, an effect of at least 0.60, $t = 8.46$, $p < .001$. The effect must be at least 0.60 because the year over year gain is 0.30. With biennial testing the gain for each increment of practice must be at least double the annual gain. It should be noted that in models six and seven, which group mean centered practice, there was a positive effect for practice at the group levels (family and child), however this effect would not explain a within person effect, only a mean difference between children and families. For the present analysis these findings are uninformative.

Although renorming could account for issues with the original norms, it could also introduce its own biases. Most importantly, the NLSY sample is somewhat select, particularly in the earlier years (which we use to norm the test). The oldest children were born to mothers who, on average, had children at earlier ages. Children from younger mothers are known to have lower scores on average on a test such as the PIAT-Math compared to children of older mothers. This bias could tilt the scales towards younger children having more difficult norms than older children, confounding our norms with the potential Flynn effect. First, we would note that this is exactly the kind of built in effect that we are attempting to observe, so even if this contamination did occur, although it may not prove our point for the Flynn effect it does demonstrate the general principle that a norming artifact can induce apparent (but spurious) growth within person. More importantly however, when we replicated all the above analyses using the official norms (i.e., using the reported standardized scores from the NLSY database), the conclusions were identical.

5. Discussion

These findings demonstrate that in the NLSYC PIAT-M data, where sample norms were computed to eliminate any concerns over developmental changes, there is not a reliable practice effect, but there is a remaining within-person increase in scores. The conclusion, by

Table 3
Model 4, age by year interaction effect, estimation results.

	Effect		Standard error		t-statistic		*p* value	
	Original norms	NLSY norms	Original norms	NLSY norms	Original norms	NLSY norms	Original norms	NLSY norms
Intercept	92.18	104.67	0.85	586.18	107.82	−3.91	< 0.001	< 0.001
Age	4.98	−4.50	1.68	2.26	2.96	−1.99	< 0.01	< 0.05
Year	1.34	1.21	0.22	0.29	6.13	4.09	< 0.001	< 0.001
Age:Year	−1.44	−1.15	0.38	0.51	−3.84	−2.26	< 0.001	< 0.05

Table 4
Model 5, practice effect, estimation results.

	Effect		Standard Error		t-statistic		*p* value	
	Original norms	NLSY norms	Original norms	NLSY norms	Original norms	nlsy norms	Original norms	NLSY norms
Intercept	1.14	1.37	0.19	0.24	6.02	5.62	< 0.001	< 0.001
Practice	−0.31	−0.41	0.05	0.06	−6.71	−6.64	< 0.001	< 0.001

Table 5
Model 6, centered practice effect, estimation results.

	Effect		Standard error		t-statistic		*p* value	
	Original norms	nlsy norms	Original norms	NLSY norms	Original norms	NLSY norms	Original norms	NLSY norms
Intercept	−1.78	−2.50	0.71	0.92	−2.51	−2.73	< 0.05	< 0.01
WP practice	−0.36	−0.51	0.05	0.06	−7.49	−8.06	< 0.001	< 0.001
BP practice	0.65	2.57	0.32	0.42	2.03	6.11	< 0.05	< 0.001
BF practice	0.71	−0.71	0.44	0.58	1.59	−1.23	> 0.05	> 0.05

Table 6
Model 7, centered practice effect and year of observation, estimation results.

	Effect		Standard error		t-statistic		*p* value	
	Original norms	NLSY norms	Original norms	NLSY norms	Original norms	NLSY norms	Original norms	NLSY norms
Intercept	−521.80	−669.48	39.62	51.40	−13.169	−13.02	< 0.001	< 0.001
WP practice	−0.005	−0.10	0.063	0.08	−0.07	−1.17	> 0.05	> 0.05
BP practice	0.47	2.37	0.32	0.42	1.49	5.68	> 0.05	< 0.001
BF practice	1.20	−0.15	0.45	0.58	2.68	−0.25	< 0.01	> 0.05
Year	0.31	0.39	0.02	0.03	15.56	14.98	< 0.001	< 0.001

elimination, is that the Flynn Effect is built into the age normed tests with the PIAT-Math. This finding is (to our knowledge) the first direct empirical support for the hypothesis that Mingroni proposed in 2007, and that we developed conceptually and graphically above. (We note that during the review process, a reviewer discussed the logical nature of the built-in Flynn effect, and stated that "it is practically inconceivable that it has not occurred.") The implications are that, within-person, year-over-year gains on tests that are both age normed and show a Flynn Effect, are likely lower than otherwise believed. Dickinson and Hiscock (2010) showed a powerful demonstration of the magnitude of this inflation, suggesting that in their study of the WAIS up to 85% of the apparent within-person decline on IQ test raw scores between ages 20 and 70 was due to the embedded Flynn effect, and not actual within-person decline.

This inflation could have enormous implications for research on education and cognitive development. Researchers may have to re-evaluate many past findings that suggest within-child improvement in scores. Somewhat paradoxically this effect could even extend to old age, as the issues with norm scores would persist wherever there are cohort differences. This argument implies that when using cross-sectional raw scores looking at old age cognitive decline, decline may be overestimated (as demonstrated by Dickinson & Hiscock, 2010), and yet when using longitudinal normed scores a researcher may make the opposite error. At the policy level, policymakers will need to consider that efforts to increase scores within children (e.g., improve a child's test score over time) that at first appear to be beneficial may be partly or even almost entirely illusory, or, more accurately, due not (only) to their interventions but (at least partially) to the Flynn effect. (Rodgers, 2014, demonstrated a form of this illusion in relation to past birth order findings.)

There are multiple solutions to the testing problem of a built-in Flynn effect, though none are ideal. Researchers should carefully evaluate for themselves, within the goals of their research program, the best way to control for age effects. Should researchers wish to age norm their tests and hope to avoid incorporating the Flynn effect into those norms we offer some suggestions. The first is to renorm tests annually (or at least very often) so that every age group is being compared to its own cohort and not a previous cohort. This approach is likely to be prohibitively expensive (and also very time consuming), except for large testing companies. It should be noted that some of the major scholastic achievement tests (e.g., the GRE, SAT, and ACT) do appear to follow this rule, with students compared against people who took the test the same year they did. Although these are not IQ tests per se, the standard methods of renorming tests are shared between IQ and achievement tests. Furthermore, we used an achievement test in the current study; our results generally are applicable to both IQ and achievement tests.

Alternatively, instead of renorming a test annually, researchers could norm the test over the span of time representing the age groups it is meant to study. In this scheme a test covering ages 5–10 would obtain the norms for age five in year one, age six in year two etc. This plan might be less expensive than annually producing norms, however it is far more time consuming than any other suggestion we provide. It also has the added drawback that, given that the Flynn Effect causes norm obsolescence over time, the length of time that a norm will be useful is shortened by the amount of time it takes to produce the set of norms.

The final alternative is to do a mathematical adjustment in datasets with a built-in Flynn Effect of about 0.02 standard deviations (the 0.3 IQ point increase per year, divided by the 15 point standard deviation on most IQ tests) for every year of age increase. We emphasize that this adjustment only applies to the PIAT-Math during this time period. There is an obvious challenge associated with developing the appropriate adjustment for each of many different tests at many different points in time. Researchers may find that using the findings of recent meta analyses (e.g., Pietschnig & Voracek, 2015; Trahan, Stuebing, Fletcher, & Hiscock, 2014) provides another (and perhaps more precise) adjustment. This approach is virtually identical to the approach used by Dickinson and Hiscock (2010). In addition to the challenges noted

above, using this type of correction may not be accurate for a given test, and current research suggests that the Flynn Effect is not necessarily consistent across nations and may be reversing direction in some areas (e.g., Dutton et al., 2016). Further, the use of a purely linear adjustment will not precisely reflect the nature of a built-in Flynn effect to the extent that the Flynn effect is non-linear for the cohorts used to norm the test.

For research already conducted and data already collected the most practical option would appear to be a post-hoc correction of within-person data. If a test is age normed and shows the Flynn Effect, we would expect a person to show a within person increase in scores equivalent to the Flynn Effect caused by the re-norming process (i.e. an increase of approximately 0.3 IQ points per year). The problem can largely be avoided entirely if there is an appropriate control group with longitudinal data as well. In that case researchers can determine if gains on a test are "real" by comparing the within-person slope of the treatment group to the control group. However, this approach necessitates longitudinally following a control group, and simple baseline measures will be insufficient.

Other methods of norming tests (e.g., IRT methods) do not avoid the problems presented here. In IRT for example, the age related difference in theta scores in the norming sample, if that sample were cross-sectional, is a combination of both a positive developmental influence and a negative Flynn Effect influence. The effects are completely confounded in any cross-sectional data and can only be parsed using longitudinal data.

We conclude and summarize by warning researchers who use one or more tests that are re-normed at multiple ages across some time interval, and which are known to show a Flynn Effect, that there may be methodological issues associated with using those tests. We have demonstrated that the re-norming process can potentially imbed the Flynn Effect into the test because of the age differences in ability that are inherent in the individuals used to norm the tests in a cross-sectional sample. Further, we analyzed the PIAT-Math scores in the NLSY, and ruled out other interpretations of the empirical within-person effect that can be identified within those scores. This finding suggests that the Flynn Effect is indeed contained within the PIAT-Math scores themselves, due to age norming over time.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.intell.2020.101481.

References

Ang, S., Rodgers, J. L., & Wänström, L. (2010). The Flynn effect within subgroups in the US: Gender, race, income, education, and urbanization differences in the NLSY-children data. *Intelligence, 38*, 367–384.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. ArXiv:1406.5823 [Stat]. Retrieved from http://arxiv.org/abs/1406.5823.

Brand, C. (1987). Bryter still and bryter. *Nature, 328*(9), 110.

Bureau of Labor Statistics, U.S. (U.S., 2014a). Department of labor, and national institute for child health and human development. *National Longitudinal Survey of youth 1979 cohort, 1979–2014 (rounds 1–26). Produced and distributed by the center for human resource research*. Columbus, OH: The Ohio State University.

Bureau of Labor Statistics, U.S. (U.S., 2014b). Department of labor, and national institute for child health and human development. Children of the NLSY79, 1979–2014. *Produced and distributed by the center for human resource research*. Columbus, OH: The Ohio State University.

Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review, 108*(2), 346–369. https://doi.org/10.1037/0033-295X.108.2.346.

Dickinson, M. D., & Hiscock, M. (2010). Age-related IQ decline is reduced markedly after adjustment for the Flynn effect. *Journal of Clinical and Experimental Neuropsychology, 32*(8), 865–870.

Dunn, L. M., & Markwardt, F. C. (1970). *Peabody individual achievement test manual* (1st ed.). American Guidance Service, Inc.

Dutton, E., van der Linden, D., & Lynn, R. (2016). The negative Flynn effect: A systematic literature review. *Intelligence, 59*, 163–169. https://doi.org/10.1016/j.intell.2016.10.002.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*(1), 29–51. https://doi.org/10.1037/0033-2909.95.1.29.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*(2), 171.

Lynn, R. (1983). IQ in Japan and the United States shows growing disparity. *Nature, 306*, 291–292.

Lynn, R. (2009). What has caused the Flynn effect? Secular increases in the development quotients of infants. *Intelligence, 37*(1), 16–24. https://doi.org/10.1016/j.intell.2008.07.008.

Mingroni, M. A. (2007). Resolving the IQ paradox: Heterosis as a cause of the Flynn effect and other trends. *Psychological Review, 114*(3), 806–829. https://doi.org/10.1037/0033-295X.114.3.806.

O'Keefe, P., & Rodgers, J. L. (2017). Double decomposition of Level-1 variables in multilevel models: An analysis of the Flynn effect in the NSLY data. *Multivariate Behavioral Research, 52*(5), 630–647. https://doi.org/10.1080/00273171.2017.1354758.

Pietschnig, J., & Voracek, M. (2015). One century of global IQ gains: A formal meta-analysis of the Flynn effect (1909–2013). *Perspectives on Psychological Science, 10*, 282–306.

Rodgers, J. L. (1998). A critique of the Flynn effect: Massive IQ gains, methodological artifacts, or both? *Intelligence, 26*, 337–356.

Rodgers, J. L. (2014). Are birth order effects on intelligence really Flynn effects? Reinterpreting Belmont and Marolla 40 years later. *Intelligence, 42*, 128–133.

Rodgers, J. L. (2015). Methodological issues associated with studying the Flynn effect: Exploratory and confirmatory efforts in the past, present, and future. *Journal of Intelligence, 3*, 111–120.

Rodgers, J. L., & Wänström, L. (2007). Identification of a Flynn effect in the NLSY: Moving from the center to the boundaries. *Intelligence, 35*(2), 187–196. https://doi.org/10.1016/j.intell.2006.06.002.

Schaie, K. W. (1986). Beyond calendar definitions of age, time, and cohort: The general developmental model revisited. *Developmental Review, 6*(3), 252–277.

Schaie, K. W. (1994). The course of adult intellectual development. *American Psychologist, 49*(4), 304–313.

Teasdale, T. W., & Owen, D. R. (2008). Secular declines in cognitive test scores: A reversal of the Flynn effect. *Intelligence, 36*(2), 121–126. https://doi.org/10.1016/j.intell.2007.01.007.

Trahan, L. H., Stuebing, K. K., Fletcher, J. M., & Hiscock, M. (2014). The Flynn effect: A meta-analysis. *Psychological Bulletin, 140*, 1332–1360.

Wechsler, D. (1991). *The Wechsler Intelligence Scale for Children – Third Edition.* San Antonio, TX: Psychological Corporation.

Woodley, M. A. (2012). A life history model of the Lynn–Fynn effect. *Personality and Individual Differences, 53*(2), 152–156. https://doi.org/10.1016/j.paid.2011.03.028.