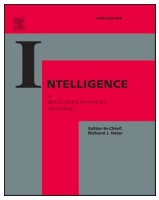
Intelligence 98 (2023) 101734

Contents lists available at ScienceDirect

Intelligence

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



Looking for Flynn effects in a recent online U.S. adult sample: Examining shifts within the SAPA Project

Elizabeth M. Dworak [a,*], William Revelle [b], David M. Condon [c]

[a] Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, United States of America
[b] Department of Psychology, Northwestern University, Evanston, IL, United States of America
[c] Department of Psychology, University of Oregon, Eugene, OR, United States of America

ARTICLE INFO

Keywords:
Flynn effect
Intelligence
Cognitive ability
International Cognitive Ability Resource
Synthetic Aperture Personality Assessment Project

ABSTRACT

Compared to European countries, research is limited regarding if the Flynn effect, or its reversal, is a current phenomenon in the United States. Though recent research on the United States suggests that a Flynn effect could still be present, or partially present, among child and adolescent samples, few studies have explored differences of cognitive ability scores among US adults. Thirteen years of cross-sectional data from a subsample of adults (n = 394,378) were obtained from the Synthetic Aperture Personality Assessment Project (SAPA Project) to examine if cognitive ability scores changed within the United States from 2006 to 2018. Responses to an overlapping set of 35 (collected 2006–2018) and 60 (collected 2011–2018) items from the open-source multiple choice intelligence assessment International Cognitive Ability Resource (ICAR) were used to examine the trends in standardized average composite cognitive ability scores and domain scores of matrix reasoning, letter and number series, verbal reasoning, and three-dimensional rotation. Composite ability scores from 35 items and domain scores (matrix reasoning; letter and number series) showed a pattern consistent with a reversed Flynn effect from 2006 to 2018 when stratified across age, education, or gender. Slopes for verbal reasoning scores, however, failed to meet or exceed an annual threshold of |0.02| SD. A reversed Flynn effect was also present from 2011 to 2018 for composite ability scores from 60 items across age, education, and gender. Despite declining scores across age and demographics in other domains of cognitive ability, three-dimensional rotation scores showed evidence of a Flynn effect with the largest slopes occurring across age stratified regressions.

1. Introduction

Labeled the Flynn effect (Herrnstein and Murray, 2010), intelligence quotient (IQ) scores substantially increased since 1932 and through the twentieth century, with differences ranging from 3.0 to 5.0 IQ points (0.20 to 0.33 SD) per decade (Flynn, 1984, 1987, 2007). These findings imply younger generations are expected to have higher IQ scores than the previous cohort. For example, if we tested a sample of Baby Boomers (born between 1946 and 1964) when they were 20 years old and compared their scores on the same test to a sample of Millennials (born between 1981 and 1996) tested at age 20, we would expect the latter group's IQ scores to be between 0.66 and 1.1 SD higher. This isn't to say that the sample of Millennials are smarter or more able than the group of Baby Boomers, but that a difference in scores exists favoring the younger generation. These results, however, should prompt other important questions – what demographic factors are contributing to the difference? Do these results generalize across adulthood? How long should the trend of increasing scores be expected to persist? Does this trend still exist in the United States?

1.1. Domains of intelligence

While brief conceptual definitions are provided regarding domains of intelligence, we recognize and caution that the distinction between any number of components of cognitive ability is difficult to disentangle from measurement and testing. This is especially true as most performance-based assessments will capture multiple, possibly overlapping, cognitive processes; thus, often making these theoretical distinctions somewhat obtuse. Regardless, we provide this information to help familiarize those less experienced with theories of intelligence and provide the necessary framework for understanding separate domains discussed throughout this manuscript.

* Corresponding author.
E-mail address: elizabeth.dworak@northwestern.edu (E.M. Dworak).

https://doi.org/10.1016/j.intell.2023.101734
Received 31 March 2022; Received in revised form 28 January 2023; Accepted 31 January 2023
Available online 2 March 2023

Researchers often differentiate intelligence as fluid and crystallized reasoning (Carroll, 1993; Cattell, 1943, 1963; Horn and Cattell, 1966; McGrew, 2009; McGrew and Wendling, 2010). Fluid intelligence or more abstract reasoning is often described as an individual's ability to abstractly reason and solve problems. On average, fluid ability peaks in adulthood at age 25 and steadily decreases (Salthouse, 2010, 2012, 2019). In comparison, crystallized intelligence is understood as knowledge that is accumulated and learned overtime. Normally, crystallized intelligence increases, on average, until age 60 and then declines (Salthouse, 2010, 2012, 2019). Given that these average peaks are based on cross-sectional data, it should be noted that variation exists around these peaks due to individual differences.

Although the fluid-crystalized model, or Cattell–Horn model, is dominant across the cognitive ability field, some researchers argue that the structure of ability is split into the two components of v:ed (verbal/educational) and k:m (spatial/mechanical) (Vernon, 1965) or three components of verbal, perceptual, and image rotation (Johnson and Bouchard, 2005). Verbal intelligence and v:ed are comparable to crystalized intelligence in that they rely on knowledge gained over time. These domains include verbal fluency, general knowledge gained through schooling, and arithmetic/numerical abilities. Perceptual intelligence and k:m are similar to fluid intelligence in that they rely on reasoning through tasks and problem solving skills such as two-dimensional spatial ability, memory, perceptual speed, and some numerical abilities. Finally, image rotation intelligence is an individual's ability to mentally rotate an object to solve a puzzle. For Vernon's verbal-perceptual model, this third category of image rotation is often grouped with k:m as it covers spatial tasks.

Despite the impressive differences in IQ scores represented by the Flynn effect, research has shown that higher IQ scores in more recent cohorts are most likely driven by boosts in fluid intelligence scores. Specifically, higher scores have been observed across nonverbal tasks such as the Raven's Progressive Matrices; which also contains a spatial component; and the more verbally dependent Wechsler Intelligence Scale for Children's (WISC) and Wechsler Adult Intelligence Scale's (WAIS) Similarities subtest (Ceci and Kanaya, 2010; Flynn, 1999; Sundet, Barlaug, and Torjussen, 2004; Weiss, 2010). Likewise, Pietschnig and Voracek (2015) found fluid intelligence scores had the largest difference from 1909 to 2013, with more recent scores being higher, compared to other domains of cognitive ability. Following the largest differences of fluid intelligence scores, studies have found that spatial/rotation ability and crystalized intelligence scores have also observed notable differences over the last century, with more recent scores being higher than older scores (Pietschnig and Voracek, 2015; Weiss, 2010); though the magnitude of higher crystalized intelligence scores for more recent cohorts are thought to have plateaued around 1987 (Pietschnig and Voracek, 2015). Taken together, these studies imply that rather than all domains of intelligence equally being on the rise, the magnitude for the differences in scores often discussed for the Flynn effect over the last century are more related to problem solving skills and abstract reasoning.

1.2. Critiques of the Flynn effect representing changes in intelligence

Though these findings are substantial, researchers such as Jensen argued these higher scores likely did not reflect true gains in intelligence but rather the Flynn effect was a difference between the latent and observed scores (Jensen, 1998; Rushton and Jensen, 2010). That is, even though observable IQ scores could be shown to be increasing, this did not necessarily indicate that latent general intelligence was also increasing. Moreover, while Jensen recognized that scores on fluid intelligence tasks were increasing, their overall *g*-ness or factor loadings on *g* did not correlate with these increases (Jensen, 1998). Thus, he posited higher observed scores were likely an artifact and either simply due to participants becoming more familiar with testing or tests losing their *g*-ness. Other researchers have critiqued Flynn's work by noting his findings were inflated and required additional corrections (Kaufman, 2010).

Rodgers (1998) questioned if higher scores for newer generations reflected a true rise in mean IQ scores or if the Flynn effect simply captured a change in variability. Specifically, he posited that a change in variance, such as decreased variance in the lower tail of the distribution and an increase in variability in the upper tail, might produce the results observed by Flynn (1984, 1987). For example, Teasdale and Owen (1989) found their results were heavily driven by increasing scores in their lower tail whereas Zhou, Zhu, and Weiss (2010) found results varied across level of ability. However, in examining Flynn's previous findings, Rodgers was unable to determine if the Flynn effect was due to a change in tail variability or differences across the distribution.

Despite these results, researchers have also posited if the observed Flynn effect could also be an artifact due to measurement invariance across different standardized test. That is, could the differences in scores by ability level be due to tests functioning differently between groups? Interested in examining this question, Benson, Beaujean, and Taub (2015) explored if various Wechsler tests were invariant and concluded that only some versions were. These findings have implications for Flynn effect results that rely on comparisons between tests that contain different items, as changes in scores cannot be distinguished between the assessment, the cohort, or their interaction.

Rodgers (1998) would ultimately propose 10 areas of research that would improve the understanding of the Flynn effect. These suggestions ranged from better understanding the Flynn effect across various background demographics and generalizability to questions about its persistence overtime and what changes or differences in IQ truly means. Though some of these questions remain unanswered, the Rodgers' outlined questions would prompt further criticisms and investigations into the overall Flynn effect.

1.3. Reverse Flynn effect

While it's appealing to think that the human IQ could be higher with each generation, Flynn (2007) admitted that these gains would not go on forever; nor did these differences in scores necessarily reflect greater mental ability of younger cohorts. Using data from the WAIS, WAIS Revised (WAIS-R), and WAIS Third Edition (WAIS-III) norming samples, Russell (2007) estimated that Full Scale IQ would likely plateau by 2024 if these samples were truly representative of the overall population. However, due to the introduction of more stringent exclusion criteria for the norming sample of the WAIS-III, Russell (2007) estimated that the scores from the WAIS-III norming sample were likely inflated compared to previous norming samples. Thus, he anticipated that the Flynn effect of Full Scale IQ scores could plateau as early as 2004.

Mirroring these estimates, research and meta-analyses over the last two decades suggest that the Flynn effect had already stagnated or begun to reverse. In a meta-analysis examining IQ scores across 31 countries from 1909 to 2013, Pietschnig and Voracek (2015) found that the magnitude of higher IQ scores observed for newer cohorts has declined. Dutton and Lynn (2013) found Finnish IQ scores had differed −2.0 IQ points (0.13 SD) from 1997 to 2009, while French IQ scores differed −3.8 IQ points (0.25 SD) from 1999 to 2009 (Dutton and Lynn, 2015); for these studies, more recent samples had lower IQ scores than previous samples. In a meta-analysis examining nine original studies that observed a reverse Flynn effect, differences ranged between −0.38 IQ points (0.03 SD) and −4.3 IQ points (0.29 SD) per decade (Dutton, van der Linden, and Lynn, 2016). Recent evidence within German-speaking countries, also suggests that the magnitude of higher visual-spatial ability scores in newer cohorts could be declining across certain regions of Europe (Pietschnig and Gittler, 2015).

1.4. Recent work in the United States and demographic differences

With previous literature primarily focused on European countries,

often due to the large-scale cognitive ability data collected by a nation's military, research examining whether the Flynn effect or its reversal is a current phenomenon in the United States is limited. The most notable studies using adult samples include the original investigations conducted by Flynn (1984, 1987, 2007, 2009) using norming data. When examining the standardized sample scores from the systematic re-norming of intelligence tests between 1932 and 1984, Flynn (1984) uncovered that scores of more recent samples were higher than those in older samples. Specifically, the results of Flynn's study indicated that the accelerated scores between cohorts were largely consistent regardless of age. Flynn (1987, 2007, 2009) would reproduce these results using new samples and data from other countries. While the results of this study inform and inspire the current research, we focus on summarizing subsequent research within the United States. These studies include a meta-analysis (Trahan, Stuebing, Fletcher, and Hiscock, 2014), research using child and adolescent samples (Ang, Rodgers, and Wänström, 2010; Giangrande, Beam, Finkel, Davis, and Turkheimer, 2022; O'Keefe and Rodgers, 2017; Platt, Keyes, McLaughlin, and Kaufman, 2019; Rodgers and Wänström, 2007; Shakeel and Peterson, 2022), and a study examining vocabulary scores in adults (Twenge, Campbell, and Sherman, 2019).

With respect to gaining a comprehensive understanding to the extent and magnitude of the Flynn effect, a meta-analysis examined studies and test manuals containing norming data collected between 1951 and 2010 within the United States and United Kingdom (Trahan et al., 2014). This study found that IQ scores were rising on average 2.31 points (0.15 SD) per decade or 2.93 points (0.195 SD) per decade after excluding older data. In addition to the overall findings, the meta-analysis found evidence the Flynn effect was consistent across age, level of ability, sample type (test manuals vs. research study), order of administration, and test pairings; though, ability level showed mixed results when limited to only examining the subsamples of modern data. In their conclusion, Trahan et al. (2014) note that the United States may not be experiencing the reversal observed in Scandinavian countries due to differences in social policies and/or educational values.

In a series of studies using data from the National Longitudinal Survey of Youth (NLSY), researchers have examined if the Flynn effect was present for children and adolescents (5- to 13-year-olds) across a series of fluid and crystallized cognitive assessments before and after controlling for maternal IQ (Ang et al., 2010; O'Keefe and Rodgers, 2017; Rodgers and Wänström, 2007). Using scores from the Wechsler Memory of Digit Span, Peabody Picture Vocabulary Test, and Peabody Individual Achievement Test subscales of Math, Reading Recognition, and Reading Comprehension, Rodgers and Wänström (2007) found that between 1986 and 2000 differences in scores generally remained stable or decreased after entering maternal IQ as a covariate. The exception to these findings was that math scores for the Peabody Individual Achievement Test were higher for newer test periods by an average of 0.23 points per year for each age group and an average of 0.30 points per year for 9- to 13-year-olds.

To further understand the results of this study, Ang et al. (2010) later investigated if demographic factors contributed to the observed rise in Peabody Individual Achievement Test Math scores using the 1986 to 2004 NLSY data. Using a series of regressions, this study examined if the Flynn effect was observed across the demographic categories of gender, race, maternal education, income, and location (urban/rural). Results indicated that regardless of demographic subsamples, increasing scores were present at varying rates for each subgroup. However, the authors note that the largest and most stable divergence in slopes across each age were for models that included mother's highest level of education; with higher educational attainment relating to greater magnitudes of coefficients. Expanding this work even further, O'Keefe and Rodgers (2017) applied double decomposition and multilevel modeling to the NLSY data collected between 1986 and 2012. The results of this study showed both within- and between-person gains in math scores with the largest difference of scores reaching 0.33 points per year; though the authors concluded that these differences in scores were predominantly observed between-family.

Building upon the work completed by O'Keefe and Rodgers (2017), Giangrande et al. (2022) used scores on various Wechsler tests (WISC, WISC-R, and WISC-III) to investigate if the Flynn effect was present for children and adolescents (7- to 15-year-olds) in the Louisville Twin Study. Using scores collected across a span of four decades, this study found evidence of a Flynn effect at both the level of within-and between-person was present in a United States sample between 1957 and 1999. Specifically, verbal, performance (i.e., visual spatial), and full-scale IQ scores showed an average difference of 0.20 SD per decade (3 points per decade; or approximately 0.02 SD per year). In another study using a collection of math and reading scores for children and adolescents (approximately for ages 9-, 13-, 15-, and 17-year-olds and those in grades 4, 8, and 12), Shakeel and Peterson (2022) examined if the Flynn effect was present across a series of large-scale achievement tests including the National Assessment of Educational Progress (NAEP), the Long-Term Trend NAEP, Trends in International Math and Science Study, Progress in International Reading Literacy Study, and the Program for International Student Assessment. This study found that the direction and magnitude of differences in scores varied by test with magnitudes of differences diminishing over time, with greater differences being present for math than reading scores. In addition to these results, Shakeel and Peterson (2022) observed that the magnitude and direction of differences often varied by age, ethnicity, and SES. As observed by Ang et al. (2010), they also found that that scores across gender were relatively consistent.

Using nonverbal scores from the Kaufman Brief Intelligence Test from between 1988 and 1989 and between 2001 and 2004, Platt et al. (2019) were unable to find a Flynn effect across their entire sample. Likewise, they did not find evidence of differences between demographic categories (i.e., gender, parental education, geographic region) for their observed effects. In further exploring the tails of fluid ability across age, this study found fluid ability were higher for younger adolescents and for those with higher ability (0.11 and 0.167 SD per decade) but were lower for older adolescents and those with lower ability (−0.076 and −0.233 SD per decade). This suggests that the Flynn effect may no longer generalize across all ages or levels of ability. Alternatively, as fluid ability remained stable across the sample these results could indicate that the Flynn effect has simply stagnated or at least plateaued in the United States. Despite these results, it's important to note that these findings are limited as fluid ability does not developmentally peak until young adulthood (Salthouse, 2010, 2012, 2019) and the authors note that changes are often limited during adolescence.

More recently, Twenge et al. (2019) found that vocabulary scores were lower for more recent adult samples, tested between 1974 and 2016, regardless of educational attainment (approximately between −0.047 SD to −0.126 SD per decade). To further understand what could account for the observed differences in vocabulary scores, they also applied a hierarchical age-period-cohort analysis controlling for education, where age was the participants age, period was the year when the participant took the measure, and cohort was the participant's year of birth (Fosse and Winship, 2019; Yang and Land, 2006). The results of this age-period-cohort analysis indicated that there were large effects for time period and age, but a lack of cohort effects. After controlling for education, age, and cohort, period effects showed approximately a 0.08 SD decrease in vocabulary scores per decade. This indicated that the decline in scores was related to a change over time that uniformly affected all groups (i.e., age, cohort, education). Cohort, on the other hand, showed a decrease, indicating that differences on vocabulary scores were less dependent on birth cohort differences. Vocabulary scores declined for participants ages 50 and older, indicating that they followed a similar maturation to crystalized intelligence. The results of this study ultimately supported previous conclusions that gains in crystalized intelligence scores may have plateaued around 1987 (Pietschnig and Voracek, 2015). Because vocabulary scores are often

classified as a measure of crystallized intelligence, it is likely that the lower vocabulary scores are contributing to the recent diminished differences of crystalized intelligence scores. While this study relied on a crystallized measure of intelligence, given the sparse research on the Flynn effect in the United States using an adult sample, this study provides insight into how at least one component of cognitive ability may currently differ with previous scores within the United States.

Taken together, this collection of studies prompts the question, will the Flynn effect observed for children and adolescents in the United States generalize to an adult sample as observed by Trahan et al. (2014)? Likewise, will previous results regarding demographic subgroups also be present when considering the education of an individual rather than parental education? Will a lack of gender difference still be present for an adult sample?

1.5. Current research

Using a set of public-domain cognitive ability measures, the present study aims to investigate the evidence for differences in cognitive ability scores in the United States between 2006 and 2018. We used 13 years of cross-sectional data from the Synthetic Aperture Personality Assessment Project (SAPA Project; Condon and Revelle, 2016; Revelle et al., 2017), a free web-based survey, to test if there was a Flynn effect for adult participants. Starting in 2006, 35 ability items, that would become part of the International Cognitive Ability Resource (ICAR; Condon and Revelle, 2014, 2016; Dworak, Revelle, Doebler, and Condon, 2021; Revelle, Dworak, and Condon, 2020), were administered. These items were used to form a composite cognitive ability score or domain scores for matrix reasoning (11 items), letter and number series (8 items), and verbal reasoning (16 items). Starting in 2011 an additional item for letter and number series and 24 three-dimensional rotation items began to be administered, with the original items, to allow for a 60-item composite score. Unlike previous Flynn effect studies, participants were disproportionately female identifying (65.03%) and between the ages of 18 and 90.

A preregistration for this study can be found on the Open Science Framework (OSF) at https://osf.io/kmgx8. All deviations from this preregistration, such as the inclusion of one month of additional data and exploratory analyses, are detailed throughout the manuscript.

2. Methods

2.1. Database

The archival data used in this study were collected through the SAPA Project (https://www.sapa-project.org/; Condon and Revelle, 2015, 2016; Condon, Roney, and Revelle, 2017; Revelle et al., 2017), a free web-based personality survey that uses stratified matrix sampling methodology to administer items. Since 2004, the SAPA Project has successfully collected cross-sectional data from over 1.5 million participants across the world. Generally, participants have found the survey through varying mechanisms such as search engines, posts on social media, or websites related to personality or psychometrics. While the breadth of what data the SAPA Project collected has grown over the last 19 years, its approach has largely remained the same in that the items administered to participants are sampled from a larger item bank using stratified matrix sampling (Revelle et al., 2017; Revelle, Dworak, and Condon, 2021; Revelle, Wilt, and Rosenthal, 2010). The data used for this study were specifically collected from April 2006 through December 2018. Although this study was preregistered to only use data collected until November 2018, more data than anticipated were available; thus, the November and December 2018 data were included in the analyses. From this larger dataset, participants for this study were included if they reported growing up in the United States and were between the ages of 18 and 90. Participants who indicated that they had previously completed the SAPA Project by responding "Yes" to the question, "Have you taken this survey before?" or had unusual response patterns (i.e., giving the same response for >8 items in a row) were excluded from the sample during data cleaning.

2.2. Participants

Participants ($N = 394{,}378$) recruited between 2006 and 2018 were used to examine the 35-item composite ability score, matrix reasoning, verbal reasoning, and the 8-item letter and number series. A detailed breakdown of participant's demographics for the overall sample and annual samples is provided in the supplementary materials. These participants were disproportionately female (65.03%; see Table S1 in the supplementary materials) and between the ages of 18 and 90 ($M = 33.72$, $SD = 15.16$, $Median = 29$). Participants under the age of 18 were excluded from the sample as this study aimed to understand differences in cross-sectional ability scores in adulthood. While an argument could be made that 18- to 25-year-olds should be excluded from the analyses due to fluid ability peaking in cross-sectional data, on average, in the mid-twenties (Salthouse, 2010, 2012, 2019), including this subset of participants allows comparisons with studies on the Flynn effect using European samples over the last 20 years. Participants above the age of 90 were automatically excluded during data cleaning procedures as these older participants are beyond the scope of the SAPA Project (Condon and Revelle, 2016). The largest proportion of these participants reported they were currently attending college (36.02%; see Table S2 in the supplementary materials). Examining annual samples indicated that education attainment was higher for participants recruited during later years of assessment (see Fig. S1 in the supplementary materials).

Participants recruited between 2011 and 2018 ($n = 303{,}540$) were subset into an additional sample to examine the 60-item composite ability scores, the 9-item letter and number series, and three-dimensional rotation; scores for these domains were only collected during this time range. This subset of participants was between the ages of 18 to 90 ($M = 35.43$, $SD = 15.80$, $Median = 31$), but slightly older than the overall 13-year sample. Like the overall sample, the subsample of participants that completed the SAPA Project survey were disproportionately female (65.03%) and a majority of participants reported that they were currently attending college (32.02%).

2.3. Measures

2.3.1. Intelligence

The SAPA Project administered intelligence items from the International Cognitive Ability Resource (ICAR; Condon and Revelle, 2014; Dworak et al., 2021; Revelle et al., 2020). ICAR is an open-source multiple choice intelligence assessment. Although originally validated against the Shipley-2 and self-reported SAT and ACT scores (Condon and Revelle, 2014), more recently a subset of 16 ICAR items, commonly referred to as ICAR16 or the ICAR sample test, were validated against the Wechsler Adult Intelligence Scale Fourth Edition (WAIS-IV). Using a small sample of students ($N = 97$), this study explored how the ICAR sample test loaded on to a Cattell-Horn-Carroll (CHC) model (Carroll, 1993; Cattell, 1943, 1963; Horn and Cattell, 1966; McGrew, 2009; McGrew and Wendling, 2010). Results indicated that the latent general factor models of the two tests correlated 0.94 (Young and Keith, 2020). Despite this validation study examining less than half of the items administered by the present study, we believe it's important to highlight Young and Keith's (2020) findings when describing each ICAR domain as research on the Flynn effect has traditionally been completed on standardized proprietary licensed assessments such as Wechsler tests.

Starting in 2006, the SAPA Project measured cognitive ability using 35 items from the ICAR. Over the next several years, Condon began to supplement the assessment resulting in cognitive ability being measured by 60 items by 2011 and nearly 1000 items by 2022. Although the larger International Cognitive Ability Resource (ICAR) contains 19 domains and 1000s of intelligence items (Dworak et al., 2021), the 60 items

administered and included in this study were representative of ICAR's four original domains: matrix reasoning, verbal reasoning, letter and number series, and three-dimensional rotation (Condon and Revelle, 2014; Revelle et al., 2021).

Ability items were administered through the SAPA Project using a multiple-choice format with eight response options for three-dimensional rotation items and six response options for matrix reasoning, verbal reasoning, and letter and number series items. Of the options, only one answer was correct. Items were scored whether items were correct (1) or incorrect (0). Items that were not administered to participants (due to the stratified matrix sampling methods used by the SAPA Project) or items skipped by participants were coded as NA and not included in the analysis. To ensure that skipping behaviors would not influence participant scores and the analyses used in this study, a Pearson correlation between number of skipped items and participant scores was run. Results of these analyses ranged from $r = -0.05$ to $r = 0.00$, indicating there was no relationship between the number of items skipped and average scores. Domain scores and overall scores were calculated by finding each participant's average score across answered items. Using a participant's average score rather than the sum of an individual's scores is more representative of a participant's performance due to the stratified matrix sampling methods used by the SAPA Project; the random sampling procedures from this method results in some participants receiving more items than other participants. Though it was not preregistered, scores for the entire sample were also standardized across the pooled 13 years of data to help with the interpretation of our findings.

Matrix reasoning. Matrix reasoning is measured using 11 items that contain 3 × 3 arrays of geometric shapes. Within this grid, one of the nine shapes is intentionally excluded. Participants are then prompted to respond by choosing which of six geometric shapes best fits within the stimuli. This task is often compared to stimuli seen in the Raven's Progressive Matrices (Condon and Revelle, 2014). Matrix reasoning items are meant to measure nonverbal reasoning, visual processing, and fluid reasoning. Of the four items validated from the ICAR sample test, matrix reasoning was found to load primarily on the CHC constructs fluid reasoning and visual-spatial processing tasks (Young and Keith, 2020). These 11 items have been administered over the 13 years of SAPA Project data included in this study.

Verbal reasoning. Verbal reasoning is measured using 16 items that use various general knowledge, logic, and vocabulary questions. Verbal reasoning items are meant to measure verbal reasoning, comprehension-knowledge, and crystallized reasoning, however, Young and Keith (2020) found the four verbal reasoning items from the ICAR sample test primarily related to the CHC construct visual-spatial processing tasks. Like matrix reasoning, these 16 items have been administered over the 13 years of SAPA Project data included in this study.

Letter and number series. For approximately 4.5 years (from April 2006 to August 2010), letter and number series was exclusively measured using 8 items. To improve this measure, collection of a ninth item began in August 2010, thus, letter and number series has been measured using 9 items for 8.5 years. Letter and number series items present participants with a sequence of digits or letters and asks the participant to choose the digit or letter that occurs next. Letter and number series items are meant to measure computational/mathematical reasoning, however, Young and Keith (2020) found that the four letter and number series items from the ICAR sample test primarily related to the CHC construct fluid reasoning tasks.

Three-dimensional rotation. Three-dimensional rotation is measured using 24 items inspired by Gittler and Glück (1998) that present participants with a marked cubic shape. Participants are then asked to choose a possible rotation of the shape. Three-dimensional rotation items measure visuospatial and mental rotation and the four items from the ICAR sample test load on the CHC construct visual-spatial processing tasks (Young and Keith, 2020). Collection of these items began in May 2011; thus, this study contains 8 years of three-dimensional rotation data.

35 ICAR items. A participant's overall ability was scored from the 35 ICAR items collected between 2006 and 2018. The 35 ICAR items are composed of the original 8 letter and number series items, 11 matrix reasoning items, and 16 verbal reasoning items. One reason to examine the 35 variables in addition to the 60 ability items is due to the history of the SAPA Project. From April 2006 to August 2010 data were only collected for the 35 items.

60 ICAR items. A participant's overall ability was scored from the 60 ICAR items collected between 2011 and 2018. The 60 ICAR items are composed of the 9 letter and number series items, 11 matrix reasoning items, 24 three-dimensional rotation items, and 16 verbal reasoning items. One reason to examine the 60 variables in addition to the 35 ability items is due to the history of the SAPA Project. In August 2010 one additional letter and number series item was added and in May 2011 24 three-dimensional rotation items were added to data collection. Thus, less data were available for the 60-item composite score of ICAR despite it being considered a better scale than the 35-item composite score.

2.3.2. Demographics

Age and estimated birth year. Age was collected by asking their participants to indicate their age. During data cleaning, participants under the age of 18 and above the age of 90 were removed. An estimated birth year was created for each participant. To do this, the participant's age was subtracted from the year the survey was taken. However, as estimated birth year could misestimate respondents by one year, this transformed variable should be interpreted with caution.

Gender. From April 2006 to February 2017 participants reported gender by choosing from a drop-down menu between male, female, and prefer not to answer. It was not until February 2017 that the option "other" was included. Because we only have data collected for this additional category ($n = 1239$; 0.31%; see Table S1 in the supplementary materials) for two years, the choice to only include binary gender in analyses was made in advance during preregistration.

Highest level of educational attainment. Education was measured using six categories (excluding associate degree and currently in graduate or professional school) from April 2006 to August 2010. From August 2010 to February 2017 educational attainment was measured with seven categories (excluding associate degree). Starting in February 2017, education was measured using eight categories. These categories allowed participants report their educational attainment as <12 years of education, high school graduate, currently in college/university, some college/university (but did not graduate), associate degree (2 year), college/university degree (4 year), currently in graduate or professional school, or graduate or professional school degree.

2.4. Planned analyses

2.4.1. Regression analyses using estimated birth year

This study preregistered to examine how average ability scores changed between estimated birth years across 13 years of SAPA data. First, estimated birth year was used in six different simple regressions to predict the average composite scores of the 35-item ICAR, the average composite scores of the 60-item ICAR, and the average domain scores of matrix reasoning, letter and number series (8 items), and verbal reasoning, and three-dimensional rotation. After these analyses, multiple regressions controlling for gender and education as covariates (separately and together) were performed to examine how well estimated birth year predicted the six ability scores. Next, an additional set of exploratory multiple regressions were run to also control the year the assessment was completed. Based on feedback from reviewers, the value of these analyses, and the potential misspecification created in how birth year was estimated, the results of the simple and multiple regressions are presented and discussed in the supplementary materials to adhere with this study's preregistration.

2.4.2. Annual scores by age

In line with previous Flynn effect research (Ang et al., 2010; Rodgers and Wänström, 2007), we also regressed each ability score on the year the assessment was completed separately for each age to examine and compare the annual trends of ability scores and their standard errors. Rather than aggregating data by cohort, this analysis allowed us to compare participants at the same age across years and determine the annual difference in scores (in standard deviations); thus, allowing for different intercepts and slopes for each age across the year of assessment. For example, we were able to compare mean scores by estimating the slope and their associated standard errors for 18-year-olds who took the ICAR items in 2006 to 18-year-olds who completed the ICAR items in 2007 and all subsequent years of testing (2008–2018). In addition to these analyses, the slopes of mean scores and their standard errors were also examined across the full sample and are discussed in the supplementary materials.

2.5. Results

2.5.1. Annual scores by age

Differences in average ability scores were examined across the 18- to 90-year-old participants by modeling a separate regression at each level of age. For this series of regressions, ability scores from one composite/domain were regressed on the year the assessment was completed using the subset data associated with that level of age; this process was completed for each of the composite and domain scores acting as the dependent variable. Doing so would provide the slopes of each level of age across the 13 or 8 years of data and the slope's associated standard error. These analyses revealed that the results may be affected by frequent missing observations and low ($n < 20$) annual observations for participants between the ages of 61 to 90 years old (histogram of ages illustrated in Fig. S2 and described in Table S17 in the supplementary materials). To adhere with this study's preregistration, a full description of the results based on analyses using the participants within the full age range (18- to 90-year-olds) are further discussed in the supplementary materials. We justify providing the results of those between 18 and 60 years old in the main text as they are less likely to be influenced by fluctuations in the annual sample size of each age (see Table S5 in the supplementary materials). Rather than detailing the slopes observed for all 43 unadjusted regressions across the two overall ability scores and five domain scores, we report the 1) average annual change in ability score, 2) the range of the slopes across the 43 ages, 3) the number of slopes that equaled or exceeded the magnitude of original Flynn effect observation or its reversal (|0.02| SD per year), 4) the direction of these slopes (negative or positive), and 5) the age level associated with each regression. Rather than listing out each individual level of age, some summaries of these regression results include age ranges; thus, indicating each level of age within the range is being described. For example, some levels of ages have been listed as 18- to 24-year-olds rather than 18-, 19-, 20-, 21-, 22-, 23-, and 24-year-olds. While we provide a summary of the results, tables of the individual regression slopes, and their associated standard errors, and the number of observations by regression are provided in Tables S3-S5 of the supplementary materials.

On average, the slopes for overall ICAR scores, measured with 35 items or 60 items, showed small annual differences with more recent participants exhibiting lower scores; albeit these averages did not exceed the threshold of |0.02| SD per year. For ICAR scores assessed with 35 items from 2006 through 2018, a small average annual slope of −0.013 SD (*Range* = [−0.037, 0.004] SD per year) was observed across all ages, however, only 12 of the slopes for 18- to 60-year-olds met or exceeded the threshold set by this study. All 12 of these slopes were negative, thus indicating participants who completed the items more recently had lower scores over this period (levels of age associated with the regression models: 18- to 24-, 51-, 54-, 55-, 57-, and 60-year-olds). Likewise, after accounting for the new item types collected in 2010 and 2011, overall ICAR scores measured with 60 items from 2011 through 2018 showed a small average annual slope of −0.009 SD (*Range* = [−0.063, 0.017] SD per year). For this measure, 11 of the slopes were greater than or equal to |0.02| SD per year, with all slopes being negative (levels of age associated with the regression models: 18- to 24-, 55- to 57-, and 59-year-olds).

Unlike the slopes for the overall ICAR scores, the results observed for the domain scores revealed a more nuanced pattern regarding the difference in annual cognitive ability scores by age. Matrix reasoning scores showed a small average annual slope of −0.009 SD from 2006 to 2018. Annual differences in scores ranged between −0.024 and 0.003 SD per year with 4 of the examined ages reaching or exceeded the established threshold. As observed with the composite scores, these 4 slopes were all negative (levels of age associated with the regression models: 18- to 21-year-olds). In contrast, the average regression slopes for verbal reasoning across the separate ages remained flat across this period (M = 0.002 SD per year). Despite showing a similar pattern to matrix reasoning, differences in verbal reasoning scores from 2006 to 2018 ranged between −0.014 and 0.015 SD per year with 0 of the slopes equaling or exceeding the magnitude of the expected Flynn effect or its reversal.

Like the results observed for matrix reasoning, 8-item letter and number series annual differences in scores ranged between −0.033 and 0.003 SD per year with a small average slope of −0.012 SD per year across the 13 years of data. Only 5 of the 43 slopes met the designated threshold, with all 5 slopes being negative (levels of age associated with the regression models: 18- to 22-year-olds). After adding one item to the letter and number series domain, 9-item letter and number series showed a similar pattern from 2011 to 2018. Specifically, the average slope of 9-item letter and number series scores remained flat, on average, across all ages (M = 0.000 SD per year). While the differences in scores ranged between −0.042 and 0.019 SD per year, 5 of the slopes continued to meet or exceed the threshold. Again, these 5 slopes were negative (levels of age associated with the regression models: 18- to 22-year-olds).

Contrary to the other domains, three-dimensional rotation showed overwhelming evidence of average increasing scores (M = 0.030 SD per year) between 2011 and 2018. Differences in three-dimensional rotation scores occurred regardless of age and ranged between 0.004 and 0.056 SD per year. For this domain of cognitive ability, 32 of the 43 slopes exceeded the threshold of 0.02 SD per year. These positive slopes were associated with the regression models for those between the ages of 24 and 54 years and for 59-year-olds.

2.6. Exploratory analyses

In addition to the preregistered analyses, follow up analyses were completed to further understand what accounted for the observed trends in the annual ability scores by age for 18- to 60-year-olds. Specifically, because differences in ability scores all appeared to follow a curvilinear pattern peaking in the thirties, we were concerned that a participant's level of education could be driving the results. Thus, the ability scores were regressed on to year of assessment separately for each age with the participant's self-reported highest level of education entered as a covariate. An additional series of regression models were then fit to the data with both highest level of education and gender entered as a covariate.

Following these analyses, we further investigated how ability scores differed across the demographic categories. For the highest level of education, we compared annual trends in ability scores by separately regressing average ability scores on the year they were taken for each level of education; those reporting they had an associate degree were excluded due to data being limited to two time points (2017 and 2018). As completed with the previous age stratified regressions, analyses were repeated with age as a covariate and then with age and gender as covariates in the regression models. This process was then repeated for

binary gender, where average ability scores were regressed on the year the assessment was taken 1) separately for each gender; 2) separately for each gender with age as a covariate; and 3) separately for each gender with age and education as a covariate.

Analyses were also completed by fitting separate regressions to grouped age ranges to provide a further understanding as to how ages with less participants (61- to 90-year-olds) may be changing in relation to a larger sample. For these analyses, ages were subset into five cohorts. The first cohort, consisting of participants between the ages of 18 and 19, was formed to provide a clear comparison to previous Flynn effect findings for this age range (Bratsberg and Rogeberg, 2018; Dutton and Lynn, 2013; Rönnlund, Carlstedt, Blomstedt, Nilsson, and Weinehall, 2013; Sundet et al., 2004; Teasdale and Owen, 2008). As those between the ages of 20 and 24 didn't align with previous studies, we subset these participants into a second group; However, this sample encapsulates a large proportion of participants who have commonly completed the SAPA Project survey. Participants between the ages of 25 and 29 years old and those between 30 and 49 years of age were subset into their own groups because they would most likely be from the same or similar birth cohorts as individuals recruited in previous Flynn effect research (Bratsberg and Rogeberg, 2018; Dutton et al., 2016). For example, participants within the 25 to 29 and 30 to 49 groups that completed the survey in 2006 should respectively have been born between 1977 and 1981 and 1957–1976; in 2018 these cohorts should have been born between 1989 and 1993 and 1969–1988. Finally, the last cohort contained participants between 50 and 90 years of age to mimic the sample initially included in Skirbekk, Stonawski, Bonsang, and Staudinger (2013). After completing these analyses for age cohort alone, the analyses were again completed with education as a covariate, and then education and gender as covariates.

2.7. Results

2.7.1. Annual scores by age controlling for education

Detailed results of the individual slopes of the 43 regressions after entering education as a covariate can be found in the supplementary materials (see Tables S6-S7). Again, rather than discussing all 43 adjusted regressions across the seven different ability scores, we highlight the 1) average annual change in ability score, 2) the range of the slopes across the 43 ages, 3) the number of slopes that equaled or exceeded the magnitude of original Flynn effect observation (|0.02| SD per year), 4) the direction of these slopes (negative or positive), and 5) what level of age the regression was associated with. Rather than listing out each individual level of age, some summaries of these regression results include age ranges; thus, indicating each level of age within the range is being described.

After adjusting the regressions for the participant's highest level of education, on average the slopes for overall ICAR scores, measured with 35 items or 60 items, became steeper. For composite ICAR scores assessed with 35 items from 2006 through 2018, the average annual slope was −0.021 SD (*Range* = [−0.037, −0.008] SD per year). Of the 43 regression models for 18- to 60-year-olds, 26 slopes met or exceeded the magnitude of |0.02| SD and were negative (levels of age associated with the regression models: 18- to 25-, 39-, 40-, 42-, 43-, 45- to 48-, 50- to 57-, 59-, and 60-year-olds). Likewise, after entering education as a covariate, composite ICAR scores measured with 60 items from 2011 through 2018 showed an average slope of −0.033 SD per year. Differences in ability scores ranged from −0.062 to −0.015 SD per year. The number age slopes that were equal or greater to the threshold increased from 11 to 37 after entering education as a covariate. As observed with the 35 ICAR items, the 37 slopes were all negative (levels of age associated with the regression models: 18- to 32-, 35- to 37-, 39-, 40-, 42- to 46-, 48-, and 50- to 60-year-olds).

After including education as a covariate, slopes for domain ability scores became steeper or flattened for some ages. The pattern of results across different ages, however, were similar to those of the unadjusted regressions. This included annual differences ranging between −0.023 to −0.002 SD per year for matrix reasoning scores ($M = -0.013$ SD per year), −0.033 to −0.005 SD per year for 8-item letter and number series scores ($M = -0.018$ SD per year), and − 0.041 to −0.002 SD per year for 9-item letter and number series scores ($M = -0.017$ SD per year). Across all three domains, the number of negative slopes increased. Of the 43 slopes, matrix reasoning had 6 negative slopes (levels of age associated with the regression models: 18- to 22-, and 50-year-olds), 8-item letter and number series scores had 12 negative slopes (levels of age associated with the regression models: 18- to 24-, 39-, 42-, 51-, 52-, and 54-year-olds), and 9-item letter and number series scores had 14 negative slopes (levels of age associated with the regression models: 18- to 25-, 27- to 29-, 45-, 52-, and 59-year-olds).

Comparable to these annual differences, the range of the slopes originally observed for verbal reasoning scores diminished (*Range* = [−0.013, 0.003]) after entering education as a covariate. Regardless of this shift in the range, slopes of verbal reasoning scores from 2006 to 2018 remained flat on average ($M = -0.006$ SD per year) and 0 of the 43 slopes were equal to or exceeded the threshold used by this study.

Despite three-dimensional rotation scores exhibiting only positive slopes when examining the unadjusted age regressions, many of the slopes decreased after controlling for education (*Range* = [−0.005, 0.038] SD per year; see Table S7 in the supplementary materials). Regardless, three-dimensional rotation scores displayed an average annual difference of 0.015 SD. In contrast with the other domains, 12 of the 43 slopes for the age stratified regressions with education entered as a covariate resulted in slopes that were equal to or greater than the Flynn effect (levels of age associated with the regression models: 29-, 32-, 33-, 35-, 36-, 37-, 40-, 41-, 44-, 47-, 49-, and 59-year-olds).

2.7.2. Annual scores by age controlling for education and gender

As the data used for this study was disproportionately female, additional analyses were completed to control for gender in addition to education. After entering gender as a covariate, small decreases to the magnitude of differences were again observed across the slopes of composite and domain scores (see Tables 1 and 2). As seen previously, 35-item and 60-item composite ICAR scores were generally lower across all ages. Specifically, the 35-item ICAR scores measured from 2006 to 2018 differed an average − 0.023 SD per year (*Range* = [−0.037, −0.012] SD per year) and the 60-item ICAR scores measured from 2011 to 2018 differed an average − 0.034 SD per year (*Range* = [−0.062, −0.016] SD per year). For the composite score using 35 items, 30 of the 43 slopes met or exceeded the cut-off magnitude of |0.02| SD per year, whereas 38 of the 43 slopes for the composite score using 60 items met or surpassed this criterion. As observed with the unadjusted and education adjusted models, all slopes were negative indicating a decline in scores (levels of age associated with the regression models predicting 35-item ICAR composite: 18- to 25-, 27-, 29-, 39- to 43-, 45- to 57-, 59-, and 60-year-olds; levels of age associated with the regression models predicting 60-item ICAR composite: 18- to 33-, 35- to 40-, 42- to 46-, and 48- to 59-year-olds).

The previous pattern of results when controlling for education held after adding gender as a covariate. On average, matrix reasoning scores ($M = -0.014$ SD per year), 8-item letter and number series scores ($M = -0.019$ SD per year), and 9-item letter and number series scores ($M = -0.018$ SD per year) showed small annual differences across their respective regression models (see Tables 1 and 2). For these domains, the rate of these differences varied by age with matrix reasoning displaying slopes between −0.023 to −0.003 SD per year, 8-item letter and number series scores exhibiting slopes between −0.033 and − 0.008 SD per year, and 9-item letter and number series scores regression slopes ranging between −0.041 and − 0.001 SD per year. Within these results, 6 of the matrix reasoning slopes were steeper than |0.02| SD per year (levels of age associated with the regression models: 18-, 19-, 21-, 22-, 24-, and 50-year-olds), whereas 15 of the 43 slopes met or exceeded this threshold for letter and number series measured with 8 items (levels of

Table 1
Annual differences in ICAR scores between 2006 and 2018 for 18- to 60-year-olds adjusted for education and gender.

Age	ICAR35	95% CI	MR	95% CI	VR	95% CI	LNS (8)	95% CI
Mean slope	**−0.023**	[−0.035, −0.011]	−0.014	[−0.024, −0.004]	−0.007	[−0.015, 0.001]	−0.019	[−0.029, −0.009]
18	**−0.035**	[−0.037, −0.032]	**−0.021**	[−0.024, −0.018]	−0.013	[−0.016, −0.010]	**−0.031**	[−0.034, −0.028]
19	**−0.037**	[−0.040, −0.034]	**−0.023**	[−0.027, −0.020]	−0.013	[−0.016, −0.009]	**−0.033**	[−0.037, −0.030]
20	**−0.029**	[−0.033, −0.026]	−0.019	[−0.022, −0.016]	−0.010	[−0.013, −0.006]	**−0.026**	[−0.029, −0.022]
21	**−0.036**	[−0.040, −0.033]	**−0.022**	[−0.026, −0.019]	−0.013	[−0.017, −0.010]	**−0.030**	[−0.034, −0.027]
22	**−0.034**	[−0.038, −0.030]	**−0.021**	[−0.025, −0.017]	−0.013	[−0.017, −0.009]	**−0.029**	[−0.034, −0.025]
23	**−0.028**	[−0.032, −0.023]	−0.018	[−0.022, −0.014]	−0.008	[−0.012, −0.004]	**−0.023**	[−0.028, −0.019]
24	**−0.026**	[−0.030, −0.021]	**−0.020**	[−0.024, −0.015]	−0.007	[−0.011, −0.002]	**−0.022**	[−0.026, −0.017]
25	**−0.023**	[−0.027, −0.018]	−0.015	[−0.020, −0.010]	−0.007	[−0.011, −0.002]	−0.019	[−0.024, −0.015]
26	−0.019	[−0.024, −0.014]	−0.015	[−0.020, −0.010]	−0.004	[−0.009, 0.001]	−0.015	[−0.020, −0.010]
27	**−0.021**	[−0.026, −0.016]	−0.012	[−0.017, −0.007]	−0.005	[−0.010, 0.000]	−0.019	[−0.024, −0.014]
28	−0.018	[−0.023, −0.013]	−0.012	[−0.017, −0.007]	−0.003	[−0.008, 0.002]	−0.017	[−0.021, −0.012]
29	**−0.022**	[−0.027, −0.017]	−0.015	[−0.020, −0.010]	−0.005	[−0.010, 0.000]	**−0.020**	[−0.026, −0.015]
30	−0.016	[−0.021, −0.011]	−0.010	[−0.015, −0.004]	−0.001	[−0.006, 0.005]	−0.013	[−0.018, −0.008]
31	−0.017	[−0.023, −0.012]	−0.011	[−0.017, −0.005]	−0.003	[−0.009, 0.002]	−0.014	[−0.019, −0.008]
32	−0.019	[−0.025, −0.014]	−0.014	[−0.019, −0.008]	−0.003	[−0.008, 0.003]	−0.015	[−0.021, −0.009]
33	−0.012	[−0.018, −0.007]	−0.010	[−0.016, −0.004]	0.001	[−0.005, 0.007]	−0.008	[−0.014, −0.003]
34	−0.014	[−0.019, −0.008]	−0.006	[−0.012, 0.000]	−0.002	[−0.008, 0.004]	−0.015	[−0.021, −0.009]
35	−0.014	[−0.020, −0.008]	−0.003	[−0.009, 0.003]	−0.002	[−0.008, 0.003]	−0.015	[−0.021, −0.009]
36	−0.013	[−0.019, −0.008]	−0.008	[−0.014, −0.002]	0.000	[−0.006, 0.006]	−0.015	[−0.021, −0.008]
37	−0.018	[−0.024, −0.012]	−0.010	[−0.016, −0.004]	0.000	[−0.006, 0.006]	**−0.020**	[−0.026, −0.013]
38	−0.015	[−0.021, −0.009]	−0.005	[−0.011, 0.001]	−0.003	[−0.009, 0.003]	−0.017	[−0.023, −0.010]
39	**−0.024**	[−0.031, −0.018]	−0.015	[−0.022, −0.008]	−0.007	[−0.014, −0.001]	**−0.021**	[−0.028, −0.015]
40	**−0.023**	[−0.029, −0.016]	−0.019	[−0.025, −0.012]	−0.005	[−0.012, 0.001]	−0.018	[−0.024, −0.011]
41	**−0.020**	[−0.027, −0.014]	−0.007	[−0.015, 0.000]	−0.008	[−0.015, −0.001]	−0.019	[−0.026, −0.012]
42	**−0.024**	[−0.031, −0.018]	−0.015	[−0.022, −0.008]	−0.011	[−0.018, −0.004]	**−0.022**	[−0.029, −0.015]
43	**−0.021**	[−0.028, −0.014]	−0.008	[−0.015, −0.001]	−0.011	[−0.018, −0.004]	−0.017	[−0.024, −0.010]
44	−0.017	[−0.024, −0.010]	−0.005	[−0.013, 0.002]	−0.005	[−0.013, 0.002]	−0.014	[−0.022, −0.007]
45	**−0.024**	[−0.031, −0.017]	−0.019	[−0.026, −0.012]	−0.006	[−0.014, 0.001]	−0.019	[−0.026, −0.012]
46	**−0.029**	[−0.036, −0.022]	−0.019	[−0.026, −0.011]	−0.012	[−0.019, −0.004]	−0.018	[−0.026, −0.011]
47	**−0.021**	[−0.028, −0.014]	−0.010	[−0.018, −0.002]	−0.008	[−0.015, 0.000]	−0.017	[−0.024, −0.009]
48	**−0.025**	[−0.032, −0.017]	−0.011	[−0.019, −0.003]	−0.010	[−0.017, −0.002]	**−0.020**	[−0.028, −0.013]
49	**−0.021**	[−0.029, −0.014]	−0.014	[−0.022, −0.006]	−0.007	[−0.015, 0.001]	−0.015	[−0.023, −0.007]
50	**−0.026**	[−0.033, −0.018]	**−0.021**	[−0.028, −0.013]	−0.006	[−0.014, 0.001]	−0.015	[−0.022, −0.007]
51	**−0.032**	[−0.040, −0.024]	−0.015	[−0.024, −0.007]	−0.014	[−0.022, −0.006]	**−0.026**	[−0.034, −0.017]
52	**−0.022**	[−0.030, −0.015]	−0.014	[−0.022, −0.005]	−0.005	[−0.013, 0.003]	**−0.024**	[−0.032, −0.015]
53	**−0.023**	[−0.031, −0.014]	−0.012	[−0.021, −0.003]	−0.005	[−0.014, 0.003]	−0.018	[−0.026, −0.009]
54	**−0.029**	[−0.037, −0.020]	−0.019	[−0.028, −0.010]	−0.010	[−0.019, −0.001]	**−0.020**	[−0.029, −0.011]
55	−0.024	[−0.033, −0.015]	−0.014	[−0.024, −0.005]	−0.012	[−0.021, −0.003]	−0.017	[−0.026, −0.008]
56	**−0.022**	[−0.032, −0.012]	−0.018	[−0.029, −0.008]	−0.003	[−0.013, 0.008]	−0.015	[−0.025, −0.004]
57	**−0.027**	[−0.037, −0.017]	−0.018	[−0.029, −0.008]	−0.007	[−0.017, 0.003]	−0.018	[−0.028, −0.008]
58	−0.019	[−0.030, −0.008]	−0.006	[−0.018, 0.005]	−0.010	[−0.022, 0.001]	−0.014	[−0.025, −0.002]
59	**−0.020**	[−0.032, −0.009]	−0.014	[−0.025, −0.002]	−0.003	[−0.015, 0.009]	−0.019	[−0.031, −0.007]
60	**−0.021**	[−0.033, −0.010]	−0.014	[−0.026, −0.002]	−0.008	[−0.020, 0.004]	−0.011	[−0.023, 0.001]

Note. Standardized ability scores were regressed on year of assessment with highest level of education and gender as covariates separately for each age. Overall and domain slopes are in SD per year. Any values equal or greater than |0.02| are bolded. ICAR = International Cognitive Ability Resource, ICAR35 = 35-item ICAR composite score, LNS (8) = 8-item Letter and Number Series, MR = Matrix Reasoning, VR = Verbal Reasoning.

age associated with the regression models: 18- to 24-, 29-, 37-, 39-, 42-, 48-, 51-, 52-, and 54-year-olds) and letter and number series measured with 9 items (levels of age associated with the regression models: 18- to 25-, 27- to 29-, 35-, 45-, 48-, and 52-year-olds). After entering gender as a covariate, regressions for verbal reasoning scores had an average slope of −0.007 SD per year (*Range* = [−0.014, 0.001]). As observed with the unadjusted and education adjusted models, 0 of the 43 slopes exceeded the established threshold.

Unlike the other ICAR domain scores, three-dimensional rotation scores continued to show positive slopes (*Range* = [−0.006, 0.038] SD per year) after entering gender and education as covariates into the regressions. On average, differences in annual three-dimensional rotation scores were 0.013 SD. Of the 43 slopes, only 10 were large enough to meet or exceed the differences in score observed by Flynn (1984), with all 10 slopes being positive (levels of age associated with the regression models: 29-, 32-, 33-, 35-, 37-, 40-, 41-, 44-, 47-, and 59-year-olds).

2.7.3. Annual scores by education with and without covariates

Differences in average ability scores were examined by regressing overall and domain scores on the year of assessment separately for each level of educational attainment. After completing these unadjusted analyses, the regressions were rerun with age entered as a covariate, and then age and gender entered as a covariate. Individual slopes and their standard errors for each model are provided in Tables S8 and S9 the supplementary materials. As the option to report education level as "associates degree (2yr)" was not added until February 2017, those with an associate degree were excluded from analyses as reporting the difference between two time points could be misleading.

The unadjusted slopes for overall ICAR scores, measured with 35 items of 60 items, showed difference in annual ability scores exceeding the threshold of |0.02| SD per year, with more recent scores being lower than prior scores. For ICAR scores assessed with 35 items from 2006 through 2018, a small average annual difference of −0.029 SD (*Range* = [−0.041, −0.018] SD per year; see Fig. 1) was observed across all levels of education, however, only the slopes for individuals with <12 years of education, high school graduates, those currently in college/university, those with some college/university experience without graduating, and those that completed a graduate or professional school degree met or exceeded the threshold set by this study. After adding age as a covariate, the magnitude of the annual differences reduced for composite ICAR scores measured with 35 items (*M* = −0.024; *Range* = [−0.038, −0.013] SD per year). As such, the slope for individuals who completed a graduate or professional school degree became flatter and was no longer less

Table 2
Annual differences in ICAR scores between 2011 and 2018 for 18- to 60-year-olds adjusted for education and gender.

Age	ICAR60	95% CI	LNS (9)	95% CI	R3D	95% CI
Mean slope	**−0.034**	[−0.057, −0.011]	−0.018	[−0.038, 0.002]	0.013	[−0.007, 0.032]
18	**−0.062**	[−0.067, −0.057]	**−0.039**	[−0.044, −0.034]	0.011	[0.006, 0.016]
19	**−0.062**	[−0.067, −0.056]	**−0.041**	[−0.047, −0.036]	0.004	[−0.001, 0.010]
20	**−0.061**	[−0.067, −0.056]	**−0.034**	[−0.040, −0.029]	0.009	[0.003, 0.015]
21	**−0.056**	[−0.062, −0.051]	**−0.029**	[−0.036, −0.023]	0.004	[−0.002, 0.011]
22	**−0.055**	[−0.062, −0.048]	**−0.035**	[−0.042, −0.028]	−0.001	[−0.008, 0.006]
23	**−0.047**	[−0.055, −0.040]	**−0.028**	[−0.036, −0.021]	0.005	[−0.004, 0.013]
24	**−0.046**	[−0.054, −0.038]	**−0.031**	[−0.040, −0.023]	0.009	[0.000, 0.018]
25	**−0.037**	[−0.046, −0.029]	**−0.026**	[−0.034, −0.018]	0.011	[0.002, 0.020]
26	**−0.036**	[−0.045, −0.027]	−0.018	[−0.027, −0.010]	0.014	[0.004, 0.024]
27	**−0.034**	[−0.043, −0.025]	**−0.023**	[−0.031, −0.014]	0.012	[0.002, 0.022]
28	**−0.037**	[−0.046, −0.028]	**−0.023**	[−0.032, −0.013]	0.006	[−0.004, 0.017]
29	**−0.036**	[−0.046, −0.027]	**−0.030**	[−0.040, −0.021]	**0.024**	[0.013, 0.035]
30	**−0.031**	[−0.041, −0.022]	−0.014	[−0.024, −0.005]	0.017	[0.006, 0.028]
31	**−0.025**	[−0.035, −0.015]	−0.010	[−0.020, 0.001]	0.007	[−0.005, 0.018]
32	**−0.028**	[−0.039, −0.018]	−0.018	[−0.029, −0.008]	**0.022**	[0.010, 0.034]
33	**−0.020**	[−0.031, −0.009]	−0.005	[−0.015, 0.006]	**0.028**	[0.016, 0.041]
34	−0.019	[−0.030, −0.007]	−0.009	[−0.020, 0.003]	0.016	[0.003, 0.029]
35	**−0.034**	[−0.045, −0.022]	**−0.021**	[−0.033, −0.010]	**0.026**	[0.013, 0.039]
36	**−0.023**	[−0.035, −0.012]	−0.007	[−0.018, 0.005]	0.018	[0.004, 0.031]
37	**−0.021**	[−0.033, −0.009]	−0.019	[−0.031, −0.007]	**0.027**	[0.013, 0.041]
38	−0.019	[−0.031, −0.007]	−0.011	[−0.024, 0.001]	0.011	[−0.003, 0.025]
39	**−0.025**	[−0.038, −0.013]	−0.014	[−0.027, −0.001]	0.018	[0.003, 0.033]
40	**−0.031**	[−0.043, −0.019]	−0.012	[−0.024, 0.001]	**0.029**	[0.014, 0.043]
41	−0.018	[−0.031, −0.005]	−0.017	[−0.030, −0.003]	**0.032**	[0.017, 0.047]
42	**−0.033**	[−0.046, −0.020]	−0.019	[−0.033, −0.006]	0.012	[−0.003, 0.026]
43	**−0.031**	[−0.045, −0.018]	−0.014	[−0.028, 0.000]	0.011	[−0.005, 0.026]
44	**−0.021**	[−0.035, −0.006]	−0.012	[−0.026, 0.003]	**0.027**	[0.010, 0.043]
45	**−0.035**	[−0.048, −0.022]	**−0.021**	[−0.035, −0.007]	−0.006	[−0.021, 0.009]
46	**−0.036**	[−0.051, −0.022]	−0.018	[−0.033, −0.003]	0.014	[−0.003, 0.030]
47	−0.016	[−0.030, −0.003]	−0.003	[−0.018, 0.011]	**0.038**	[0.022, 0.053]
48	**−0.033**	[−0.048, −0.018]	**−0.021**	[−0.036, −0.005]	0.008	[−0.009, 0.026]
49	**−0.021**	[−0.037, −0.006]	−0.007	[−0.022, 0.009]	0.019	[0.002, 0.037]
50	**−0.034**	[−0.048, −0.019]	−0.016	[−0.031, −0.001]	0.011	[−0.005, 0.028]
51	**−0.039**	[−0.054, −0.023]	−0.016	[−0.034, 0.001]	0.017	[−0.001, 0.035]
52	**−0.033**	[−0.049, −0.017]	**−0.022**	[−0.039, −0.005]	0.012	[−0.007, 0.031]
53	**−0.037**	[−0.054, −0.021]	−0.013	[−0.031, 0.004]	0.000	[−0.019, 0.019]
54	**−0.042**	[−0.059, −0.026]	−0.019	[−0.037, −0.001]	0.005	[−0.014, 0.025]
55	**−0.039**	[−0.056, −0.022]	−0.006	[−0.024, 0.012]	0.003	[−0.016, 0.022]
56	**−0.040**	[−0.058, −0.022]	−0.010	[−0.030, 0.009]	−0.004	[−0.026, 0.017]
57	**−0.032**	[−0.051, −0.014]	−0.006	[−0.026, 0.014]	0.005	[−0.015, 0.025]
58	**−0.026**	[−0.047, −0.006]	−0.001	[−0.023, 0.021]	0.012	[−0.011, 0.035]
59	**−0.036**	[−0.057, −0.014]	−0.019	[−0.042, 0.004]	**0.022**	[−0.003, 0.046]
60	−0.018	[−0.040, 0.004]	−0.003	[−0.026, 0.021]	0.006	[−0.018, 0.030]

Note. Standardized ability scores were regressed on year of assessment with highest level of education and gender as covariates separately for each age. Overall and domain slopes are in SD per year. Any values equal or greater than |0.02| are bolded. ICAR = International Cognitive Ability Resource, ICAR60 = 60-item ICAR composite score, LNS (9) = 9-item Letter and Number Series, R3D = Three-Dimensional Rotation.

than −0.02 SD per year; thus, only leaving 4 slopes that met or exceeded our threshold. Next, gender was entered as a covariate into the regression for each education level. The magnitude and direction of the slopes after adding gender as a covariate were almost identical to those observed to the models without gender for overall ICAR measured with 35 items ($M = -0.024$; *Range* = [−0.038, −0.014] SD per year). Given these results, the slopes for the 4 groups with less than a four-year college degree all exhibited annual differences in overall ability scores that exceeded the threshold of |0.02| SD.

Accounting for the new item types, composite ICAR scores measured with 60 items from 2011 through 2018 showed a similar but more extreme pattern for the slopes of the unadjusted regressions ($M = -0.050$; *Range* = [−0.072, −0.032] SD per year). For the unadjusted regression models stratified by highest level of education, annual differences for all 7 slopes were negative and surpassed the threshold of |0.02| SD. After entering age as a covariate, the pattern of annual differences held for the 7 slopes ($M = -0.042$; *Range* = [−0.065, −0.020] SD per year). However, entering both age and gender as covariates marginally reduced the slopes ($M = -0.042$; *Range* = [−0.064, −0.019] SD per year) so that the slopes for participants with four-year college degrees no longer met or exceeded this study's threshold.

Similar to the results observed for each age regression, annual differences in ability scores varied by domain. Slopes for the unadjusted matrix reasoning regressions ($M = -0.020$; *Range* = [−0.027, −0.015] SD per year) and slopes for matrix reasoning scores where age was a covariate ($M = -0.016$; *Range* = [−0.022, −0.008] SD per year) only exceeded the |0.02| SD per year threshold for individuals with <12 years of education, high school graduates, and those currently in college/university. Once gender was added as a covariate to the regression ($M = -0.016$; *Range* = [−0.023, −0.008] SD per year), only the slopes for high school graduates and those currently in college/university continued to meet or exceed this specified cut-off.

Regardless of measuring letter and number series with 8 items (from 2006 to 2018) or 9 items (from 2011 to 2018), the slopes of the unadjusted regressions showed, on average, that participants who completed letter and number series items more recently had lower scores than those recruited earlier in the survey ($M_{LNS8} = -0.024$; $Range_{LNS8}$ = [−0.036, −0.012]; $M_{LNS9} = -0.029$; $Range_{LNS9}$ = [−0.050, −0.014] SD per year). Across these 7 slopes, only 4 slopes met or exceeded the threshold for the education levels of <12 years of education, high school graduates, those currently in college/university, and those with some college/university experience without graduating. These results and the magnitude of the slopes were still present after adding age as a covariate into the regression ($M_{LNS8} = -0.020$; $Range_{LNS8}$ = [−0.033, −0.009]; M_{LNS9} =

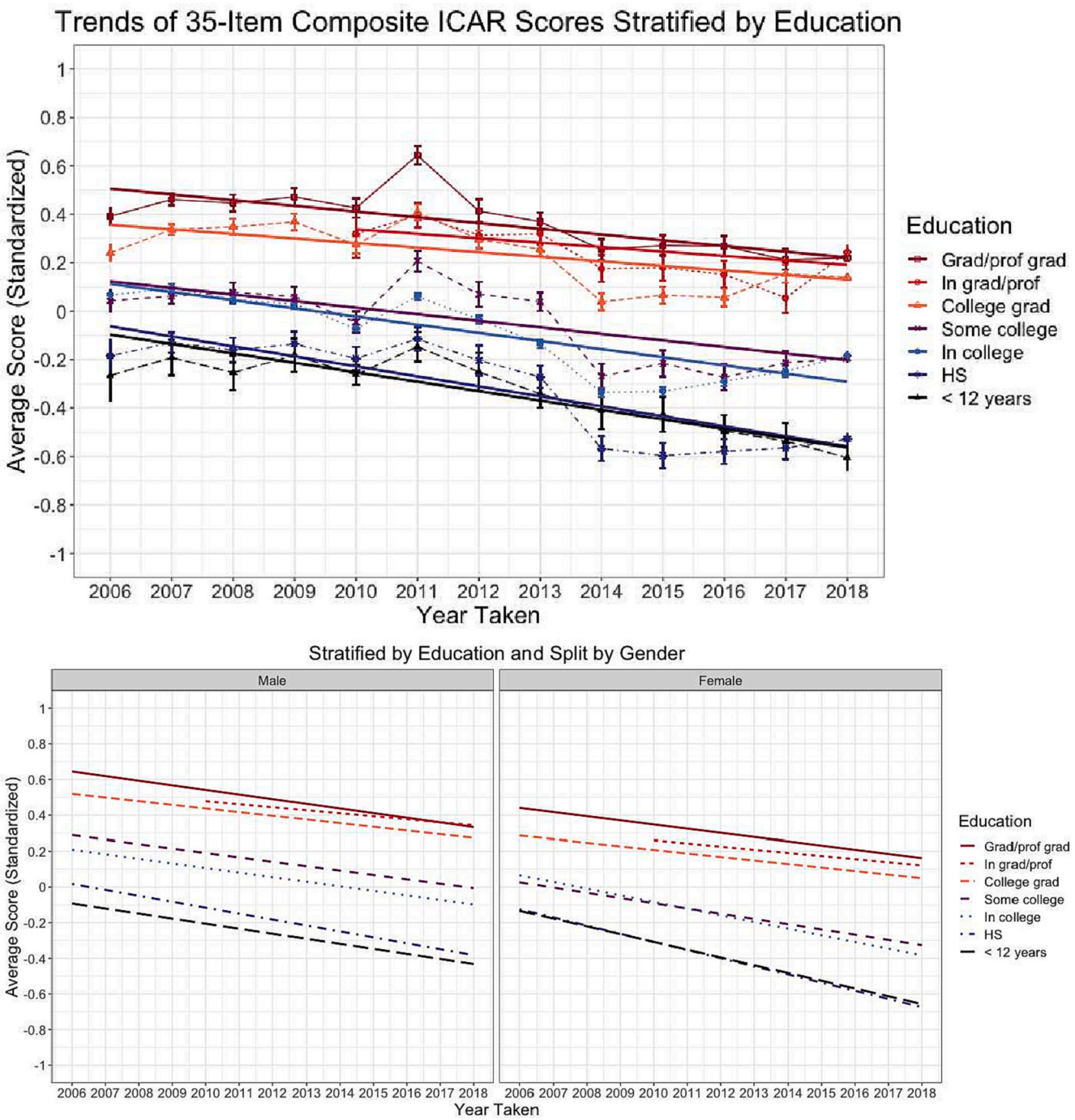


Fig. 1. Trends of 35-item composite ICAR scores stratified by education.

Note. Data collection for the category "currently in graduate or professional school" did not start until August 2010. The dashed lines in the top graph connect the average standardized score and its associated standard error for each year and level of education. The solid lines in the top graph represent the associated slope of the average standardized score for each level of education. The lines in the bottom graph are the associated slope of the average standardized score for each level of education split between male (left) and female (right) participants. ICAR = International Cognitive Ability Resource, Grad/prof grad = Graduate or professional degree, In grad/prof = Currently in graduate or professional school, College grad = College graduate, Some college = Some college, did not graduate, In college = Currently attending college, HS = High school graduate, <12 years = <12 years of education.

−0.023; $Range_{LNS9}$ = [−0.041, −0.006] SD per year) and after including both age and gender as covariates (M_{LNS8} = −0.020; $Range_{LNS8}$ = [−0.034, −0.009]; M_{LNS9} = −0.023; $Range_{LNS9}$ = [−0.040, −0.005] SD per year).

Finally, despite the unadjusted regression of verbal reasoning scores (M = −0.007; $Range$ = [−0.016, 0.001] SD per year) having predominantly negative slopes across the levels of education and three-dimensional rotation scores (M = 0.009; Range = [−0.001, 0.016] SD per year) having positive slopes, 0 slopes met or exceeded the specified threshold. This trend continued after adding age to the separate regressions (M_{VR} = −0.007; $Range_{VR}$ = [−0.015, 0.004]; M_{R3D} = 0.009; $Range_{R3D}$ = [−0.001, 0.016] SD per year) and after both age and gender were included as covariates (M_{VR} = −0.007; $Range_{VR}$ = [−0.016, 0.005]; M_{R3D} = 0.011; $Range_{R3D}$ = [−0.001, 0.018] SD per year).

2.7.4. Annual scores by gender with and without covariates

Differences in average ability scores were examined by regressing overall and domain scores on the year of assessment separately for each level of gender (binary). After completing these unadjusted analyses, the regressions were rerun with age entered as a covariate, and then age and education entered as a covariate. Individual regression slopes and their standard errors for each model are provided in Tables S10 and S11 of the supplementary materials. As the option to report gender as "Other" was not added until February 2017, those with that did not report their gender as male or female were excluded from analyses as reporting the difference between two time points could be misleading.

The unadjusted slopes for overall ICAR scores, measured with 35 items of 60 items, showed differences in annual ability scores that did not exceed the threshold established by this study. For ICAR scores assessed with 35 items, the average annual difference across gender was −0.014 SD (Male = −0.010; Female = −0.017 SD per year) for the unadjusted regressions, whereas the average annual difference was −0.016 SD (Male = −0.018; Female = −0.014 SD per year) for overall ICAR scores measured with 60 items. Once age was added as a covariate in the regression, these slopes became steeper, such that the average slope was −0.017 SD per year (Male = −0.014; Female = −0.020 SD per year) for overall ICAR scores measured with 35 items and − 0.020 SD per year (Male = −0.023; Female = −0.017 SD per year) for overall ICAR scores measured with 60 items. Including education as a covariate in the regression with age, the slopes became steeper with an average difference of −0.024 SD per year (Male = −0.021; Female = −0.027 SD per year) for overall ICAR scores measured with 35 items and − 0.040 SD per year (Male = −0.040; Female = −0.041 SD per year) for overall ICAR scores measured with 60 items.

In terms of the domain scores, unadjusted regressions showed similar patterns across gender with only three-dimensional rotation scores exceeding the magnitude observed by the Flynn effect ($M = 0.022$; Male = 0.021; Female = 0.024 SD per year). Although the other unadjusted slopes did not exceed the threshold, similar differences were observed for the scores for males and females for matrix reasoning ($M = -0.013$; Male = −0.011; Female = −0.015 SD per year), verbal reasoning ($M = -0.006$; Male = 0.009; Female = 0.004 SD per year), and letter and number series measured using 8 items ($M = -0.013$; Male = −0.009; Female = −0.017 SD per year) or 9 items ($M = -0.005$; Male = −0.005; Female = −0.004 SD per year). While including age as a covariate in the separate regressions for each gender increased the magnitude of differences observed for the scores, slopes continued to not meet or exceed the specified threshold for most of the domain scores. On average, matrix reasoning scores ($M = -0.012$; Male = −0.010; Female = −0.013 SD per year), 8-item letter and number series scores ($M = -0.015$; Male = −0.011; Female = −0.019 SD per year), and 9-item letter and number series scores ($M = -0.008$; Male = −0.008; Female = −0.008 SD per year) showed small annual differences. Contrary to these slopes, verbal reasoning scores had no average difference ($M = 0.000$; Male = 0.001; Female = −0.002 SD per year) and three-dimensional rotation scores had notable annual differences ($M = 0.026$; Male = 0.022; Female = 0.029 SD per year) between newer and older participants.

After including education as a covariate with age for the separate regressions across gender, the magnitude of the differences for three-dimensional rotation diminished to an average slope of 0.013 SD per year (Male = 0.010; Female = 0.015 SD per year) and the slopes for letter and number series scores became steeper to an average of −0.020 SD per year (Male = −0.016; Female = −0.024 SD per year) when using 8 items and − 0.022 SD (Male = −0.020; Female = −0.025 SD per year) when using 9 items over a shorter time period. The slopes of the matrix reasoning scores ($M = -0.014$; Male = −0.013; Female = −0.016 SD per year) and verbal reasoning scores ($M = -0.006$; Male = −0.004; Female = −0.008 SD per year), however, only showed small annual differences with the inclusion of the two covariates.

2.7.5. Annual scores by age cohort

Annual differences in composite and domain scores for individual regressions can be found in the supplementary materials (see Tables S12–S13). As a reminder, these scores may differ from those observed looking at individual age slopes as the grouped cohorts of age include 61- to 90-year-old participants and grouping across ages provides a larger group sample size; thus, decreasing the size of the standard error. Composite ability scores from 35 items showed differences across age groups between −0.037 to −0.005 SD per year ($M = -0.021$ SD per year) from 2006 to 2018 when examining age alone in relation to the year the assessment was taken with 3 of the 5 slopes exceeding the cut-off score (levels of age group associated with the regression models: 18–19, 20–24, and 50–90). The range grew in magnitude to differences between −0.036 to −0.017 SD per year entering education as a covariate ($M = -0.027$ SD per year), however, the number of slopes exceeding the threshold remained the same. Finally, including both education and gender as covariates into the regressions resulted in differences between −0.036 to −0.019 SD per year ($M = -0.027$ SD per year). After accounting for these covariates, 4 of the slopes for the five age cohorts exceeded the rate of change observed by the Flynn effect (level of age group associated with the regression models: 18–19, 20–24, 25–29, and 50–90).The domain scores showed respective differences across all age groups ranging from −0.023 to −0.004 SD per year for matrix reasoning scores ($M = -0.015$ SD per year) and − 0.033 to −0.007 SD per year for 8-item letter and number series scores ($M = -0.020$ SD per year). After controlling for education as a covariate, matrix reasoning scores generally dropped to range between differences of −0.023 to −0.010 SD per year ($M = -0.017$ SD per year). This range of slopes decreased between differences of −0.023 to −0.011 SD per year ($M = -0.018$ SD per year) after gender was also entered as a covariate. A similar pattern was observed for 8-item letter and number series scores as the results from the analysis controlling for education ranged between differences of −0.032 to −0.015 SD per year ($M = -0.023$ SD per year) and − 0.032 to −0.016 SD per year ($M = -0.024$ SD per year) after education and gender were controlled for. Across all the unadjusted and adjusted models containing covariates, only 3 of the 5 slopes exceeded the established threshold for differences in matrix reasoning and 8-item letter and number series scores (levels of age group associated with the regression models: 18–19, 20–24, and 50–90). Unlike the other domains, slopes across age cohorts never met or exceeded this study's threshold. Using the unadjusted regressions, verbal reasoning scores initially showed differences between −0.013 to 0.008 SD per year ($M = -0.002$ SD per year). Once education was controlled for, all slopes became negative with verbal reasoning scores showing differences between −0.013 to −0.002 SD per year ($M = -0.007$ SD per year) and between −0.013 to −0.005 SD per year ($M = -0.008$ SD per year) once gender was added as a covariate.

Analyses of composite and domain scores administered from 2011 to 2018 yielded similar results. Specifically, composite ability scores from 60 items showed annual differences that exceeded the established threshold for all but one of the slopes (levels of age group associated with the regression models: 18–19, 20–24, 25–29, and 50–90). The slopes of these overall scores ranged from −0.062 to 0.006 SD per year ($M = -0.028$ SD per year). As observed with verbal reasoning, slopes became steeper after education was entered as a covariate; resulting in the slopes for all age groups to exceed the threshold of $|0.02|$ SD per year. For this set of regressions, the slopes ranged between −0.062 and − 0.026 SD per year ($M = -0.045$ SD per year). Including gender as a covariate with education further increased the magnitude of the differences observed in the previous composite score models to range between −0.062 to −0.027 SD per year ($M = -0.046$ SD per year). When examining domain scores for age groups over this shorter period, 9-item letter and number series scores annual differences ranged between −0.042 and 0.009 SD per year ($M = -0.015$ SD per year) for unadjusted regression models. Of the 5 slopes, only 2 slopes exceeded the established threshold (levels of age group associated with the regression

models: 18–19 and 20–24). After including education as a covariate in these regressions, differences in letter and number series scores measured with 9 items increased in magnitude with one of the slopes flipping from positive to negative. For education adjusted models, annual differences ranged between −0.040 to −0.013 SD ($M = -0.027$ SD per year) and the number of slopes that met or exceeded the Flynn effect threshold increased from to 2 slopes to 4 slopes (levels of age group associated with the regression models: 18–19, 20–24, 25–29, and 50–90). Models where both gender and education were entered as covariates yielded similar results with slopes ranging between annual differences of −0.040 to −0.014 SD ($M = -0.027$ SD per year). For this model, the same 4 slopes showed magnitudes that exceeded the established threshold for this study.

In contrast with the other domains of cognitive ability exhibiting lower scores for more years of assessment, unadjusted models for three-dimensional rotation scores regressed across the different age groups showed positive slopes that ranged between 0.007 and 0.040 SD per year ($M = 0.021$ SD per year). However, despite the positive direction of these slopes, only 2 of the 5 slopes exceeded the established threshold (levels of age group associated with the regression models: 25–29 and 30–49). Including education as a covariate in the regression resulted in the magnitude of the annual differences to decrease to 0.001 to 0.019 SD per year ($M = 0.010$ SD per year) with 0 of the slopes exceeding the established threshold. Likewise, the slopes from the models adjusting for both education and gender decreased in magnitude to range between 0.001 and 0.017 SD per year ($M = 0.009$ SD per year).

3. Discussion

The present study aimed to examine if a Flynn effect or a reverse Flynn effect was a phenomenon within a large sample of adults from the United States between 2006 and 2018. To our knowledge, this is one of the first studies of this size to examine differences in ability scores with an adult United States sample during the twenty-first century. The results of the analyses completed with composite cognitive ability scores and domain scores had five primary findings: 1) There was no evidence of a Flynn effect across composite ability scores but possible evidence of a reversal; 2) one domain showed possible evidence for a Flynn effect, one domain showed no differences, and the remaining domains showed evidence of a reversal of varying magnitudes; 3) lower average scores were frequently observed for more recent participants across all levels of education; 4) differences in scores were similar across gender; 5) the greatest differences in annual scores were observed for 18- to 22-year-olds and individuals with less than a 4-year college degree.

Regardless of using a composite score assessed with 35 items (from 2006 through 2018) or 60 items (from 2011 through 2018), differences meeting or exceeding the magnitude of a Flynn effect, or its reversal, were not present across the full sample of 18- to 60-year-olds when examining unadjusted regressions stratified by age. When magnitudes of the observed slopes did exceed the anticipated threshold, the trends were consistently negative; thus, indicating more recent scores were lower than preceding scores. After rerunning the age stratified regression models with education as a covariate and both education and gender as covariates, the number and magnitude of differences increased such that most of the slopes for the age stratified regressions were equal to or exceeded −0.02 SD per year. Repeating these analyses using cohorts of participants, rather than individual levels of age, yielded similar outcomes. These results are directly at odds with the differences of 0.20 to 0.33 SD per decade observed by Flynn (1984, 1987, 2007) and the 0.15 to 0.195 SD per decade observed by Trahan et al. (2014), where more recent norming samples had higher IQ scores. Likewise, these differences do not reflect more recent results using child and adolescent samples (Ang et al., 2010; Rodgers and Wänström, 2007; Shakeel and Peterson, 2022). Instead, these findings are consistent with studies that have found a reversal of the Flynn effect (Dutton and Lynn, 2013, 2015; Rönnlund et al., 2013; Sundet et al., 2004; Teasdale and Owen, 2008; Woodley and Meisenberg, 2013). In particular, the range of slopes for the 35 and 60-item composite scores overlapped and exceeded the differences of −0.03 to −0.29 SD per decade observed by Dutton et al. (2016).

Like the composite scores, differences were not consistent across the full sample of 18- to 60-year-olds or across domains of cognitive ability scores. Taken together, the 0.20 to 0.33 SD per decade gains observed by Flynn (1984, 1987, 2007) were not substantially present for the ICAR domain scores. Rather, slopes from stratified age regressions of matrix reasoning, letter and number series, and verbal reasoning scores support a reversal or stagnation of the Flynn effect. Given that matrix reasoning and letter and number series scores can be used to assess components of fluid reasoning, one interpretation of these results is that average fluid ability scores are lower for more recent participants within the sample. This evidence is a stark contrast to previous research showing positive differences in ability scores were largely driven by nonverbal fluid tasks like Raven's Progressive Matrices (Ceci and Kanaya, 2010; Flynn, 2007; Neisser, 1997; Sundet et al., 2004; Weiss, 2010).

Contrary to these other results, domain scores for three-dimensional rotation exhibited differences such that more recently recruited participants had higher scores than prior participants. Using the unadjusted age stratified regressions, a majority of the slopes for three-dimensional rotation scores met or exceeded those originally observed by Flynn, however, the magnitude and number of slopes that met or surpassed this threshold decreased after adding covariates to the model. The results in this study align with findings by Rönnlund et al. (2013) that visual-spatial tests exhibited some of the largest gains across test domains. The observed differences for three-dimensional rotation seem to counter some of the recent findings that positive differences in spatial ability scores have decelerated (Pietschnig and Voracek, 2015) or that more recent spatial ability scores are lower than previously tested scores (Pietschnig and Gittler, 2015); albeit Pietschnig and Gittler (2015) only examined German-speaking countries.

As the number and size of slopes that exceeded the magnitude of the Flynn effect or its reversal frequently changed after adding education or education and gender as covariates, we also explored if the results were consistent across demographic categories. While differences for composite and domain scores were relatively consistent between gender and highest level of education, the magnitude of these differences varied. In particular, the rate of differences in scores were generally greater for individuals with less than a 4-year college degree than those with a college degree or higher. As observed with the regressions stratified by age, the coefficients exceeding the magnitude of the Flynn effect were present for composite (measured with 35 or 60 items), matrix reasoning, and letter and number series scores were lower for more recent participants than proceeding participants. Slopes for verbal reasoning and three-dimensional rotation scores, however, never met or exceeded the criteria established by this study. As a limited number of studies have considered the role of education or parental education in the differences of various domains of IQ scores over time, our results do not directly align with previous studies. This is partially due to the direction of the coefficients and differences observed across the levels of educational attainment, but also likely due to the differences in the assessment being used. Notably, Ang et al. (2010) found that the rate of their observed Flynn effect in Peabody Individual Achievement Test Math scores was greater for children and adolescents with more educated mothers. Platt et al. (2019), on the contrary, found no differences in Kaufman Brief Intelligence Test matrices scores when examining parental education. Finally, Twenge et al. (2019) reported that despite observing negative slopes across all levels of education, that differences in vocabulary scores were greatest for those with a college degree or higher. Taken together, our results are arguably most consistent with those found by Ang et al. (2010) as the differences observed in our study indicated that those with higher levels of education were at least buffered from the decreasing rates of scores.

For gender, the magnitude of differences for female participants was

marginally greater, on average, than male participants after including the covariates in the regression models. Nevertheless, coefficients across the two regression models often did not meet or exceed annual differences of |0.20| SD until after models were adjusted for both age and education. Like the results observed for age stratified and education stratified regressions, composite scores, matrix reasoning scores, and letter and number series scores showed evidence of a reverse Flynn effect. Despite examining an adult sample, the similar magnitude and direction of coefficients between male and female participants generalize to previous research examining data from children and adolescents (Ang et al., 2010; Platt et al., 2019; Shakeel and Peterson, 2022).

The overall results of the present study were mixed. While composite ability scores and scores for certain domains of cognitive ability (matrix reasoning and letter and number series) showed patterns consistent with a reverse Flynn effect across age and demographic variables, three-dimensional rotation scores showed evidence of a Flynn effect when stratified by age or gender. Taken together, these results reinforce the importance of Flynn effect research examining scores at both the composite and domain level. As this study did not test for specific factors that could be driving the differences in mean ability scores over time, we cannot make definitive statements as to what caused the varying directions or magnitudes of slopes. However, we reflect on what conflicting differences in cognitive ability scores represent in relation to previously debated causal factors for the Flynn effect.

There is a plethora of theories as to both why the Flynn effect and its reversal or diminished differences are occurring. Though Pietschnig and Voracek (2015) provides a succinct overview, posited causal hypotheses are briefly described here. The main factors believed to contribute to rising IQ scores range from environmental or biological effects to more hybrid and health related factors. Of these theories, Pietschnig and Voracek (2015) note that only a few factors, such as nutrition (Colom, Lluis-Font, and Andrés-Pueyo, 2005; Lynn, 2009), education (Teasdale and Owen, 2005), and test-taking behavior (Must and Must, 2013; Pietschnig, Tran, and Voracek, 2013), could account for both the gains and stagnation observed with the Flynn effect. Specifically, previously observed gains in mean ability scores cannot produce indefinite growth or are eventually restricted by a ceiling effect (Pietschnig and Voracek, 2015; Pietschnig, Voracek, and Gittler, 2018). Recent work by Bratsberg and Rogeberg (2018) examining biological and environmental hypotheses, however, suggests that in addition to changes in the caliber and exposure to education and nutrition, that worse health and increased media exposure could also account for the reversal of the Flynn effect. As the present study found differential slopes in mean cognitive ability scores, it seems unlikely that that quality of nutrition or health would account for conflicting differences among the three-dimension rotation tasks and the remaining tasks. Rather, we would expect to see a reverse Flynn effect across all domains if the differences were due to changes in nutrition or worsening health as cognitive processes impacted by these variables are likely overlapping. This, however, was not observed in the present study as the signs and magnitudes of the slopes varied between domains. As this study did not examine variables related to nutrition or health, further analyses would be required to rule out these factors from the current study. Regardless, we believe education, test-taking, and media exposure emerge as potential moderators for explaining the observed gains in three-dimensional rotation scores and declines or stagnation in matrix reasoning, letter and number series, verbal reasoning, and composite ability scores.

While Pietschnig (2016) and Pietschnig and Voracek (2015) suggest that technology is not likely a singular causal contributor to IQ gains and their declines due to the emergence of widespread access to digital devices following the first studies describing stagnation and declines of ability scores, it is not impossible to posit that greater exposure to media and video games could be buffering declines for visual-spatial oriented tasks compared to fluid ability or reasoning tasks as these tasks are more salient in this type of media (Clark, Lawlor-Savage, and Goghari, 2016). This is not to say that positive coefficients observed in differences of three-dimensional scores are due to playing video games, as much as the relationship between playing video games and spatial ability (Sala, Tatlidil, and Gobet, 2018) may be acting as a greater protective factor for lower scores that could be observed in this domain of ability. This theory, however, would again require additional testing to understand if it has a moderating role within the sample.

As the present study explored the differences in scores across levels of educational attainment and the highest level of education has increased across the testing period of the SAPA Project sample, our results suggest the causal hypothesis that exposure to education accounts for the direction and strength of the Flynn effect (Bratsberg and Rogeberg, 2018; Pietschnig and Voracek, 2015) was not observed within this sample. Rather, exposure to education may only be protective for certain age groups. Not only did the present study find that the steepest negative slopes of composite or domain scores occurred for individuals with less than a 4-year college degree, the largest differences for age stratified regressions after controlling for educational attainment were exhibited for those between the ages of 18 and 22. While these findings complement previous research with 18- to 20-year-old conscripts (Bratsberg and Rogeberg, 2018; Dutton and Lynn, 2013; Sundet et al., 2004; Teasdale and Owen, 2008) and a subsample of 18-year-old study participants within United States (Platt et al., 2019), exposure to education has not been able to explain the differential gains and declines across fluid and crystallized IQ scores observed in previous research (Pietschnig and Voracek, 2015). However, it could be the case that our results indicate a change of quality or content of education and test-taking skills within this large United States sample. As scores were lower for more recent participants across all levels of education, this might suggest that either the caliber of education has decreased across this study's sample and/or that there has been a shift in the perceived value of certain cognitive skills (Clark et al., 2016).

Resembling this causal argument, Flynn (2007) proposed that increased fluid IQ scores could simply be due to society deeming it as valuable. Applying this logic, one could speculate that skills related to matrix reasoning, letter number and series, and verbal reasoning are less valued by society than they were when the original Flynn effect was observed within the United States. Regardless, it should be remembered that the results of this study and differences in ability scores measured by the Flynn effect and its reversal in general may not equate to real gains or declines in intelligence.

3.1. Limitations

Despite including 394,378 participants with varying levels of education and gender, the sample used for our analyses may be unrepresentative of the United States population. In particular, demographics of these subsamples are sometimes under- or oversampled (i.e., over half of the participants identified as female). Thus, the SAPA Project suffers as findings from its data may not be generalized to its target population or subpopulations without appropriate weighting. Future work should repeat the analyses after applying post-stratification methods, such as raking, using available demographics before any results are discussed in terms of how these scores relate to shifts in mean cognitive ability scores within the United States. Alternatively, post-hoc weighted sampling by geographic location (ZIP Code or state) could also be used to ensure the data approximate regional distributions within the United States.

Although post-stratification would allow for the overall sample and subsamples to be more representative of the distributions of the United States population, it should also be recognized that the SAPA Project relies on participants finding or seeking out the survey; meaning members of the population do not have an equal probability of being recruited into the survey (Condon, 2018). Thus, selection bias has likely been introduced into the sample due to those voluntarily taking the survey being non-representative of the target population (Lohr, 2010). This significantly differs from previous Flynn effect studies that relied on systematic norming data collected by proprietary licensed measures

using probability sampling or population-based conscript data where participation was required. However, despite these norming data using stratified proportional sampling to match the United States Census in terms of sex, ethnicity, geographic region, and highest level of educational attainment (Wechsler, 2008), these methods do not eliminate all selection bias for a given sample (Lohr, 2010).

Rodgers (1998) makes a similar point in his proposed list of questions and notes that those exploring the Flynn effect should reflect on their selection methods. Thus, we contemplate how selection bias and sampling in the SAPA Project could be influencing the current study. The largest inconsistency over the 13-year sample is that the 2018 sample is significantly larger than the previous years. Likewise, participants recruited in 2018 had a higher average and median level of educational attainment and age than previous annual samples. As the coefficients for regression models stratified by education did not exhibit the same magnitude of declines for those with a 4-year college degree or higher as those with less than a college degree (4-yr), the difference in sampling might account for why the slopes for higher levels of education were flatter, on average. Another inconsistency was that there was a disproportionate number of 20- year-olds recruited in 2010 (approximately 6000 to 7000 more than the observed in other samples). While it doesn't explicitly appear that this oversampling of 20-year-olds for this one year of data directly influenced the regression coefficients associated with this age across the 13 years of data, it cannot not be completely ruled out. Further sampling inconsistencies over the 13 years of data include the 2006 and 2007 samples having fewer male participants observed in previous years and more male than female participants in 2010. Despite the magnitude of differences across scores being similar between male and female participants, discrepant annual sampling could be limiting the trends observed in this study and how they generalize to the United States population.

As sampling demographics and sizes were inconsistent across the 13 years of data and the SAPA Project depends on individuals who are interested in taking an online survey, it might also be the case that those interested in taking an online personality survey have changed. In the early years of recruitment for the SAPA Project, it is likely that a large proportion of individuals who originally took the survey were either directed to it by an instructor, heard about it at a research conference, or found it through websites/sources associated with academia. As the annual sample sizes have increased in more recent years and the SAPA Project has discussed in more public outlets such social media (mistressredditor, 2013; Murray, 2017; SAPAPsych, 2014a, 2014b) or online articles (Guarino, 2018), it's plausible that newer annual samples are more "average" or normal representation than those recruited during the SAPA Project's former years.

Beyond inconsistencies in demographics across the sample, another factor that could be accounting for lower scores for more recent participants could be due to a decline in motivation. As the SAPA Project is advertised as a personality survey, individuals seeking out the SAPA Project may not be fully engaged with items not measuring temperament at the capacity as they are with more typically considered personality items. As performance is a function of both ability and motivation, participants not trying as hard on ICAR items might also help explain why a reverse Flynn effect was observed despite more recent samples having greater proportions of participants with higher education. This lack of motivation, however, would fail to explain why scores for the most difficult ICAR domain, three-dimensional rotation, were higher for the most recent samples.

4. Concluding remarks

This study set out to investigate if a Flynn effect or a reverse Flynn effect was a phenomenon in a large United States sample recruited between 2006 and 2018. Regardless of education, gender, and age, lower annual scores were observed for composite cognitive ability measured by 35 items, and the matrix reasoning and letter and number series scores measured across the 13 years of assessment. These differences were replicated across the 60-item composite ability scores from 2011 to 2018, however, three-dimensional rotation scores measured during this 8-year period showed evidence of a Flynn effect of varying magnitudes across 18- to 60-year-olds. The largest differences in mean ability scores were often observed for participants between the ages of 18 to 22. Beyond age, a reverse Flynn effect was also present across all levels of educational attainment, with the rate of decreasing scores being steeper for those with less than a 4-year college degree. While additional work needs to be done to further incorporate other demographic measures from the SAPA Project, the current study indicates that the Flynn effect and its associated reversal may no longer generalize across all ages or levels of education. It also underlines the need for further research using large adult samples to understand if the Flynn effect or if its reversal is a phenomenon in the United States during the twenty-first century.

Funding

The data for this study and item development of the International Cognitive Ability Resource were partially supported by a grant from the National Science Foundation (SMA-1419324).

Declaration of Competing Interest

We have no known conflict of interest to disclose.

A preregistration for this study can be found on the Open Science Framework (OSF) at https://osf.io/kmgx8/. All deviations from this preregistration are detailed throughout the manuscript.

Data availability

The data for this study were collected under Northwestern University's Institutional Review Board (#STU00202975) and are part of a larger ongoing project that has been reviewed by the University of Oregon's Institutional Review Board (#08212019.031). Data are periodically published in the public domain (Condon and Revelle, 2016; Condon, Roney, and Revelle, 2017). All reasonable requests for access to the data will be met through contact with the last author.

Acknowledgements

The lead author would like to thank Dr. Crystal Steltenpohl and Dr. Jordan Wagge for helping to reconcile deviations from this study's preregistration and their support in encouraging open science practices.

Appendix A. Supplementary materials

Supplementary materials to this article, such as analyses, tables, and figures, can be found online at https://doi.org/10.1016/j.intell.2023.101734.

References

Ang, S., Rodgers, J. L., & Wänström, L. (2010). The Flynn effect within subgroups in the U.S.: Gender, race, income, education, and urbanization differences in the NLSY-children data. *Intelligence, 38*(4), 367–384. https://doi.org/10.1016/j.intell.2010.05.004

Benson, N., Beaujean, A. A., & Taub, G. E. (2015). Using score equating and measurement invariance to examine the Flynn effect in the Wechsler Adult Intelligence Scale. *Multivariate Behavioral Research, 50*(4), 398–415. https://doi.org/10.1080/00273171.2015.1022642

Bratsberg, B., & Rogeberg, O. (2018). Flynn effect and its reversal are both environmentally caused. *Proceedings of the National Academy of Sciences, 115*, 6674–6678. https://doi.org/10.1073/pnas.1718793115

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin, 40*, 153–193. https://doi.org/10.1037/h0059973

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1–22.

Ceci, S. J., & Kanaya, T. (2010). "Apples and oranges are both round": Furthering the discussion on the Flynn effect. *Journal of Psychoeducational Assessment, 28*, 441–447. https://doi.org/10.1177/0734282910373339

Clark, C. M., Lawlor-Savage, L., & Goghari, V. M. (2016). The Flynn effect: A quantitative commentary on modernity and human intelligence. *Measurement: Interdisciplinary Research and Perspectives, 14*, 39–53. https://doi.org/10.1080/15366367.2016.1156910

Colom, R., Lluis-Font, J. M., & Andrés-Pueyo, A. (2005). The generational intelligence gains are caused by decreasing variance in the lower half of the distribution: Supporting evidence for the nutrition hypothesis. *Intelligence, 33*, 83–91. https://doi.org/10.1016/j.intell.2004.07.010

Condon, D. M. (2018). The SAPA Personality Inventory: An empirically-derived, hierarchically-organized self-report personality assessment model. *PsyArXiv*. https://doi.org/10.31234/osf.io/sc4p9

Condon, D. M., & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence, 43*, 52–64. https://doi.org/10.1016/j.intell.2014.01.004

Condon, D. M., & Revelle, W. (2015). Selected personality data from the SAPA-Project: On the structure of phrased self-report items. *Journal of Open Psychology Data, 3*, Article e6. https://doi.org/10.5334/jopd.al

Condon, D. M., & Revelle, W. (2016). Selected ICAR data from the SAPA-Project: Development and initial validation of a public-domain measure. *Journal of Open Psychology Data, 4*(1), Article e1. https://doi.org/10.5334/jopd.25

Condon, D. M., Roney, E., & Revelle, W. (2017). A SAPA Project update: On the structure of phrased self-report personality items. *Journal of Open Psychology Data, 5*(1), 3. https://doi.org/10.5334/jopd.32

Dutton, E., & Lynn, R. (2013). A negative Flynn effect in Finland 1997–2009. *Intelligence, 41*, 817–820. https://doi.org/10.1016/j.intell.2013.05.008

Dutton, E., & Lynn, R. (2015). A negative Flynn effect in France (1999 to 2008–9). *Intelligence, 51*, 67–70. https://doi.org/10.1016/j.intell.2015.05.005

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

Dworak, E. M., Revelle, W., Doebler, P., & Condon, D. M. (2021). Using the International Cognitive Ability Resource as an open source tool to explore individual differences in cognitive ability. *Personality and Individual Differences, 169*, Article 109906. https://doi.org/10.1016/j.paid.2020.109906

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

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171–191. https://doi.org/10.1037/0033-2909.101.2.171

Flynn, J. R. (1999). The discovery of IQ gains over time. *American Psychologist, 54*, 5–20.

Flynn, J. R. (2007). *What is Intelligence?: Beyond the Flynn Effect*. Cambridge University Press.

Flynn, J. R. (2009). The WAIS-III and WAIS-IV: Daubert motions favor the certainly false over the approximately true. *Applied Neuropsychology, 16*(2), 98–104. https://doi.org/10.1080/09084280902864360

Fosse, E., & Winship, C. (2019). Analyzing age-period-cohort data: A review and critique. *Annual Review of Sociology, 45*, 467–492. https://doi.org/10.1146/annurev-soc-073018-022616

Giangrande, E. J., Beam, C. R., Finkel, D., Davis, D. W., & Turkheimer, E. (2022). Genetically informed, multilevel analysis of the Flynn effect across four decades and three WISC versions. *Child Development, 93*, e47–e58. https://doi.org/10.1111/cdev.13675

Gittler, G., & Glück, J. (1998). Differential transfer of learning: Effects of instruction in descriptive geometry on spatial test performance. *Journal for Geometry and Graphics, 2*, 71–84.

Guarino, B. (2018, September 17). Scientists identify four personality types. *Washington Post*. https://www.washingtonpost.com/science/2018/09/17/scientists-identify-four-personality-types/?noredirect=on&utmterm=.2d12d0ef3c14.

Herrnstein, R. J., & Murray, C. (2010). *Bell curve: Intelligence and Class Structure in American Life*. Simon and Schuster.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57*, 253–270.

Jensen, A. R. (1998). *The g Factor: The Science of Mental Ability*. Praeger.

Johnson, W., & Bouchard, T. J. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence, 33*, 393–416. https://doi.org/10.1016/j.intell.2004.12.002

Kaufman, A. S. (2010). "In what way are apples and oranges alike?" A critique of Flynn's interpretation of the Flynn effect. *Journal of Psychoeducational Assessment, 28*, 382–398. https://doi.org/10.1177/0734282910373346

Lohr, S. L. (2010). *Sampling: Design and Analysis* (2nd ed.). Brooks/Cole.

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

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*, 1–10. https://doi.org/10.1016/j.intell.2008.08.004

McGrew, K. S., & Wendling, B. J. (2010). Cattell–Horn–Carroll cognitive-achievement relations: What we have learned from the past 20 years of research. *Psychology in the Schools, 47*, 651–675. https://doi.org/10.1002/pits.20497

mistressredditor. (2013, October 26). Cutting edge personality test measures the Big Five, Small Six, and Great Three factors of personality. [Online forum post]. *Reddit*. https://www.reddit.com/r/psychology/comments/1pa169/cuttingedgepersonalitytestmeasuresthebig/.

Murray, C. A. (). Want to kill part of the day and also contribute to science? A full-scale personality inventory is waiting for you. *http://sapa-project.org* [Tweet] *Twitter* [@charlesmurray] https://twitter.com/charlesmurray/status/902230016859262977.

Must, O., & Must, A. (2013). Changes in test taking patterns over time. *Intelligence, 41*, 780–790. https://doi.org/10.1016/j.intell.2013.04.005

Neisser, U. (1997). Rising scores on intelligence tests. *American Scientist, 85*, 440–447.

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. (2016). The Flynn effect: Technology may be part of it, but is most certainly not all of it. *Measurement, 14*, 70–73. https://doi.org/10.1080/15366367.2016.1171612

Pietschnig, J., & Gittler, G. (2015). A reversal of the Flynn effect for spatial perception in German-speaking countries: Evidence from a cross-temporal IRT-based meta-analysis (1977–2014). *Intelligence, 53*, 145–153. https://doi.org/10.1016/j.intell.2015.10.004

Pietschnig, J., Tran, U. S., & Voracek, M. (2013). Item-response theory modeling of IQ gains (the Flynn effect) on crystallized intelligence: Rodger's hypothesis yes, Brands' hypothesis perhaps. *Intelligence, 41*, 791–801. https://doi.org/10.1016/j.intell.2013.06.005

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. https://doi.org/10.1177/1745691615577701

Pietschnig, J., Voracek, M., & Gittler, G. (2018). Is the Flynn effect related to migration? *Politische Psychologie*, 267–283.

Platt, J. M., Keyes, K. M., McLaughlin, K. A., & Kaufman, A. S. (2019). The Flynn effect for fluid IQ may not generalize to all ages or ability levels: A population-based study of 10,000 US adolescents. *Intelligence, 77*, Article 101385. https://doi.org/10.1016/j.intell.2019.101385

Revelle, W., Condon, D. M., Wilt, J., French, J. A., Brown, A., & Elleman, L. G. (2017). Web-and phone-based data collection using planned missing designs. In *The SAGE Handbook of Online Research Methods* (pp. 578–594). SAGE Publications Ltd. https://doi.org/10.4135/9781473957992.n33.

Revelle, W., Dworak, E. M., & Condon, D. M. (2020). Cognitive ability in everyday life: The utility of open source measures. *Current Directions in Psychological Science, 29*, 358–363. https://doi.org/10.1177/0963721420922178

Revelle, W., Dworak, E. M., & Condon, D. M. (2021). Exploring the persome: The power of the item in understanding personality structure. *Personality and Individual Differences, 169*, Article 109905. https://doi.org/10.1016/j.paid.2020.109905

Revelle, W., Wilt, J., & Rosenthal, A. (2010). Individual differences in cognition: New methods for examining the personality-cognition link. In A. Gruszka, G. Matthews, & B. Szymura (Eds.), *Handbook of Individual Differences in Cognition: Attention, Memory and Executive Control* (pp. 27–49). Springer. https://doi.org/10.1007/978-1-4419-1210-7_2.

Rodgers, J. L. (1998). A critique of the Flynn effect: Massive IQ gains, methodological artifacts, or both? *Intelligence, 26*, 337–356. https://doi.org/10.1016/S0160-2896(99)00004-5

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

Rönnlund, M., Carlstedt, B., Blomstedt, Y., Nilsson, L. G., & Weinehall, L. (2013). Secular trends in cognitive test performance: Swedish conscript data 1970–1993. *Intelligence, 41*, 19–24. https://doi.org/10.1016/j.intell.2012.10.001

Rushton, J. P., & Jensen, A. R. (2010). The rise and fall of the Flynn effect as a reason to expect a narrowing of the black–white IQ gap. *Intelligence, 38*, 213–219. https://doi.org/10.1016/j.intell. 2009.12.002

Russell, E. W. (2007). The Flynn effect revisited. *Applied Neuropsychology, 14*, 262–266. https://doi.org/10.1080/09084280701719211

Sala, G., Tatlidil, K. S., & Gobet, F. (2018). Video game training does not enhance cognitive ability: A comprehensive meta-analytic investigation. *Psychological Bulletin, 144*, 111–139. https://doi.org/10.1037/bul0000139

Salthouse, T. A. (2010). Selective review of cognitive aging. *Journal of the International Neuropsychological Society, 16*, 754–760. https://doi.org/10.1017/S1355617710000706

Salthouse, T. A. (2012). Consequences of age-related cognitive declines. *Annual Review of Psychology, 63*, 201–226. https://doi.org/10.1146/annurev-psych-120710-100328

Salthouse, T. A. (2019). Trajectories of normal cognitive aging. *Psychology and Aging, 34*, 17–24. https://doi.org/10.1037/pag0000288

SAPAPsych. (2014, March 25). [Academic] Help me get more participants for my senior thesis and receive data on your personality profile in return! All welcome, ages 14 and up. should take around 20 minutes. [Online forum post]. *Reddit*. from https://www.reddit.com/r/SampleSize/comments/21cz1k/academichelpmegetmorepartipantsformy/.

SAPAPsych. (2014, April 5). Cutting edge personality test (measures Big Five, Small Six, and Great Three factors of personality) with new and improved personality feedback. We took your comments and revised our test! [Online forum post]. *Reddit*. https://www.reddit.com/r/psychology/comments/22a6ku/cuttingedgepersonalitytestmeasuresbigfive/.

Shakeel, M. D., & Peterson, P. E. (2022). A half century of progress in US student achievement: Agency and Flynn effects, ethnic and SES differences. *Educational Psychology Review, 34*, 1255–1342. https://doi.org/10.1007/s10648-021-09657-y

Skirbekk, V., Stonawski, M., Bonsang, E., & Staudinger, U. M. (2013). The Flynn effect and population aging. *Intelligence, 41*, 169–177. https://doi.org/10.1016/j.intell.2013.02.001

Sundet, J., Barlaug, D., & Torjussen, T. (2004). The end of the Flynn effect? A study of secular trends in mean intelligence test scores of Norwegian conscripts during half a century. *Intelligence, 32*, 349–362. https://doi.org/10.1016/S0160-2896(04)00052-2

Teasdale, T. W., & Owen, D. R. (1989). Continuing secular increases in intelligence and a stable prevalence of high intelligence levels. *Intelligence, 13*, 255–262.

Teasdale, T. W., & Owen, D. R. (2005). A long-term rise and recent decline in intelligence test performance: The Flynn effect in reverse. *Personality and Individual Differences, 39*, 837–843. https://doi.org/10.1016/j.paid.2005.01.029

Teasdale, T. W., & Owen, D. R. (2008). Secular declines in cognitive test scores: A reversal of the Flynn effect. *Intelligence, 36*, 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*(5), 1332–1360. https://doi.org/10.1037/a0037173

Twenge, J. M., Campbell, W. K., & Sherman, R. A. (2019). Declines in vocabulary among American adults within levels of educational attainment, 1974–2016. *Intelligence, 76*, Article 101377. https://doi.org/10.1016/j.intell.2019.101377

Vernon, P. E. (1965). Ability factors and environmental influences. *American Psychologist, 20*, 723–733. https://doi.org/10.1037/h0021472

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale—Fourth Edition Technical and Interpretive Manual*. Pearson.

Weiss, L. G. (2010). Considerations on the Flynn effect. *Journal of Psychoeducational Assessment, 28*, 482–493. https://doi.org/10.1177/0734282910373572

Woodley, M. A., & Meisenberg, G. (2013). In the Netherlands the anti-Flynn effect is a Jensen effect. *Personality and Individual Differences, 54*, 871–876. https://doi.org/10.1016/j.paid.2012.12.022

Yang, Y., & Land, K. (2006). A mixed models approach to the age-period-cohort analysis of repeated cross-section surveys, with an application to data on trends in verbal test scores. *Sociological Methodology, 36*, 75–97. https://doi.org/10.1111/j.1467-9531.2006.00175.x

Young, S. R., & Keith, T. Z. (2020). An examination of the convergent validity of the ICAR16 and WAIS-IV. *Journal of Psychoeducational Assessment, 38*, 1052–1059. https://doi.org/10.1177/0734282920943455

Zhou, X., Zhu, J., & Weiss, L. G. (2010). Peeking inside the "black box" of the Flynn effect: Evidence from three Wechsler instruments. *Journal of Psychoeducational Assessment, 28*, 399–411. https://doi.org/10.1177/0734282910373340