The Generality of Educational Effects on Intelligence: A Replication

Jordan Lasker[1, *]

Emil O. W. Kirkegaard[2]

[1] Texas Tech University

[2] Ulster Institute for Social Research

* jlasker@ttu.edu

ORCIDs: JL - 0000-0002-5143-2191; EK – 0000-0001-5607-0321

Keywords: Intelligence, education, cognitive enhancement, signaling theory, human capital, higher education

Abstract

Recent research has provided strong demonstrations to the effects that education improves scores on intelligence tests. We tested whether the improvements elicited by education were consistent with raised intelligence or enhancements to specific skills involved in intelligence testing. We used the structural equation models from Ritchie, Bates & Deary (2015) on a longitudinal sample of over 4,000 American men who took an intelligence test near the end of high school and then took another around 37 years of age. Our results were consistent with theirs in that we found that the effect of education on intelligence test scores was not an improvement to intelligence itself, but instead was relegated to improvements to specific skills. Our results support the notion that education is not a source of enhanced intelligence, but it can help specific skills.

Introduction

Education is related to intelligence[1], whether education is conceptualized as scores on measures of academic achievement (Zaboski et al., 2018), grades (Cucina et al., 2016), or simple years of education completed (Strenze, 2007). There is considerable evidence that intelligence is a causal factor in educational attainment. For example, intelligence measured early in life predicts subsequent educational attainment (Butler et al., 1985; Deary et al., 2007; Fergusson et al., 2005), when comparing identical twins the more intelligent twin tends to achieve a higher level of education (Johnson et al., 2006; Sandewall et al., 2014; Stanek et al., 2011), and when children are more intelligent than their parents the same is true and vice-versa for children who perform worse than their parents (McGue et al., 2020). On the other hand, there is a great deal of evidence that education improves scores on intelligence tests (Ritchie & Tucker-Drob, 2018) and there are trends in educational attainment that transcend differences in intelligence (McGue et al., 2022).

Bidirectional causation for the link between education and intelligence is both possible and plausible: longitudinal studies, twin-, sibling-, and parent-controls, and a variety of behavior-genetic models support causal effects of intelligence on educational attainment, while natural experiments aplenty have shown that intelligence test scores are positively affected by increased schooling. Intelligent people may seek more cognitively demanding and stimulating environments, pushing them towards higher education, while individuals who are more educated may be more cognitively stimulated and thus develop greater intelligence. However, the bidirectional nature of the linkage between intelligence and education is contentious (Deary & Johnson, 2010).

An important angle for investigating this link is to assess if education affects changes in intelligence or in scores alone. This distinction is important, as scores on intelligence tests are trivially malleable. A researcher interested in increasing test scores could undertake an intervention to provide test takers with answers to test questions, considerably elevating their scores without raising their intelligence whatsoever. This distinction hinges on the difference between latent variables underlying test performance – like intelligence – and the scores used to measure them (Borsboom, 2006). If an intervention affects an intelligence test score, it is not immediately apparent if that means intelligence *the construct* has been impacted.[2]

The typical course of action when evaluating the effect of something on intelligence is to compare scores before and after an exposure (Haier, 2014). Such a procedure does not actually constitute evidence brought to bear on the question of whether and how much intelligence is affected by something. More methodologically inclined researchers have sought to use the

[1] Also referred to as *g*, or general intelligence.

[2] When an intervention affects a construct and thus test scores are consistent with all causes of performance covariation being mediated by said construct, that condition is known as measurement invariance.

effects of cognitive training, brain lesions, and educational interventions, analyzed at the latent level and with respect to various theoretically-informed models, to probe the nature of intelligence, assess its structure, and to evaluate if intelligence itself is altered when scores are (Protzko, 2017; Protzko et al., 2021; Protzko & Colom, 2021). Research on educational effects on intelligence has typically embodied the former variety of investigation, with focus levied on scores rather than constructs. There is, however, one exception, a study by Ritchie, Bates & Deary (2015).

Ritchie, Bates & Deary (henceforth *RBD*) leveraged a large (n = 1,091) longitudinal sample whose intelligence was measured at ages 11 and 70 to assess whether years of education affected intelligence or specific abilities, like logical memory, spatial reasoning, or vocabulary. If education affected intelligence, it was noted, effects "would be apparent in all the cognitive capacities associated with *g*, and, thus, should raise all mental abilities in proportion to their loading on *g*." (ibid., p. 574). The method these researchers used to test how educational effects on intelligence worked was elegant in its simplicity. They posited three competing models: one in which education affected intelligence alone, one where it affected intelligence *and* performance on specific tests, and one where it only affected specific tests and intelligence was not impacted. They found that a model in which only specific tests were influenced fit best.

Their method was admirable, and their results deserve to be replicated to further clarify the nature of the complex relationship between intelligence and education, but data that can be applied to that end is rare. Moreover, their study can be criticized on several grounds. Firstly, their measure of intelligence at age 11 was a singular score rather than several scores, so their measure of early intelligence is not latent intelligence, it approximates it. Secondly, their sample was based out of the United Kingdom, so it is unknown if it will generalize to different cultures. And third, standards for sample sizes have grown in recent years due to the influence low statistical power has on the likelihood a study's results can be replicated; by more contemporary standards, some could argue their sample was too small, and the effects of education – specific or general – may have thus been discounted because of their statistical nonsignificance.

The Present Study

Here, we report a large, well-measured, international replication of RBD's seminal work in a CDC archival dataset (the Vietnam Experience Study, or VES) containing data from over 4,000 men who were given a battery of tests in early adulthood (19.92 years, SD = 1.72) and reassessed some twenty odd years later with a larger battery (37.43 years, SD = 2.52). The unique qualities of this dataset – large sample size, its longitudinal nature, its socially and developmentally meaningful timing, and its excellent test batteries – afforded us an excellent opportunity to investigate the stability of intelligence in the transition from early adulthood to midlife, the effects of education on intelligence, and the contribution of educational effects on intelligence to another measure (or component) of social status, income.

Our first research question regarded the stability of intelligence. We wanted to know how stable intelligence was going from early adulthood to midlife. We contrasted three models for this. In each model, intelligence, as measured at the outset of data collection, and as measured at the follow-up were modeled, with one parameter distinguishing the models: the correlation between intelligence at different timepoints. The first model forced intelligence to be perfectly stable (i.e., Pearson's $r = 1$), the next to be perfectly unstable ($r = 0$), and in the last, the parameter was freed and allowed to take any value.

Research on the stability of intelligence across the lifespan has almost universally focused on the stability of fullscale IQ (FSIQ) scores, individual scale scores, or principal component scores (Ahmed et al., 2020; Deary, 2014; Gow, 2016; Larsen et al., 2008; Lechner et al., 2021; Mansukoski et al., 2019; Schalke et al., 2013; Watkins & Canivez, 2004; Watkins & Smith, 2013). The exceptions are notable, as they reveal a major problem with approaches that do not use latent variables: due to some combination of unreliability and dimensionality, the stability of intelligence has been underestimated in studies that only used observed scores (Rönnlund et al., 2015; Yu et al., 2018).

To help alleviate this gap, we also compared the stability of our FSIQs to the stability of latent intelligence. We were able to provide considerable analytic leverage to this question because, while the tests administered in early life and midlife were not all identical, some of them were. The common tests between time periods were highly correlated with one another (same test r's= 0.785 and 0.842), so we can take the knowledge that composite score reliability exceeds the reliability of individual subtests and use these as lower bounds for the reliability of our FSIQ measure to test whether differences in FSIQ and latent intelligence stability could plausibly be the result of random error. This procedure does affirm the consequent, so we consider them to be of secondary importance and worthy of serious qualification. Related to this analysis, in the process of the convergent validation of test scores, Floyd et al. (2013, p. 397) noted that IQ scores – which primarily represent intelligence – were less correlated than factors representing latent intelligence, but the discrepancy was too large to have been accounted for by reliability, implicating a role for non-general dimensionality in FSIQs that, we believe, is likely to also play a confounding role in analyses of the stability of intelligence, since it is also known that intelligence is more stable than specific abilities (Breit et al., 2021; Larsen et al., 2008; Plomin et al., 1994; Watkins & Canivez, 2004), and the influence of both is subsumed into broad sumscores like FSIQs.

Our second research question was about our exact replication of RBD's (2015) study. We wished to assess whether educational effects on intelligence test scores were due to one of three possibilities that they also tested. The first, their model A, is one in which education affects intelligence directly, and that is how education affects intelligence test results. The second, their model B, is one in which education affects *both* intelligence and specific tests, such that education is both a broad cognitive enhancer and a pathway to enhancements in terms of specific skills like vocabulary or arithmetic. The third, their model C, is one in which education

does not impact intelligence but, instead, impacts only specific subtests. This model features improved intelligence test scores, but not improved intelligence. We mimicked their procedure and hypotheses exactly: "We tested which models had better fit and predicted that, if education improves intelligence by raising *g*, either or both of models A and B would have significantly better fit to the data than Model C." (ibid., p. 575).

Our third and final research question was an attempt to answer the fifth question asked by Ritchie & Tucker-Drob (2018). This question is as follows: "[H]ow important are these [causal] effects [of education on intelligence test scores]? There is strong evidence from industrial and organizational psychology and cognitive epidemiology studies that IQ is associated with occupational, health, and other outcomes, but to our knowledge, no studies have explicitly tested whether the additional IQ points gained because of education themselves go on to improve these outcomes. A quasiexperimental study… found that raising the school-leaving age improved not only IQ but also a variety of indicators of health and well-being. It is possible that the educational benefits to the upstream variables were partly mediated via the IQ increases (or vice versa), but this would need explicitly to be investigated." (ibid., p. 10).

To address this problem, we considered it within the context of a structural equation model involving intelligence measured early in life, education, and intelligence measured later in life. Educational effects would be allowed to land where they would – on specific tests, intelligence, or any combination thereof – but, in one model, the direct path from education to income would be present, and in the other, it would remain unmodeled. If the regression coefficients with respect to income for variables affected by education remain unchanged regardless of model, then the effect of education on income is unmediated by its cognitively salubrious effects. If, on the other hand, educational effects on income are mediated by its effects on intelligence or specific tests, then the explicit inclusion of education should substantially reduce the degree to which affected variables relate to income.

This assessment is considered less strongly than the other two because each variable could be related to income independently and the effects of education may be minute, so our ability to render a powerful conclusion here is more greatly circumscribed due to the need for higher statistical power to reliably provide any answer. Moreover, income, though highly correlated from year-to-year in adulthood, was based on income in the year prior to testing. We are only presenting this analysis as a first attempt at assessing whether educational effects on socioeconomic attainment, in the form of income, are mediated by cognitive enhancement.

Method

Participants

Participants were members of the VES, the full details of which can be obtained from the CDC.[3] The VES is a population-representative cohort, excepting individuals who scored below the 10th percentile in initial testing. Members of a randomly selected subset of the larger sample of 15,288 men were subjected to extensive medical examination, interviews, and neuropsychological evaluations and studied specifically to assess the effects of service in the Vietnam War, since a large portion of that sample (2,490) were Vietnam veterans, and the rest (1,972) were not. As noted by the CDC (footnote 3), there were few differences between these groups excepting problems including "depression, anxiety, and combat-related post-traumatic stress disorder." Since the sample was screened with intelligence tests and individuals below the 10th percentile were subsequently removed from it, everyone had full testing data for the first period, but there was nevertheless a negligible percentage of missingness for the follow-up sample that was usually subtest-specific, although missingness was unrelated to initial scoring.

Measures

Intelligence Testing. The sample was administered five tests on their introduction into the military and a further fourteen at the follow-up. The five introductory tests were the Army Classification Battery (ACB) verbal subtest, the ACB arithmetic subtest, the Pattern Analysis Test, the General Information Test (GIT), and the Armed Forces Qualification Test (AFQT). The individual items for the AFQT's subtests were not available, but items for the GIT were. The fourteen follow-up tests were the ACB verbal and arithmetic subtests, the Wide-Range Achievement Test, the Word List Generation Test, California Verbal Learning Test, Wisconsin Card Sorting Test, Paced Auditory Serial Addition Test, the Wechsler Adult Intelligence Scale – Revised (WAIS-R) block design and general information scales, the left- and right-handed Grooved Pegboard Task, and the direct, immediate, and delayed versions of the Rey-Osterrieth complex figure drawing (CFD) task. Because they were so highly associated with one another, we combined the immediate and delayed CFD scores. These variables have been described by the CDC and other publications (Larsen et al., 2008; Lasker et al., 2021).

Educational Attainment. Participants provided the number of years of education they completed during their follow-up interview.

Income. Participants provided their combined family gross income for the calendar year immediately prior to the telephone interview that preceded psychological and medical testing.

[3] https://www.cdc.gov/nceh/veterans/default1c.htm.

Analyses

The **lavaan R** package (Rosseel et al., 2022) was used to estimate and compare all of our structural equation models, including those derived from RBD's(2015) study. Intelligence was identified at both timepoints through standardization of the latent variances for the configural and baseline models. For model fit comparisons, we focused primarily on the χ^2 exact fit test, since this is the most powerful test of model fit (Ropovik, 2015), but we also referenced other measures of model fit, with our preferences ordered such that BIC was preferred to AIC, which ranked above CFI, which we preferred to RMSEA, since this is the typical ordering of how sensitive these fit indices are to detecting misfit. We dropped nonsignificant paths from the model ($p > .05$) and kept them dropped on an individual basis if the subsequent change in model fit was also significant ($p < .05$) to avoid post-selection biasing to our other p-values, and taking after RBD (2015), we did not display nonsignificant paths in the diagrams of our models.

Results

Descriptive statistics are provided in Table 1 alongside a correlation matrix of all variables included in our models. We replicated the positive manifold of intercorrelations among our intelligence tests (r's between .115 and .824, p's <.0001). Correlations with education were universally positive (.134 – .555), and correlations with income were also universally positive (.165 – .392). Educational and adult income correlations with early cognitive test results (.241 – .532) were notable since scoring predated income and most of higher education. The income-education correlation was significant and positive (.349).

[Table 1]

We performed factor analyses of the tests in each era. There were only five introductory tests, so we could not discover any coherent group factors, but a model with g alone fit well (CFI = .991, RMSEA = .084, SRMR = .018)[4]. This fit can be compared to the dynamic fit cutoffs for a single-factor model (CFI = .981, RMSEA = .105, SRMR = .028, N = 4,355), and the dynamic fit cutoffs work for the later tests (CFI = .961, RMSEA = .048, SRMR = .036, N = 4,426), which had an empirical fit that was sufficiently close (CFI = .964, RMSEA = .059, SRMR = .037). To maintain comportment with prior work on this subject (especially RBD's), we elected to test these single-factor models, but note that the results differed little if residual covariances were modeled as

[4] This model had one residual covariance, between the AFQT and PA tests (r = .447). The dynamic fit cutoffs if two-thirds of items had a residual correlation greater than .3 were CFI = .968, RMSEA = .141, and SRMR = .037. For the later test model, dynamic fit cutoffs assumed a third of variables had residual correlations greater than .3, while the model required residual covariances between CD and CC and GPTL and GPTR and, for reasons of theoretical coherence and lack of fitting alternative factor models with more group factors, we had an additional nine residual covariances. A model with more group factors was not possible beyond reasons of fit, because some had only two indicators, which made them effectively identical to residual covariances. The dynamic fit cutoffs if two-thirds or all items had residual covariances greater than .3 were CFI = .925 and .896, RMSEA = .069 and .084, and SRMR = .046 and .051.

group factors, as had to be done to obtain them, for reasons of model fit and the high generality of the test batteries.

The three types of models we tested are taken from RBD, but slightly modified because instead of a manifest variable in the first period, we have a latent variable. These models are illustrated in Figure 1. The first models we tested were the models of the stability of intelligence which resembled those, except without education. The fits for those models can be found in Table 2. The model where intelligence was neither perfectly stable over time nor unrelated between time points fit well, while the others fit horribly. The correlation between intelligence measured at introduction and follow-up was .945 ($p < .0001$).[5] FSIQ correlations, however, were smaller, at .817 which, when corrected for an underestimated reliability of .8, still significantly differed from the latent correlation (.913, CI: .903 – .927).

[Figure 1]

[Table 2]

Next, we pursued our replication of RBD. In each model, the path from early intelligence to later intelligence was significant and positive (r's between .901 and .941), as was the path from early intelligence to education (.564 – .574). Model A is Figure 2, Model B is Figure 3, and Model C is Figure 4. Residual covariances were not pictured but are available in the analysis codebook. The model fits for these models are contained in Table 3.

Model A clearly fit worse than either of models B or C, with a difference of more than 200 AIC and BIC, marginally worse CFI, and 300 additional χ^2 and only seven additional degrees of freedom. The comparison between models B and C required more delicate consideration. Firstly, their conclusions are identical with respect to intelligence: Model C did not include a path from education to intelligence, and it was nonsignificant in Model B (p = .156). Moreover, Model B fit worse in terms of a χ^2 test (p = .0055) and had seven additional AIC. Due to its five additional degrees of freedom, its BIC was considerably better, but for models that are not directly nested, this is not as meaningful as AIC. Model C had to be selected over Model B because (1) Model B did not correspond to its theoretical description anyway, and (2) it fit worse by the best metrics for discriminating model specifications.

In following RBD's methods, we also carried out their three supplementary analyses. First, we reinstated all drop paths and performed our model comparisons again. Second, we

[5] This analysis is also interesting in so far as it showcases the extent to which different test batteries produce measurements of the same general factor of intelligence, which is interesting and a confounder to stability. Overcoming this difference is difficult because of the limited number of shared tests between timepoints, but we can partially address it by assessing the correlation of general factors composed of common subtests. We did this and the correlation was 1.041, i.e., evidence for an ultra-Heywood case. Moreover, the correlations with these tests removed in aggregate, and from the endpoint only were .927, and .940, respectively. This .940 was significantly different from .945.

dropped all our residual covariances and ran the models again. Third, we adjusted all our later subtests for early intelligence and dropped the path from early to later intelligence. All three methods conformed to our initial pattern of results, much like RBD's situation. Our conclusion from this analysis was identical to theirs: "In all cases, then, the fit of Model C – which did not contain a path from educational duration to the *g* factor of intelligence – was superior, consistent with the position that education has domain-specific, and not domain-general, effects on intelligence." (RBD, p. 578).

Our final analysis was to assess the effects of education on a measure of participants' income. This comparison was aimed at testing whether the relationship between education and socioeconomic status was mediated by cognitive enhancements. Since education does not enhance intelligence, intelligence must be kept in the model. Because, due at least to job market signaling, education should be the dominant signal for educational effects, the test is whether a model with education depletes the validity of the subtests it affected. Table 4 contains these results, and our codebook explores other specifications, all of which produced consistent qualitative conclusions. Namely, education still affects income net of any cognitive enhancement, and what subtests validity there exists for predicting income is substantially unaffected by the inclusion of education in the model. The lone exception was a marginally significant subtest whose effect became nonsignificant when education was included. With all subtests and education removed, the beta for intelligence is .425, with everything but education removed, it is .347, and with both education and intelligence modeled but subtests removed, their respective betas are .176 and .321. To the extent our test battery indexes specific abilities, we can say that the socioeconomically beneficial effects of education are independent of either general or specific cognitive enhancements, and that both education and intelligence have independent predictive validity.

Discussion

Our aims were threefold: Firstly, we wanted to establish how stable intelligence was over an important part of the human lifespan: the transition from early adulthood to midlife. Secondly, we wanted to perform a replication study of Ritchie, Bates & Deary's seminal 2015 work delving into the effect of education on cognitive development. And finally, we wished to know to what extent the effect of education on socioeconomic attainment in the form of income depended on cognition enhancement. We were fortunate to have measures of attained adult income around the time that quantity typically becomes stable in the United States, to have such a large sample, and to have such excellent intelligence measures for those purposes.

The stability of intelligence over time in this data was exceptional for typical stability studies which exclusively use manifest variables. The latent stability exceeded previously published estimates from this dataset that used exploratory factor analysis for factor computation (Larsen et al., 2008). The likeliest reason for this is probably that EFA errs more than CFA, although the estimand for both is the same in this case.

We noted that typical studies use manifest scores. In some cases, like RBD's, they use a mixture of manifest and latent variables. We propose that, in general, the stability of intelligence will be estimated such that studies using only manifest variables will estimate the lowest stability, followed by studies with mixed variable types, and finally, by studies using latent variables exclusively. We had the opportunity to generate that pattern, finding that manifest-manifest r = .817, manifest-latent r = .915 between early FSIQ and later intelligence and r = .859 between early intelligence and later FSIQ, and latent-latent r = .945. Even considerably overcorrecting FSIQs for measurement error left us with greater latent stability. Stability was basically unchanged removing common subtests.

The level of achieved stability provides substantial construct validity evidence for intelligence. One requirement for construct validity and the interpretation of test-retest reliability as measurement of a common construct at different times is that there is longitudinal invariance which, due to the similarity of our measured constructs across time points, seems to be high enough for that to be a tenable suggestion. This is important because, unless intelligence is causal, it would not be meaningful to assess how influenced it was; construct validity is about causality, and one of its underassessed requirements probably worked out in our data. However, because our tests differed, we could not directly assess longitudinal invariance.

We replicated RBD's results, and thus, their conclusions as well: "[E]ducation's ability to raise intelligence test scores is driven by domain-specific effects that do not show 'far transfer' to general cognitive ability." (Ritchie, Bates & Deary, 2015, pp. 578-579). This puts education into a class alongside every other known intervention that works for raising scores on intelligence tests, from cognitive training in general (Sala et al., 2019; Sala & Gobet, 2017) to notorious interventions like the Dual N-Back (Branwen, 2012) or even phenomena like the Flynn effect (Beaujean & Sheng, 2010; Beaujean, 2006; Beaujean & Osterlind, 2008; Beaujean & Sheng, 2014; Benson et al., 2015; Fox & Mitchum, 2013, 2014; Must et al., 2009; Must & Must, 2013, 2018; Pietschnig et al., 2013; Pietschnig & Gittler, 2015; Shiu et al., 2013; Wai & Putallaz, 2011; Wicherts et al., 2004; Woodley et al., 2014) or adoption (Jensen, 1997; te Nijenhuis et al., 2015). That is, though they succeed in raising scores, they fail to raise intelligence.

A now-antiquated method, Jensen's method of correlated vectors, commonly known as MCV, was recently invoked in a relevant debate. In a series of papers, te Nijenhuis et al. (te Nijenhuis, van der Boor, et al., 2019; te Nijenhuis, Choi, et al., 2019) and Flynn (Flynn, 2019b, 2019a) debated Jensen effects – positive associations between various variables and *g* loadings – and te Nijenhuis et al. concluded that the relationship between a vector of education relationships with various tests and those tests' *g* loadings was null, and thus education was not a Jensen effect. Our own data support the opposite conclusion, as we found an extremely strong Jensen effect (r = .921, Tucker's ϕ = .986). At the same time, we found that education did not affect *g*. The difference between an actual test via structural equation modeling and correlating vectors without a statistical test is very marked in this example. For answering questions about what affects *g*, MCV ought to be substituted with SEM when possible.

Explaining why MCV errs with education data seems straightforward. Education should be expected to be most strongly related to verbal and otherwise scholastic tests, which also tend to be more *g* loaded (Kan et al., 2013). The problem with MCV that can emerge in meta-analyses of the relationship between *g* and education, or other criteria in the absence of measurement bias is mostly in that MCV is upwardly bounded by a variety of factors, including sample size, the standard deviation of loadings, the size of group differences, and so on (Dutton & Kirkegaard, 2021; Sorjonen et al., 2017). Therefore, low statistical power is a major concern with MCV. Importantly, MCV applied to the loadings from our final models would not be appropriate because of post-selection inference issues with the resulting estimates of effects in addition to low power, so MCV does not offer additional analytic leverage to SEMs. The conclusions that can be rendered with the two methods applied to the description of things that affect scores changes have no inherent relationship.

Because of the well-known finding that most of the validity of intelligence tests comes from intelligence and specific abilities have limited predictive validity (Ree & Carretta, 2022), it should be *a priori* questionable whether the effects of education on specific abilities should predict higher socioeconomic status. There is considerable evidence that education affects socioeconomic attainment (Sandewall et al., 2014), but also strong evidence against mandatory schooling reforms boosting socioeconomic attainment (Clark & Cummins, 2020) and some evidence that mandatory education has not boosted economic growth (Edwards, 2018). The reason for the former effect – which does appear causal – is hotly debated. The two dominant hypotheses, human capital and signaling (Caplan, 2018), predict, respectively, that education influences socioeconomic attainment via fostering skills or through indicating skills to employers without actually affecting them. Our data do not allow us to differentiate these hypotheses cleanly but, like McGue et al., (2022), our results clearly favor signaling. In our data, education did not provide broad cognitive improvements, nor were its specific effects mediators of the effect of education on income. To our awareness, this is the first attempt to provide a quantitative hint towards Ritchie & Tucker-Drob's (2018) fifth question: "[h]ow important are [the cognitively beneficial] effects [of education]?" (p. 10). For income, our answer is virtually nil.

The observed pattern of results was robust to the use of early intelligence rather than later intelligence, albeit with the WGI effect nonsignificant in the absence of education. We also briefly looked at several health outcomes but produced mostly null relationships and highly inconsistent effects. Because we never had plans to attempt to pull these outcomes into the nomothetic web of education, we also omitted them from analysis.

Limitations and Future Directions

The stability of intelligence was almost certainly underestimated in our dataset. The reason for this is psychometric sampling error. The first-order model of intelligence that we used pollutes intelligence with whatever specific variance is overrepresented among the

manifest variables. Using a higher-order model helps to obviate this issue, as it reduces the relative quantities of different forms of specific variance, in effect delivering a "purer" estimate of intelligence. As noted by Floyd et al. (2013, p. 385): "A higher order factor should control for overrepresentation and underrepresentation of measures of the same Stratum I or II ability that may contribute to psychometric sampling error."

If we had enough variables for a higher-order model – meaning enough variables for many coherent group factors at both timepoints – it is likely that stability would have been higher for two reasons. First, because intelligence is known to be more stable than specific abilities, and second, because the content differences over time may have led to unrepresented variance in one or the other period being subsumed into intelligence, artificially reducing their association. A clear direction for future studies, then, is to assess the long-term stability of intelligence measurements with an intentional effort to reduce the pernicious influence of psychometric sampling error.

An obvious qualification of our validity results for education is that we may have lacked the power to detect the potentially subtle effects of educational benefits to specific skills to income (Ree & Carretta, 2011), but the signs of some of the effects were in the wrong directions in the first place, so this is difficult to humor. Regardless, it appears that, at a minimum, the socioeconomically beneficial effect of education is not totally mediated by subtests it affects, and that their predictive validity is largely independent of education. Just as well, it seems the predictive validity of education is largely distinct from its effects on specific skills. We may have had insufficient breadth to find specific skill mediators of educational effects on attained income though. For example, if spatial or mechanical skills, noted for their job especial academic and job market relevance (Berkowitz & Stern, 2018; Prada & Urzúa, 2014; Wai et al., 2009), were insufficiently modeled, we would not have had the ability to assess how much those mediated educational effects. However, there is some evidence that students high in spatial ability perform *worse* in school, as they consider it to be less interesting than the world of work (Gohm et al., 1998). Over certain timespans, outcomes like income may appear to be negatively related to education because of the age-related tradeoff between the two that could result. Moreover, selection of this sort may generate apparent negative effects of education in general that confound inference with longitudinal data like ours that only has two measurement periods.

We must also embrace several of the caveats from RBD. Namely, developmental, and occupational effects on intelligence and specific skills could not be disentangled and may have been related, although not apparently though an income-generating pathway. Additionally, the breadth of cognitive tests must be increased from our five at induction into the sample and thirteen in the end and the number of group factors should greatly increase, so that effects on reliable, broad indicators of ability net of intelligence can be investigated. It is likely that a bifactor model would help with this, since it could increase the typically abysmal reliability of group factors (Benson et al., 2018) by incorporating variance from local independence violations (Reise et al., 2007) which may, in fact, be reliable group factor variance. However, if that causal

model does not stand up when tested with appropriate data (e.g., Franić et al., 2013), then that *should not* be done. Our estimated educational effects may have also been more limited than those found by RBD because they were measured at a point that was nearer to the typical termination of education than theirs, so effects may have, to some extent, already accrued to the scores measured at sample induction.

Conclusions

We assessed the stability of intelligence at the latent level, the specificity versus generality of educational effects on intelligence test results, and the mediation of educational effects on socioeconomic attainment by its effects on specific tests. Our first result was that intelligence was very stable in the transition from early adulthood to middle-age. Our second result was that a model of the effects of education in which its effects were specific to certain tests rather than affecting intelligence directly fit our data best. And our third result, was that educational effects on socioeconomic attainment were largely independent of its effects on intelligence test results. Our findings are consistent with other latent-level investigations of the stability of intelligence measured across the lifespan (Rönnlund et al., 2015; Yu et al., 2018), Ritchie, Bates & Deary's (2015) analysis of the generality of the effects of education on intelligence test results, and a growing literature showing that education affects greater socioeconomic attainment regardless of how smart you are (McGue et al., 2022).

Code Availability: All code required to replicate these analyses is available online at https://rpubs.com/JLLJ/VESEDU. Data is available at https://ves.emilkirkegaard.dk/data/.

Conflict of Interest Statement: There were no conflicts of interest involved in this study.

Funding Statement: This research received no funding.

Table 1. Descriptive Statistics and Correlations for Study Variables

Variable	*n*	Time	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1. ACB Verbal	4,384	Introduction	107.16	22.26	—																			
2. ACB Arithmetic	4,385	Introduction	104.43	22.01	.699	—																		
3. PA [a]	4,386	Introduction	104.32	22.64	.516	.576	—																	
4. GIT [a]	4,376	Introduction	102.06	18.43	.659	.589	.467	—																
5. AFQT [a]	4,441	Introduction	0.14	0.81	.714	.737	.728	.645	—															
6. WRAT [a]	4,460	Follow-up	61.17	14.73	.746	.589	.412	.517	.578	—														
7. CVLT [a]	4,462	Follow-up	11.06	2.33	.317	.331	.264	.250	.312	.309	—													
8. WCST [a]	4,462	Follow-up	0.79	0.17	.327	.360	.331	.282	.368	.292	.192	—												
9. WBD [a]	4,462	Follow-up	10.52	2.64	.437	.502	.634	.418	.629	.382	.269	.356	—											
10. WGI [a]	4,462	Follow-up	10.07	2.80	.725	.635	.482	.582	.626	.652	.329	.330	.453	—										
11. GPT-R [a]	4,450	Follow-up	-73.66	11.82	.204	.201	.261	.174	.253	.196	.117	.186	.301	.173	—									
12. GPT-L [a]	4,448	Follow-up	-77.38	13.77	.208	.212	.263	.192	.269	.204	.115	.196	.307	.186	.634	—								
13. PASAT [a]	4,450	Follow-up	108.84	50.72	.408	.521	.371	.365	.432	.417	.289	.285	.388	.366	.226	.216	—							
14. CD [a]	4,462	Follow-up	32.73	3.31	.290	.333	.380	.254	.372	.269	.205	.291	.398	.278	.223	.223	.247	—						
15. CC [a]	4,462	Follow-up	0*	0.98	.309	.343	.464	.316	.461	.275	.329	.274	.502	.351	.201	.232	.288	.489	—					
16. WLGT [a]	4,462	Follow-up	35.12	10.92	.443	.370	.289	.310	.360	.504	.278	.209	.281	.414	.168	.157	.357	.176	.220	—				
17. ACB Verbal	4,462	Follow-up	116.52	23.04	.824	.658	.484	.620	.670	.766	.333	.361	.453	.719	.220	.226	.440	.325	.322	.463	—			
18. ACB Arithmetic	4,462	Follow-up	104.56	24.40	.642	.785	.545	.548	.688	.585	.356	.396	.532	.622	.240	.236	.562	.384	.394	.365	.691	—		
19. Education	4,376	Follow-up	4.34	1.52	.360	.367	.241	.303	.323	.273	.165	.212	.222	.312	.169	.165	.268	.186	.175	.189	.341	.392	—	
20. Income	4,460	Follow-up	13.29	2.30	.532	.487	.343	.377	.430	.511	.202	.241	.275	.555	.134	.149	.304	.209	.222	.338	.506	.467	.349	—

[a] Pattern Analysis Test, General Information Test, Armed Forces Qualifying Test, Wide-Range Achievement Test, California Verbal Learning Test, Wisconsin Card Sorting Test, WAIS-R Block Design, General Information, Paced Auditory Serial Addition Test, Grooved Pegboard Right, Left, Complex Figure Drawing Direct, Later, Word List Generation Test.

* 2.004e-16.

Table 2. The Stability of Intelligence over Time

Model	X^2/df	CFI	RMSEA (95% CI)	SRMR	AIC/BIC
No Relationship	8439.536/120	.819	.127 (.124 - .129)	.273	182974/183299
Empirical	2101.090/119	.957	.062 (.060 - .064)	.042	176637/176969
Identity	3615.279/120	.924	.082 (.080 - .084)	.151	178150/178475

Note: The "No Relationship" model forced the correlation between intelligence measured early and later in life to be zero, the "Empirical" model allowed it to be freely estimated, and the "Identity" model constrained it to one.

Table 3. The Effect of Education on Intelligence

Model	Description	X^2/df	CFI	RMSEA (95% CI)	SRMR	AIC/BIC
A	Education to Intelligence	2456.636/135	.952	.063 (.061 - .065)	.043	187090/187441
B	Education to Intelligence and Specific Tests	2171.200/128	.958	.061 (.059 - .063)	.040	186819/187214
C	Education to Specific Tests	2154.666/123	.958	.062 (.060 - .064)	.039	186812/187239

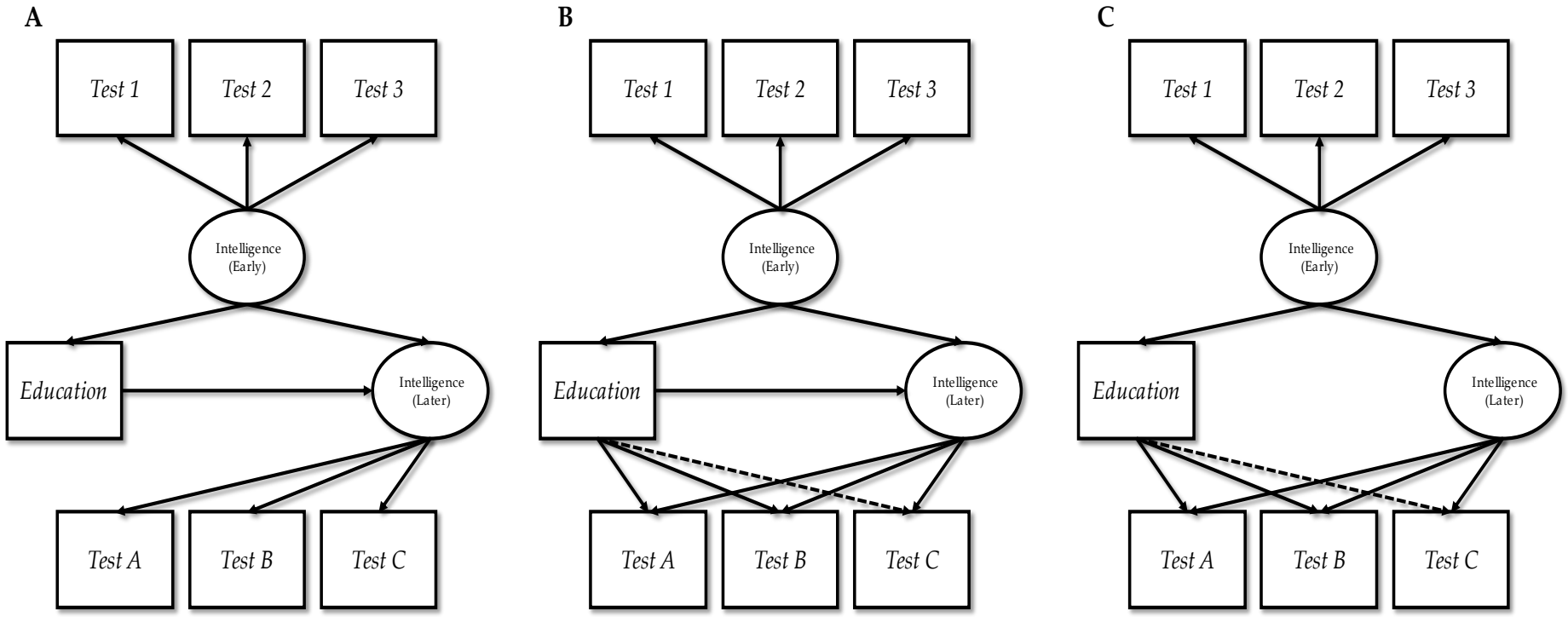
Note: Model B only differed from Model C in terms of the specific tests affected because the path from education to intelligence was nonsignificant (r = .017, CI: -.006 – .040, p = .156).

Table 4. Effects on Income in Different Models

Model Variable	Beta	*p*
No Education		
Intelligence	.681	<.001
WRAT	-.079	.001
WBD	-.104	<.001
WGI	-.077	.005
CC	-.033	.047
WLGT	-.025	.137
ACVL	-.080	.016
ACAL	.053	.090
Education Included		
Intelligence	.605	<.001
WRAT	-.109	<.001
WBD	-.086	<.001
WGI	-.124	<.001
CC	-.028	.086
WLGT	-.032	.057
ACVL	-.067	.040
ACAL	.050	.104
Education	.203	<.001

Note: Bolded subtests had inconsistently significant effects. Later intelligence used.

Figure 1. Tested Theoretical Models



As in RBD, each model predicts that early-life intelligence affects later-life intelligence. Model A implies effects of education on intelligence. Model B implies effects of education on both intelligence and at least one and potentially all – as indicated by dashed lines – specific subtests. Model C implies effects on at least one and potentially all specific subtests.

Figure 2. Model A

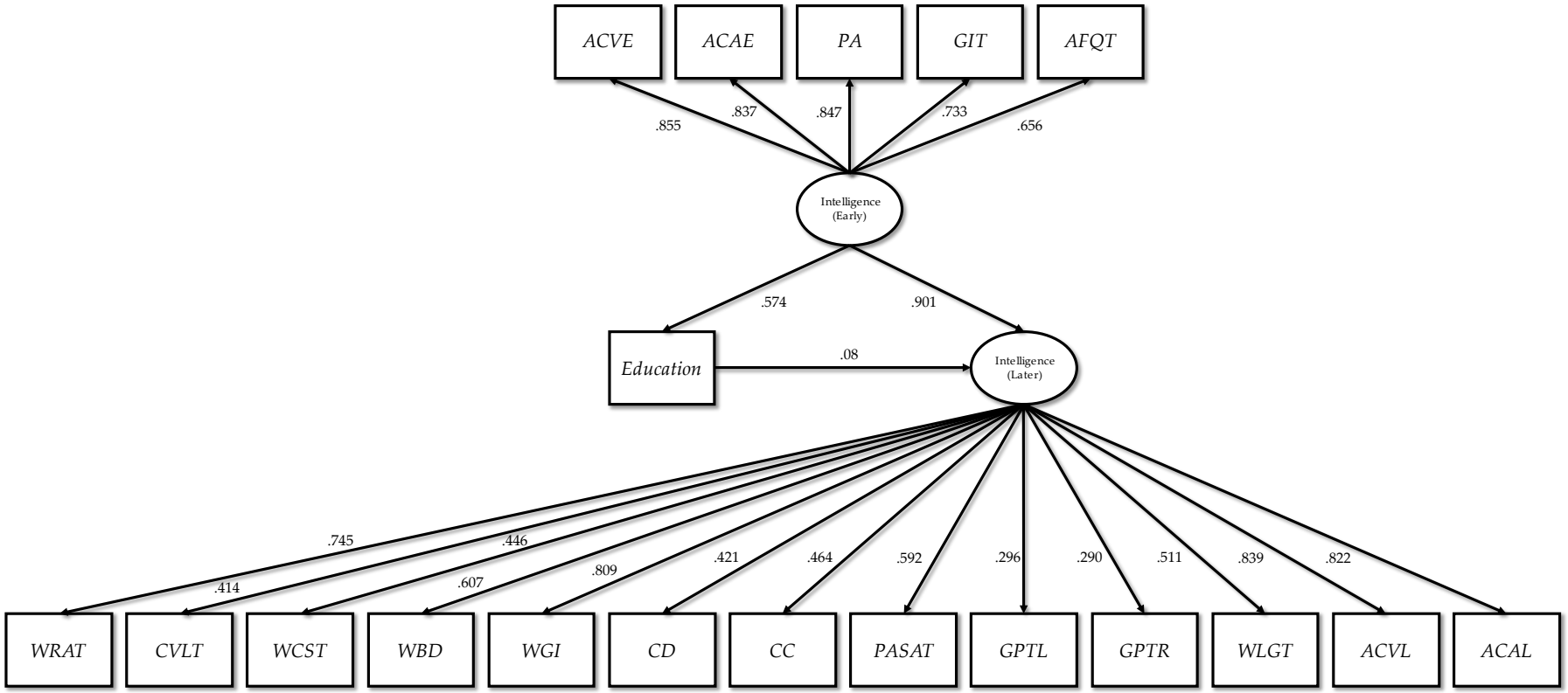


Figure 3. Model B

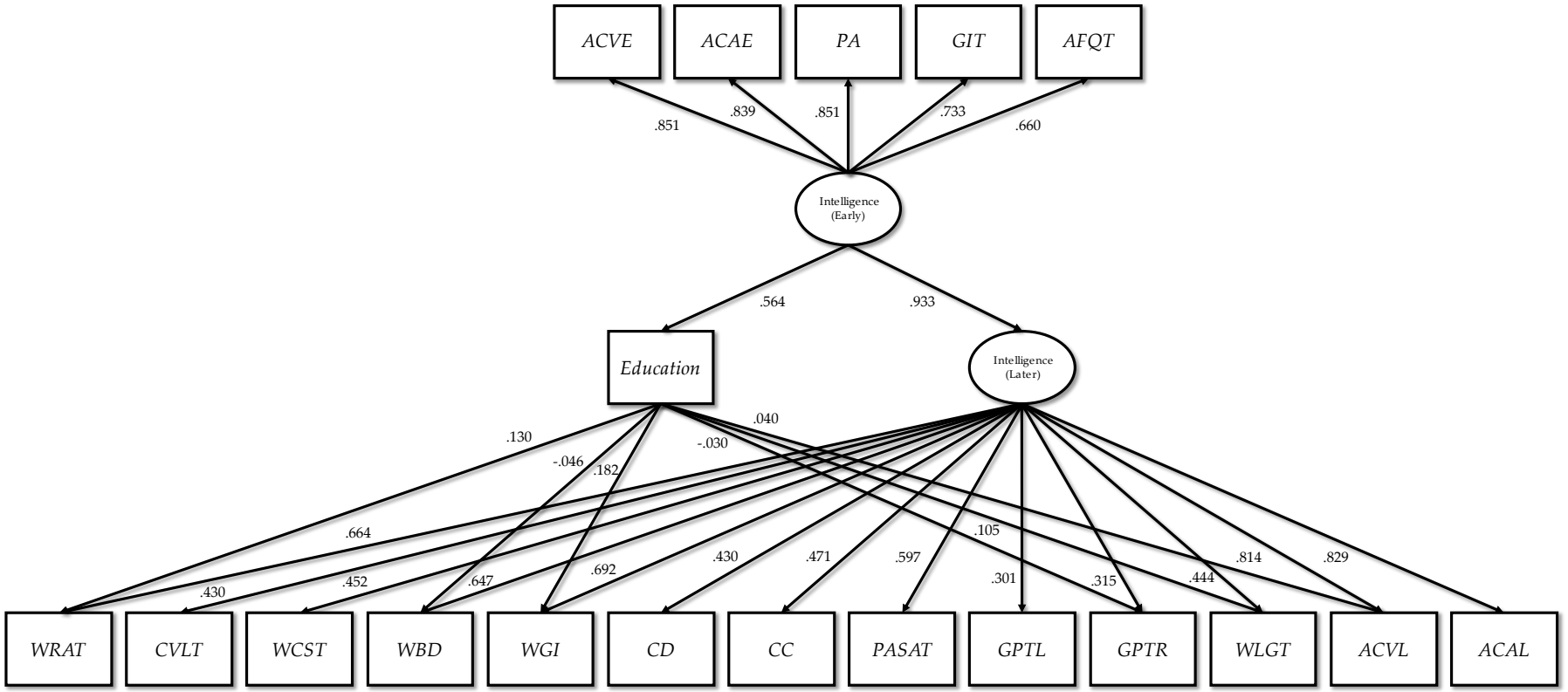
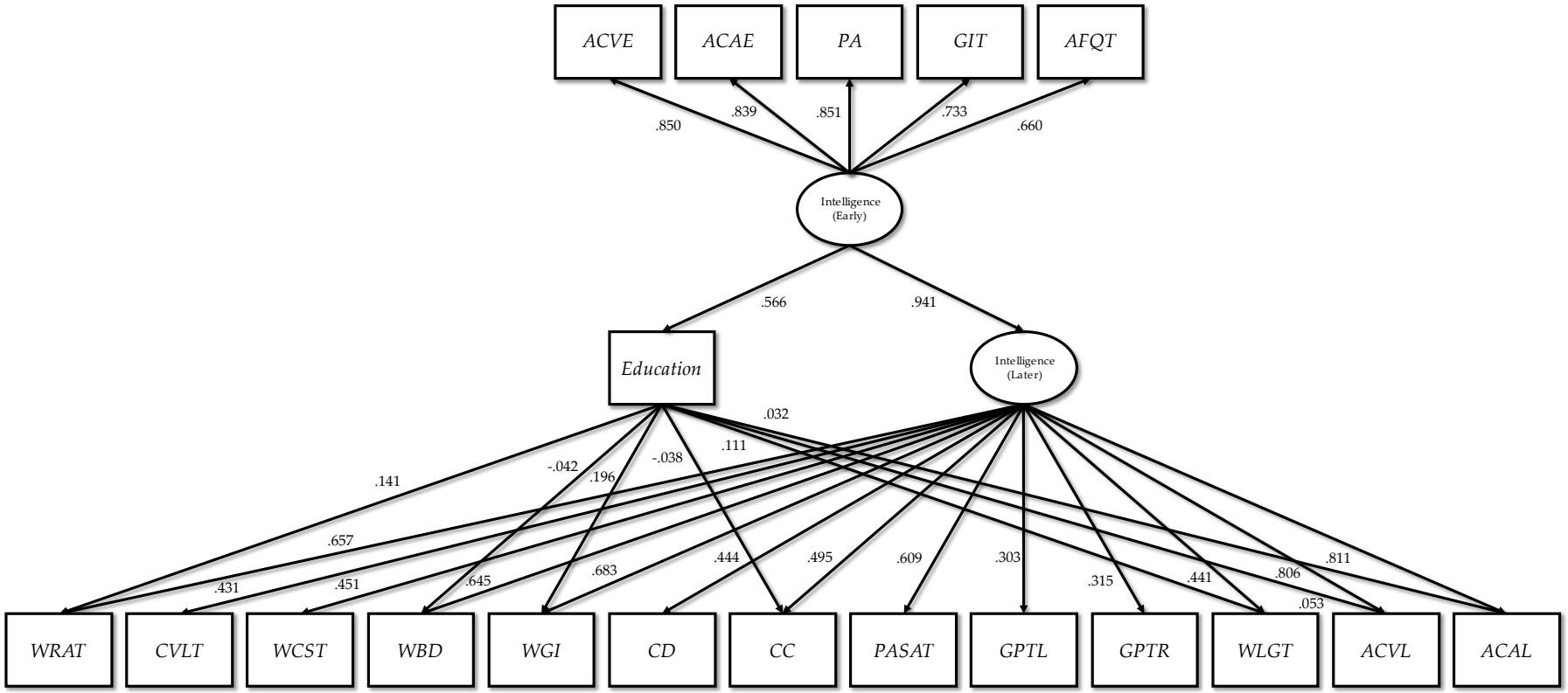


Figure 4. Model C



References

Ahmed, A., Kramer, M. S., Bernard, J. Y., Perez Trejo, M. E., Martin, R. M., Oken, E., & Yang, S. (2020). Early childhood growth trajectory and later cognitive ability: Evidence from a large prospective birth cohort of healthy term-born children. *International Journal of Epidemiology*, *49*(6), 1998–2009. https://doi.org/10.1093/ije/dyaa105

Alexander Beaujean, A., & Sheng, Y. (2010). Examining the Flynn Effect in the General Social Survey Vocabulary test using item response theory. *Personality and Individual Differences*, *48*(3), 294–298. https://doi.org/10.1016/j.paid.2009.10.019

Beaujean, A. (2006). *USING ITEM RESPONSE THEORY TO ASSESS THE LYNN-FLYNN EFFECT* [PhD, University of Missouri-Columbia]. https://web.archive.org/web/20190124005432/http://www.azmonyar.com/DownloadPDF/89592758.pdf

Beaujean, A. A., & Osterlind, S. J. (2008). Using Item Response Theory to assess the Flynn Effect in the National Longitudinal Study of Youth 79 Children and Young Adults data. *Intelligence*, *36*(5), 455–463. https://doi.org/10.1016/j.intell.2007.10.004

Beaujean, A., & Sheng, Y. (2014). Assessing the Flynn Effect in the Wechsler Scales. *Journal of Individual Differences*, *35*(2), 63–78. https://doi.org/10.1027/1614-0001/a000128

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

Benson, N. F., Beaujean, A. A., McGill, R. J., & Dombrowski, S. C. (2018). Revisiting Carroll's survey of factor-analytic studies: Implications for the clinical assessment of intelligence. *Psychological Assessment*, *30*(8), 1028–1038. https://doi.org/10.1037/pas0000556

Berkowitz, M., & Stern, E. (2018). Which Cognitive Abilities Make the Difference? Predicting Academic Achievements in Advanced STEM Studies. *Journal of Intelligence*, *6*(4), 48. https://doi.org/10.3390/jintelligence6040048

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425. https://doi.org/10.1007/s11336-006-1447-6

Branwen, G. (2012). *Dual n-Back Meta-Analysis*. https://www.gwern.net/DNB-meta-analysis

Breit, M., Scherrer, V., & Preckel, F. (2021). Temporal stability of specific ability scores and intelligence profiles in high ability students. *Intelligence*, *86*, 101538. https://doi.org/10.1016/j.intell.2021.101538

Butler, S. R., Marsh, H. W., Sheppard, M. J., & Sheppard, J. L. (1985). Seven-year longitudinal study of the early prediction of reading achievement. *Journal of Educational Psychology*, *77*(3), 349–361. https://doi.org/10.1037/0022-0663.77.3.349

Caplan, B. (2018). *The Case against Education: Why the Education System Is a Waste of Time and Money* (Illustrated edition). Princeton University Press.

Clark, G., & Cummins, N. (2020). *Does Education Matter? Tests from Extensions of Compulsory Schooling in England and Wales 1919-22, 1947, and 1972* (SSRN Scholarly Paper ID 3688207). Social Science Research Network. https://papers.ssrn.com/abstract=3688207

Cucina, J. M., Peyton, S. T., Su, C., & Byle, K. A. (2016). Role of mental abilities and mental tests in explaining high-school grades. *Intelligence*, *54*, 90–104. https://doi.org/10.1016/j.intell.2015.11.007

Deary, I. J. (2014). The Stability of Intelligence From Childhood to Old Age. *Current Directions in Psychological Science*, *23*(4), 239–245.

Deary, I. J., & Johnson, W. (2010). Intelligence and education: Causal perceptions drive analytic processes and therefore conclusions. *International Journal of Epidemiology*, *39*(5), 1362–1369. https://doi.org/10.1093/ije/dyq072

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*(1), 13–21. https://doi.org/10.1016/j.intell.2006.02.001

Dutton, E., & Kirkegaard, E. (2021). The Negative Religiousness-IQ Nexus is a Jensen Effect on Individual-Level Data: A Refutation of Dutton et al.'s 'The Myth of the Stupid Believer.' *Journal of Religion and Health*. https://doi.org/10.1007/s10943-021-01351-1

Edwards, J. (2018). A replication of 'Education and catch-up in the Industrial Revolution' (American Economic Journal: Macroeconomics, 2011). *Economics*, *12*(1), 20180003. https://doi.org/10.5018/economics-ejournal.ja.2018-3

Fergusson, D. M., John Horwood, L., & Ridder, E. M. (2005). Show me the child at seven II: Childhood intelligence and later outcomes in adolescence and young adulthood. *Journal of Child Psychology and Psychiatry*, *46*(8), 850–858. https://doi.org/10.1111/j.1469-7610.2005.01472.x

Floyd, R. G., Reynolds, M. R., Farmer, R. L., & Kranzler, J. H. (2013). Are the General Factors From Different Child And Adolescent Intelligence Tests the Same? Results From a Five-

Sample, Six-Test Analysis. *School Psychology Review*, *42*(4), 383–401. https://doi.org/10.1080/02796015.2013.12087461

Flynn, J. R. (2019a). A final reply to te Nijenhuis et al. (2019). *Journal of Biosocial Science*, *51*(6), 920–921. https://doi.org/10.1017/S0021932019000282

Flynn, J. R. (2019b). A response to te Nijenhuis et al. (2019). *Journal of Biosocial Science*, *51*(6), 913–916. https://doi.org/10.1017/S0021932019000270

Fox, M. C., & Mitchum, A. L. (2013). A knowledge-based theory of rising scores on "culture-free" tests. *Journal of Experimental Psychology: General*, *142*(3), 979–1000. https://doi.org/10.1037/a0030155

Fox, M. C., & Mitchum, A. L. (2014). Confirming the Cognition of Rising Scores: Fox and Mitchum (2013) Predicts Violations of Measurement Invariance in Series Completion between Age-Matched Cohorts. *PLoS ONE*, *9*(5), e95780. https://doi.org/10.1371/journal.pone.0095780

Franić, S., Dolan, C. V., Borsboom, D., Hudziak, J. J., van Beijsterveldt, C. E. M., & Boomsma, D. I. (2013). Can genetics help psychometrics? Improving dimensionality assessment through genetic factor modeling. *Psychological Methods*, *18*(3), 406–433. https://doi.org/10.1037/a0032755

Gohm, C. L., Humphreys, L. G., & Yao, G. (1998). Underachievement among Spatially Gifted Students. *American Educational Research Journal*, *35*(3), 515–531. https://doi.org/10.2307/1163447

Gow, A. J. (2016). Intelligence and Aging. In N. A. Pachana (Ed.), *Encyclopedia of Geropsychology* (pp. 1–13). Springer Singapore. https://doi.org/10.1007/978-981-287-080-3_261-1

Haier, R. E. (2014). Increased intelligence is a myth (so far). *Frontiers in Systems Neuroscience*, *8*. https://doi.org/10.3389/fnsys.2014.00034

Jensen, A. R. (1997). Adoption data and two g-related hypotheses. *Intelligence*, *25*(1), 1–6. https://doi.org/10.1016/S0160-2896(97)90003-9

Johnson, W., McGue, M., & Iacono, W. G. (2006). Genetic and environmental influences on academic achievement trajectories during adolescence. *Developmental Psychology*, *42*(3), 514–532. https://doi.org/10.1037/0012-1649.42.3.514

Kan, K.-J., Wicherts, J. M., Dolan, C. V., & van der Maas, H. L. J. (2013). On the Nature and Nurture of Intelligence and Specific Cognitive Abilities: The More Heritable, the More Culture Dependent. *Psychological Science*, *24*(12), 2420–2428. https://doi.org/10.1177/0956797613493292

Larsen, L., Hartmann, P., & Nyborg, H. (2008). The stability of general intelligence from early adulthood to middle-age. *Intelligence*, *36*(1), 29–34. https://doi.org/10.1016/j.intell.2007.01.001

Lasker, J., Nyborg, H., & Kirkegaard, E. O. W. (2021). *Spearman's Hypothesis in the Vietnam Experience Study and National Longitudinal Survey of Youth '79*. PsyArXiv. https://doi.org/10.31234/osf.io/m4yn9

Lechner, C. M., Gauly, B., Miyamoto, A., & Wicht, A. (2021). Stability and change in adults' literacy and numeracy skills: Evidence from two large-scale panel studies. *Personality and Individual Differences*, *180*, 110990. https://doi.org/10.1016/j.paid.2021.110990

Mansukoski, L., Hogervorst, E., Fúrlan, L., Galvez-Sobral, J. A., Brooke-Wavell, K., & Bogin, B. (2019). Instability in longitudinal childhood IQ scores of Guatemalan high SES

individuals born between 1941-1953. *PLOS ONE*, *14*(4), e0215828. https://doi.org/10.1371/journal.pone.0215828

McGue, M., Anderson, E. L., Willoughby, E., Giannelis, A., Iacono, W. G., & Lee, J. J. (2022). Not by g alone: The benefits of a college education among individuals with low levels of general cognitive ability. *Intelligence*, *92*, 101642. https://doi.org/10.1016/j.intell.2022.101642

McGue, M., Willoughby, E. A., Rustichini, A., Johnson, W., Iacono, W. G., & Lee, J. J. (2020). The Contribution of Cognitive and Noncognitive Skills to Intergenerational Social Mobility. *Psychological Science*, *31*(7), 835–847. https://doi.org/10.1177/0956797620924677

Must, O., & Must, A. (2013). Changes in test-taking patterns over time. *Intelligence*, *41*(6), 780–790. https://doi.org/10.1016/j.intell.2013.04.005

Must, O., & Must, A. (2018). Speed and the Flynn Effect. *Intelligence*, *68*, 37–47. https://doi.org/10.1016/j.intell.2018.03.001

Must, O., te Nijenhuis, J., Must, A., & van Vianen, A. E. M. (2009). Comparability of IQ scores over time. *Intelligence*, *37*(1), 25–33. https://doi.org/10.1016/j.intell.2008.05.002

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: Rodgers' hypothesis yes, Brand's hypothesis perhaps. *Intelligence*, *41*(6), 791–801. https://doi.org/10.1016/j.intell.2013.06.005

Plomin, R., Pedersen, N. L., Lichtenstein, P., & McClearn, G. E. (1994). Variability and stability in cognitive abilities are largely genetic later in life. *Behavior Genetics*, *24*(3), 207–215. https://doi.org/10.1007/BF01067188

Prada, M. F., & Urzúa, S. S. (2014). *One Size does not Fit All: Multiple Dimensions of Ability, College Attendance and Wages* (Working Paper No. 20752; Working Paper Series). National Bureau of Economic Research. https://doi.org/10.3386/w20752

Protzko, J. (2017). Effects of cognitive training on the structure of intelligence. *Psychonomic Bulletin & Review*, *24*(4), 1022–1031. https://doi.org/10.3758/s13423-016-1196-1

Protzko, J., & Colom, R. (2021). Testing the structure of human cognitive ability using evidence obtained from the impact of brain lesions over abilities. *Intelligence*, *89*, 101581. https://doi.org/10.1016/j.intell.2021.101581

Protzko, J., Nijenhuis, J. te, Ziada, K., Metwaly, H. A. M., & Bakhiet, S. (2021). *What to do Without a Control Group: You have to go latent, but not all latents are equal*. PsyArXiv. https://doi.org/10.31234/osf.io/vymp3

Ree, M. J., & Carretta, T. R. (2011). The Observation of Incremental Validity Does Not Always Mean Unique Contribution to Prediction. *International Journal of Selection and Assessment*, *19*(3), 276–279. https://doi.org/10.1111/j.1468-2389.2011.00556.x

Ree, M. J., & Carretta, T. R. (2022). Thirty years of research on general and specific abilities: Still not much more than g. *Intelligence*, *91*, 101617. https://doi.org/10.1016/j.intell.2021.101617

Reise, S., Morizot, J., & Hays, R. (2007). The Role of the Bifactor Model in Resolving Dimensionality Issues in Health Outcomes Measures. *Quality of Life Research : An*

International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, *16 Suppl 1*, 19–31. https://doi.org/10.1007/s11136-007-9183-7

Ritchie, S. J., Bates, T. C., & Deary, I. J. (2015). Is Education Associated With Improvements in General Cognitive Ability, or in Specific Skills? *Developmental Psychology*, *51*(5), 573–582. https://doi.org/10.1037/a0038981

Ritchie, S. J., & Tucker-Drob, E. M. (2018). How Much Does Education Improve Intelligence? A Meta-Analysis. *Psychological Science*, *29*(8), 1358–1369. https://doi.org/10.1177/0956797618774253

Rönnlund, M., Sundström, A., & Nilsson, L.-G. (2015). Interindividual differences in general cognitive ability from age 18 to age 65years are extremely stable and strongly associated with working memory capacity. *Intelligence*, *53*, 59–64. https://doi.org/10.1016/j.intell.2015.08.011

Ropovik, I. (2015). A cautionary note on testing latent variable models. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01715

Rosseel, Y., Jorgensen, T. D., Rockwood, N., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M., Scharf, F., & Du, H. (2022). *lavaan: Latent Variable Analysis* (0.6-11) [Computer software]. https://CRAN.R-project.org/package=lavaan

Sala, G., Aksayli, N. D., Tatlidil, K. S., Tatsumi, T., Gondo, Y., & Gobet, F. (2019). Near and Far Transfer in Cognitive Training: A Second-Order Meta-Analysis. *Collabra: Psychology*, *5*(1), 18. https://doi.org/10.1525/collabra.203

Sala, G., & Gobet, F. (2017). Does Far Transfer Exist? Negative Evidence From Chess, Music, and Working Memory Training. *Current Directions in Psychological Science*, *26*(6), 515–520. https://doi.org/10.1177/0963721417712760

Sandewall, Ö., Cesarini, D., & Johannesson, M. (2014). The co-twin methodology and returns to schooling—Testing a critical assumption. *Labour Economics*, *26*, 1–10. https://doi.org/10.1016/j.labeco.2013.10.002

Schalke, D., Brunner, M., Geiser, C., Preckel, F., Keller, U., Spengler, M., & Martin, R. (2013). Stability and change in intelligence from age 12 to age 52: Results from the Luxembourg MAGRIP study. *Developmental Psychology*, *49*(8), 1529–1543. https://doi.org/10.1037/a0030623

Shiu, W., Beaujean, A. A., Must, O., te Nijenhuis, J., & Must, A. (2013). An item-level examination of the Flynn effect on the National Intelligence Test in Estonia. *Intelligence*, *41*(6), 770–779. https://doi.org/10.1016/j.intell.2013.05.007

Sorjonen, K., Aurell, J., & Melin, B. (2017). Predicting group differences from the correlation of vectors. *Intelligence*, *64*, 67–70. https://doi.org/10.1016/j.intell.2017.07.008

Stanek, K. C., Iacono, W. G., & McGue, M. (2011). Returns to Education: What Do Twin Studies Control? *Twin Research and Human Genetics*, *14*(6), 509–515.

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence*, *35*(5), 401–426. https://doi.org/10.1016/j.intell.2006.09.004

te Nijenhuis, J., Jongeneel-Grimen, B., & Armstrong, E. L. (2015). Are adoption gains on the g factor? A meta-analysis. *Personality and Individual Differences*, *73*, 56–60. https://doi.org/10.1016/j.paid.2014.09.022

te Nijenhuis, J., van der Boor, E., Choi, Y. Y., & Lee, K. (2019). Do schooling gains yield anomalous Jensen effects? A reply to Flynn (2019) including a meta-analysis. *Journal of Biosocial Science*, *51*(6), 917–919. https://doi.org/10.1017/S002193201900021X

te Nijenhuis, J. te, Choi, Y. Y., Hoek, M. van den, Valueva, E., & Lee, K. H. (2019). Spearman's hypothesis tested comparing Korean young adults with various other groups of young adults on the items of the Advanced Progressive Matrices. *Journal of Biosocial Science*, *51*(6), 875–912. https://doi.org/10.1017/S0021932019000026

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, *101*(4), 817–835. https://doi.org/10.1037/a0016127

Wai, J., & Putallaz, M. (2011). The Flynn effect puzzle: A 30-year examination from the right tail of the ability distribution provides some missing pieces. *Intelligence*, *39*(6), 443–455. https://doi.org/10.1016/j.intell.2011.07.006

Watkins, M. W., & Canivez, G. L. (2004). Temporal Stability of WISC-III Subtest Composite: Strengths and Weaknesses. *Psychological Assessment*, *16*(2), 133–138. https://doi.org/10.1037/1040-3590.16.2.133

Watkins, M. W., & Smith, L. G. (2013). Long-term stability of the Wechsler Intelligence Scale for Children—Fourth Edition. *Psychological Assessment*, *25*(2), 477–483. https://doi.org/10.1037/a0031653

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., van Baal, G. C. M., Boomsma, D. I., & Span, M. M. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence*, *32*(5), 509–537. https://doi.org/10.1016/j.intell.2004.07.002

Woodley, M. A., te Nijenhuis, J., Must, O., & Must, A. (2014). Controlling for increased guessing enhances the independence of the Flynn effect from g: The return of the Brand effect. *Intelligence*, *43*, 27–34. https://doi.org/10.1016/j.intell.2013.12.004

Yu, H., McCoach, D. B., Gottfried, A. W., & Gottfried, A. E. (2018). Stability of intelligence from infancy through adolescence: An autoregressive latent variable model. *Intelligence*, *69*, 8–15. https://doi.org/10.1016/j.intell.2018.03.011

Zaboski, B. A., Kranzler, J. H., & Gage, N. A. (2018). Meta-analysis of the relationship between academic achievement and broad abilities of the Cattell-horn-Carroll theory. *Journal of School Psychology*, *71*, 42–56. https://doi.org/10.1016/j.jsp.2018.10.001