Intelligence and Age at First Intercourse: Cause or Confound?

S. Mason Garrison and Joseph Lee Rodgers

Vanderbilt University

Author Note

S. Mason Garrison, Department of Psychology and Human Development, Vanderbilt University; Joseph Lee Rodgers, Department of Psychology and Human Development, Vanderbilt University.

This manuscript is based on longitudinal data from the National Longitudinal Survey of Youth 1979 (NLSY79), part of the National Longitudinal Surveys (NLS) program. This data is publicly available; a non-exhaustive bibliography of articles using these surveys can be found at `https://nlsinfo.org/bibliography-start` This material is based upon work that has been supported by the National Institute of Health under Grant No. (R01-HD065865) and the National Science Foundation Graduate Research Fellowship Program under Grant No. (DGE-1445197), and various means of institutional support from the following universities: University of British Columbia, University of Oklahoma & Vanderbilt University. The aforementioned funding sources did not have any input in the production of this article.

Correspondence concerning this article should be addressed to S. Mason Garrison, Department of Psychology and Human Development, Vanderbilt University, Nashville, TN. Contact: s.mason.garrison@gmail.com

Abstract

Last compile was 2015/10/30 at 14:19:19

Keywords: Age at first intercourse; Cross-sectional data; Intelligence; Quasi-Experimental

Intelligence and Age at First Intercourse: Cause or Confound?

Anecdotal evidence from the popular media, such as MTV's reality television franchise, *16 and Pregnant*, suggests that teenage promiscuity is on the rise. Academic evidence confirms such anecdotes; age at first intercourse (AFI) is indeed declining and has so for some time (Bozon, 2003; Finer, 2007). Early AFI is associated with a plethora of negative downstream consequences, including lower education attainment (Harden, 2012; Spriggs & Halpern, 2008; Wellings et al., 2001), failure to meet education and career goals (Halpern, Joyner, Udry, & Suchindran, 2000), increased risk of teenage pregnancy (Leitenberg & Saltzman, 2000; Wellings et al., 2001), and increased rates of sexually transmitted infections (Kaestle, Halpern, Miller, & Ford, 2005). Moreover, beyond the obvious benefit of avoiding those negative outcomes, delaying AFI is associated with greater relationship satisfaction, perception of increased attractiveness, and higher household income (Harden, 2012). Because the aforementioned consequences are severe and long-reaching, psychology has begun to explore potential causal mechanisms of early AFI. Indeed, the field has found a consistent correlate in the literature – intelligence.

Higher levels of intelligence are positively associated with delaying first intercourse (Halpern et al., 2000; Mott, 1983; Paul, Fitzjohn, Herbison, & Dickson, 2000; Woodward, Fergusson, & Horwood, 2001). Specifically, it seems that intelligent individuals delay intercourse to "safeguard" their futures (Kirby, 2002b; Manlove, 1998; Raffaelli & Crockett, 2003). They perceive the risks associated with early intercourse, (e.g., pregnancy, STIs) to have career–shattering outcomes (Halpern et al., 2000; Harden & Mendle, 2011). Although this correlate holds promise – much of the field has treated this finding as causal and non-spurious. Yet, there is a fundamental confound in the existing literature that makes it impossible to infer causality.

Practically, all of the AFI-intelligence literature has used between family analyses. In all such analyses, gene and environmental influences, such as education and maternal intelligence are hopelessly confounded (See Harden, 2014). By ignoring such confounds, results are uninterpretable and risk misattributions of causality (Rowe & Rodgers, 1997;

Rutter, 2007). Indeed, both intelligence and AFI are highly heritable and have sizable shared environmental variances (Harden & Mendle, 2011; Harden, 2014; Plomin & Spinath, 2004). Thus, we need to critically evaluate whether intelligence is a cause of AFI or merely a theoretically attractive confound.

Cause or Confound?

There are numerous theories on the motivations behind adolescents initiation of first intercourse (See Rodgers, 1996 or Buhi & Goodson, 2007 for a review), and even more specific antecedents (Buhi & Goodson, 2007; Kirby, 2002a; B. C. Miller et al., 1997; Santelli & Beilenson, 1992). Many of these theories either emphasize biology/genetics, where typical adolescent development through puberty (and various hormone changes) drives the interest in sexual behavior (W. B. Miller et al., 1999; Udry, 1979), or social/environmental processes, such as Social Learning (DiBlasio & Benda, 1990; Hogben & Byrne, 1998), where social norms alter the likelihood of early sexual behavior; or Social Control theory (Hirschi, 2002), where societal convention reduces the likelihood that individuals will act on their naturally deviant behavior. Under these environmentally centric theories the underlying biology is either ignored or actively resisted (in the case of Social Control theory, while under many of the genetic centric theories, the environmental components are ignored.

Recently, there have been numerous articles advocating integrative models (See Harden, Mendle, Hill, Turkheimer, & Emery, 2008 and Harden, 2014). The integrative Biopsychosocial Model acknowledges both genetic and environmental contributions to human behavior (Engel, 1977; Petersen, 1987; Rodgers, Rowe, & Buster, 1999). Indeed, biology, psychology, and society jointly influence adolescents' decisions to engage in sexual intercourse (Meschke, Zweig, Barber, & Eccles, 2000; Zimmer-Gembeck & Helfand, 2008). Even though this paper focuses on a single predictor – intelligence, we are doing so within the broader context.

Intelligence as the Cause

We've previously mentioned that the short-term risks of early AFI are overwhelming negative, whereas the rewards for delay are equally positive. These consequences extend into adulthood – early AFI is associated with adult delinquency (Harden et al., 2008), anti-social behavior, and substance abuse (Boislard & Poulin, 2011), while those who delayed had higher household incomes in adulthood (Harden, 2012). It is intuitively appealing to believe that intelligent individuals are more likely to observe this high risk, low reward trade off, and act upon such observations by delaying intercourse. Accordingly, intelligent individuals perceive the consequences of early AFI to have career–shattering outcomes (Halpern et al., 2000; Harden & Mendle, 2011).

Indeed, the literature is consistent with this theory. Those with higher educational goals delay their first intercourse (Boislard & Poulin, 2011; Schvaneveldt, Miller, Berry, & Lee, 2001), while those who had previously reported higher goals, but engaged in early sexual intercourse reduced their goals (Schvaneveldt et al., 2001). Beyond academic goals, those with a greater affinity for risk and those who perceive benefits from teen-pregnancy are more likely to engage in risky sexual activities (Raffaelli & Crockett, 2003). A greater understanding of the risks associated with sexual intercourse, such as HIV transmission, is also associated with delayed AFI (C. Mathews et al., 2009).

Smarter adolescents are more likely to report delayed intercourse (Halpern et al., 2000; Mott, 1983; Paul et al., 2000; Woodward et al., 2001). Beyond intercourse, smarter individuals appear to postpone all sexual/romantic activity (Halpern et al., 2000). Such blanket delays may be a proactive attempt to avoid first intercourse precursors. Thus, many researchers have concluded that "[h]igher intelligence operates as a protective factor against early sexual activity during adolescence, and lower intelligence, to a point, is a risk factor." (Halpern et al., 2000)[pg., 213].

However, Halpern et al. (2000) and many of the other studies we have referenced above(e.g., C. Mathews et al., 2009; B. C. Miller et al., 1997; Paul et al., 2000) have used between family designs, typically cross-sectional in nature. Such designs cannot distinguish between processes that act to create differences between families and

processes that create differences among family members(Lahey & D'Onofrio, 2010). Thus the previous studies do no provide conclusive evidence that intelligence is the causal influence behind the AFI-intelligence relationship.

Intelligence as a Confound

A equally valid family of explanations exist in which intelligence is not the driver of the AFI-intelligence relationship, merely a theoretically attractive confound. Instead, various confounds including family level selection effects, or third variables at the individual level could be causing the relationship. Indeed many such findings that link intelligence with various outcomes are the product of misattributing between family confounds to individual level causes.

The relationship between birth order and intelligence is a classic example of this misattribution (See Rodgers, Cleveland, van den Oord, & Rowe, 2000, Rodgers, 2014, or Damian & Roberts, 2015). Between family studies have consistently found that first born children have higher IQs than later born children (Belmont & Marolla, 1973; Zajonc, 1976), and that first borns are higher achievers (Clark & Rice, 1982; Galton, 1875). Yet within family studies have just as consistently found zero relationship (Berbaum & Moreland, 1980; Retherford & Sewell, 1991; Rodgers et al., 2000). Moreover, when within and between analyses are simultaneously conducted, the methodological source of the IQ-birth order effect are clearly revealed – between family differences in family size (Black, Devereux, & Salvanes, 2011; Rodgers et al., 2000; Wichman, Rodgers, & Maccallum, 2006, 2007). Potential causes of this confound include parental IQ and SES[1](Page & Grandon, 1979; Rodgers et al., 2000). See Anastasi (1956) for an insightful overview, written prior to the IQ-birth order debate.[2]

Between family influences such as SES and maternal intelligence could drive the

[1]Selection effects based on SES should not to be confused with the confluence/resource dilution model (Blake, 1981; Zajonc & Bargh, 1980)

[2]"Parenthetically, it may be added that studies on the relation of birth order to intellectual and other psychological characteristics have frequently yielded ambiguous and inconsistent results because of the failure to take family size into account."(Anastasi, 1956, pg 201)

relationship. Socioeconomic status is associated with the onset of first intercourse (Lammers, Ireland, Resnick, & Blum, 2000) and correlated with intelligence (Murray, 1998; Neisser et al., 1996; Strenze, 2007). Parental intelligence and parental education are also linked with child intelligence (Bouchard, Jr., 2004; Devlin, Daniels, & Roeder, 1997; Mercy & Steelman, 1982), and pose very viable alternative explanations in which parents are the one dissuading their children from engaging in early intercourse. For example, daughters whose mothers communicated frequently about the risk associated with sexual intercourse were less likely to have unprotected sex and engaged in sex less frequently (Hutchinson, Jemmott, Jemmott, Braverman, & Fong, 2003). Thus it could be that intelligent mothers, not intelligent children, are the ones recognizing the consequences of early intercourse and acting accordingly.in order to truly understand the causal nature of intelligence on Age at First Intercourse, we need to be able to untangle between and within family processes.

Prior Within Family Analyses

The authors are aware of two studies which explicitly untangled between and within family influences on the AFI-intelligence relationship (Harden & Mendle, 2011; Meredith, 2013).[3] Harden & Mendle (2011) used 536 same-sex twin pairs from the Add Health Study to "test[] whether relations between intelligence, academic achievement and age at first sex were due to unmeasured genetic and environmental differences between families." Twins who differed in their intelligence or their academic achievement did not differ in their age at first intercourse. They concluded that "the association between intelligence and age at first sex could be attributed entirely to unmeasured environmental differences between families."

[3]Technically, three studies – Nedelec, Schwartz, Connolly, & Beaver (2012) conducted an extensive exploratory analysis of MZ twin pairs from the Add Health Study. They used intelligence difference scores to predict various social outcomes. They generally found null results in their small samples (N ranged from 48 to 166 pairs). Their sample is an underpowered subset of the same sample that Harden & Mendle (2011) used.

Current Study

To summarize, the current study examines the relationship between intelligence and age at first intercourse, using siblings and their children from a multi-generational nationally representative sample. This examination extends the intelligence literature in several key ways. We (1) tested whether the relationship between intelligence and age at first intercourse was consistent using between and within family analyses; (2) evaluated the alternative explanation that maternal intelligence influences child AFI; and (3) replicated these findings using assessments of intelligence from other ages.

We made the following predictions, based primary upon Harden & Mendle (2011) and Meredith (2013):

Between Families,

1. Does Gen2 intelligence predict Gen2 AFI?: We expect intelligence to be associated with age of first intercourse because there is a sizable body of literature reporting that result (Kirby, 2002b; Manlove, 1998; Raffaelli & Crockett, 2003).

2. Does Gen1 intelligence predict Gen2 AFI?: We also expect maternal intelligence to be associated with age of first intercourse because the heritability of intelligence is quite high (Bouchard, Jr., 2004; Devlin et al., 1997). If intelligence does causally influence AFI we would expect that the cross-generational association between AFI and intelligence would be considerably weaker, but existent. However, if the intelligence-AFI relationship is the product of between family confounds, then we would expect that the cross-generational association between AFI and intelligence would be stronger than the within generation association because maternal intelligence would be more closely linked with household SES and various parental causes. Comparably sized effects would also be consistent with a between family confound. Given that Harden & Mendle (2011) and Meredith (2013) found no within family effect for intelligence, we expect that maternal intelligence will have a comparable or larger effect on between family AFI than child intelligence.

Within Families,

3. Does Gen2 intelligence predict Gen2 AFI?: No, we do not expect to find all

within family differences in intelligence and AFI, given that neither Harden & Mendle (2011) nor Meredith (2013) reported an effect.

4. Does Gen1 intelligence predict Gen2 AFI?: Unknown: it is possible that maternal intelligence will have an effect, as such a link would explain the between family effects as well as many of the alternative household-level influences.

5. Is the relationship consistent across methods?: Doubtful, we do not expect the results to be consistent across methods because both Harden & Mendle (2011) and Meredith (2013) found no within-familiy effect, while the traditional findings from between family studies find an effect (Kirby, 2002b; Manlove, 1998; Raffaelli & Crockett, 2003).

Method

Research Design

We adapted Kenny and colleagues (2001; 2006) reciprocal standard dyad model to facilitate sibling comparisons. Sibling-based quasi-experimental models are particularly effective at incorporating genetic and environmental design elements (Lahey & D'Onofrio, 2010; Rutter, 2007).

$$\mathrm{Y_{i\Delta}} = \beta_0 + \beta_1 \bar{\mathrm{Y}}_\mathrm{i} + \beta_2 \bar{\mathrm{X}}_\mathrm{i} + \beta_3 \mathrm{X_{i\Delta}} \tag{1}$$

where,

$$Y_{i1} = \max(Y_{ij});\ Y_{i2} = \min(Y_{ij});\ Y_{i\Delta} = Y_{i1} - Y_{i2};\ X_{i\Delta} = X_{i1} - X_{i2} \tag{2}$$

In this model, the relative difference in kin outcomes ($\mathrm{Y_\Delta}$) is predicted from the mean level of the outcome ($\mathrm{Y_{mean}}$), the mean level of the predictor ($\mathrm{X_{mean}}$), and the between-kin predictor difference ($\mathrm{X_\Delta}$). The mean levels support causal inference through at least partial control for genes and shared environment. Therefore, we simultaneously evaluate the individual difference ($\mathrm{X_\Delta}$) and the joint contribution of genes and shared environment ($\mathrm{Y_{mean}}$ & $\mathrm{X_{mean}}$).

More broadly, this model allows us to explicitly untangle between and within family influences. If there is a true causal effect between the individual difference and

the outcome (in our case – intelligence and AFI respectively), then we would expect kin differences in intelligence to be significantly associated with kin differences in the outcome. If the effect is a spurious effect – the function of between family differences – then we would expect to find no significant relationship between the differences in the outcome with the differences in the predictor.

Sample

The National Longitudinal Survey of Youth 1979 (NLSY79) is a nationally representative household probability sample, jointly sponsored by the US Department of Labor and US Department of Defense. In 1980, 12,686 adolescents were surveyed from 8,770 households on a battery of measures. The initial survey consisted of three subsamples:

- a cross–sectional probability sample of 6,111 non-institutionalized adolescents residing in the United States on December 31^{st} of 1978,
- an over–sampled civilian subsample of 5,295 racial minorities and disadvantaged whites, and
- a representative sample of 1,280 youths serving in the US Military on September 30^{th}, 1978.

In the civilian samples, subjects' birthdates ranged from January 1, 1957 to December 31, 1964, and were between the ages of 14 and 21 on 31^{st} of 1978, whereas military subject's birthdates ranged from January 1, 1957 to December 31, 1961, and were likewise between 17 and 21 years old. Participants were surveyed annually until 1994, and then surveyed biannually to the present day. Two waves of planned attrition occurred. After the 1984 interview, all but 201 randomly selected members of the military sample were dropped. After the 1990 interview, all 1,643 disadvantaged whites from the oversample were dropped. More information, such as the data and documentation can be found on the Bureau of Labor Statistics (BLS) website: `http://www.bls.gov/nls/nlsy79.htm`.

In 1986, the biological children of the female NLSY79 participants were surveyed, resulting in the NLSY79 Children and Young Adults (NLSY79–CYA) survey. These 11,512 participants are also surveyed on a biannual basis. Accordingly, participants in the NLSY79 will be periodically referred to as the Generation 1 (Gen1) sample, whereas the NLSY79–CYA will be referred to as the Generation 2 (Gen2) sample.

Tetrads

Mother-Child-Aunt-Nibling (MCAN) tetrads were created using the NLSY Kinship Links (Rodgers et al., 2015) and supporting R package (Beasley et al., 2015). The oldest two female kin (Mother, Aunt) were selected from each household; additional female Generation 1 kin were excluded. Three tetrad designs were employed, in which the genders of Generation 2 were the defining feature:

- Mother-Daughter-Aunt-Niece (MDAN) included the oldest female child from each of the sisters,
- Mother-Son-Aunt-Nephew (MSAN) included the oldest male child from each of the sisters, and
- Mother-Child-Aunt-Nibling (MCAN) included the first born child from each of the sisters.

All outcomes were standardized by gender prior to tetrad creation. Table 3 on page 27 reports descriptive statistics for all relevative variables used throughout this paper by whether the respondent has a sibling in the sample.

Age at First Intercourse

Generation 1. NLSY-79 subjects were surveyed about their AFI over a maximum of three time-points (1983, 1984, 1985). In theory, subjects were only to be asked in later waves, if subjects had not reported an AFI in the 1983 wave. However, in practice, many subjects were surveyed twice. Female participants were asked additional information (Year of First Intercourse, Month of First Intercourse) in waves 1984 and

1985. Because subjects were surveyed repeatedly, we used this opportunity to estimate the reliability of self-reported AFI as well as the reliability of the AFI difference scores. In Table 6 on page 30, the lower triangle reports the correlations of self-reported AFI across 1983-1985; the diagonal indicates the number of respondents reporting AFI for that year, and upper triangle indicate the number of respondents that reported AFI for both respect years. Stars indicate significant at the .01 level. The test-retest correlations are high ($r > .75$) across all viable pairings, suggesting that our subjects are consistently reporting their AFIs.

Gen2. Over the life-time of the NLSY-CYA survey, participants were surveyed about their AFI. The exact phrasing of the question varied by administration. Between 1988 and 2000, subjects were asked for age, year, and month of first intercourse. After 2000, subjects were only asked their age. The reason for this change is unknown. However, the first author suspects that the change had to do with the fact that the modal response for month was consistently: "Don't Know". Indeed, only 1147 subjects reported a viable month of first intercourse.

Regardless, we calculated AFI as follows, using SAS University Edition SAS Institute Inc (2015). First, we transformed year of first intercourse into age. If subjects reported both age and year within the same survey, we averaged the age scores. Across all survey years, we identified the minimum AFI and maximum AFI for each subject. Then we took the average of those two scores. Given that the expected AFI of a subject $\neq$ the reported AFI, we added 1 to the Maximum AFI. Therefore, if the subject only reported one instance of AFI, their AFI would now reflect their expected AFI. For example, a subject who reports AFI at 16 could be 16 years and 1 day old OR 16 years and 364 days old. Thus the expected value for 16 is in fact 16.5. We calculated AFI in this manner because we wished to include the maximum amount of information without ignoring the expected value problem with self-reported age. Using this method, the average Gen2 AFI was 16.01 years (sd = 2.30; n = 6288).[4]

[4]Taking the average of all AFIs (without addressing expected value), results in 15.49 (sd = 2.30; n = 6288). Adding in expected value of .5 changes this value to 15.99.

After transforming all AFI scores, we recoded all impossible AFIs as missing. We considered a score to be impossible if the reported AFI that exceeded participant's age at time of survey ($\overline{\text{AFI}} = 15.99$, sd = 2.30, n = 6235). Next, we excluded all AFIs below age 12 (16.14, sd = 2.10, n = 6087). Finally we excluded subjects who reported AFI prior to menstruation (16.16, sd = 2.09, n = 6047). We excluded those below age 12 because those responses likely are the result of misunderstanding or non-consensual sexual activity, while we excluded those with premenstrual AFI because of we were only interested in post-pubescent sexual activity. AFI varied by gender and race. Most notably, women reported AFIs that were 6 months later than men, and black men reported the lowest AFI (15 yrs). For a complete breakdown, see Table 3 on page 27 and see Figure 1 on page 38.

Measures

Generation 1

The Armed Services Vocational Aptitude Battery (ASVAB; Form 8A; Palmer, Hartke, Ree, Welsh, & Valentine, Jr., 1988) was administered to Gen1 participants during the summer and fall of 1980 (U.S. Department of Defense, 1982), and was used to establish national norms for the Department of Defense (Waters, Laurence, Camara, & Green, 1987). The Armed Forces Qualification Test (AFQT) is derived from the ASVAB, and used as a measure of general trainability (Maier & Sims, 1986). It is a composite of four subscales: Arithmetic Reasoning (AR; 30 items), Math Knowledge (MK; 25 items), Paragraph Comprehension (PC; 15 items), and Word Knowledge (WK; 35 items). Arithmetic Reasoning targets the ability to solve word problems. Math Knowledge also tests quantitative ability, by assessing knowledge of high school level mathematics, with special emphasis on algebra, fractions, and geometry. The remaining subscales focus on verbal ability, and are sometimes referred to as the Verbal Composite (VE). Specifically, Word Knowledge tests the subjects' knowledge of the meaning of words within a given context, whereas Paragraph Comprehension targets a subject's ability to understand the meanings of paragraphs. Other administrations of the pencil

and paper ASVAB reveal that all the AFQT subscales have high internal consistency ($\alpha_{AR} = .91$; $\alpha_{WK} = .92$; $\alpha_{PC} = .81$; $\alpha_{MK} = .87$; Kass, Mitchell, Grafton, & Wing, 1982). Reported reliability of the AFQT (8A) ranges from .87 to .93 (Palmer et al., 1988).

Methods of calculating the AFQT have varied throughout the ASVAB's administrative lifetime (Mayberry & Hiatt, 1992). For pencil and paper administrations, standard scores were created for each of the subscale scores ($\bar{x} = 50$, sd = 10), and then combined into a standard score. Then, the AFQT standard score is derived from the following formula:

$$\text{AFQT} = \text{AR} + \text{MK} + 2\text{VE}, \tag{3}$$

$$\text{where VE} = \text{PC} + \text{WK}. \tag{4}$$

This score is then converted into a percentile, which determines an applicant's basic qualification for enlistment. All applicants must earn a score at or above the 10^{th}percentile(Defense Manpower Data Center, 2012). Each branch has its own minimum score, ranging from 31 to 36 (U.S. Department of the Army, 2013; U.S. Coast Guard, 2004), and each branch uses different linear combinations of these subtests to determine an applicant's eligibility for specialty positions. Additionally, multiple researchers have used the AFQT standard score as a proxy for general intelligence (g) (Herrnstein & Murray, 1994; Der, Batty, & Deary, 2009). Indeed, the military has found that the AFQT correlated 0.8 with the Wechsler Adult Intelligence Scale (WAIS; McGrevy, Knouse, & Thompson, 1974). Moreover, the AFQT consistently predicts outcomes traditionally associated with intelligence(Welsh, Kucinkas, & Curran, 1990), including grades (Wilbourn, Valentine, Jr., & Ree, 1984; J. J. Mathews, 1977).

Generation 2

Administration of ability measures has varied considerably across the lifecourse of the NLSY-CYA survey (See Table 2.12 from Center for Human Resources Research, 2009 for a summary). However, the vast majority of subjects have completed the following test batteries:

- Peabody Individual Achievement Test (PIAT; Dunn & Markwardt, 1970):

  - Math Subtest (84 items),
  - Reading Recognition Subtest (84 items),
  - Reading Comprehension Subtest (84 items),

- The Peabody Picture Vocabulary Test-Revised (PPVT-R; Form L; Dunn & Dunn, 1981; 175 items), and
- Wechsler Intelligence Scales for ChildrenRevised (WISC-R; Wechsler, 1974) Digit Span Subscale (28 items).

Although individual item level data was available for all of the aforementioned tests, conducting a unidimensional 2-PL is not a viable means of estimating general ability because of the nature of test construction and administration. The PIATs and PPVT-R were administered to subjects in an adaptive manner. The starting items on the PIAT Math and PVVT-R were determined by age, whereas the starting items for the remaining PIAT subtests were determined based on PIAT Math performance. Moreover, administration of a given test were terminated when a subject reached a "ceiling." For example, testing was terminated for the PIAT Math if a subject incorrectly answered 5 of the most recent 7 questions (See Baker, Keck, Mott, & Quinlan, 1993 for a thorough overview of NLSY-CYA test administration protocols). In essence, this administration procedure results in a tremendous amount of non-randomly missing data.

Although the administration created non-randomly missing data, the standard scores of the PVVT-R, PIATs, and WISC-R Digit Span themselves are valid and very reliable assessments of cognitive ability (Mott & Baker, 1995). Accordingly, we elected to use the standard scores of all the Gen2 ability measures already mentioned. However, subjects were surveyed on a biannual basis. Thus we could not use cognitive tests at a given age. Instead, we aggregated scores across a 4 year window, and targeted ages 9 and 10. We targeted 9.5 because all cognitive tests were administered within the 8–11 age window, we wanted to maximize the number of subjects with viable ability scores, and we wanted to ensure temporal precedence with respect to AFI. In the case of missing subtests, we allowed age 11 scores to replace age 9 scores, and age 8 scores to

replace age 10 scores. By employing a 4 year window, all subjects had an equal chance of replacing the primary test administration. Our replacement strategy ensured that the average age of testing matched the average of our targeted ages.

Measurement. A unidimensional confirmatory factor analytic model was run in M*plus* 7.31 (Muthén & Muthén, 2014), and used a robust maximum likelihood estimator (MLR). There were 8,254 useable observations in 3,742 clusters. A single factor solution fit the model decently (RMSEA = .101, p(RMESA < .05) = 0; CFI = .973; TLI = .946, SRMR = .027). Table 4 on page 28 contains a full summary of the model fit statistics, and Table 5 on page 29 contains the factor loadings.

Replicability & Reliability. Given the recent concerns about replicatability in psychology (Open Science Collaboration, 2015), we repeated our aggregates of Gen2 intelligence, centered at ages 10.5 and 11.5, and replicated all of our analyses. These replications can be found in the Appendices A and B, respectively. Appendix A begins on page 52 and appendix B begins on page 60. The test-retest reliabilities of Gen2 intelligence across our three aggregations is reported in the lower triangle of Table 8 on page 30. The diagonal indicates the number of respondents with intelligence aggregations for that year, and upper triangle reveal the number of respondents with viable scores for both respective ages. Stars indicate significant at the .01 level. The test-retest correlations are very high (r > .90) across all pairings, suggesting that our method captures consistent (but not identical) measures of intelligence across ages. Additional analyses examining the reliability of intelligence difference scores are reported in a later section.

Reliability of Difference Scores

Our design assumes that the difference scores of our measures are reliable. INSERT MORE ON THIS. We've reported the test-retest reliability of Gen2 intelligence and Gen1 AFI in earlier sections. Here, we report the test-retest reliability of the differences of those measures across kin.

Estimated Reliabilities. Sibling differences in AFI as reported in 1983, 1984, and 1985 were strongly correlated with each other (See Table **??** on page ??).

Comparing sibling differences in AFI as reported in 1983 and 1984 (n = 783 pairs) we found a strong correlation (r = .76). The sample of sibling pairs with complete information in 1985 was too small (n = 12) to compare to the other two years. Regardless, sibling differences in self-reported AFI appear reliable. Although we could not calculate test-retest reliabilities for Gen2, we have no reason to believe that those differences would fundamentally differ from Gen1's reports.

Cousin differences in intelligence as assessed at ages 9.5, 10.5, and 11.5 were correlated using three different linking methods (Mixed, Daughers, Sons). Table 9 on page 31 reports the correlated differences of the first borns of each sister, Table 10 on page 31 reports the correlated differences of the first born girls, and Table 11 on page 31 reports the correlated differences of the first born sons. Reliabilities across linking methods was consistent and high (min r = .86; max r = .95). However, again, we were unable to calculate the test-retest difference score reliablities for Gen1.

Calculated Reliabilities. Nonetheless, we were able to confirm that difference scores for both generations were reliable for the measures we could estimate. For the remainder, we have calculated the reliability of the difference scores using the following equation (Lord, 1963):

$$\rho_{dd'} = \frac{\sigma_x^2 \rho_{xx'} + \sigma_y^2 \rho_{yy'} - 2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 - 2\rho_{xy}\sigma_x\sigma_y} \tag{5}$$

where,

- $\rho_{dd'}$ is the reliability of the difference score,
- σ_x^2 is the variance of kin_1's score,
- $\rho_{xx'}$ is the reliability of kin_1's score,
- σ_y^2 is the variance of kin_2's score,
- $\rho_{yy'}$ is the reliability of kin_2's score,
- ρ_{xy} is the correlation between kin_1's and kin_2's scores.

Accordingly, we can substitute the following values into equation 5 to calculate the difference score reliability for Generation 1 intelligence. where,

- σ_x^2 and σ_y^2 are 1.1881 (from Table 1 on page 26)
- $\rho_{xx'}$ and $\rho_{yy'}$ are .87 (the low end of AFQT reliability reported in Palmer et al., 1988),
- ρ_{xy} is the correlation between sisters is .67.

```
sigmax = sqrt(1.1881)

sigmay = sigmax

rxx = .87

ryy = rxx

rxy=.67


(sigmax*sigmax*rxx +sigmay*sigmay*ryy -2*rxy*sigmax*sigmay) /
  (sigmax*sigmax + sigmay*sigmay -2*rxy*sigmax*sigmay)
```

```
## [1] 0.606
```

The calculated reliability of Generation 1's differences in AFQT was 0.606, acceptable, but lower than the empirical correlation we derived for cousin differences.

We can also calculate the difference score reliability for Generation 2 AFI, by substituting the following values into equation 5, where,

- σ_x^2 and σ_y^2 are 4.41 (from Table 2 on page 26)
- $\rho_{xx'}$ and $\rho_{yy'}$ are .76 (from Table 6 on page 30),
- ρ_{xy} is the AFI correlation between first born cousins is .099.(from Figure 3 on page 40)

```
sigmax = sqrt(4.41)

sigmay = sigmax

rxx = .76

ryy = rxx

rxy=.099

(sigmax*sigmax*rxx +sigmay*sigmay*ryy -2*rxy*sigmax*sigmay) /
  (sigmax*sigmax + sigmay*sigmay -2*rxy*sigmax*sigmay)
```

```
## [1] 0.734
```

Gen2 Mean AFI difference scores were also reliable (r = 0.734) and comparable to Generation 1 sibling differences.

Results

We examined the relationship between AFI and intelligence using two designs: between and within families. The results are organized into those two designs. The between family analyses report the relationships between the average AFI and various measures of ability. The within family analyses attempt to replicate the between family findings by testing whether differences in AFI can be explained by differences in various measures of ability. If there is a causal relationship between intelligence and AFI then differences in AFI will be significantly associated with differences in ability. If the relationship is the result of between family confounds, such as shared environmental influences, then differences in AFI will not be significantly associated with differences in ability, and accordingly, AFI cannot be caused by intelligence.

Between Family

First, we examined the between family results. We tested whether the family average of Gen2 AFI could be predicted by the family averages of Gen1 ability and of Gen2 ability. We evaluated the influences both independently and simultaneously. All ability scores have been standardized by generation ($\overline{g} = 0$, sd $= 1$), prior to averaging by household. AFI scores have been standardized by gender ($\overline{\text{AFI}} = 0$, sd $= 1$), prior to averaging by household.

Gen1 Mean Intelligence → Gen2 Mean AFI. Gen1 sister averages of standardized AFQT scores were used to predict Gen2 averages of gender standardized AFI. Table 12 on page 32 displays the results by Gen2 linking. The Mixed model reports the averages of the first borns of each sister (n = 342), the Daughters model reports the averages of the first born girls (n = 264), and the Sons model reports the averages of the first born sons (n = 282). All three models reveal similar results. A one unit increase in the average standardized intelligence of the children's mothers predicted $\approx .013$ increase in average Gen2 AFI. The adjusted R^2 varied slightly by Gen2 linking (Mixed = .087, Daughters = .097, Sons = .103).

Gen2 Mean Intelligence → Gen2 Mean AFI. Gen2 cousin averages of standardized ability scores were used to predict Gen2 averages of gender standardized AFI. Table 13 on page 33 displays the results by Gen2 linking. The Mixed model reports the averages of the first borns of each sister (n = 344), the Daughters model reports the averages of the first born girls (n = 267), and the Sons model reports the averages of the first born sons (n = 283). All three models reveal similar results. A one unit increase in the average standardized intelligence of the children predicted $\approx .075$ increase in average Gen2 AFI. The adjusted R^2 varied slightly by Gen2 linking (Mixed = .014, Daughters = .016, Sons = .009).

Joint Mean Intelligence → Gen2 Mean AFI. Results from the Gen1 sister averages of standardized AFQT scores and Gen2 cousin averages of standardized ability scores predictions of Gen2 averages of gender standardized AFI are displayed in Table 14 on page 34. Again, three models based on Gen2 linking are displayed: Mixed

(n = 337), Daughters(n = 260), and the Sons(n = 278). All three models reveal similar results. Gen1 intelligence was significantly associated with Gen2 AFI ($p < .01$), while Gen2 intelligence was not significantly associated with Gen2 AFI. A one unit increase in the average standardized intelligence of the children's mothers predicted $\approx .014$ increase in average Gen2 AFI. The adjusted R^2 varied slightly by Gen2 linking (Mixed = .086, Daughters = .097, Sons = .100), but each were practically identical to the Mean Gen1 Intelligence models.

Within Family

We attempted to replicate the between family analyses reported in the previous subsection, using within family analyses. Using the discordant sibling model, we predicted the differences in Generation 2 AFI as a function of differences in intelligence, controlling for means of the outcomes and predictors. We ran three series of models, where we examined the individual and then joint influence of Gen1 intelligence and Gen2 intelligence. Moreover, within each series we included three Generation 2 linking method variants, just as we did in the between family analyses: Mixed model reports the differences of the first borns of each sister, the Daughters model reports the differences of the first born girls, and the Sons model reports the differences of the first born sons.

G1 δ Intelligence $\rightarrow$ Gen2 Dif AFI. Generation 1 sister differences in standardized AFQT scores were used to predict Gen2 differences of gender standardized AFI, controlling for Generation 1 sister averages of standardized AFQT scores and Gen2 averages of gender standardized AFI. Table 15 on page 35 displays the results by Generation 2 linking method. The Mixed model reports the averages and differences of the first borns of each sister (n = 336), the Daughters model reports the averages and differences of the first born girls (n = 258), and the Sons model reports the averages and differences of the first born sons (n = 278). All three models reveal similar results. Generation 2 averages of gender standardized AFI were significant predictors of Gen2 differences in gender standardized AFI ($p < .01$), across all three linking methods. A one unit increase in the average gender standardized AFI predicted ≈ 0.34 increase in

average Gen2 AFI difference, controlling for all over variables in the model.

In the Sons model, the Generation 1 sister average of standardized AFQT scores was a significant predictor of differences in Gen2 AFI ($p < .01$). A one unit increase in the average standardized intelligence of the children's mothers predicted $\approx .0083$ decrease in the AFI difference between siblings. All other variables were not significant, including all kin difference variables. The adjusted R^2 varied slightly by Generation 2 linking method (Mixed = .066, Daughters = .072, Sons = .106).

Gen2 Dif Intelligence → Gen2 Dif AFI. Gen2 cousin differences in standardized ability scores were used to predict Gen2 differences of gender standardized AFI, controlling for Gen2 cousin averages of standardized ability scores and gender standardized AFI. Table **??** on page ?? displays the results by Generation 2 linking method. The Mixed model reports the averages and differences of the first borns of each sister (n = 291), the Daughters model reports the averages and differences of the first born girls (n = 223), and the Sons model reports the averages and differences of the first born sons (n = 238). All three models reveal similar results. Gen2 averages of gender standardized AFI were significant predictors of Generation 2 differences in gender standardized AFI ($p < .01$), across all three linking methods. A one unit increase in the average gender standardized AFI predicted ≈ 0.38 increase in average Gen2 AFI difference, controlling for all over variables in the model.

In the Sons model, the Generation 2 cousin average of standardized ability scores was a significant predictor of differences in Generation 2 AFI ($p < .05$). A one unit increase in the average standardized intelligence of the children predicted $\approx .107$ decrease in the AFI difference between siblings. All other variables were not significant, including all kin difference variables. The adjusted R^2 varied slightly by Generation 2 linking method (Mixed = .103, Daughters = .121, Sons = .132).

Joint Dif Intelligence → Gen2 Dif AFI. Generation 1 sister differences in standardized AFQT scores and Gen2 cousin differences in standardized ability scores were used to predict Generation 2 differences of gender standardized AFI, controlling for Generation 1 sister averages of standardized AFQT scores, Gen2 cousin averages of

standardized ability scores, and Gen2 cousin averages of gender standardized AFI. Table 17 on page 37 displays the results by Generation 2 linking method. The Mixed model reports the averages and differences of the first borns of each sister (n = 285), the Daughters model reports the averages and differences of the first born girls (n = 217), and the Sons model reports the averages and differences of the first born sons (n = 235). All three models reveal similar results. Gen2 averages of gender standardized AFI were significant predictors of Generation 2 differences in gender standardized AFI ($p < .01$), across all three linking methods. A one unit increase in the average gender standardized AFI predicted ≈ 0.38 increase in Generation 2 AFI difference, controlling for all over variables in the model.

All other variables were not significant, including all kin difference variables. The adjusted R^2 varied slightly by Generation 2 linking method (Mixed = .090, Daughters = .105, Sons = .131).

Discussion

This article presents the relationship between AFI and intelligence using two difference designs: between- and within-family. The between-family design allowed us to replicate previous researchers who used a cross-sectional sample. The within-family design allowed us to evaluate intelligence differences within the family to address issues of causality. The results revealed a stark contrast between the two methods.

Between vs. Within

Between. Notably, the between-family analyses showed a relationship between intelligence and AFI. Thus, we were able to replicate the findings of various researchers (Halpern et al., 2000; Mott, 1983; Paul et al., 2000; Woodward et al., 2001), and confirm hypotheses 1 and 2. However the relationship between AFI and intelligence was stronger between maternal intelligence and child AFI than between the child's own intelligence and child AFI, which suggests that family-level variables rather than individual-level intelligence is source of the relationship. If intelligence had causally influenced AFI we would seen a considerably weaker cross-generational association

between AFI and intelligence. Instead we find that the within generation association is the weaker effect, suggesting that AFI is not causally influenced by intelligence.

An alternative interpretation of this finding could be that maternal intelligence is driving the effect. Smarter mothers might be more effective at inducing their children to delay intercourse – perhaps by effectively conveying the riskiness of sexual intercourse (Hutchinson et al., 2003; C. Mathews et al., 2009). Considering that intelligence is highly heritable (Bouchard, Jr., 2004) and thus highly correlated across generations, this alternative explaination would still be consistent with the traditional between family findings, which do not control for maternal intelligence (Halpern et al., 2000; Mott, 1983; Paul et al., 2000; Woodward et al., 2001).

Within. In the within-family analyses, the effect vanishes for both maternal intelligence and child intelligence. The smarter of the Generation 2 children was not more likely to delay intercourse. Moreover, in spite of our finding that the Generation 1 intelligence was a relatively strong predictor of Generation 2 AFI, we did not find that differences in Generation 1 intelligence are associated with differences in Generation 2 AFI. Thus, the alternative explaination for the between-family results we posed in the previous paragraph cannot be the case. For, if Generation 1 intelligence was driving the effect, we would have found a significant association, which we did not.

Concluding Remarks

Rodgers et al. (2008) looked at the relationship between IQ and education as they influenced age at first birth in Danish twin data. Their conclusion:

> [V]ariance in AFB emerges from [IQ and education] differences between families, not differences between sisters within the same family.

We have exactly the same result in the current study. Notably, the IQ differences between siblings are relatively small; is this important? Under a purely genetic model, Rodgers & Rowe (1987), estimated the average absolute deviation in IQ among random pairs to be 17.1 IQ points, compared to 12.1 for full siblings. But cousins, who also are in our study, deviate on average by 16.1 IQ points, nearly as much as random pairs.

We believe that differences within the family simply are not important for defining AFI outcome differences. This finding matches the Harden & Mendle (2011) biometrical analysis of Add Health, where they found that only shared environmental influences mattered those would manifest in between-family differences, but not in within-family differences.

Further, we're not convinced that in these between-family analyses, intelligence is the actual cause of AFI differences if so, we think they would perhaps diffuse a bit, but would still show up in within-family analyses Rather, we think maternal and child IQ are indirect measures of many other household features, any one of which may be more proximal as the causal explanation income, parental education, family interaction. Or, the whole package of these features may stand in for a general environmental factor, a "little e," which indexes the quality of home environment a composite of parental income, intelligence, education, family interaction.