Intelligence 38 (2010) 367–384

Contents lists available at ScienceDirect

Intelligence

The Flynn Effect within subgroups in the U.S.: Gender, race, income, education, and urbanization differences in the NLSY-Children data

SiewChing Ang[a], Joseph Lee Rodgers[a,*], Linda Wänström[b]

[a] Department of Psychology, University of Oklahoma, United States
[b] Department of Statistics, University of Stockholm, Sweden

ARTICLE INFO

Article history:
Received 6 July 2009
Received in revised form 21 May 2010
Accepted 22 May 2010
Available online 20 June 2010

Keywords:
Flynn Effect
National Longitudinal Survey of Youth
NLSY-Children data
PIAT-Math
Secular changes in intelligence

ABSTRACT

Although the Flynn Effect has been studied widely across cultural, geographic, and intellectual domains, and many explanatory theories have been proposed, little past research attention has been paid to subgroup differences. Rodgers and Wänström (2007) identified an aggregate-level Flynn Effect (FE) at each age between 5 and 13 in the Children of the National Longitudinal Survey of Youth (NLSYC) PIAT-Math data. FE patterns were not obtained for Reading Recognition, Reading Comprehension, or Digit Span, consistent with past FE research suggesting a closer relationship to fluid intelligence measures of problem solving and analytic reasoning than to crystallized measures of verbal comprehension and memory. These prior findings suggest that the NLSYC data can be used as a natural laboratory to study more subtle FE patterns within various demographic subgroups. We test for subgroup Flynn Effect differences by gender, race/ethnicity, maternal education, household income, and urbanization. No subgroups differences emerged for three demographic categories. However, children with more educated (especially college educated) mothers and/or children born into higher income households had an accelerated Flynn Effect in their PIAT-M scores compared to cohort peers with lower educated mothers or lower income households. We interpret both the positive and the null findings in relation to previous theoretical explanations.

1. Introduction

The Flynn Effect (FE) refers to an increase in intelligence scores across time. The effect was first reported by Lynn (1982) and especially by Flynn (1984), and is commonly known as the Flynn Effect (though it also been referred to as the Lynn–Flynn Effect; see Rushton, 1997). Flynn (2006) observed that increasing IQ scores were documented much earlier, though without much broad attention (e.g., Smith, 1942; Tuddenham, 1948). Despite substantial modern attention to the FE in the intelligence literature (see, e.g., Neisser, 1998 and Flynn, 2007 for summaries), surprisingly little attention has been given to certain critical features of the FE, including how it operates among different population subgroups (Rodgers, 1998; Rodgers & Wänström, 2007; Sundet, Borren & Tambs, 2008). This paper examines how the FE performs within several important demographic subgroups – gender, race/ethnicity, urbanization, mother's education, and household income – using the PIAT-Math subscale from the Children of the National Longitudinal Survey (NLSYC) data.

Many Flynn Effect studies have been reported on various populations worldwide (Colom, AndresPueyo & JuanEspinosa, 1998; Colom, JuanEspinosa & García, 2001; Daley, Whaley, Sigman, Espinosa & Neumann, 2003; Flynn, 1987; Lynn, 1982; Lynn & Hampson, 1986; Rodgers & Wänström, 2007; Sundet, Barlaug & Torjussen, 2004; Teasdale & Owen, 1989, 2000). Although IQ scores have risen systematically, it is questionable whether general intelligence (g) itself has changed (Jensen, 1998; Kane & Oakland, 2000; Must, Must & Raudik, 2003; Flynn, 1987, 2006). Further, the FE appears to operate more strongly within the fluid intelligence domain, and not as much within crystallized intelligence (Flynn, 1987, 2006; Jensen, 1991; Loehlin, 1996; Lynn, 2009; but also see Flynn, 2009b, who identified Flynn Effects on the vocabulary subscale of the WAIS).

[*] Corresponding author.
E-mail address: jrodgers@ou.edu (J.L. Rodgers).

0160-2896/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2010.05.004

Fluid intelligence refers to problem solving or reasoning, whereas crystallized intelligence refers to language acquisition and knowledge attainment (Horn & Catell, 1966).

Many causal explanations of the FE have been proposed, though consensus has not emerged, and some controversy has ensued in regards to certain theories. Flynn Effect scholars have attributed the FE to massive environmental changes, such as the proliferation of movies, television, video games and computers (Greenfield, 1998); urbanization (Flynn 1998; Williams, 1998); increased access to education (Flynn, 1998; Greenfield, 1998); general school factors (Schooler, 1998; Williams, 1998); preschool education (Teasdale & Berliner, 1991); the development of new math education/curriculum (Blair, Gamsonb, Thorne & Baker, 2005); the development of quality and quantity of schools, teachers and teacher-training programs with increased educational funding (Greenfield, 1998; Williams, 1998); and changes in parental styles (Williams, 1998). Mahlberg (1997) proposed a FE explanation based on a cultural memory interpretation. Studies have also connected the FE with changes in family size (resulting in increased resources per child; Sundet et al., 2008), and increased parental educational levels (Ceci, 1996; Grissmer, Kirby, Berends and Williamson, 1994; Sundet et al). Nutritional mechanisms have provided the basis for another class of popular theories (Lynn, 1989, 1990; Martorell, 1998; Schoenthaler, Amos, Eysenck, Peritz and Yudkin, 1991); Flynn (2009a) provided evidence in opposition to a nutritional interpretation. Steen (2009) favored a "rising tide" hypothesis associated with medical improvements, "because of intervention into medical conditions that depressed intelligence in the past" (p. 129).

Various other theories have used processes that were not purely environmental. Lynn (1998) suggested there may exist a genetic component to the FE. An interpretation by Dickens and Flynn (2001) involved an evocative gene–environment interaction model. Mingroni (2007) suggested that heterosis (also known as hybrid vigor or outbreeding) was a plausible explanation (also see Jensen, 1998).

Others have suggested that IQ increases are artifactual. This class of explanations includes indications of factorial invariance violations across time/ages of the measures used to document the FE, identified by Wicherts et al (2004) and Beaujean & Osterlind (2008). (Both of these studies also identified real ability changes as well, however.) Testing issues have also been identified, including increased attention to speeded tests (Brand, 1996; Brand, Freshwater and Dockrell, 1989), and test norming processes (Kanaya, Scullin & Ceci, 2003). Practice effects have been implicated (Williams, 1998), along with other environmental factors (Blake, 1989).

Finally, Jensen (1998) suggested a multiplicity hypothesis, in which many small factors (some likely unspecified) have combined to create the FE. A particular attraction of this interpretation is that it provides explanation for the consistency of the FE in many cultures. The pace and direction of the FE should change substantially over time if one or a very few explanations were accounting for it and if those influences themselves changed over time. The multiplicity hypothesis would explain the persistence of the FE in many geographic settings, even as there have been temporal shifts in a number of its individual putative causes.

Although various causal explanations for the FE have been proposed in the literature, more empirical understanding of the FE is needed before the theories to explain the FE are fully informed by the data patterns. Rodgers (1998) outlined 10 research questions to be answered to support causal evaluation of the FE, one of which focused on identifying differences across demographic categories. A decade later, few of these questions have been yet addressed, and the status of the FE across demographic subgroups is still unclear. Flynn (1998) and Greenfield (1998) studied income differences. Rushton (1999, 2000) considered the FE in relation to race differences. Research by Teasdale and Owen (1989) and by Colom, Lluis-Font and Andres-Pueyo (2005) identifying differential FE in the lower tails of the intelligence distributions had implications for subgroup differences (though indirect). Little other research addressing, for example, basic gender, race, and education differences has been conducted.

The current study is a replication and expansion of the work done by Rodgers and Wänström (2007). Using the NLSYC data from 1986 to 2000, they compared various cognitive assessments at each age from five to 13 years old, using all biological children born to a 1978 sample of mothers whose ages were 36–43 years old on December 31, 2000 (birth cohorts of 1957–1964). The cognitive assessments included were the Wechsler Memory for Digit Span test, the Peabody Picture Vocabulary Test, and three Peabody Individual Achievement (PIAT) Test subscales: PIAT-Math (PIAT-M), PIAT-Reading Recognition and PIAT-Reading Comprehension. Results indicated that the PIAT-M showed the largest FE (the mean slope per year was .30 for the oldest five ages) compared to other subscales. This effect was statistically significant for almost all ages even after controlling for mother's IQ (which was included to adjust for selection bias due to younger mothers having had disproportionately more children). For most other subscales, the FE decreased or disappeared entirely after adjusting for maternal IQ. This result was predicted, because the PIAT-M is much closer to a measure of fluid intelligence, whereas the other tests more strongly reflect crystallized intelligence.

Although the Flynn Effect has been empirically shown in the NLSYC PIAT-M scores, a number of interesting questions remain unanswered. One question, motivated by the critique in Rodgers (1998) over a decade ago, is: Are there differences in these patterns within the subpopulation groups, such as gender, race/ethnicity, education, income, or urbanization? Specifically, this paper focuses on the exploration of the FE within each of these categories over an 18-year period, from 1986 to 2004.

2. Methods

2.1. Sample and designs

The original National Longitudinal Survey of Youth (NLSY79), which contains the mothers of the NLSYC children sample that we use in this research, was based on a household probability sample of 12,686 adolescents aged 14–21 on Dec 31, 1978 (http://www.bls.gov/nls/nlsy79.htm). Since 1986, on a biannual basis, all biological children born to the female NLSY79 respondents have been surveyed, including the administration of cognitive assessments. By 2004, the NLSY79 females were between 39 and 47 years old, and had given birth to 15,359 children. These children's ages ranged between newborn to the mid 30's. Our design involves using the longitudinal structure of the NLSYC to compare children of the same age across time—five-year-olds in 1984 to five-year-olds in 1985, …, to five-year-olds in 2004; six-year-olds in 1984 to six-year-olds in 1985, …, to six-year-olds in 2004; etc. through age 13. We imposed a requirement of a sample size of at least 30 in each age group for at least 6 consecutive years to ensure stable comparisons, and this requirement results in complete data for children between 5 and 13 years old. We note that a small proportion of the observations used in the mean comparisons to evaluate the Flynn Effect include siblings. For example, a family with a seven-year-old in 1985 who had a sibling aged seven in 1987 would contribute within-family variance to the comparison for seven-year-olds from 1985 to 1987. Each comparison in our study contains a small fraction of such within-family variance. We did not exclude the siblings, for several reasons. First, many of the explanations listed in the introduction would apply to sibling as well as non-sibling comparisons. Second, in terms of past empirical results, Rodgers (1998) discounted within-family explanations of the Flynn Effect, based on past birth order and early FE findings, suggesting that including siblings could create a downward bias in estimates of the FE. However, a recent study by Sundet, Eriksen, Borren and Tambs (2010) linked within-family patterns in Norwegian data to a relatively small but significant Flynn Effect, suggesting that the FE can at least potentially be observed in within-family patterns. It appears unlikely that the inclusion of within-family variance would create upward bias in estimating the Flynn Effect. If anything, the inclusion of siblings is likely to be either neutral or a conservative effect in FE estimation. We do note, however, that the standard errors associated with the FE estimates could contain some bias because of the resulting correlated errors; this potential for slightly liberal statistical tests will be treated in the methods section.

As the mothers of the children were based on a household probability sample of the U.S., sampling weights are available in the NLSY79 and NLSYC datasets to adjust for the sampling design and attrition. These weights are applied to the NLSYC to obtain population estimates. They are used throughout the analyses presented in this study, and so results can be considered to generalize to the population of U.S. children born to mothers who were adolescents (aged 14–21) at the end of 1978. Due to the large sample sizes in the overall NLSYC sample, statistical analyses based on most relevant subgroups of the sample have substantial statistical power.

The validity of the NLSYC is enhanced because all children were tested with the same cognitive ability assessments. We examine the FE across an 18-year period, from 1986 to 2004, and define nine replication samples for ages from 5 to 13. The same design was used in Rodgers and Wänström (2007), which was based on NLSYC data from 1986 to 2000.

2.2. Measures

2.2.1. Outcome variable

The outcome variable is the normed, age-standardized scores of the PIAT-M for children ages 5–13 (see Rodgers, Rowe & May, 1994; Rodgers, Cleveland, van den Oord and Rowe, 2000; and Wichman, Rodgers and MacCallum, 2006 for previous studies, among many others, using the PIAT-M scores in research on childhood intelligence). The PIAT-M consists of 84 items, used for all ages from five to 13. The 84 items are ordered in difficulty, from basic quantitative matching items used as a baseline for five-year-olds, to complex algebra and geometry items used for older children. An example of an early item (Item #1) is “Point to the ‘4’ in the display.” An example of a midrange item (Item #42) is “Which girl has the most money? Ann has 12 nickels; Sue has 70 cents; Mary has 3 quarters; Alice has 8 dimes.” An example of a late item (Item #70) is “Which term is represented by the ratio a^5/a^2? a^3, a^7, a^{10}, or $a^{5.2}$?” In the administration of the PIAT-M test, children started at a PIAT-M item appropriate to their age, and then moved back to previous age starting points if they missed baseline items. Otherwise they moved forward and answered each item. Five consecutive correct answers established the baseline of their math achievement, with the basal defined as the first of those five items. The PIAT-M was finished when the child made five mistakes out of the most recent seven items (the last item of which is defined as the ceiling). The non-normalized raw score is the ceiling minus the number of incorrect responses between the basal and the ceiling scores. These scores were age-standardized against national norms by the Center for Human Resource Research, who manages the NLSYC data.

We note that the PIAT-M items undoubtedly behave differently for different ages. For example, the sample item above involving coins (Item #42) would be handled virtually automatically by older children and adults, but would require explicit computation and real-world reasoning by a 5- or 6-year-old. A different study we are conducting that is in progress involves obtaining item content ratings by educational experts to assess the item content at different ages.

2.2.2. Predictor and control variables

The first variable we define as a predictor is the survey year in which the PIAT-M scores were collected (called Year) from 1986 to 2004, a total of 10 years on a biannual cycle. Because of the biannual data collection cycle, each unit of the Year coding actually corresponds to two years of time. This time variable is used to measure the potential FE. Furthermore, because most of our analyses involve binary variables (such as gender), Year is centered around its mean before it is entered into the regression interaction (see Aiken and West, 1991).

The other predictors involve the population subgroups such as children's gender, race/ethnicity (reported by the mother), maternal education, total household income, and urbanization (whether the respondent resided in rural or urban areas). Race/ethnicity is recoded into three dummy variables: Hispanic (1) vs. Non-Hispanic (0), African American (1) vs. Non-African American (0), and Caucasian/Asian/Native American (Caucasian/AA/NA) (1) vs. Others (0). The last of these race categories contains predominantly Caucasian respondents, with a few Asian American and Native Americans as well. We note that these race (and other) categories were defined within the NLSYC data files, and are not variables over which we had control.

We measured mother's education and income using two coding schemes. First, mother's education (MED) was defined as the quantitative measure of years of total education, with a range of 0–20. Thus, MED = 12 would indicate that the mother had finished high school (but no further); MED = 16 that she had finished college. This variable was centered before being entered. Because of the particular salience of going to college in the U.S., we also defined a dummy variable for Mother's education by coding education into 0 if the mother had an MED value of 12 or less and 1 if she had at least some college (MED = 13 or greater).

We also used two different coding schemes for total household income. We note that income generally increases as the survey year increases (although our adjustment for maternal IQ, discussed later in this section, partials out some of this increase because of the correlation between maternal IQ and education/income). The first coding scheme simply used the reported total household income (centered). Within each survey year, the median incomes were \$13,200 (1986), \$15,900 (1988), \$22,000 (1990), \$24,700 (1992), \$28,000 (1994), \$35,000 (1996), \$38,000 (1998), \$46,000 (2000), \$53,000 (2002), \$59,500 (2004). Because all of the other variables had dichotomized versions, we also defined an income dummy variable as the median split within the sample for each survey year (which adjusted for the increased incomes over year).

2.3. Statistical analysis

To evaluate the Flynn Effect, we defined regression models in which PIAT-M scores within a particular age are predicted from Year, with the unstandardized regression slope of this variable estimating the size of the Flynn Effect. To account for change per year, these slopes must be divided by two, because one unit on the Year variable actually measured two-year intervals. Then, we added to the models the quantitative and dummy variables accounting for subgroup differences. Finally, the critical measures were the constructed interaction terms between the demographic predictor and the centered Year variable. When we identified significant demographic by year interactions, we also defined several models with higher-way interactions to further probe the nature of the relationships.

Reviewers noted the substantial number of statistical tests in our tables of results, and suggested a substantial adjustment to our alpha level. However, the Type I Error Rates is not as inflated as it may appear. The goals of our paper have nothing to do with evaluating differences among these demographic categories in terms of PIAT-M performance; extensive literature exists on each of those questions already. Nor are we concerned with whether there is a Flynn Effect in these data; that has already been established (Rodgers & Wänström, 2007). By far the majority of the statistical tests reported in Tables 1 and 2 are in relation to these already studied domains. The relevant statistical tests for this paper are those associated with the interactions (these are bolded within the tables). Our general strategy is to define our basic research question in relation to each age and each demographic category, and the Type I Error rate is set in relation to each research question. We will report significance levels for $\alpha = .05$, $\alpha = .01$, and $\alpha = .001$.

When we find statistical results that suggest stability across ages in the interaction tests, in those cases we will then shift our focus to effect sizes and graphical presentation for interpretational purposes, as suggested by Wilkinson & Task Force on Statistical Inference, APA Board of Scientific Affairs (1999). The unstandardized regression coefficients are (unstandardized) effect size indicators within our study. Those effect sizes and the consistency of patterns across ages are what will drive our interpretations; the statistical tests are a screen for general stability.

Table 1
Main effects and interactions on PIAT-Math normed scores for year and maternal education by age, controlling for mother's IQ, with year and education variables centered; left hand statistics are for binary HS versus college education measure, right hand statistics are for quantitative grades completed measure (interaction tests of differential Flynn Effect bolded).

	Parameter	N	Slope Est.	SE	t
5 years old	Year	3210	0.54/0.56	0.10/0.10	$5.13^{***}/5.51^{***}$
	Education		0.35/0.48	0.61/0.13	$0.57/3.68^{***}$
	Year*Education		**−0.14/−0.03**	**0.21/0.04**	**−0.65/−0.67**
6 years old	Year	3280	0.31/0.36	0.08/0.08	$3.78^{***}/4.50^{***}$
	Education		1.33/0.57	0.50/0.11	$2.68^{**}/5.24^{***}$
	Year*Education		**0.06/−0.01**	**0.16/0.03**	**0.34/−0.31**
7 years old	Year	3284	0.27/0.30	0.08/0.07	$3.57^{***}/4.07^{***}$
	Education		1.42/0.37	0.44/0.10	$3.19^{**}/3.82^{***}$
	Year*Education		**0.43/0.11**	**0.15/0.03**	$\mathbf{2.78^{**}/3.49^{***}}$
8 years old	Year	3201	0.49/0.51	0.09/0.09	$5.59^{***}/5.89^{***}$
	Education		1.06/0.72	0.52/0.12	$2.06^{*}/6.07^{***}$
	Year*Education		**0.22/0.04**	**0.18/0.04**	**1.20/1.06**
9 years old	Year	3196	0.90/0.91	0.10/0.10	$9.29^{***}/9.45^{***}$
	Education		0.22/0.18	0.58/0.13	0.39/1.37
	Year*Education		0.61/0.05	0.20/0.04	$\mathbf{3.09^{**}/1.19}$
10 years old	Year	3018	0.66/0.68	0.10/0.10	$6.72^{***}/6.98^{***}$
	Education		0.96/0.44	0.58/0.13	$1.66/3.26^{***}$
	Year*Education		**0.31/0.01**	**0.20/0.04**	**1.51/0.35**
11 years old	Year	2939	0.82/0.79	0.10/0.10	$7.93^{***}/7.76^{***}$
	Education		−0.20/0.36	0.63/0.14	$-0.31/2.59^{**}$
	Year*Education		**0.41/0.02**	**0.22/0.04**	**1.90/0.35**
12 years old	Year	2678	0.65/0.63	0.11/0.10	$6.08^{***}/6.06^{***}$
	Education		0.79/0.60	0.63/0.15	$1.25/4.12^{***}$
	Year*Education		**0.46/0.05**	**0.22/0.04**	$\mathbf{2.09^{*}/1.02}$
13 years old	Year	2520	0.45/0.38	0.12/0.12	$3.86^{***}/3.31^{**}$
	Education		−1.58/0.04	0.73/0.16	$-2.18^{*}/0.28$
	Year*Education		**0.91/0.13**	**0.25/0.05**	$\mathbf{3.69^{***}/2.53^{*}}$

Note. $^{*}p \leq 0.05$; $^{**}p \leq 0.01$; $^{***}p \leq 0.001$. N = total observations used in the regression models from 1986 to 2004.

Table 2
Main effects and interactions on PIAT-Math normed scores for year and household income by age, controlling for mother's IQ, with year and household income variables centered; left-hand statistics are for binary income variable, right-hand statistics are for quantitative income measures (interaction tests of differential Flynn effect bolded).

Dependent	Parameter	N	Slope Est.	SE	t
5 years old	Year	2132	0.66/0.40	0.25/0.11	2.65^{**}/3.49^{***}
	Income		2.16/0.001	0.630/0.005	3.42^{***}/3.06^{**}
	Year*Income		**0.11/0.005**	**0.22/0.002**	**0.49/2.35^{*}**
6 years old	Year	2766	0.42/0.38	0.08/0.09	5.00^{***}/4.28^{***}
	Income		2.95/0.000	0.49/0.002	5.97^{***}/1.02
	Year*Income		**0.04/0.000**	**0.17/0.001**	**0.24/−.35**
7 years old	Year	2822	0.40/0.32	0.08/0.08	4.97^{***}/3.84^{***}
	Income		1.54/0.000	0.45/0.002	3.44^{***}/−0.37
	Year*Income		**0.18/0.003**	**0.16/0.001**	**1.16/2.22^{*}**
8 years old	Year	2716	0.56/0.49	0.09/0.10	6.08^{***}/5.14^{***}
	Income		3.71/0.002	0.51/0.004	7.29^{***}/3.47^{***}
	Year*Income		**0.13/0.000**	**0.18/0.002**	**0.71/−0.47**
9 years old	Year	2722	0.91/0.82	0.10/0.10	8.98^{***}/7.84^{***}
	Income		1.54/0.001	0.55/0.004	2.77^{***}/2.51^{*}
	Year*Income		**0.36/0.003**	**0.20/0.002**	**1.77/1.87**
10 years old	Year	2570	0.21/0.70	0.34/0.11	0.62/6.55^{***}
	Income		3.83/0.003	0.56/0.007	6.85^{***}/4.26^{***}
	Year*Income		**0.29/0.000**	**0.20/0.002**	**1.43/-2.33^{*}**
11 years old	Year	2472	0.84/0.80	0.11/0.11	7.78^{***}/7.26^{***}
	Income		2.17/0.001	0.59/0.004	3.67^{***}/2.20^{*}
	Year*Income		**−0.24/0.002**	**0.21/0.002**	**−1.13/1.00**
12 years old	Year	2247	0.70/0.59	0.11/0.11	6.28^{***}/5.19^{***}
	Income		2.54/0.001	0.60/0.004	4.25^{***}/2.40^{*}
	Year*Income		**0.53/0.003**	**0.22/0.001**	**2.41^{*}/1.93**
13 years old	Year	2132	0.43/0.36	0.12/0.12	3.5^{***}/2.90^{***}
	Income		0.75/0.002	0.67/0.005	1.12/3.00^{**}
	Year*Income		**0.53/0.005**	**0.24/0.002**	**2.18^{*}/2.74^{**}**

Note. $^{*}p \leq 0.05$; $^{**}p \leq 0.01$; $^{***}p \leq 0.001$. N is the total observations used in the regression models from 1986 to 2004.

2.4. Adjustment to support internal validity

Obviously, children born in 1986 on average had younger mothers than children born in 1988, and so on through 2004; children of younger mothers are overrepresented throughout the sample, with this type of selection bias gradually disappearing over time. Mother's IQ was moderately positively correlated with age at first birth (AFB) in the NLSYC data in the mid-1990s, $r_{IQ,AFB} = .27$ (Rowe & Rodgers, 2002). Rodgers and Wänström (2007) examined the maternal IQ and AFB in 2002 NLSYC data and found $r_{IQ,AFB} = .43$. The selection bias caused by this confound is well-known, and has been treated in past NLSYC research (e.g., Rodgers & Wänström). To adjust for this confound, we use mothers' IQ as measured in the NLSY79 with the Armed Forces Qualification Test (AFQT). The AFQT is an IQ test from the Armed Services Vocational Aptitude Battery. The AFQT scores for the NLSY79 respondents were measured in 1980 when the mothers were 15–23 years old.

We used age-normed scores from the PIAT-M (normed against a national standardization sample), and thus the math scores of the children have been adjusted to equate for age differences among the children respondents themselves. Rodgers and Wänström (2007) reported that the difference between using the normed scores or the raw scores was empirically trivial, as the correlations between the two sets of scores across ages and assessment were reported to be above 0.90, sometimes as high as 0.99. This is methodologically predictable, because our design compares children at fixed ages (but different time points), so that the age norming procedure is not very relevant to the current study.

Children's IQ gains will be presented both adjusting for and not adjusting for mother's IQ, with emphasis and ultimate focus on the former. Concerns that that children's IQ might be correlated with mother's IQ and thus some of the contributing variance might be taken away were addressed and resolved by Rodgers and Wänström (2007).

3. Results

3.1. Descriptive results

Fig. 1 plots the PIAT-M normed scores across all 9 ages by demographic subgroups from 1986 to 2004; there we see an unambiguous Flynn Effect for all subgroup categories at all ages. Thus, the overall Flynn Effects by age identified by Rodgers and Wänström (2007) replicate within subgroups of all five demographic categories. Notably higher PIAT-M scores were observed for the race category of Caucasian/AA/NA, for children whose mothers attended at least some college, and for children from higher income households. There were smaller and less consistent differences (that were, nevertheless, interpretable) for the gender and urbanization categories. We now report results of regression analyses formally evaluating the stability of these patterns and identifying effect sizes, with focus on testing for differential FE across demographic categories.

3.2. Regression analysis by age with year and demographic predictors

3.2.1. Gender

When controlling for mother's IQ, seven of nine age groups showed a significant IQ increase across years (all except for 5- and 6-year-olds), replicating the general FE found earlier. Six of 9 age groups (ages 5, 6, 9, 11, 12 and 13) showed a significant Gender difference. Females ages 5 and 6 scored higher than males on the average, and for ages 9, 11, 12 and 13, males scored significantly higher than females, a replication of a common and often-studied pattern. In the critical set of tests of the gender by Year interaction, no significant Year by Gender interactions were identified. Fig. 2 shows the means by gender for ages 5–13, adjusted for mother's IQ, obtained from the LSMEANS statement in SAS (which computes estimated means statistically adjusted for control variables). We concluded that there were no statistically stable Gender by Year interactions.

3.2.2. Race/ethnicity

Three separate analyses were run using each race dummy variable, to test for each race effect individually. Controlling for mother's IQ, all slopes for Year were significant across 9 age categories in these three pairs of ethnic comparisons,

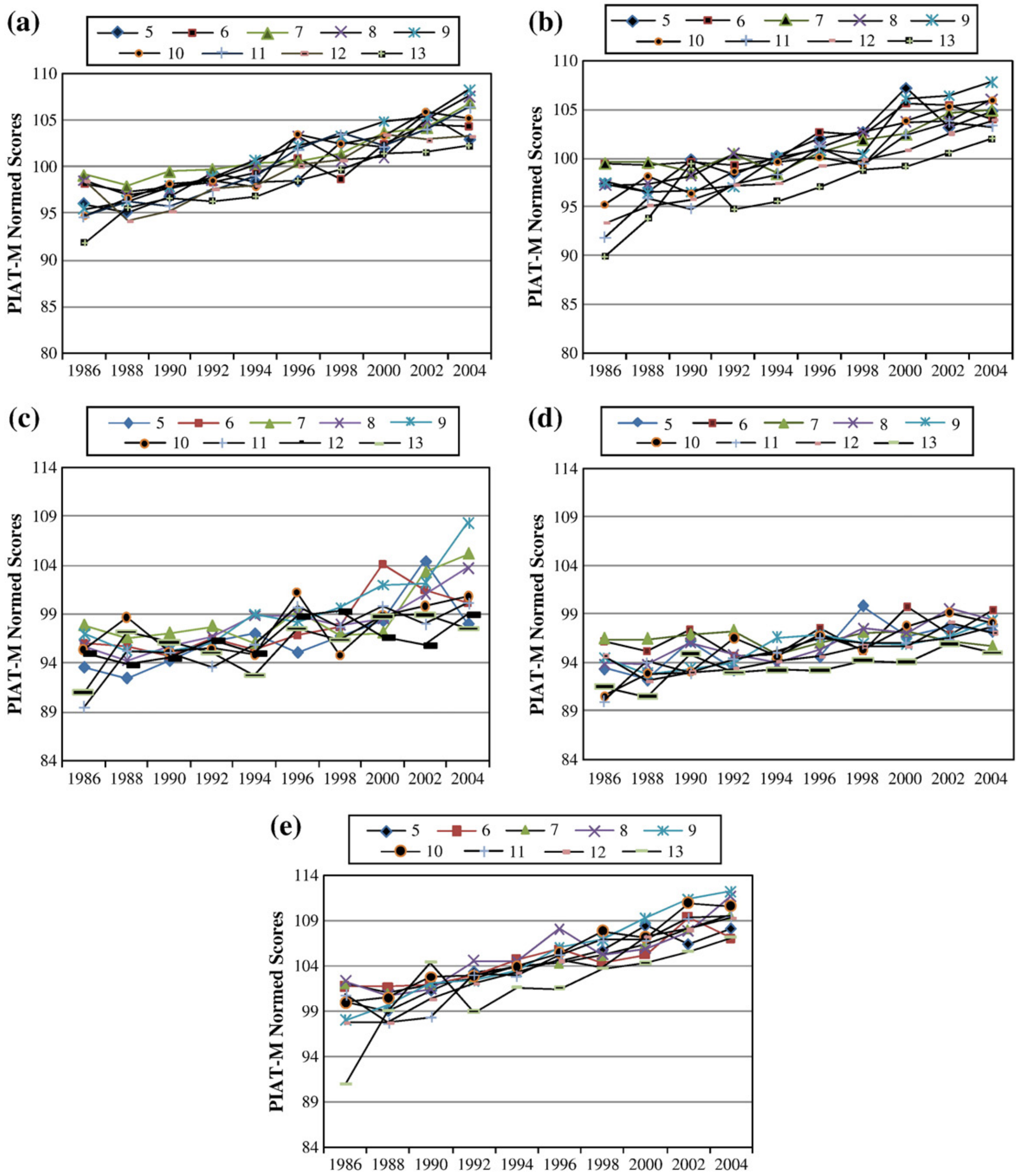

Fig. 1. PIAT-M Normed Score Means, 1986–2004: (a) for males ages 5–13; (b) for females ages 5–13; (c) for Hispanics Ages 5–13; (d) for African Americans Ages 5–13; (e) for Non-African-Americans-Non-Hispanics Ages 5–13; (f) for Children Ages 5–13 whose Mothers had High School or Less Education; (g) for Children Ages 5–13 whose Mothers had Some College or Completed College Education; (h) for Children Ages 5–13 with Lower Household Income; (i) for Children Ages 5–13 with Higher Household Income; (j) for Children Ages 5–13 in Rural Area, 1986–2004; (k) for Children Ages 5–13 in Urban Area, 1986–2004.

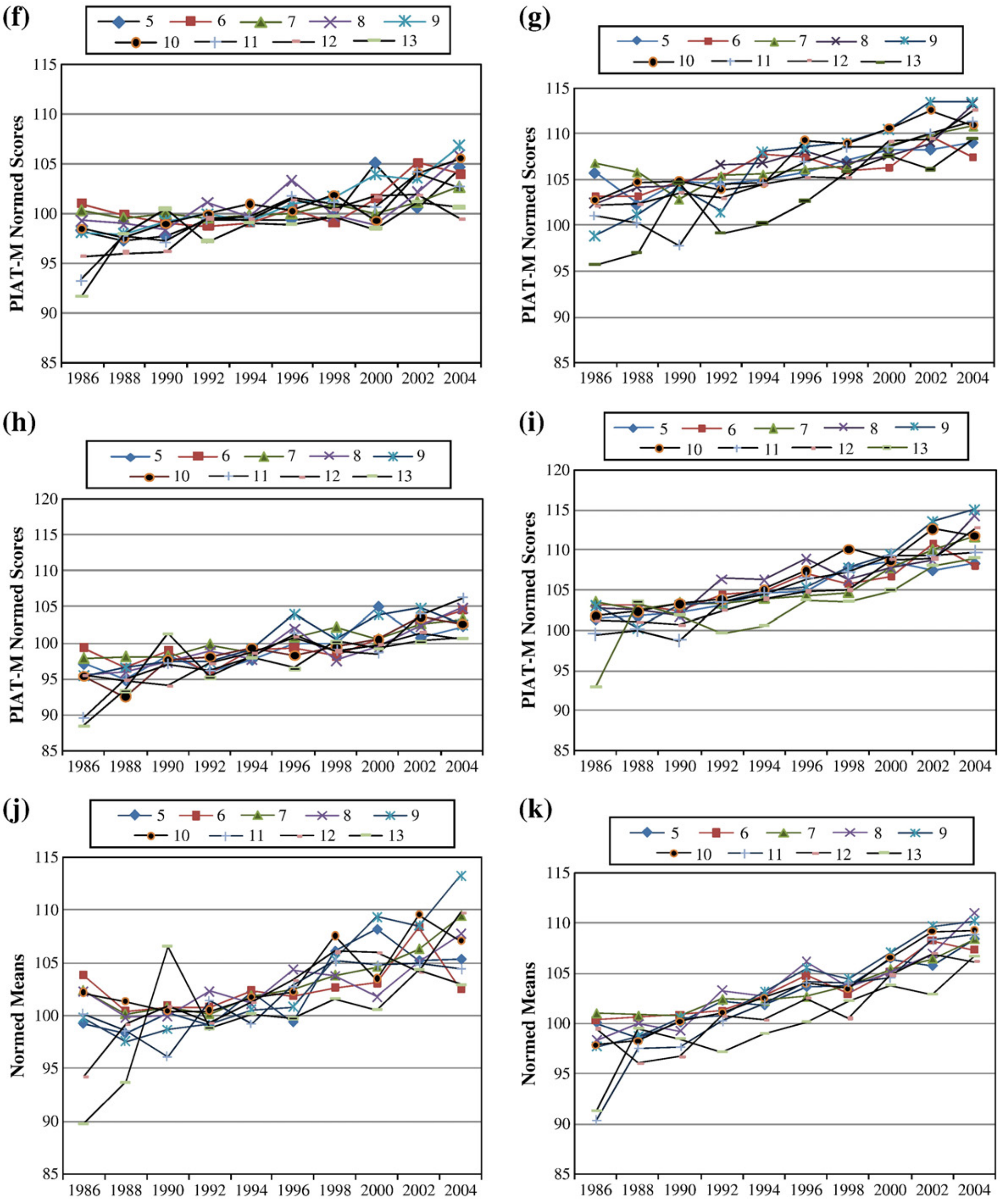

Fig. 1 (*continued*).

again replicating the general FE found previously in all age categories. The main effects for race/ethnicity indicated that 9 of 9 age groups reported significantly higher PIAT-M means for Caucasian/NA/AA than Hispanics or African Americans, and higher PIAT-M scores for Hispanics than African Americans in most instances. There were two significant interactions across the three analyses for nine ages. Adjusted means (from the LSMEANS statement in SAS, adjusting for mother's IQ) are shown in Fig. 3. In summary, all slopes for Year and race/ethnic groups were significant, and 2 of 27 interactions (nine ages for three models) for Year by race/ethnicity were significant.

We also defined a statistical model that included two race variables simultaneously, to evaluate the overall race effect. These models included an interaction term for each of the race dummies by Year. Across the 18 interaction tests, one was significant at $\alpha = .05$. We concluded that there were no statistically stable Race by Year interactions.

3.2.3. Maternal education

The first set of results we present are for maternal education coded in a binary form, distinguishing children with mothers who finished high school or less from those who went to at least some college. Adjusting for maternal IQ, slopes for Year and Education were significant for all age categories, showing that children of mothers with higher education had higher PIAT-M scores, and replicating the FE for all ages and maternal education categories. In the critical test of the interactions, for 4 of 9 age groups (ages 7, 9, 12 and 13) we found statistically stable results at $\alpha=.05$. (Before adjusting for maternal IQ, six of 9 interaction tests were significant, including all 5 oldest groups ages 9–13). The results are shown bolded in the left hand side of each entry in Table 1, and shown graphically in Fig. 4.

These patterns suggest that further evaluation of the graphical and effect size patterns is justified. The positive regression coefficients for all age categories except for age 5 show that children whose mothers had a college degree had higher a higher FE than children whose mothers had less education, and this pattern can be observed in Fig. 4, especially for the older ages. In studying the effect sizes, the average slope per year for children whose mothers were older and had college education was 0.24 higher than children with older but less educated mothers for the oldest 7 age groups (the average of the regression coefficients across age, divided by two to account for the yearly change). It is notable that the pattern appears to maintain for all of the older ages—both those that were statistically significant and those that were not. In other words, the effect sizes are slightly smaller at some ages than others, but the general pattern is strong and consistent.

When we coded the maternal education variable as a quantitative indicator of number of years of school attended by the mother (and included mothers IQ), the pattern maintained but weakened slightly. Significant results remained for ages 7 and 13. The effect size indicators are in the right hand side of each entry in Table 1. The regression coefficients as effect sizes were positive for all of the seven older ages, again supporting a higher FE for children with higher educated mothers. (It should be noted that the centered quantitative education variable was constructed to

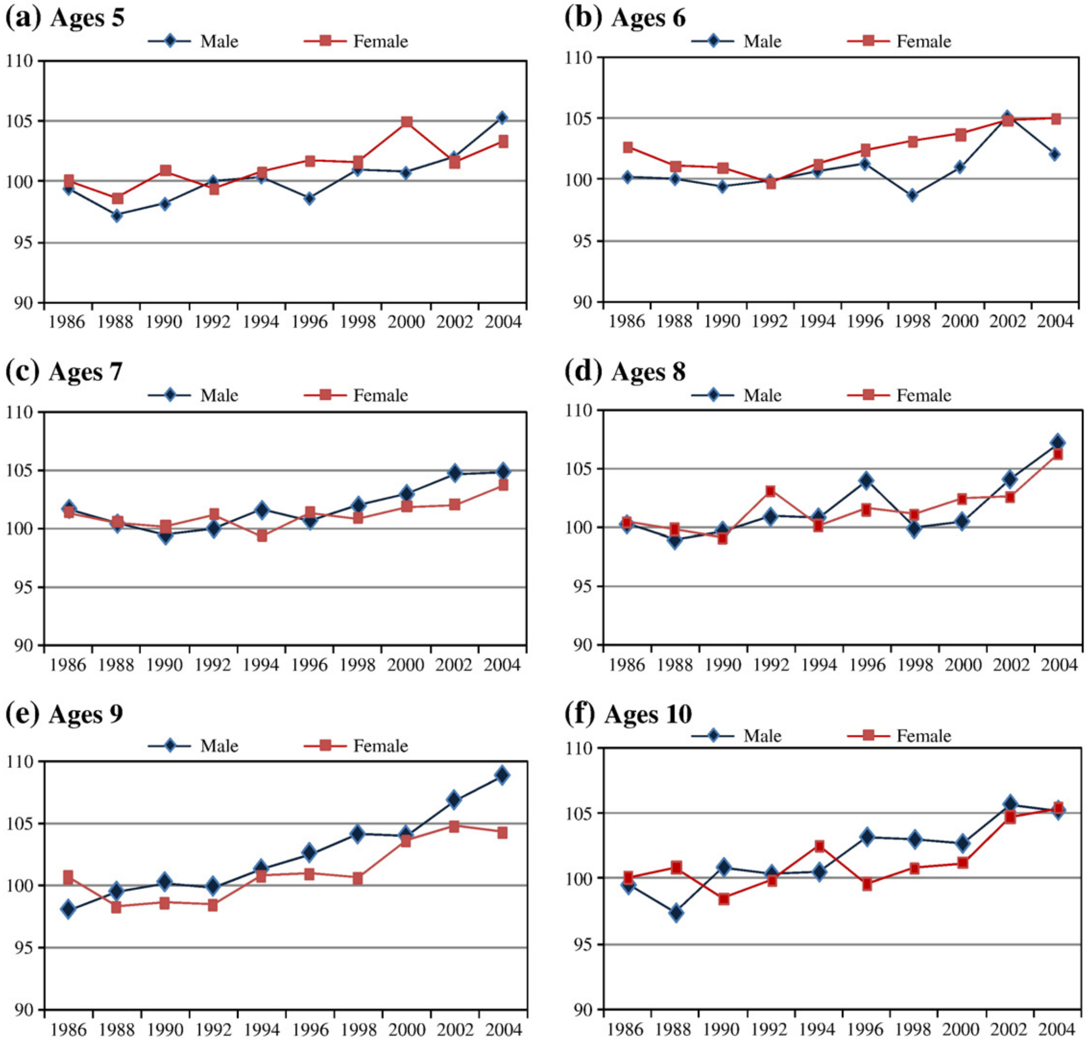

Fig. 2. Adjusted PIAT-M Normed Mean Scores for Ages 5–13 (Controlling for Maternal IQ) by Gender, 1986–2004.

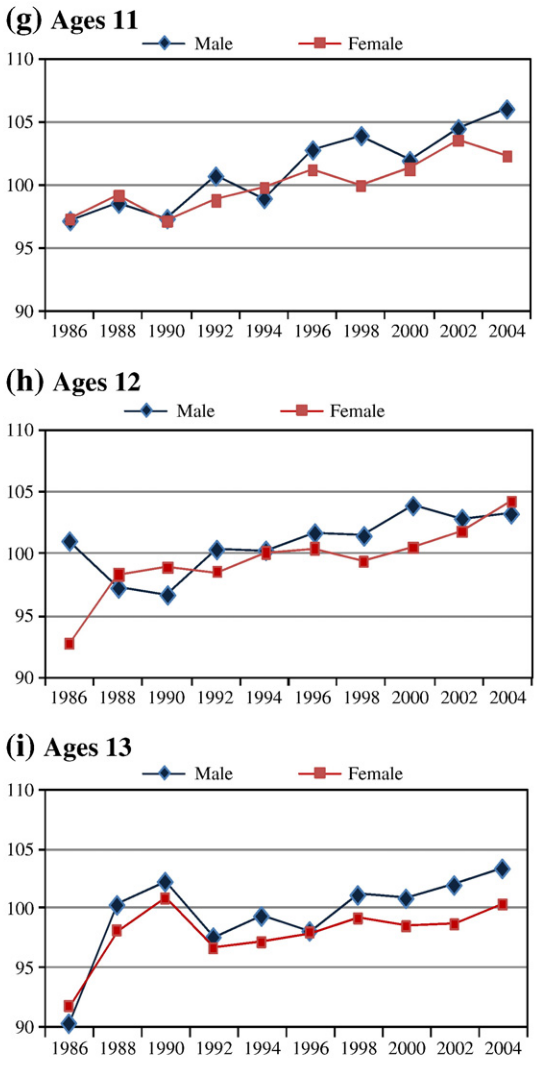

Fig. 2 (*continued*).

have a mean of zero, and as a result the regression slopes for the interactions are defined in a metric that produced small regression estimates and standard errors).

3.2.4. Total household income

Household income was also measured in two ways, as a binary variable using the median of each year, and as a centered quantitative variable measuring overall household income (in thousands of dollars). As before, we report the results of the binary coding first. Regression models including mother's IQ showed 8 of 9 slopes for Year were significant (all except age 10), replicating the overall FE. Main effects for household income were also significant in 8 of 9 age replications, except for age 13. In all instances, children whose household income was higher than the median scored higher in PIAT-M compared to their peers whose household had lower income. In the critical tests for the interaction effect, in the two oldest age groups the interaction effect was significant (three of the interactions were significant for the models unadjusted for maternal IQ). As with maternal education, children raised in higher income homes experienced a differential increased FE compared to their counterparts raised by mothers in lower household incomes. Fig. 5 shows the adjusted means by Income for ages 5–13, and particular attention should be paid to ages 12 and 13 as prototypes of this pattern.

As with maternal education, these household income results support further attention to the effect sizes. All slope estimates were positive for each age except for one (age 11). The average interaction slope per year for the oldest 7 age groups was 0.13 (as before, the average of the oldest seven slope estimates, divided by two to adjust into the yearly metric), suggesting that .13 is added to the FE slope each year for children raised in higher income homes.

When we used the quantitative coding for income, significant interactions were obtained for ages 5, 7, and 13 (and 7 of the nine were in the direction implied by the divergence pattern described above). These patterns were generally consistent in form across ages (inspection of Fig. 5 shows that even several of the nonsignificant age categories demonstrate the prototype pattern).

3.2.5. Urbanization

Two groups were defined in relation to urbanization, rural and urban. Adjusting for maternal IQ, the FE was replicated for 7 of 9 Year slopes across age categories (except for ages 6 and 8). Two of 9 urbanization slopes were significant, with higher scores for urban children at age 8 and higher scores for rural children at age 12. Three of 9 interaction terms were significant; for ages 6 and 8 the urban Flynn Effect slope was higher, for age 9 the rural FE slope was higher. These patterns are portrayed in Fig. 6. Because the stability of the slopes is not maintained across ages, we do not inspect the effect sizes in any detail, but rather conclude that there were not consistent FE differentials across urbanization categories like those we found for maternal education and household income.

3.2.6. Education and income combined

The two demographic variables that showed relatively consistent and meaningful divergence within categories were the education and income variables. As a result, we defined models that included both of these variables to further assess the regression patterns. We note that a correlation exists between these two IV's, and so they are certainly not accounting for separate and independent sources of variables. We used the centered quantitative form of the maternal education and income variables for this analysis. We estimated the overall model with the education and income variables, the education and income variables interacted with Year, and then the three-way interaction of education, income, and Year. In cases when the three-way interaction was not significant, we re-estimated the model dropping the three-way interaction.

For six of the nine ages there were meaningful and statistically significant results at $\alpha = .05$ in the interaction tests. For ages 8 and 12, we found a significant three-way interaction, suggesting different Flynn Effect patterns as a result of education and income combined. For ages 5, 10, and 13, we found only a significant Year by income interaction. For age 7, we found only a significant maternal education interaction. Neither variable appeared to be especially dominant in explaining the differential Flynn Effect, and we are left with the conclusion that both maternal education and household income are indicators of the process that appears to account for Flynn Effect differences.

4. Discussion

In this discussion, we will first summarize the findings from our empirical study. Following, we will interpret the results in relation to past literature, with particular attention to the implications of our results for several of the past explanatory theories. Next, we identify threats to validity. Finally, we will identify the progress in understanding the Flynn Effect that emerges from our study.

First, we replicated the Flynn Effect identified by Rodgers and Wänström (2007). We used the same data, but extended the 1986 to 2000 time period they used to cover the years from 1986 to 2004. The overall FE in PIAT-Math scores continued at a slightly increased rate through the additional four years. Second, each gender, each race, each maternal education category, each income category, and both urban and rural data showed a marked FE within our data (see Fig. 1). This finding resolves one outstanding question, whether the FE is almost fully accounted for by a particular demographic subset. The answer to this question is clearly "no," because the FE obtained in all of the many demographic subgroups we examined. Third, the patterns expected (based on prior literature) in mean differences across demographic subgroups were obtained in our data: higher PIAT-M scores for the Caucasian/NA/AA categogry, for females in earlier ages and males in later ages, for children of higher educated parents, and for children in higher income families. These findings can be considered validity checks, to determine that our data patterns match those from many previous research studies.

Our primary set of findings concern the tests for interaction effects, which evaluate a differential FE across demographic categories. To motivate the considerable interest in the answer to this question, it is important to note that simply finding FE patterns in each subgroup does not document that they are of the same magnitude in each

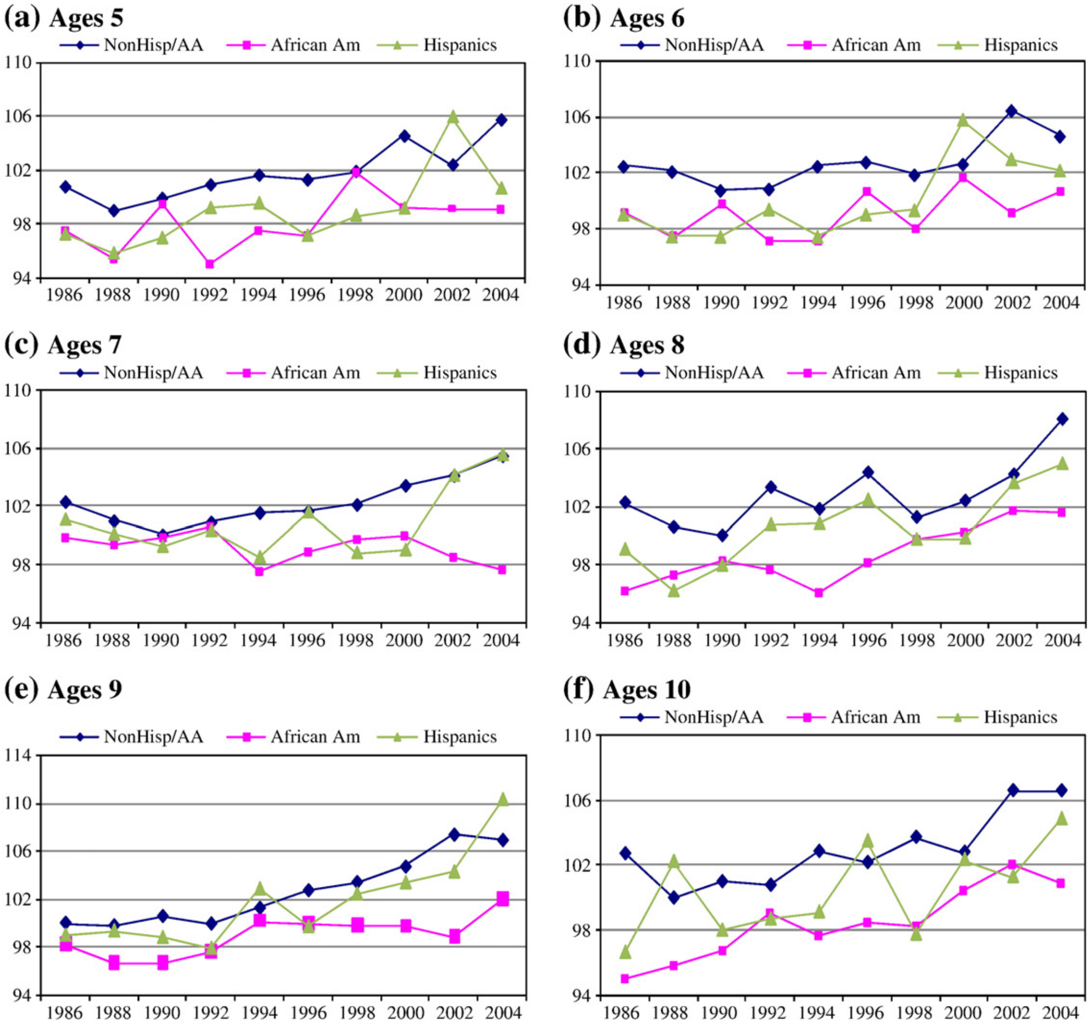

Fig. 3. Adjusted PIAT-M Normed Mean Scores for Year and Race/ Ethnicity (Controlling for Maternal IQ) by Ages 5–13, 1986–2004.

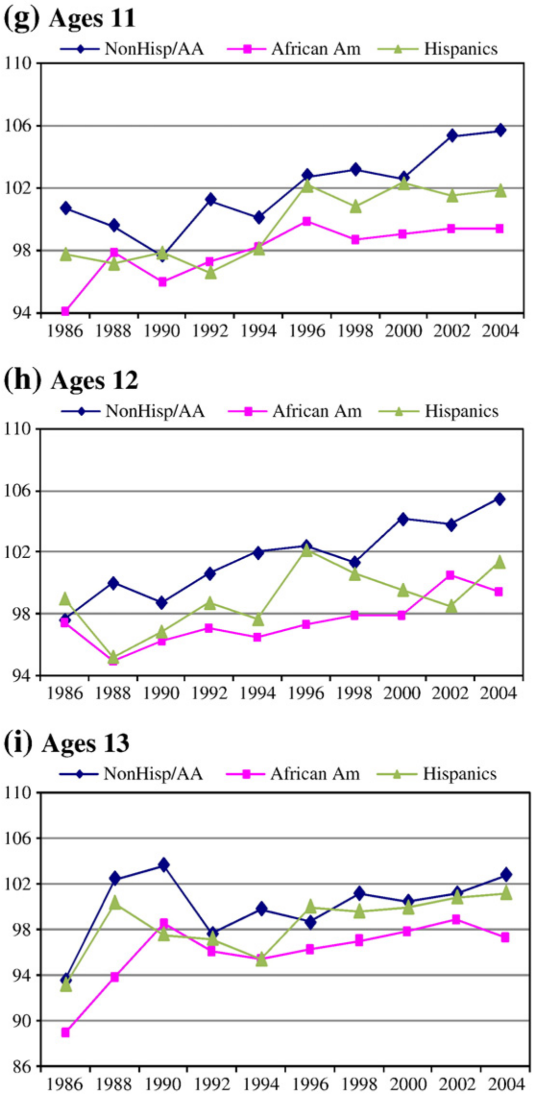

Fig. 3 (*continued*).

subgroup; the interaction tests and the coefficients associated with those evaluate this question. We found no consistent Year by Gender, Year by Race, or Year by Urbanization interaction patterns. However, these are not uninteresting null findings. As we will discuss, the absence of such patterns has implications for past theoretical explanations.

The most important and interesting findings from our empirical study are a number of interactions that are consistent across ages for the Year by Maternal Education, and the Year by Household Income tests. These patterns suggest that the rate of increase in the PIAT-M scores is steeper in the higher educated and higher income categories than for the lower educated and lower income categories. We note that although not all the tests were significant, the direction of the effect sizes were impressive in their consistency. For example, Fig. 4 can be inspected to get a sense of this consistency for the maternal education categories. Starting with age seven, for every age there is a point at which the FE diverges, showing a steeper FE pattern for the children of more highly educated mothers.

For recent summaries and theories, a great deal of past research has documented gender differences in math and quantitative reasoning ability (see Ceci, Williams and Barnett, 2009; Halpern et al, 2007; Johnson, Carothers & Deary, 2008; Lynn, Allik and Irwing, 2004). To summarize a complex but fairly consistent set of patterns, females typically outperform males in childhood, with a crossover occurring shortly before puberty, and males outperform females by the early teens (see Lubinski & Humphreys, 1990, for broader specification). Further, male cognitive ability scores in general – including math ability – are typically more variable than those for females (Johnson et al., 2008; Lehre, Lehre, Laake and Danbolt, 2009). We replicated those findings within our results. For example, Fig. 2 clearly shows the crossover effect between ages five and 13, and standard errors for males were larger in general than for females. But the *change* in PIAT-M scores for males and females tracked one another. Any theory that would predict patterns in relation to gender is informed by this finding. For example, there are obviously hormonal differences between genders (see Geschwind & Behan, 1982; Gardner, 1993; Benbow, 1988). Though hormonal differences have not to our knowledge been implicated in theoretical explanations of the FE, if someone posited that the Flynn Effect had a hormonal (e.g., testosterone) basis, our negative results would tend to rule out that explanation. Similarly, females have been identified as more oriented toward schoolwork, especially in childhood (Kenny-Benson, Pomerantz, Ryan and Patrick, 2006). A theory that the FE is caused primarily by changes over time in schoolwork would be difficult to reconcile with our findings. If schoolwork has a causal effect on the FE, younger females should have higher FEs than younger males; they did not (see Fig. 2). On the other hand, theories that imply approximately equal Flynn Effects across gender would be supported by our findings, including the nutrition hypothesis, educational processes, testing artifacts, medical interpretations, and heterosis.

Similarly, race/ethnicity differences themselves were consistent with past findings, in that the race differences were ordered with Caucasian/NA/AA respondents' scores higher than those of Hispanics' scores, which were in turn higher than those of African Americans. There have been gradual shifts in the size of the race differences in math and other ability measures. For example, Hauser (1998) and Grissmer, Williamson, Kirby and Berends (1998) documented convergence of the race difference in data from the National Assessment of Educational Progress. Until the current study, this finding could be explained by a differential Flynn Effect in which minority scores increased at a steeper rate. However, we found no interaction in our data; the three different race categories each showed substantial FE's, but they also tracked closely to the same consistent increase. The absence of race differences in FE patterns also has implications for the various other theories. If FE patterns in the NLSY-Children emerged from within the family, or were related to average family size (e.g., Sundet et al., 2008), ethnic differences in family culture and family size could potentially create differential FE patterns; but those differences were not observed. If average educational quality is lower for minorities, this could lead to differential FE patterns; again, this finding did not obtain. As for gender, theories that are silent with regards race differences in FE patterns are consistent with the current findings, including the nutrition hypothesis, testing artifacts, and heterosis.

The absence of a difference in FE patterns between rural and urban settings also has interesting implications. Perhaps the theory most linked to this evaluation is the heterosis hypothesis (Mingroni, 2007). This hypothesis proposes that the FE is caused by general increased geographical breadth in the mating pool, which should lead to increased genetic variability and a resulting positive selection pressure on intelligence (and other traits as well; see Mingroni). It seems likely that those living in an urban setting would travel more, and would be more likely to migrate (both internally within the U.S., and internationally), and to communicate (and obtain mates) across distance and across cultures. The absence of any FE differences in urbanization categories does not appear to support the heterosis hypothesis. Other theories silent with respect to urbanization differences would include the nutrition hypothesis, educational processes, medical interpretations, and testing artifacts.

The presence of a consistent and significant set of interactions for Year by Maternal Education, and Year by Household Income – suggesting differential FE across subgroups – are the two strongest and most interesting positive findings in our study (although the analysis including both in the same model was suggestive that these two measures may be indicators of an underlying common cause). The patterns in Figs. 4 and 5 and Tables 1 and 2 show a steeper Flynn Effect for children of mothers with higher education and higher income households. The pattern is identified primarily for the older children in this sample. The implication is that the Flynn Effect is magnified in the upper half of the distribution, that children of higher ability parents and those living in homes with higher household incomes had PIAT-M scores that increased faster during the 1986–2004 period than children

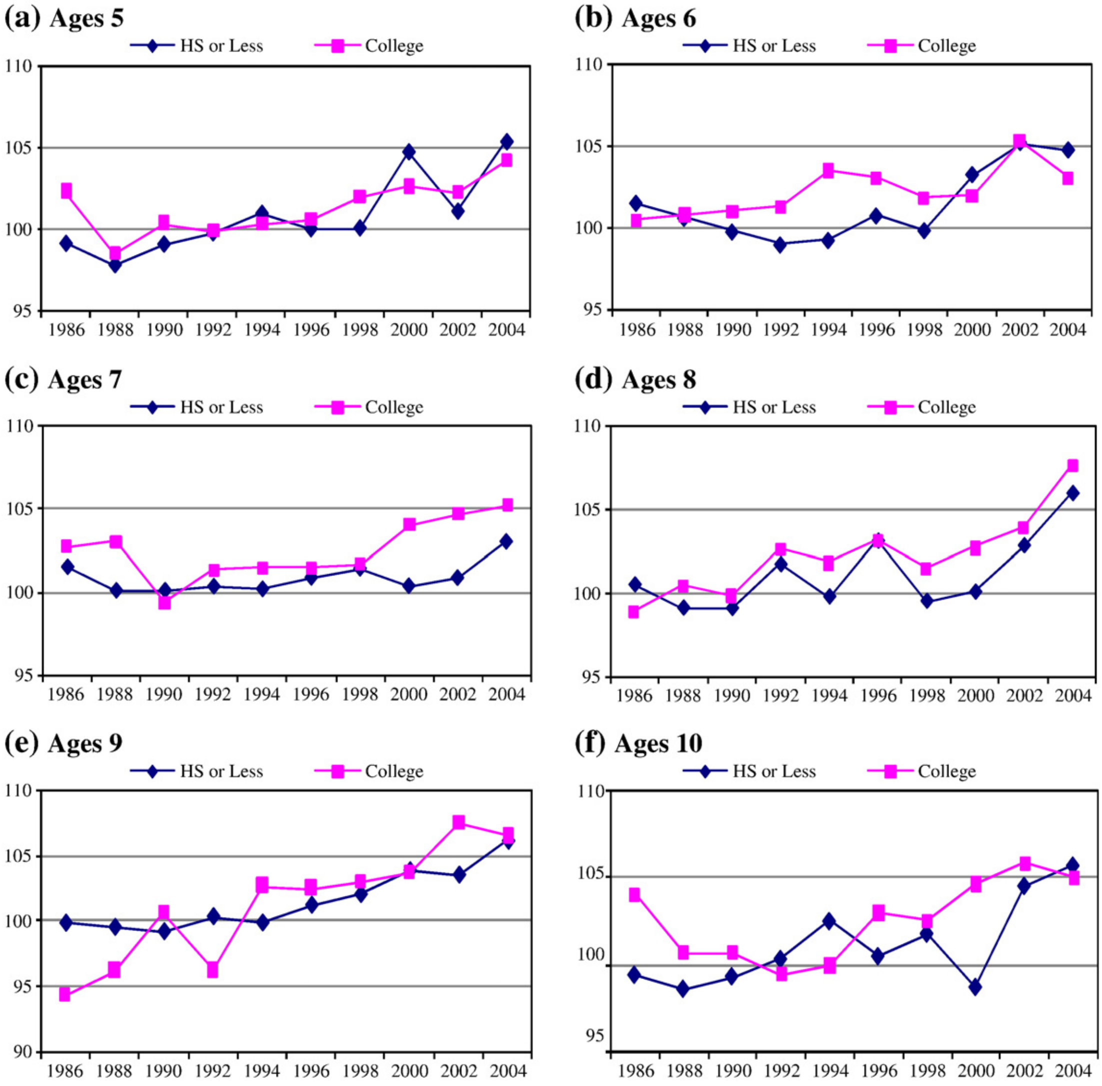

Fig. 4. Adjusted PIAT-M Normed Mean Scores for Year and Maternal Education (after Controlling for Maternal IQ) by Ages 5–13, 1986–2004.

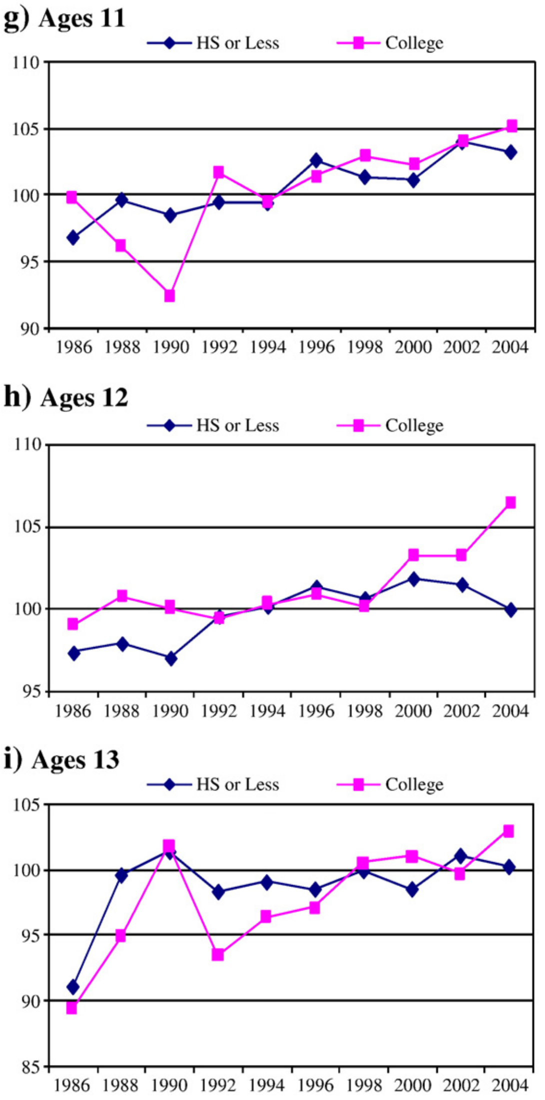

Fig. 4 (*continued*).

with lower educated women and those from lower income households. There are several important implications of these findings.

First, this pattern is inconsistent with the few previous studies that looked at different parts of the distribution. Teasdale and Owen (1989) used Danish draft data, and found a Flynn Effect that was concentrated in the lower half of the distribution. They attributed their results to educational changes. Colom et al. (2005) used male data from schoolchildren in Barcelona, and found that the Flynn Effect increases from 1970 to 1999 were concentrated in the lower and middle parts of the intelligence distribution, and suggested support for the nutrition hypothesis. In both of these studies, not only did the FE concentrate in the lower half of the ability distribution, there was virtually no FE in the upper half. This is in contrast to our results in two ways, first because we found higher FE in the children of higher educated mothers (who are themselves likely to be of higher ability), second because there was still a substantial FE in both halves of the distribution. These observations would focus our causal explanations for the FE on differences that would manifest in different parts of the Danish and Spanish distribution compared to American data, particularly the upper part of the distribution.

In addition, differences in the intelligence measures used in the studies could also have caused these cross-cultural differences. Teasdale and Owen used a battery of tests that included letter matrices, verbal analogies, number series, and geometric series. These overlap with the PIAT-M in the quantitative domain, but also include verbal (crystallized) intelligence components more so than the PIAT-M. Similarly, Colum et al. used the Pressey Graphic Test, a visio-spatial reasoning test, which has four components, verbal instruction, reasoning, classification, and spatial perception. These domains appear to include both fluid and crystallized intelligence components; the last three domains overlap with the PIAT-M scores, though there is likely a stronger loading for the Pressey on crystallized intelligence than for the PIAT-M because of the first component. Of course other cultural differences are quite plausible causes of these FE differences, but to focus on the tests themselves suggests that the use of a broader span intelligence test results in FE results focused in the lower part of the distribution, whereas the more pure fluid measure used in the current study causes a broader FE result that is more strongly focused in the upper part of the distribution. These results are difficult to reconcile, and will need additional inspection and careful thought.

Second, and critically, the portion of this effect is net of any race, gender, or urbanization effects. There are income and education differences across the race categories; but those did not show up as differential FE in those analyses. Thus, the cause of the FE differential in the education/income categories has to be interpreted as occurring in the domain that does not covary with race differences in education/income, an observation that focuses the interpretation considerably. What types of education/income differences are there that would show up the same in all race categories? One example would be general improvements in education methods, as long as those methods are ones applied consistently across all educational levels. Another would be a generally positive improvement, for all levels of society, in standard of living. A more specific realization of this explanation would be television. Most homes in the U.S. – without regard for income or education differences – expose their children to television at increasing rates. Medical improvements that accrue approximately equivalently across race and urban categories would also be implicated (see Steen, 2009, for further discussion of medical improvements); one example would be the almost universal treatment of water with fluoride to prevent tooth decay, which is realized in many settings equivalently across race and gender categories.

Third, in relation to the various explanations that have been suggested for the Flynn Effect, the maternal education finding is strongly supportive of the value of education as part of the explanatory structure. However, we also note that maternal education and household income are each a proxy for a number of other variables. Other variables that might be indirectly measured by these two variables include parenting style, parental intelligence, occupational prestige, and maternal age at first birth (AFB; see Rodgers et al., 2008, and Neiss et al., 2002, for further development of the link between education and AFB). More specifically, we would expect higher educated mothers to pay more attention to nutrition; to provide better educational support for their children by sending them to better schools and contributing to education at home; and to provide support for technological processes that would facilitate learning and achievement. Thus, these various theoretical explanations are supported by this finding. It is also plausible that higher educated mothers would have children more likely to go to school far away, to travel, and to communicate cross-culturally, supporting the heterosis hypothesis as well. However, we note that the level at which the heterosis hypothesis operates is completely different from the other explanations. Rather than a direct link from (for example) better education in support of math knowledge, the heterosis hypothesis works through a background process driven by genetic changes, and is therefore much more difficult to test and evaluate. Most of these other variables, however, would likely show differences by race and/or urbanization category as well.

Our results do cast doubt on some of the artifactual explanations, at least as a full and complete explanation. Research by Beaujean and Osterlind (2008) fitting item response theory (IRT) models to the PIAT-M data in the NLSY supports that the Flynn Effect still exists in these data, even when item invariance is estimated and accounted for (indeed, invariance in item difficulty is part of the explanation of the Flynn Effect). Further, increased testing and testing practice effects are unlikely to explain away our results, showing such dramatic improvements across time on exactly the same instrument, especially for younger ages which have not been previously tested to a great extent. Norming changes do not contribute to our findings, as discussed above. Nor are changing fertility patterns across this relatively short 18-year time period likely to provide much explanatory value.

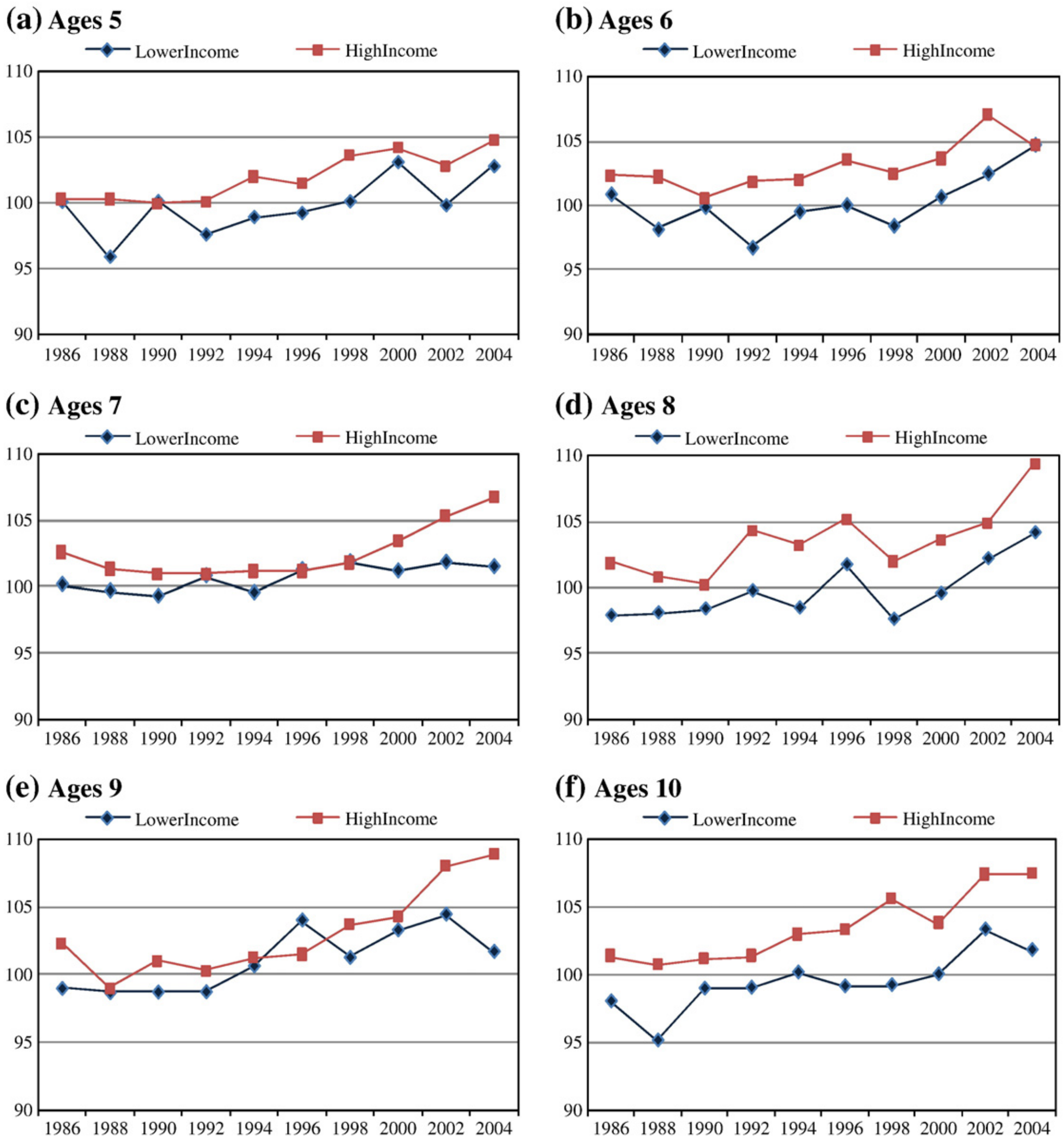

Fig. 5. Adjusted PIAT-M Normed Mean Scores for Year and Household Income (after Controlling for Maternal IQ) by Ages 5–13, 1986–2004.

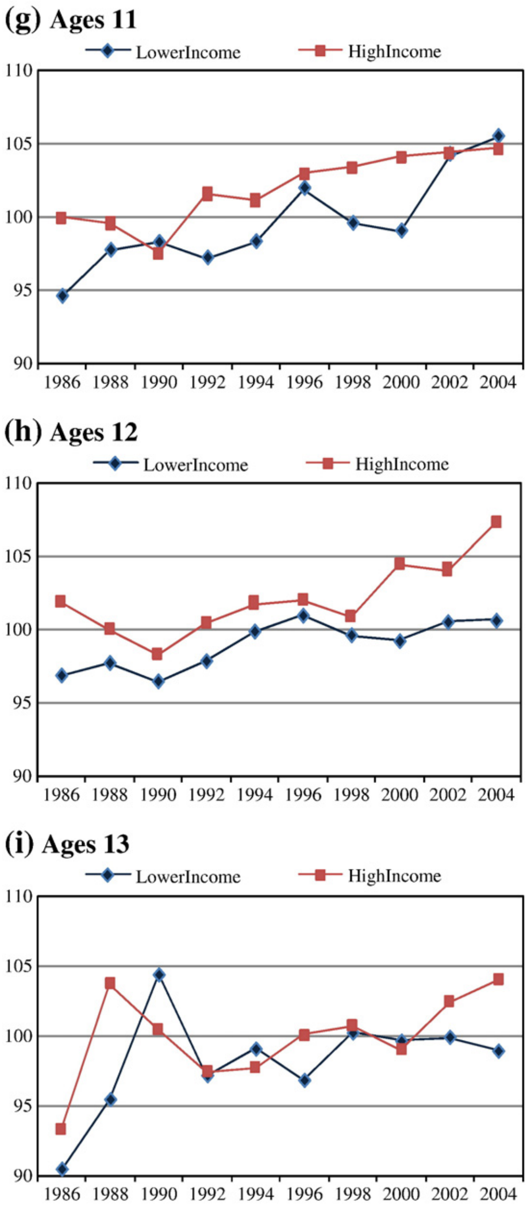

Fig. 5 (*continued*).

Virtually all past FE studies have based their findings on national-level testing results, in which norms are relevant. Many, in fact, have to adjust their findings in relation to re-norming national tests, which occur at regular intervals. Others have the difficulty of at least slightly different ages. Those, like most of the Scandinavian studies in which data were collected during military conscription processes, naturally have slightly different ages of testing, though this concern is mitigated because the test-takers are approaching adulthood. We have a strong advantage in using the NLSYC data, first because we have relatively precise control over age issues, second because we have nine replications of our study at each age from five to 13, and third because this fixed-age design controls for age confounding, which makes the norming process to adjust for age differences virtually irrelevant. (We note, nevertheless, that the age norms that are used with the NLSYC data allow comparisons across our different age groups.) An important advantage of our study is that the results are more directly interpretable in relation to a substantive test-taking process. Given the clear link to the specific PIAT-M items, we can consider what might cause children (especially older children) to improve so dramatically.

Of course if nutrition is increasing systematically in the background, this would plausibly cause improvements. However, it is difficult to understand how a nutrition explanation would focus so explicitly on supporting only math improvement (though some researchers have focused on the parts of the brain that could be differentially affected; e.g., Blair et al, 2005). It is also difficult to reconcile a nutrition explanation with the consistency of the FE in many settings. For example, is it logical to think that there really are facilitating nutrition effects occurring in each and every year (in the aggregate), as patterns in the figures suggest they have been? We find the nutrition hypothesis to be highly plausible, and one likely cause of the FE patterns, but we also do not find it to be convincing as a single explanation.

On the other hand, explanations that focus on education improvement do plausibly stand up to the challenge of these patterns, at least as they are broadly considered. The whole educational establishment is working to improve the personnel, methods, technology, and the whole infrastructure that supports learning, and this process is ongoing and consistent (though a critic might plausibly doubt the successful implementation of the process at such a consistent level as reflected in the FE). Further, math education has seen concerted attention during the period from which our data are obtained. Flynn (2006) also found educational improvement to be a likely explanatory mechanism.

Along with educational improvements, general increases in quality of life, including medical improvements, public health initiatives, housing quality, communication improvements (including the internet and television), and other factors that correspond to general cultural progress also appears consistent with these patterns (Steen, 2009). This explanatory category has the advantage of multiple components, consistent with Jensen's multiplicity hypothesis, and helps explain how the FE has been consistent across periods when many of the explanatory processes themselves were changing.

There are two follow-up studies that are in process and nearing completion, results of which we will report in future papers. First, we are developing a much more in-depth distributional study. We can explicitly identify the different parts of the children's ability distribution, and we will focus much more minutely on the various distributional features than just to divide the distribution in half in future research. Secondly, we have completed an item-level psychometric analysis (Ang, 2008) of the PIAT-M items in relation to the Flynn Effect patterns shown within subgroups in the current study, to identify what kinds of items produce the FE in the NLSYC data (see Beaujean & Osterlind, 2008, for a different item-level study of the NLSYC data). This item-level study includes linking content ratings of the PIAT-Math items to item difficulty estimates of differential item functioning (DIF) using a 2 parameter IRT model fit to the PIAT-Math data (see Beaujean & Osterlind for details). This follow-up study, along with Beaujean and Osterlind (also see Wicherts et al, 2004) provides a methodologically important step forward in understanding Flynn Effect patterns, by accounting for the potential for lack of item invariance as a cause of the Flynn Effect. The current study is based on the "standard methodology" in which the Flynn Effect is measured in relation to the mean difference in raw scores. The additional studies described in this paragraph provide important contextualizing for these and other past patterns.

There are of course other threats to validity associated with our study, though we hasten to add that there are relatively few compared to many previous FE studies. The weaknesses include the use of two-year measurement intervals, and the natural measurement error that is associated with testing administration. These are balanced against our use of a set of children whose mothers were recruited through a national probability sample, the use of comparable and high-quality instrumentation, and empirical and theoretical background that provided strong motivation for where to look and how to design our study. We also note recent attention to working on item-level FE studies, including a focus on differential item functioning (DIF) and invariance issues (Wicherts et al, 2004), to which we intend to contribute future research. Our focus on raw scores within the current study (as well as virtually all past FE studies) may ultimately be anachronistic, as more sophisticated psychometric procedures are brought to bear on explaining FE patterns. In the context of threats to validity within the current study, one such threat is the possibility of cohort DIF, which could be caused by shifting meaning of items across cohorts.

In conclusion, we now know new and important features of the Flynn Effect, at least the FE found in U.S. data from the past two decades or so. The effect itself is strong and consistent, but we found no differential gender or race FE, nor was there much of a differential urbanization status identified. The positive finding of a differential FE in relation to maternal education (and at a smaller level, household income) at the older ages is suggestive of

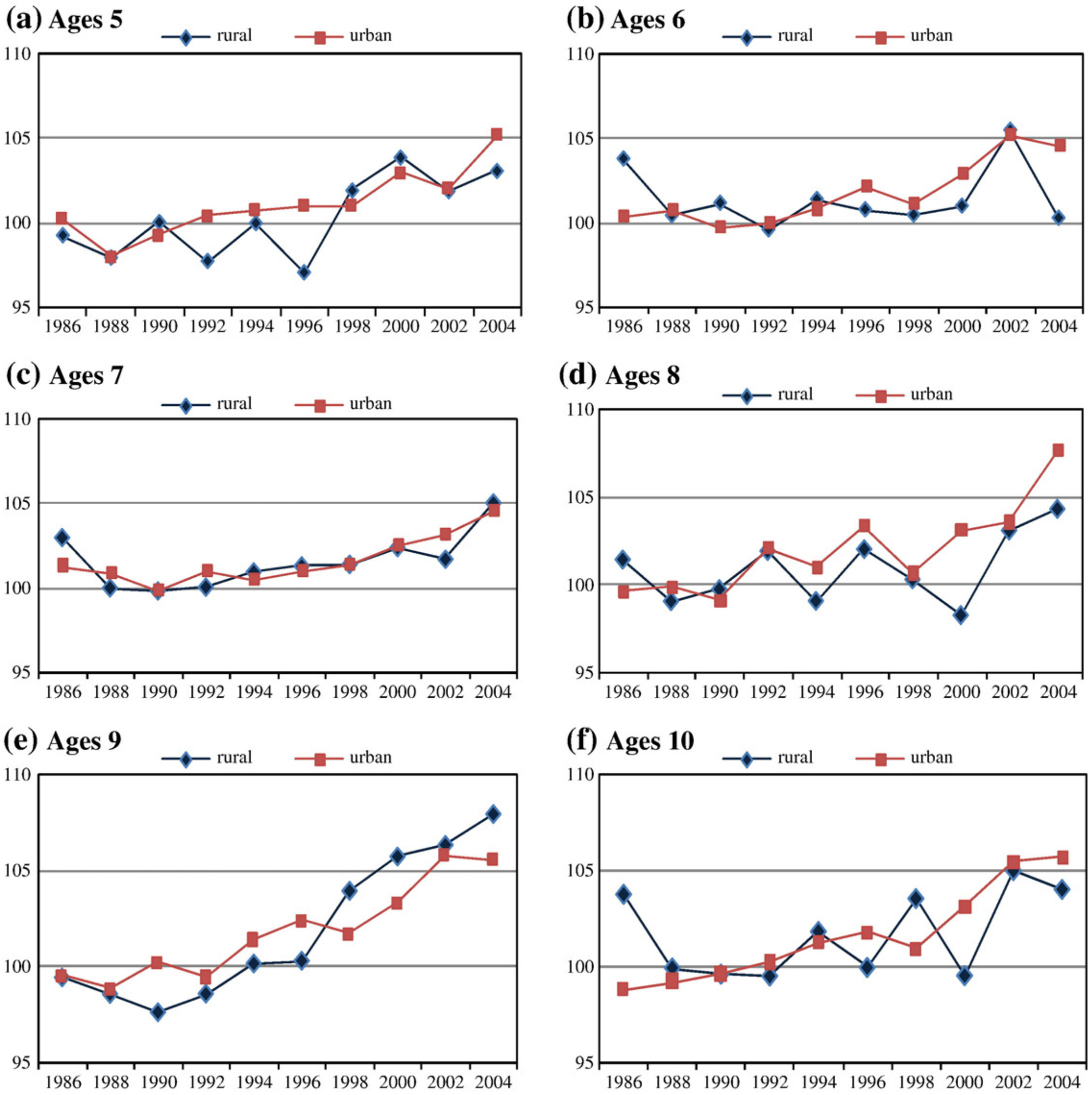

Fig. 6. Adjusted PIAT-M Normed Mean Scores for Year and Urbanization (after Controlling for Maternal IQ) by Ages 5–13, 1986–2004.

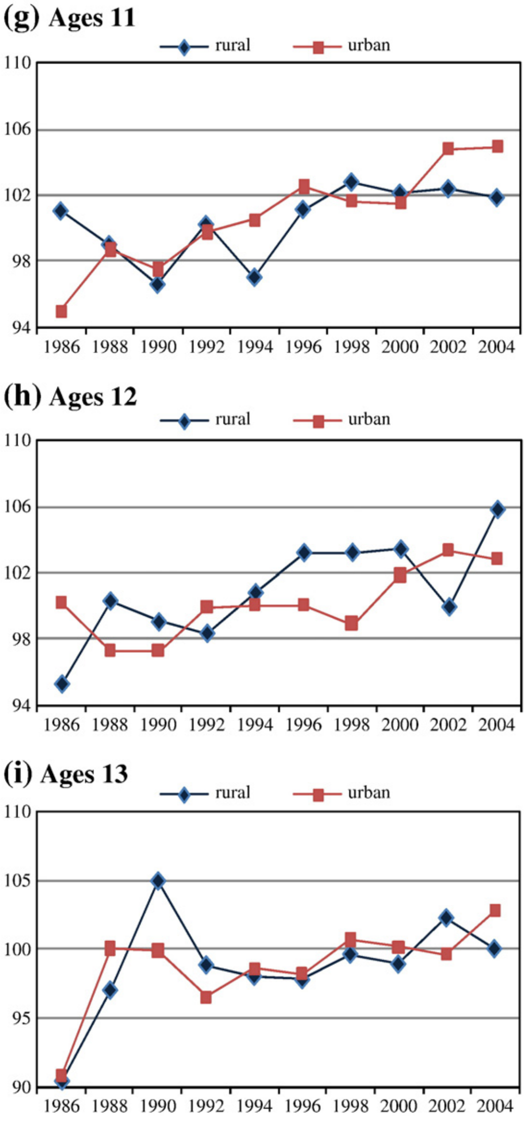

Fig. 6 (*continued*).

some of the dynamics of the process leading to the Flynn Effect. However, we do not consider our findings to be confirmatory in any sense. Rather, as suggested in Rodgers (1998), these findings should be helpful in relation to future designs to help establish the elusive causal factor(s) that are explanatory of this fascinating process.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* London: Sage Publication.

Ang, S. (2008). The Flynn Effect within groups and PIAT math items: Moving from the general to the specific. Doctoral dissertation, University of Oklahoma Department of Psychology.

Beaujean, A. A., & Osterlind, S. J. (2008). Using item response theory to assess the Flynn Effect in the National Longitudinal Survey of Youth 79 Children and Young Adults data. *Intelligence*, *36*, 455—463.

Benbow, C. P. (1988). Sex differences in mathematical reasoning ability in intellectually talented preadolescents: Their nature, effects and possible causes. *Behavioral and Brain Sciences*, *11*, 169—232.

Blake, J. (1989). *Family size and achievement.* Berkeley: University of California Press.

Blair, C., Gamsonb, D., Thorne, S., & Baker, D. (2005). Rising mean IQ: Cognitive demand of mathematics education for young children, population exposure to formal schooling, and the neurobiology of the prefrontal cortex. *Intelligence*, *33*, 93—106.

Brand, C. R. (1996). *The g factor: General intelligence and its implications.* Chichester: Wiley.

Brand, C. R., Freshwater, S., & Dockrell, W. B. (1989). Has there been a "Massive" rise in IQ levels in the west? *Irish Journal of Psychology*, *10*, 388–394.

Ceci, S. J. (1996). *On intelligence: A biocecological treatise on intellectual development*, Expanded ed. Cambridge, MA: Harvard University Press.

Ceci, S. J., Williams, W. M., & Barnett, S. M. (2009). Women's underrepresentation in science: Sociocultural and biological considerations. *Psychological Bulletin*, *135*, 218—261.

Colom, R., Andre´s Pueyo, A., & Juan-Espinosa, M. (1998). Generational IQ gains: Spanish data. *Personality and Individual Differences*, *25*, 927—935.

Colom, R., Juan-Espinosa, M., & García, L. F. (2001). The secular increase in test scores is a 'Jensen effect'. *Personality and Individual Differences*, *30*, 553—559.

Colom, R., Lluis-Font, J. M., & Andres-Pueyo, A. (2005). The generational intelligence gains are caused by decreasing variance in the lower half of the distribution: Supporting evidence for the nutrition hypothesis. *Intelligence*, *33*, 83—91.

Daley, T. C., Whaley, S. E., Sigman, M. D., Espinosa, M. P., & Neumann, C. (2003). IQ on the rise: The Flynn effect in rural Kenyan children. *Psychological Science*, *14*, 215—219.

Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review*, *108*, 346—369.

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

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*(2), 171—191.

Flynn, J. R. (1998). IQ gains over time: Toward finding the causes. In Ulric Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 25—66). Washington DC: American Psychological Association.

Flynn, J. R. (2006). IQ gains, WISC subtests, and fluid g: g-theory and the relevance of Spearman's hypothesis to race. In G. R. Bock, J. A. Goode, & K. Webb (Eds.), *The nature of intelligence* (pp. 202—227). New York: Wiley.

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

Flynn, J. R. (2009a). Requiem for nutrition as the cause of IQ gains: Ravens gains in Britain 1938–2008. *Economics and Human Biology*, *7*, 18—27.

Flynn, J. R. (2009b). The WAIS III and WAIS IV: Daubert motions favor the certainly false over the approximately true. *Applied Neuropsychology*, *16*, 98—104.

Gardner, H. (1993). *Frames of mind.* New York: Basic Books.

Geschwind, N., & Behan, P. (1982). Left-handedness: Association with immune disease, migraine, and developmental learning disorder. *Proceedings of the National Academy of Sciences*, *79*, 5097—5100.

Greenfield, P. M. (1998). The cultural evolution of IQ. In Ulric Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 81—124). Washington DC: American Psychological Association.

Grissmer, D. W., Kirby, S. N., Berends, M., & Williamson, S. (1994). *Student achievement and the changing American family.* Santa Monica, CA: RAND Institute on Education and Training.

Grissmer, D. W., Williamson, S., Kirby, S. N., & Berends, M. (1998). Exploring the rapid rise in Black achievement scores in the United States (1970–1990). In Ulric Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 25—66). Washington DC: American Psychological Association.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, *8*, 1—51.

Hauser, R. M. (1998). Trends in Black–White test-score differentials: Uses and misuses of NAEP/SAT data. In Ulric Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 25—66). Washington DC: American Psychological Association.

Horn, J. L., & Catell, R. B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology*, *57*, 253—270.

Jensen, A. R. (1991). Speed of cognitive processes: A chronometric anchor for psychometric tests of g. *Psychological Test Bulletin*, *4*, 59—70.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport: Praeger.

Johnson, W., Carothers, A., & Deary, I. J. (2008). Sex differences in variability in general intelligence: A new look at the old question. *Perspectives on Psychological Science*, *6*, 518—531.

Kanaya, T., Scullin, M. H., & Ceci, S. J. (2003). The Flynn Effect and U.S. policies: The impact of rising IQ scores on American society via mental retardation diagnoses. *American Psychologist*, *58*, 778—790.

Kane, H., & Oakland, T. D. (2000). Secular declines in Spearman's g: Some evidence from the United States. *Journal of Genetic Psychology*, *161*, 337−345.

Kenny-Benson, G. A., Pomerantz, E. M., Ryan, A. M., & Patrick, H. (2006). Sex differences in math performance: The role of children's approach to schoolwork. *Developmental Psychology*, *42*, 11−26.

Lehre, A. C., Lehre, K. P., Laake, P., & Danbolt, N. C. (2009). Greater intrasex phenotype variability in males than in females is a fundamental aspect of the gender differences in humans. *Developmental Psychobiology*, *51*, 190−206.

Loehlin, J. C. (1996). Environment and intelligence: A comment. In D. Detterman (Ed.), *Current topics in human intelligence. The environment*, *vol. 5*. (pp. 151−156)Norwoord, NJ: Ablex.

Lubinski, D., & Humphreys, L. G. (1990). A broadly-based analysis of mathematical giftedness. *Intelligence*, *14*, 327−355.

Lynn, R. (1982). IQ in Japan and the United States shows a growing disparity. *Nature*, *306*, 291−292.

Lynn, R. (1989). A nutrition theory of the secular increase in intelligence: Positive correlations between height, head size, and IQ. *British Journal of Educational Psychology*, *59*, 372−377.

Lynn, R. (1990). The role of nutrition in secular increases in intelligence. *Personality and Individual Differences*, *11*, 273−285.

Lynn, R. (1998). In support of the nutrition theory. In Ulric Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 207−218). Washington DC: American Psychological Association.

Lynn, R. (2009). Fluid intelligence but not vocabulary has increased in Britain, 1979–2008. *Intelligence*, *2009*, 249−255.

Lynn, R., Allik, J., & Irwing, P. (2004). Sex differences on three factors identified in Raven's standard progressive matrices. *Intelligence*, *32*, 411−424.

Lynn, R., & Hampson, S. (1986). The rise of national intelligence: Evidence from Britain, Japan and the U.S.A. *Personality and Individual Difference*, *7*, 23−32.

Mahlberg, A. (1997). The rise in IQ scores. *American Psychologist*, *52*, 71.

Martorell, R. (1998). Nutrition and the worldwide rise in IQ scores. In Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 183−206). Washington, DC: American Psychological Association.

Mingroni, M. A. (2007). Resolving the IQ paradox: Heterosis as a cause of the Flynn Effect and other trends. *Psychological Review*, *114*(3), 806−829.

Must, O., Must, A., & Raudik, V. (2003). The Flynn effect for gains in literacy found in Estonia is not a Jensen effect. *Personality and Individual Differences*, *34*, 1287−1292.

Neiss, M., Rowe, D. C., & Rodgers, J. L. (2002). Does education mediate the relationship between IQ and age of first birth? A behavior genetic analysis. *Journal of Biosocial Science*, *34*, 259−275.

Neisser, U. (Ed.). (1998). *The rising curve: Long-term gains in IQ and related measures*. Washington, DC: American Psychological Association.

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

Rodgers, J. L., Cleveland, H. H., van den Oord, E., & Rowe, D. C. (2000). Resolving the debate over birth order, family size, and intelligence. *American Psychologist*, *55*, 599−612.

Rodgers, J. L., Kohler, H. -P., McGue, M., Behrman, J., Petersen, I., Bingley, P., et al. (2008). Education and IQ as direct, mediated, or spurious influences on female fertility outcomes: Linear and biometrical models fit to Danish twin data. *American Journal of Sociology*, *114*(Supplement), S202−S232.

Rodgers, J. L., Rowe, D. C., & May, K. (1994). DF analysis of NLSY IQ/achievement data: Nonshared environmental influences. *Intelligence*, *19*, 157−177.

Rodgers, J. L., & Wänström, L. (2007). Identification of a Flynn Effect in the NLSY: Moving from the center to the boundaries. *Intelligence*, *35*, 187−196.

Rowe, D. C., & Rodgers, J. L. (2002). Expanding variance and the case of historical changes in IQ means: A critique of Dickens and Flynn (2001). *Psychological Review*, *109*, 759−763.

Rushton, J. P. (1997). Review essay. *Society*, *34*, 78−82.

Rushton, J. P. (1999). Secular gains in IQ not related to the g factor and inbreeding depression—Unlike Black–White differences: A reply to Flynn. *Personality and Individual Differences*, *26*, 381−389.

Rushton, J. P. (2000). Flynn effects not genetic and unrelated to race differences. *American Psychologist*, *55*, 542−543.

Schoenthaler, S. J., Amos, S. P., Eysenck, H. J., Peritz, E., & Yudkin, J. (1991). Controlled trial of vitamin-mineral supplementation: Effects on intelligence and performance. *Personality and Individual Differences*, *12*, 351−362.

Schooler, C. (1998). Environmental complexity and the Flynn Effect. In Ulric Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 67−80). Washington DC: American Psychological Association.

Smith, S. (1942). Language and nonverbal test performance of racial groups in Honolulu before and after a 14-year interval. *Journal of General Psychology*, *26*, 51−92.

Steen, R. G. (2009). Human intelligence and medical illness. New York: Springer.

Sundet, J. M., Borren, I., & Tambs, K. (2008). The Flynn effect is partly caused by changing fertility patterns. *Intelligence*, *36*, 183−191.

Sundet, J. M., Barlaug, D. G., & Torjussen, T. M. (2004). The end of the Flynn effect? A study of secular trends in mean intelligence test scores of Norwegian conscripts during half a century. *Intelligence*, *32*, 349−362.

Sundet, J. M., Eriksen, W., Borren, I., & Tambs, K. (2010). The Flynn effect in sibships: Investigating the role of age differences between siblings. *Intelligence*, *38*, 38−44.

Teasdale, T. W., & Berliner, P. (1991). Experience of kindergartens in relation to educational level and cognitive abilities in adulthood: A geographical analysis. *Scandinavian Journal of Psychology*, *32*, 336−343.

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

Teasdale, T. W., & Owen, D. T. (2000). Forty year secular trends in cognitive ability. *Intelligence*, *28*, 115−120.

Tuddenham, R. D. (1948). Soldier intelligence in World Wars I and II. *American Psychologist*, *3*, 54−56.

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., Van Baal, G. C., Boomsma, D. I., et al. (2004). Are intelligence measurement invariant over time? Investigating the Flynn effect. *Intelligence*, *32*, 509−537.

Wichman, A. L., Rodgers, J. L., & MacCallum, R. C. (2006). A multilevel approach to the relationship between birth order and intelligence. *Personality and Social Psychology Bulletin*, *32*, 117−127.

Wilkinson, L. & Task Force on Statistical Inference, APA Board of Scientific Affairs. (1999). Statistical methods in Psychology journals: Guidelines 1178 and explanations. American Psychologist, 54, 595−604.

Williams, W. M. (1998). Are we raising smarter children today? School and home-related influences on IQ. In Ulric Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 125−154). Washington DC: American Psychological Association.