Breastfeeding and the Child Cognitive Outcomes: A Propensity Score Matching Approach

Miao Jiang · E. Michael Foster · Christina M. Gibson-Davis

Published online: 17 September 2010
© Springer Science+Business Media, LLC 2010

Abstract To estimate the effect of breastfeeding initiation and duration on child development outcomes, 3,271 children and their mothers participating in the Child Development Supplement of the Panel Study of Income Dynamics provide data for these analyses. Main outcomes include Woodcock Johnson Psycho-Educational Battery-Revised (WJ-R) test score (letter word, passage comprehension, applied problem, and broad reading), and Wechsler Intelligence Scale for Children-Revised (WISC-R) test score at the 2002 survey. Controlled variables include family, maternal, and child characteristics, many of which can be traced back to the year the child was born. The analytic technique is propensity score matching with multiple imputations. After using propensity scores to adjust for confounding factors, breastfeeding initiation showed statistically significant effects but the practical scale remains small. Breastfeeding duration showed a non-linear effect on those outcomes and most of the effects are not significant. The effects of breastfeeding on child’s cognitive outcomes are modest in practical terms. The non-linear effects suggest that selection into breastfeeding may account for the increased score of children who are breastfed.

Keywords Breastfeeding · Child cognitive outcomes · Propensity score · Multiple imputations · Confounding · Effect size

Introduction

While breastfeeding is widely regarded as providing numerous benefits for child and maternal health [1], its effect on children’s cognitive development remains a matter of debate. Breastfeeding is thought to directly affect child cognition through the superior nutrients found in breast milk; breast milk contains long-chain polyunsaturated fatty acids (LCPUFAs) [2–7]. LCPUFAs, particularly n-3 docosahexaenoic (DHA) and n-6 arachidonic acid (AA), are necessary for retinal and neural development [8–11], and animal studies indicate that severe LCPUFA deprivation harms cognitive and behavioral performance [12]. Evidence showing a causal role between LCPUFAs and enhanced human neural functioning, however, is inconclusive [10, 12–14].

Likewise, the association between breastfeeding and cognition also has been mixed. Some find positive effects of breastfeeding [4, 15–20], and a meta-analysis of 20 studies find that on average, breastfeeding increases cognition by five points for low birth weight infants and three points for normal weight infants [21]. (These estimates imply small effect sizes.) Two recent reviews [22, 23] conclude that, based on evidence from more rigorous studies, breastfeeding does improve cognition, particularly for children who are low birth weight. Others find that these effects are spurious, explained by confounding factors, such as genetic and socioenvironmental factors [24–26]. Furthermore, researchers disagree on the effects of duration of breastfeeding on child outcomes—some find a significant relationship [19, 21, 27, 28]; others, a non-linear dose–response relationship [29]; and still others, no significant relationship after adjusting for confounders [30].

Much of the ambiguity in prior research reflects difficult challenges, especially those of identifying the causal effect...
of any risk factor outside of random assignment. In observational analyses, researchers face difficult choices involving how to condition on observed differences between women who do and do not breastfeed and how to assess the possibility of unobserved confounding. For example, recent research indicates that maternal IQ and the home stimulation environment predict breastfeeding status better than even poverty or education [24, 25], yet few studies include these two measures as covariates. In addition, many studies have relied on clinical based samples [7, 19, 24] or have only considered low birth weight infants [16, 31]. Such samples suffer from a range of problems involving both internal and external validity [32]. For example, conditioning on low birth weight likely creates a spurious negative correlation between breastfeeding and (unobserved) positive determinants of the health of women and children, thus exaggerating the effect of breastfeeding. (For details, see Greenland [33] or Pearl [34].) Other issues include non-standardized measures of breastfeeding, small sample sizes, and retrospective collection of data.

This article examines the effect of breastfeeding initiation and duration on child cognitive outcomes. Our goal is to address the methodological problems afflicting earlier studies. For example, we use propensity score matching to correct for selection on observables in estimates of the association between breastfeeding and child outcomes. This technique represents an alternative approach to ordinary regression for adjusting for observed confounding factors. Furthermore, we examine outcomes for a nationally representative sample. Thus, our results can be more easily generalized to the overall population than can results from previous studies with restricted samples. Moreover, we examine five measures of cognitive development. This range of outcomes allows us to identify an overall pattern of benefits of breastfeeding on child’s cognition. The comprehensiveness of the data, including measures of maternal IQ and home cognitive stimulation, enables us to better account for the effect of observed confounders that are likely to bias our estimates.

Methods

Study Sample

The Panel Study of Income Dynamics (PSID) includes extensive data on the income and well-being of a representative sample of U.S. families over nearly four decades. The study covers a range of topics, including family composition change, food and housing expenditures, employment, income, health, and welfare. Data were collected annually from 1968 to 1997 and biannually after 1997. The number of families involved has grown from around 5,000 in 1968 to 8,000 as of 2005 [35]. These data are the source of much of what is known about poverty, income instability and family well-being in the U.S. population [36]. The PSID, for example, was the first survey to document that poverty status was not always permanent.

The Child Development Supplement (CDS) is a special component of PSID. In 1997, PSID randomly selected 3,563 children from 2,394 PSID families and collected information on a broad aspect of their lives, from cognitive, behavioral, and health status to family and neighborhood environment. Those children were born between 1984 and 1997. In 2002–2003, these families were re-interviewed and data were successfully collected on 2,907 children [37]. The CDS provides researchers with comprehensive, nationally representative, and longitudinal data on children and their caregivers [35].

Our study is based on data from 3,271 CDS children and their families. Of the original 3,563 children, 292 children were excluded as follows: (1) they did not live with their biological mother at the time of the 1997 interview ($n = 271$); (2) they were missing information on breastfeeding status ($n = 11$); (3) they were missing information on race and ethnicity ($n = 5$); and (4) they were missing information on date of birth ($n = 5$). For any other information missing, we imputed those values using multiple imputations as described below.

Variables

Outcomes

Woodcock Johnson Psycho-Educational Battery-Revised (WJ-R) [38] measures achievement in several domains of intellectual ability, including current developmental status, degree of mastery in reading and mathematics, and group standing (either age or grade group) [39]. CDS includes three subsets to measure reading and math skills: the letter word, passage comprehension, and applied problems sets, as well as two summation measures for broad reading and broad math.

Wechsler Intelligence Scale for Children-Revised (WISC-R) [40] tests for the child’s ability to process information as it directly measures the short-term working memory in terms of how many sequential digits the respondent can repeat backward or forward. CDS includes digit-forward, digit-backward, and total score. This analysis uses the total score.

These test measures are available for both the 1997 and the 2002 surveys (except broad math, which was only administered in the 1997 survey), but we use the 2002 measures for two reasons. First, the variables indicating the stimulating environment at home were collected at the same time as the 1997 outcome measures. Therefore,
reduce the risk of endogeneity, we rely on outcomes observed at the later time point. Second, there are less missing values in the 2002 outcomes because more children had reached either 3 or 6 years old to be eligible to take the cognitive tests.

Breastfeeding

Our independent variables of primary interest are initiation and duration of breastfeeding. Both questions were asked in the 1997 CDS interview. Breastfeeding initiation is a dummy variable with “1” indicating ever breastfeeding. Breastfeeding duration is grouped into four categories: never breastfed; less than 6 months; 7–12 months; and more than 12 months. We examine alternative cutoffs under the sensitivity analysis.

Covariates

Based on previous literature and the availability of the PSID and CDS data, we select the following covariates to be included in our analyses. For child covariates, we include the child’s age (in months, continuous), race and ethnicity [white, non-Hispanic (omitted category); black, non-Hispanic; and other], sex, number of siblings, whether the child was first born to the mother, whether the child was born preterm (e.g., gestational age is less than 37 weeks), whether the child was born small for gestational age (SGA) (e.g., weighing less than a specified percentile of birth weight for a given gestational age, we used a gender-specific SGA measure from Alexander and colleagues [41]), the mother-rated child’s health at birth as compared to other babies (1: worse; 2: same; 3: better; treated as continuous variable), and the Home Observation for Measurement of the Environment (HOME) scale [42]. Maternal characteristics include IQ (taken from the passage completion test of Woodcock Johnson test-Revised administered to primary care givers in the 1997 CDS interview), education [high school dropout; completed high school; attended college (omitted category)], her age at the time of the child’s birth, her self-rated health status (from 1 to 5: poor, fair, good, very good, and excellent, treated as continuous variable), whether she was enrolled in Women, Children, and Infant (WIC) program or Medicaid during pregnancy, and whether she was working, married, or the head of the household (e.g., the only adult) the year of the child’s birth. Family level characteristics include the household’s income from the year of the child’s birth.

Sampling weights are available in the data. These weights are not included in the propensity score equation, as the equation has already included most of the covariates that are used to calculate the weights (e.g., race). In addition, preliminary analysis reveals that including the survey weight in the propensity score equation does not make any real difference in predicting the breastfeeding status. The 2002 survey weight is, however, used in the multiple imputations as a covariate to provide more information for filling the missing values of the data.

Statistical Analysis

Propensity Score Matching

A propensity score is the conditional probability of assignment to treatment given a set of covariates. Rosenbaum and Rubin [43] demonstrate that given the propensity score, treatment assignment and the observed covariates are conditionally independent. They further demonstrate that if the treatment is ignorable given the covariate vector X, then it is also ignorable given the propensity score. This assumption of “ignorability” is the key in propensity score matching (as well as in ordinary least square regression). It means conditional on observed characteristics, treatment assignment is effectively randomized. Under this assumption, one can estimate the treatment effect by conditioning on a balanced propensity score. Therefore, propensity score summarizes the multi-dimension covariate vector into a single measure.

Propensity score matching has several advantages relative to ordinary regression. First, regression generally assumes linearity between the covariates and the outcome of interest. When this assumption does not hold, the model may be mis-specified, leading to over- or understate of the program effect. (Model mis-specification creates a form of omitted variable bias.) Propensity score matching relaxes the linearity assumption, thus is less sensitive to the functional form between the covariates and the outcome.

Second, when the distributions of the covariates among the treated and the comparison cases do not overlap, regression uses extrapolation to form the predicted counterfactual values for those cases unique to one group or the other. This procedure can lead to misleading estimates of the treatment effect if the model’s functional form is not correct [44]. Propensity score matching, on the other hand, identifies treated cases (in our case, breast-feeding mothers) that have no comparable comparison cases (they have a propensity score close to 1). These mismatched cases can then be excluded from the analysis.

Third, when the treatment effects varies across subgroups of individuals (i.e., moderated by observed covariates), regression produces an estimated effect that is not equal to the effect of treatment on the treated; neither can it be converted to the effect of treatment on the untreated, or the effect of treatment in the population. For more discussion, see [45].

Note however, like regression, propensity score matching relies on the key assumption of no unobserved
confounders [46–49]. We conduct the sensitivity analysis examining this issue.

The Propensity Score Model for the Breastfeeding Initiation

First, we calculate the propensity score—the predicted probability of breastfeeding initiation for each individual—using a probit model. The next step is to check the covariates balance. The propensity score represents a “balancing score”—when matching cases on the propensity score, the distribution of the covariates between those who did and did not breastfeed should be the same [43]. This property, however, holds only as sample sizes approach infinity and so is only approximate in any given sample. As a result, a good practice involves comparing the distribution of covariates between groups adjusted for propensity score. If the covariates do not balance, then one should modify the propensity score equation [50] and re-match the cases. While over-fitting the data in this manner would not be desirable for many purposes (e.g., such as testing the hypotheses about the determinants of breast-feeding), modifying the propensity score in this way ensures that the propensity score better captures the between-group differences in the covariates.

To improve the covariates balance, we ran race/ethnicity-specific propensity score models defined by black, non-Hispanic (n = 1299) and non-black (white, non-Hispanic (n = 1,547) and other (n = 425)). Because the number is not big enough to run separate model for the “other” group, we still include a dummy variable for race = other in the model.

The covariates balance was further improved by including square terms for some of the continuous predictors such as age, and family income. Note each child still only has one propensity score, which is calculated using the prediction model corresponding to his or her specific gender-race subgroup.

With the propensity score, one can then estimate the effect of breastfeeding. For each treatment case, we first obtain a counterfactual outcome by calculating a weighted average outcome of all non-breastfeeding children. The closer a non-breastfeeding child’s propensity score is to the treatment case, the greater the weight. The econometrics literature suggests that this method is especially effective in estimating the counterfactual [51]. Then the mean difference between the actual and the counterfactual outcome of the treatment cases is the average effect of breastfeeding for a breastfeeding child.

In addition, we calculate the effect size, which is the standardized mean difference for the treatment effects regardless of the sample size [52]. This measure reflects the magnitude of the estimated effects of breastfeeding on child cognitive outcomes more accurately because we have multiple outcomes and the statistical significance subjects to the issue of multiple testing. (We adjusted the P value using Bonferroni correction.) Cohen has defined, with caution, an effect size of 0.2 as “small”, 0.5 as “medium”, and 0.8 as “large” [52]. We adopt these conventions here.

The Propensity Score Model for the Breastfeeding Duration

Until recently, propensity score methods have been limited to the binary treatment situation (e.g., single treatment and comparison group). Imbens, however, extends the method to the multi-group situation where treatment can have several levels (ordered or not) [53]. His work builds on the fact that propensity score matching essentially involves a weighting scheme [54]. Like the Horvitz-Thompson estimator [55], we can get an average effect by weighing observations by the inverse of the probability of treatment level (extent of breastfeeding) actually observed [53].

We follow Imbens’ procedure to examine the relationship between duration of breastfeeding and child outcomes. The propensity score for the four categories of duration is calculated from a multinomial logit model. The resulting coefficient estimates are used to calculate the propensity score for each duration category. Then we run ordinary regression that included three dummy variables on breastfeeding duration (the “never breastfeeding” category serving as the omitted group), with the inverse of the propensity score for each category as the weights.

Handling of Missing Data

The analysis relies on data collected at multiple interviews. This feature is a strength of the study—data are available on family income in the actual year the birth occurred. However, the rate of missing data is relatively high. Therefore, our analyses rely on multiple imputations (MI), which essentially created multiple datasets with filled-in estimates for the missing values. To maximize the use of information available, we include child level weights of 2002 and two other IQ measures collected from the 1972 PSID interview [56] as additional covariates in the imputation. We estimated four imputations and combined the estimates using Rubin’s rules [57].

The analyses were conducted using PSMATCH2 [58] as well as ICE and MICOMBINE commands [59] embedded in Stata SE 10 [60].

Except the descriptive statistics, all other results are pooled estimates across the four imputations using Rubin’s rule.
Results

Descriptive Statistics

Table 1 provides descriptive statistics for our study sample. The number of observations represents what is available in the data. It varies across the rows of the table—some variables were missing much more often than others. (Descriptive statistics for the data set after imputations are available from the authors.)

At baseline (the 1997 survey), the children were on average 89 months old (7.5 years old), the youngest being 4 months old, and the oldest 15 years old. Half were girls; 40% were African Americans, and 41% were the first child of the family (see Table 1). 8% of children were born prematurely, and 15% were born small for gestational age.

Also at baseline, one in seven mothers did not complete high school, and one-third were unmarried when the child was born. Mothers were on average 27 years old at the time of giving birth. A little more than one-third of them received Medicaid during the pregnancy and 43% of them participated in the WIC program. 46% of the mothers reported having ever breastfed their babies.

Unadjusted comparisons suggest that in terms of the outcomes and the sociodemographic characteristics, children who were breastfed have advantages over those who were not. Breastfed children had higher scores on WJ-R and WISC-R tests, were disproportionately white, less likely to be born early (6% vs. 10%) or born small for gestational age (10% vs. 20%), and had a more cognitively stimulating home environment. Furthermore, the mothers of breastfed children tend to be older, with higher IQs and higher levels of educational achievement. Additionally, mothers who breastfed, in the year of the child’s birth, had higher annual family incomes ($43,695 vs. $28,357), were more likely to be married (77% vs. 50%), and less likely to be enrolled in the WIC or Medicaid program.

Checking the Balance of Covariates

Figure 1 illustrates how the covariate balance is examined using one outcome, the letter word test. We plot the t statistics for comparing the group-means on every covariate. Any points outside of the region between the two vertical dashed lines are significant at $P = 0.05$ levels.

The unadjusted comparison of the covariates shows a number of covariates were significantly different between the two groups (see points for “Unmatched”). These covariates include, for example, family income at the year of child’s birth, mother’s education, WIC enrollment, and the proportion of mothers who were married at the time of the birth. On the other hand, after matching, all the t statistics fell within the non-significant region, indicating that all the covariates are now balanced. Covariate balance checking for other outcomes is available from the authors.

The Effects of Breastfeeding Initiation

Table 2 presents the estimated effects of breastfeeding initiation on child outcomes. The left panel is a simple unadjusted between-group comparison. The right panel compares children who were breastfed to their propensity-score-matched controls. For each model, the table presents the effect, the standard error, and the significance of the $P$-value. Chance findings with five outcomes are possible. Therefore, we only report significance levels at 0.01 or 0.001 level.

First consider the unmatched sample (the first column for each of the outcomes). As we have seen in the descriptive statistics, children who were breastfed, as compared to children who were not, had higher scores on intellectual tests when assessed in the 2002 survey. All these coefficients indicate a beneficial effect of breastfeeding, and they are all statistically significant. The differences, when expressed as effect sizes, are quite large: children who were breastfed had the applied problem scores that were .57 of a standard deviation higher than children who were not breastfed. All of the other scores, with the exception of the WISC, were at least a third of a standard deviation higher for breastfed children. To put these effects in context, these differences are larger than those reported, on the same measures of achievement, between poor and non-poor children [61]. Effect sizes of this magnitude strike us as implausibly large.

On the other hand, in the propensity-score-matched comparison, many of the effects shown in the unadjusted comparison are greatly reduced. For example, the effect on the passage comprehension test score is reduced from 6.83 to 2.18, that for the broad reading score being reduced from 8.26 to 3.58. After adjusting for multiple testing, three out of the five effects remain significant, the effect sizes are smaller and more plausible, with only one effect size (on applied problem) being larger than one-fifth of a standard deviation. All of these effect sizes are considered small under Cohen’s rules [52].

The Effect of Breastfeeding Duration

Table 3 describes the effects of breastfeeding duration. The table presents the mean outcome for each category, the $P$-value for comparing each of the other three categories with the reference category (never breastfeeding), and the effect size.

These results show a non-linear relationship between the outcomes and the breastfeeding duration. Compared to children who were never breastfed, children who were breastfed for 1–6 months had significantly higher scores on
Table 1: Descriptive statistics for the study sample

<table>
<thead>
<tr>
<th></th>
<th>All cases</th>
<th>Breasftfeeding</th>
<th>No breastfeeding</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obs⁰</td>
<td>Mean/ proportion</td>
<td>SD</td>
</tr>
<tr>
<td>Ever breastfeeding</td>
<td>3271</td>
<td>0.46</td>
<td>0.50</td>
</tr>
<tr>
<td>Length of breastfeeding (in months)</td>
<td>3271</td>
<td>3.7</td>
<td>6.21</td>
</tr>
</tbody>
</table>

### Outcomes measured in 2002

#### Woodcock Johnson test-revised

<table>
<thead>
<tr>
<th></th>
<th>Obs</th>
<th>Mean/ proportion</th>
<th>SD</th>
<th>Obs</th>
<th>Mean/ proportion</th>
<th>SD</th>
<th>Obs</th>
<th>Mean/ proportion</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Letter word</td>
<td>2444</td>
<td>103.9</td>
<td>19.0</td>
<td>1136</td>
<td>108.5</td>
<td>18.6</td>
<td>1308</td>
<td>99.9</td>
<td>18.4</td>
</tr>
<tr>
<td>Passage comprehension</td>
<td>2361</td>
<td>103.5</td>
<td>17.4</td>
<td>1099</td>
<td>107.2</td>
<td>17.8</td>
<td>1262</td>
<td>100.3</td>
<td>16.4</td>
</tr>
<tr>
<td>Applied problem</td>
<td>2436</td>
<td>102.3</td>
<td>17.0</td>
<td>1133</td>
<td>107.2</td>
<td>16.8</td>
<td>1303</td>
<td>98.0</td>
<td>16.0</td>
</tr>
<tr>
<td>Broadening</td>
<td>2452</td>
<td>99.6</td>
<td>27.1</td>
<td>1139</td>
<td>104.1</td>
<td>27.8</td>
<td>1313</td>
<td>95.8</td>
<td>25.9</td>
</tr>
<tr>
<td>WISC⁰ digit span test</td>
<td>2434</td>
<td>14.5</td>
<td>4.5</td>
<td>1131</td>
<td>14.7</td>
<td>4.6</td>
<td>1303</td>
<td>14.3</td>
<td>4.3</td>
</tr>
</tbody>
</table>

#### Covariates

<table>
<thead>
<tr>
<th></th>
<th>Obs</th>
<th>Mean/ proportion</th>
<th>SD</th>
<th>Obs</th>
<th>Mean/ proportion</th>
<th>SD</th>
<th>Obs</th>
<th>Mean/ proportion</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>3271</td>
<td>0.49</td>
<td>0.50</td>
<td>1512</td>
<td>0.50</td>
<td>0.50</td>
<td>1759</td>
<td>0.49</td>
<td>0.50</td>
</tr>
<tr>
<td>Age (months)</td>
<td>3271</td>
<td>89.3</td>
<td>46.0</td>
<td>1512</td>
<td>86.5</td>
<td>46.3</td>
<td>1759</td>
<td>91.8</td>
<td>45.5</td>
</tr>
<tr>
<td>Number of siblings</td>
<td>3225</td>
<td>1.4</td>
<td>1.2</td>
<td>1499</td>
<td>1.4</td>
<td>1.2</td>
<td>1726</td>
<td>1.4</td>
<td>1.2</td>
</tr>
<tr>
<td>First born</td>
<td>3198</td>
<td>0.41</td>
<td>0.49</td>
<td>1477</td>
<td>0.42</td>
<td>0.49</td>
<td>1721</td>
<td>0.39</td>
<td>0.49</td>
</tr>
<tr>
<td>Black</td>
<td>3271</td>
<td>0.40</td>
<td>0.49</td>
<td>1512</td>
<td>0.20</td>
<td>0.40</td>
<td>1759</td>
<td>0.57</td>
<td>0.50</td>
</tr>
<tr>
<td>White</td>
<td>3271</td>
<td>0.47</td>
<td>0.50</td>
<td>1512</td>
<td>0.62</td>
<td>0.48</td>
<td>1759</td>
<td>0.34</td>
<td>0.47</td>
</tr>
<tr>
<td>Other race/ethnicity group</td>
<td>3271</td>
<td>0.13</td>
<td>0.34</td>
<td>1512</td>
<td>0.18</td>
<td>0.38</td>
<td>1759</td>
<td>0.09</td>
<td>0.28</td>
</tr>
<tr>
<td>HOME measure⁰</td>
<td>3271</td>
<td>17.8</td>
<td>3.1</td>
<td>1512</td>
<td>18.2</td>
<td>3.1</td>
<td>1759</td>
<td>17.4</td>
<td>3.0</td>
</tr>
<tr>
<td>Born preterm</td>
<td>3270</td>
<td>0.08</td>
<td>0.28</td>
<td>1511</td>
<td>0.06</td>
<td>0.24</td>
<td>1759</td>
<td>0.10</td>
<td>0.30</td>
</tr>
<tr>
<td>Born small for gestational age</td>
<td>3271</td>
<td>0.15</td>
<td>0.36</td>
<td>1512</td>
<td>0.10</td>
<td>0.30</td>
<td>1759</td>
<td>0.20</td>
<td>0.40</td>
</tr>
<tr>
<td>Child’s health status at birth⁰</td>
<td>3260</td>
<td>2.21</td>
<td>0.57</td>
<td>1508</td>
<td>2.29</td>
<td>0.59</td>
<td>1752</td>
<td>2.13</td>
<td>0.55</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Obs</th>
<th>Mean/ proportion</th>
<th>SD</th>
<th>Obs</th>
<th>Mean/ proportion</th>
<th>SD</th>
<th>Obs</th>
<th>Mean/ proportion</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mother (at 1997 survey)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WIC participation</td>
<td>3180</td>
<td>0.43</td>
<td>0.50</td>
<td>1442</td>
<td>0.29</td>
<td>0.45</td>
<td>1738</td>
<td>0.55</td>
<td>0.50</td>
</tr>
<tr>
<td>Mom’s age giving birth (years)-20</td>
<td>3270</td>
<td>7.3</td>
<td>5.8</td>
<td>1512</td>
<td>8.5</td>
<td>5.6</td>
<td>1758</td>
<td>6.4</td>
<td>5.8</td>
</tr>
<tr>
<td>IQ</td>
<td>2523</td>
<td>30.7</td>
<td>5.4</td>
<td>1195</td>
<td>32.5</td>
<td>5.0</td>
<td>1328</td>
<td>29.0</td>
<td>5.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Obs</th>
<th>Mean/ proportion</th>
<th>SD</th>
<th>Obs</th>
<th>Mean/ proportion</th>
<th>SD</th>
<th>Obs</th>
<th>Mean/ proportion</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-school dropout</td>
<td>3271</td>
<td>0.14</td>
<td>0.35</td>
<td>1512</td>
<td>0.08</td>
<td>0.26</td>
<td>1759</td>
<td>0.20</td>
<td>0.40</td>
</tr>
<tr>
<td>High school graduate</td>
<td>3271</td>
<td>0.35</td>
<td>0.48</td>
<td>1512</td>
<td>0.27</td>
<td>0.44</td>
<td>1759</td>
<td>0.42</td>
<td>0.49</td>
</tr>
<tr>
<td>College or more</td>
<td>3271</td>
<td>0.38</td>
<td>0.49</td>
<td>1512</td>
<td>0.49</td>
<td>0.50</td>
<td>1759</td>
<td>0.28</td>
<td>0.45</td>
</tr>
<tr>
<td>Missing education</td>
<td>3271</td>
<td>0.13</td>
<td>0.33</td>
<td>1512</td>
<td>0.17</td>
<td>0.37</td>
<td>1759</td>
<td>0.09</td>
<td>0.29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Obs</th>
<th>Mean/ proportion</th>
<th>SD</th>
<th>Obs</th>
<th>Mean/ proportion</th>
<th>SD</th>
<th>Obs</th>
<th>Mean/ proportion</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mother (year of child birth)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family income ($)</td>
<td>2751</td>
<td>35176</td>
<td>36171</td>
<td>1223</td>
<td>43695</td>
<td>35789</td>
<td>1528</td>
<td>28357</td>
<td>26014</td>
</tr>
<tr>
<td>Married</td>
<td>3271</td>
<td>0.63</td>
<td>0.48</td>
<td>1512</td>
<td>0.77</td>
<td>0.42</td>
<td>1759</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>Employed</td>
<td>2527</td>
<td>0.47</td>
<td>0.50</td>
<td>1150</td>
<td>0.52</td>
<td>0.50</td>
<td>1377</td>
<td>0.43</td>
<td>0.50</td>
</tr>
<tr>
<td>Head of household⁰</td>
<td>2603</td>
<td>0.15</td>
<td>0.36</td>
<td>1179</td>
<td>0.08</td>
<td>0.27</td>
<td>1424</td>
<td>0.21</td>
<td>0.41</td>
</tr>
<tr>
<td>Self-rated health status⁰</td>
<td>2522</td>
<td>4.1</td>
<td>0.9</td>
<td>1164</td>
<td>4.2</td>
<td>0.9</td>
<td>1358</td>
<td>4.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Received medicaid during pregnancy</td>
<td>3175</td>
<td>0.35</td>
<td>0.48</td>
<td>1439</td>
<td>0.24</td>
<td>0.43</td>
<td>1736</td>
<td>0.44</td>
<td>0.50</td>
</tr>
</tbody>
</table>

⁰ Number of observations with non-missing values

⁰⁰ Wechsler Intelligence Scale for Children-Revised

⁰¹ Home Observation for Measurement of the Environment

⁰² / Worse, 2 Same, 3 Better (as compared with other babies born at birth)

⁰³ Indicates that mother was the only adult in the household

⁰⁴ / Poor, 2 Fair, 3 Good, 4 Very good, 5 Excellent
the applied problem tests; for the other outcomes or durations, there is no significant effect found. Moreover, for both passage comprehension and broad reading, the children who were never breastfed compare favorably to those who were breastfed for more than 12 months. For example, for children who were not breastfed had average passage comprehension scores of 98.1; on the same measure, children who were breastfed for more than a year had average scores of 96.7.

These results are meaningful because they suggest that longer spells of breastfeeding are uncorrelated with increases in the measures of achievement. Therefore, for breastfeeding to be causally related to higher cognitive scores, it must be the case that the cognitive benefits of breastfeeding arise only if the child is breastfed within the first 6 months of life. While this explanation is plausible—there may be a window of opportunity when breastfeeding can do the most good—these results are also consistent with an explanation in which positive selection into breastfeeding explains the higher scores.

Sensitivity Analysis

We address several limitations of our analyses and methodology. First, bias is possible because of the long lag between birth and the survey. Such measurement error involving breastfeeding would bias estimates downward [62]. We limit our analysis of the effect of breastfeeding initiation to children who aged 7 or younger when the mother was interviewed in 1997. Two outcomes, the letter word and the applied problem test scores remain significant after the matching. But the effect sizes (0.24 and 0.30, respectively) are still modest. This result is available from the first author.
Table 3 Effects of breastfeeding duration analysis

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Sample size</th>
<th>Propensity score adjusted with multiple imputations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Never</td>
</tr>
<tr>
<td>Woodcock Johnson test-revised</td>
<td>1759</td>
<td>895</td>
</tr>
<tr>
<td>Letter word</td>
<td>Mean</td>
<td>102.67</td>
</tr>
<tr>
<td></td>
<td>Effect size*</td>
<td>0.13</td>
</tr>
<tr>
<td>Passage comprehension</td>
<td>Mean</td>
<td>98.07</td>
</tr>
<tr>
<td></td>
<td>Effect size</td>
<td>0.04</td>
</tr>
<tr>
<td>Applied problem</td>
<td>Mean</td>
<td>100.81</td>
</tr>
<tr>
<td></td>
<td>Effect size</td>
<td>0.18</td>
</tr>
<tr>
<td>Broad reading</td>
<td>Mean</td>
<td>99.06</td>
</tr>
<tr>
<td></td>
<td>Effect size</td>
<td>0.05</td>
</tr>
<tr>
<td>WISC™ digit span test</td>
<td>Mean</td>
<td>14.40</td>
</tr>
<tr>
<td></td>
<td>Effect size</td>
<td>0.02</td>
</tr>
</tbody>
</table>

*a Effect size represents the difference in means divided by the standard deviation for the “never” group
*b Wechsler Intelligence Scale for Children-Revised
*c P-values refer to the null hypothesis that the mean for this level of breastfeeding is equivalent to that for the “never” category. * Significant at 5%; ** significant at 1%; *** significant at 0.1%

Second, we use alternative cutoffs for breastfeeding durations. For example, we divide the duration into less than 3 months, 3–12 months, and more than 12 months, and found a significant increase in the applied problem scores for the less-than-3-month category. But none of the other categories shows statistically significant effects. The relationship across duration is still non-linear, and all the effect sizes are less than 0.20. This result is available from the first author.

Third, both regression and the propensity score matching assume that unobserved confounding does not exist [50]. We examine the potential effects of such confounding if they were to exist in a regression framework. Suppose the true relationship between the child cognition and breastfeeding is as follows:

\[ Y_i = \beta E_i + \gamma H_i + \epsilon_i \]

\( Y_i \) is the child's cognitive score, \( E_i \) is the child's breastfeeding status, \( H \) is an unobserved confounder beyond all the covariates we have included in the analysis, \( \beta \) is the (true) effect of breastfeeding on the cognitive score. The resulting estimate of \( \beta (\hat{\beta}) \) is the sum of the true treatment effect and some effect caused by the unobserved confounder:

\[ \hat{\beta} = \beta + \frac{\delta H_i}{\delta E_i} = \beta + \frac{\gamma}{\epsilon} \]

The bias depends on the relationship between \( H \) and the outcome (\( \gamma \)) and between \( H \) and the treatment (\( 1/\epsilon \)). Note that this exercise does not involve actually generating \( H \). The intent of the analysis is to quantify a qualitative concern, “how large would any remaining bias have to be to change the study's key figures?”

Following the literature, we assume \( H \) is dichotomous, and for convenience it indicates whether people are healthy or not (over and above the covariates we have included). For further convenience, we assume this unobserved confounder predicts the treatment perfectly. In our case, all healthy women would breastfeed their children while all unhealthy women would not. In essence, \( P(E = 1 \mid H = 1) = P(E = 0 \mid H = 0) = 1 \), and this effectively sets \( \epsilon \) to 1. Then the remaining bias \( \gamma \) strictly a function of the link between \( H \) and the outcome, and the effect of this bias is expressed in the same metric as the effect of the exposure/treatment.

How big would \( \gamma \) have to be to change the interpretation of our findings? In our case, the key issue is whether the bias would make a relationship appear significant when it is not. In effect, the question is whether the bias pulled the observed t-statistic more than \( t_{\text{est}} = 1.96 \) standard errors away from zero. For a positive true effect, for example,

\[ \gamma_{\text{critical}} = \hat{\beta} - 1.96 \times \hat{\epsilon} \]

\( \gamma_{\text{critical}} \) is the smallest value of gamma to produce an estimate that is just statistically significant at 0.05 level (while in fact the true effect is not significant).

One can see that the size of this bias depends on both the magnitude of the estimated effect as well as its precision. A larger, more precise effect will be more robust to possible unobserved confounding.

Using the standard normal distribution and the estimated standard error, we can calculate the critical value of gamma...
above. For example, in the propensity-score-matched sample, the estimated effect of breastfeeding initiation on child’s letter word test is 3.43 with a standard error of 1.05 (see Table 2). The critical value of gamma for a significant effect with the above standard error is 3.43 - 1.96 x 1.05 = 1.38. So the unobserved confounding would have to have an absolute effect of 1.38 or larger to change the statistical significance of our result. This estimate is about 40% of the observed effect and suggests that the unobserved confounding would have to be rather large. Note that this effect has to exist over and above the balancing of all the observed characteristics. Though there is no way to know for sure (as with any sensitivity analysis), it appears our finding is rather robust to unobserved confounding. On the other hand, the magnitude of bias for the WISC test needs to be as small as 0.01 (or 1% of the observed effect) to flip the significance of the result. In general, these findings suggest that the two strongest relationships in the table are unlikely to be caused by modest confounding. (The full set of estimates is available from the first author.)

Discussion

Previous research has indicated that in unadjusted comparisons with children who are formula fed, children who are breastfed score higher on measures of cognitive functioning [16, 17, 63, 64]. The reasons for this difference, however, are unclear. Breast milk contains nutrients that promote optimum cognitive development; alternatively, the benefits could reflect not breastfeeding but the characteristics of the mothers who breastfeed. The distinction is the key: in the latter case, promoting breastfeeding will not confer those advantages to the women and their children.

Our analyses suggest that adjusting for observed confounding substantially reduces the apparent benefits of breastfeeding initiation. Although the effects remain statistically significant for at least two of the outcomes, the adjusted difference between breastfed and non-breastfed children translates into an effect size between one-tenth and one-fifth of a standard deviation, an effect which is small according to Cohen's effect size rules.

More tellingly, we did not find that children who were breastfed for longer periods of time had higher cognitive scores than children who were never breastfed. In the duration analyses, positive effects on the outcomes were found for children who were breastfed less than six months when compared to children who were not breastfed. But no beneficial effects of breastfeeding was found when a child was breastfed for more than six months. Moreover, children who were never breastfed compare favorably on some outcomes to those who were breastfed for more than 12 months (e.g., passage comprehension and broad reading test). These results suggest that there may be some critical window (e.g., the first six months of life) in early infant development during which breastfeeding can affect cognitive outcomes; after that window has closed, breastfeeding is not associated with better cognitive scores. This pattern also hints at unobserved differences—perhaps the women who start but then discontinue breastfeeding are disadvantaged in ways not captured here.

Our findings do suffer from limitations. First, data on breastfeeding practice are based on maternal recall. We address this issue in several ways. First, we compare the prevalence of breastfeeding initiation in this sample to what was reported in the National Health and Nutrition Examination Survey during the same period (NHANES 1988–1994) [65] and found them comparable. For example, the overall breastfeeding prevalence, that for non-Hispanic white, and that for non-Hispanic black was 53.6, 60.2, and 26.3% for NHANES sample. [66] The corresponding prevalence was 46, 61.0, and 22.7% for the PSID sample. The amount of breastfeeding was also close in the two samples. Overall, 22.4 and 8.9% of the NHANES III sample were still breastfeeding at 6 and 12 months [66]. Those for the PSID sample were 23.9 and 10.9%, respectively. The prevalence for premature and full-term babies was similar among the two samples, too. This shows two things. First, our sample was compatible with what might be expected for a nationally representative sample. Second, these comparisons suggest that recall problems may be rather small. This consistency with other figures is reassuring.

In addition, when we restrict our sample to younger children in the supplemental analysis, we find similar modest effect. Furthermore, for maternal recall to influence the estimated effect of breastfeeding, the relationship between actual and reported breastfeeding practice would have to be moderated by the outcomes of interest. We cannot rule out selective recall bias, but it is hard to imagine how that might be related to the child outcomes.

Another limitation is that our data lack a measure of breastfeeding exclusivity. According to NHANES III data, 9.5% of the children were still exclusively breastfeeding at the age of 6 months [66]. Other studies focus on the period during which breast milk is not supplemented with other, solid food [67, 68]. The effects presented here may mask larger benefits for the sub-group of participants who breastfeed exclusively. However, as an empirical matter, the impact of exclusivity is difficult to assess in observational data. It is reasonable to assume that those who breastfeed exclusively are not a representative sub-sample of those who breastfeed at all; in that case, the potential for confounding is likely even greater.

One can bound the potential effect of exclusivity by thinking of it as a form of compliance. Suppose that non-exclusive breastfeeding has no effect whatsoever. Roughly
40% of babies that are ever breastfed are breastfed through the first three months of life [69]. In that case, the effect we find here should be multiplied by 2.5 (or 1/0.40). Multiplying a small estimate still produces a rather modest effect [70]. (This effect is the so-called compliance-adjusted effect.) Note that as we shrink the size of the group for which breastfeeding may be effective (and thereby inflate the effect of breastfeeding), the public health implications are diminished.

Finally, our study suffers from the limitations inherent to observational studies. However, the putative gold standard of evaluation—a randomized trial—is not possible or even desirable for breastfeeding research. A body of research considers the circumstances under which observational analyses produces findings similar to a randomized experiment [71, 72]. That literature suggests that a well done observational study can produce findings quite similar to those of a randomized trial (where both exist). In terms of addressing selection bias, regression has limitations in their underlying assumptions and the number of variables they can match. Propensity score matching relaxes the linear assumption of the regression and simplifies the matching variables into a single predicted probability.

We do have a key assumption underlying the propensity score model, that is, there is no unmeasured confounder. However, we include a wide range of covariates and at least by the standard or prior research, our list is comprehensive. In addition, we estimate how large the potential bias due to unobserved confounding has to be to change the statistical significance of our results. This sensitivity analysis reveals that for some outcomes, such as the letter word and the applied problem tests, our results are rather robust; while for other outcomes like the passage comprehension and the WISC tests, our results are sensitive. Note that for the two outcomes that are sensitive, they are only significant at $P = 0.05$ level (instead of at 0.01 or 0.001 level). This links back to the issue of multiple testing. When using adjustment such as Bonferroni correction, these two outcomes are no longer statistically significant. Therefore, the effects of these two tests could indeed be explained by some unobserved confounding.

Do these findings mean that breastfeeding should not be encouraged? Of course not. The health benefits of breastfeeding for both the child and the mother have been well demonstrated [1]. On the other hand, women who did not breastfeed their babies (either because they cannot or they do not want to) should not be made to feel that they are damaging the cognitive potential of their children. Such guilt may be more damaging to the child, especially if it interferes with attachment and/or contributes to maternal depression [73, 74]. Nor should the benefits of breastfeeding be overstated, given the barriers to breastfeeding [75], which remain formidable in the workplace and other contexts.

Furthermore, from a broader public health perspective, the practical importance of breastfeeding on cognitive development may be small. Other factors, such as poverty, low birth weight, or maternal risky behaviors, may be more important in determining a child’s intellectual well-being [61, 76, 77]. If society wishes to increase the cognitive outcomes of at-risk children, it may be more sensible to direct attention to factors that have a larger impact on child outcomes.

Acknowledgments. This study was not funded by any governmental or non-governmental agency.

References


60. Stata statistical software: Release 10 [program]. (2007). College Station, TX.