Can the National Health and Nutrition Examination Survey III (NHANES III) data help resolve the controversy over low blood lead levels and neuropsychological development in children?

Brice M. Stone\textsuperscript{a}, Cecil R. Reynolds\textsuperscript{b,*}

\textsuperscript{a}Brice Stone Consulting and Metrica, Inc., San Antonio, TX 78217-6363, USA
\textsuperscript{b}Department of Educational Psychology, Bastrop Mental Health Associates, Texas A&M University, College Station, TX 77843-4225, USA

Accepted 31 October 2002

Abstract

The National Health and Nutrition Examination Survey III (NHANES III) was designed to provide national estimates of the health and nutritional status of the United States population aged 2 months and above. A Youth data subset includes individuals from ages 2 months to 16 years totaling 13,944 individuals. Lanphear, Dietrich, Auinger, and Cox (2000) examined these data and concluded that deficits in cognitive and academic skills associated with lead exposure occur at blood lead concentrations of less than 5 µg/dl. Attempts to replicate and extend these findings reveal serious shortcomings in the NHANES III data that center around missing data, odd distributions of blood lead levels as well as cognitive and academic scores, and potential inaccuracies in the data collection itself. A review of these issues is presented along with a series of empirical analyses of the data under multiple sets of assumptions leading to the conclusion that the NHANES III data are inherently inadequate for use in addressing neurodevelopmental issues. Policy issues and scientific issues related to cognitive and other neurodevelopmental phenomena should not be considered on the basis of the NHANES III Youth dataset.

Keywords: Intelligence; Cognition; Lead; Methodology

Lead is a heavy metal that is ubiquitous in soil throughout virtually all regions of our planet. It has had a variety of industrial and commercial uses over the years, perhaps best known as...
an additive to early gasoline mixtures, paint, solder, and even cooking utensils. Lead is taken up by the human body, primarily through inhalation or ingestion, and is stored primarily in bone tissue. Lead has no known biological benefit to living organisms (e.g., Hartman, 1995) and at some concentration in living organisms clearly becomes a neurotoxicant and at very high levels leads to seizures, cortical lesions, and death. There is considerable consensus of opinion among scientists, clinicians, and public health officials regarding the significance and impact of markedly elevated blood lead levels (BLLs), meaning those of 40 µg/dl and higher and reasonable consensus at 20 µg/dl and above (e.g., Juberg, 2000).

Briefly, the Centers for Disease Control (CDC), recommends education and follow-up testing when BLLs of 10–19 µg/dl are present, with clinical intervention directed for BLLs of 20 µg/dl and higher. At BLLs of 45 µg/dl and above, more extensive care and follow-up are necessary, with levels above 70 µg/dl representing the need for urgent care. There have been great controversies over the association and effects of BLLs of 10–19 µg/dl (e.g., Brown, 2001; Hebben, 2001; Juberg, 2000; Kaufman, 2001a, 2001b; Nation & Gleaves, 2001; Needleman & Bellinger, 2001; Wasserman & Factor-Litvak, 2001). At the extremes of these disagreements appear to be Kaufman (2001a) and Needleman and Bellinger (2001). Lanphear, Dietrich, Auinger, and Cox (2000) have created additional controversy. In their research, using the National Health and Nutrition Examination Survey III (NHANES III) Youth data, they conclude there are detrimental effects of lead on the neuropsychological development of children at blood levels of less than 5 µg/dl. The average BLL of young children (12–24 months) in the United States is 3.1 and for the overall population it is about 2.9 (Juberg, 2000). If Lanphear et al. (2000) are correct in their assertions, there are not only clinical implications for children and the practitioners who see them, but also potentially massive implications for public health policy and governmental regulation of lead sources and applications throughout society. The health and the economic issues and ramifications are enormous. And, Congress and the CDC are already considering changes to law and possibly public policy based upon these results.

Lanphear et al. (2000) used the NHANES III Youth data as the basis for their conduct of a series of what essentially are variants of multiple regression using a limited set of control variables (e.g., poverty index ratio, gender, ethnicity, serum ferritin, and serum cotinine, the latter being a metabolite and thus, a marker variable for exposure to tobacco smoke) and many variables with a large number of missing values. As we will note in more detail later, the Lanphear control variables are rather limited relative to other studies in the literature and fail to include potentially important variables in the NHANES III Youth dataset. Using age-corrected scaled scores on four cognitive measures (two from the Wechsler Intelligence Scale for Children—Revised; WISC-R, Wechsler, 1974; and the reading and Arithmetic subtests of the Wide Range Achievement Test—Revised; WRAT-R, Jastak & Wilkinson, 1984), Lanphear et al. report statistically significant, albeit small, associations with BLLs ranging below 5 µg/dl. However, the NHANES III data may not be appropriate for the design of the Lanphear et al. study and additional covariates may be required. A variety of assumptions also must be made about the very complex sampling design of the NHANES III cohort that does not appear to hold when the sample is examined.

This paper presents a review of these issues and a variety of new analyses of the NHANES III data that question its utility for evaluating complex, low incidence events and what are potentially small effects. To understand many of these issues, it is necessary first to review the
NHANES III dataset and how it was and may be used in examining the association of lead to cognitive dysfunction or related neurodevelopmental deficits in children.

1. The NHANES III data

The National Health and Nutrition Examination Survey (NHANES) is a periodic survey conducted by the National Center for Health Statistics (NCHS). **The Third National Health and Nutrition Examination Survey (NHANES III), conducted from 1988 through 1994, was the seventh in a series of these surveys based on a complex, multi-stage sample plan. It was designed to provide nationally representative estimates of the health and nutritional status of the United States’ civilian, noninstitutionalized population aged 2 months and older. Clinical examinations including various laboratory tests were conducted along with some neuropsychological testing and extensive interviews using a standardized survey developed for NHANES III.**

A four-stage sample design was used: (1) Primary Sampling Units (PSUs) comprising mostly single counties, (2) area segments within PSUs, (3) households within area segments, and (4) persons within households. The PSUs in the first stage were mostly individual counties; in a few cases, adjacent counties were combined to keep PSUs above a certain minimum size. There were 81 PSUs in the sample, selected with probability proportionate to measures of size and without replacement. The measure of size reflected the desire to oversample the minority groups in NHANES III. The area segments were stratified by percent Mexican American prior to sample selection. The subsampling rates were set to produce a national, approximately equal probability sample of households in most of the United States, with higher rates for the geographic strata with high minority concentrations.

The NHANES III Household Youth Questionnaire Data File contains all data collected during the household interviews for children and youths 2 months to 16 years of age. Demographic data, survey design variables, and sampling weights are also included for this age group. The sampling plan was designed to be overinclusive of low income and minority groups, however, weights were determined by NCHS to mimic a population proportionate representative sample.

The data that comprise this file were obtained from three separate interviews administered in the household: the Screener, the Family, and Household Youth questionnaires. The Screener was a brief interview administered to all households within selected sampling units to determine eligibility for participation in the survey. Information obtained from the Screener questionnaire was the basis for the demographic variables, and survey-related variables. The purpose of the Family questionnaire was to obtain data on educational levels, occupation, health insurance coverage, income, and food security of family members, and to record characteristics of the house itself.

The Household Youth questionnaire was administered to a proxy respondent, usually the child’s parent or guardian. Questions on topics such as birth experience, motor and social development, health services usage, dental care, selected medical conditions, school attendance, language, and vitamin, mineral, and medicine usage were asked according to age subgroups. The Household Youth questionnaire was completed for 13,944 children and youths aged 2 months to 16 years of age during the 6 years of NHANES III. Interviews were conducted in
both English and Spanish. Laboratory data were collected for a number of variables including serum lead, cotinine, and ferritin, among others.

NHANES also collected data regarding cognitive function using parts of two tests, the WISC-R and the WRAT-R. Two subtests of the WISC-R, a verbal component (Digit Span) and a performance exam (Block Design), were administered. Two subtests of the WRAT-R, math and reading, were conducted. The scores for all four subtests were derived for each child relative to his/her age group based on test-specific standardization samples created by the test developers, for example, age-corrected deviation scaled scores were used (Jastak & Wilkinson, 1984; Kramer, Allen, & Gergen, 1995; Wechsler, 1974). The WRAT math and reading scores were standardized to a mean of 100 and standard deviation of 15, while the WISC-R Block Design and Digit Span subtests' scores were standardized to a mean of 10 and standard deviation of 3.

Since NHANES III was based on a complex multi-stage sampling, designed to over-represent certain groups (see above), appropriate sample weights should be used in analyses to produce national estimates and associated variances. The final interview weights (designated as WTPFQX6) should be used for analysis of interview data alone. If interview data are combined with examination, laboratory, or dietary data, the total mobile examination center or MEC weights (WTPFEX6) or MEC + homeweights (WTPFHX6) must be used.

The Lanphear et al. (2000) study used children from NHANES III aged 6–16 years who had values for their BLLs (PBP, μg/dl), numbering 4,853. Of these 4,853, 13.33% (647) exhibited BLLs below a detectable level. The undetected BLLs were assigned a value of 0.7 by NHANES, which Lanphear et al. subsequently used in their analyses. Assigning such a constant value to such a large percentage of the sample truncates the true score distribution. The demographic characteristics of this portion of the sample suggest it is not a random group as it differs considerably from those youth with detectable BLLs on many of the demographic characteristics. This has the potential to bias the results in undetectable ways and create problems with the assumptions underlying many of the statistical analyses that would otherwise be useful with the sample and that were in fact used by Lanphear et al. In addition, not all of the 4,853 subjects had valid cognitive test scores.

The integrity of the data collected and the representativeness of the sample drawn for NHANES III are crucial to the validity of any reported results. If the data are collected in ways that produce inaccuracies, internal validity of any studies will be destroyed. To the extent the sample weights do not create a dataset that is representative of the target population (the youth population of the United States, aged 2 months to 16 years), external validity will be lacking. In either case, generalization of results outside the confines of the study sample would be inappropriate.

2. Is the NHANES III weighted laboratory sample representative of the youth population of the United States?

A representative sample should produce test scores and other laboratory data that mimic the population values for youth in the United States. A representative sample should produce accurate parameter estimates for all relevant variables and should also mimic the demographic characteristics of the target population, in this instance, American youth between the ages
Table 1
Mean scores on NHANES III cognitive variables for Lanphear et al. (2000) sample

| Group or subgroup                      | WRAT-R variable | WISC-R variable |
|----------------------------------------|-----------------|-----------------|
|                                        | Arithmetic      | Reading         | Block Design | Digit Span |
| Total, N (4,853)                       | 93.5            | 91.9            | 9.5          | 8.7        |
| African American                       | 86.1            | 91.9            | 7.2          | 8.0        |
| Hispanic                               | 86.9            | 85.4            | 8.8          | 7.5        |
| White                                  | 96.4            | 95.1            | 10.2         | 9.1        |
| Poverty index, lowest tercile          | 86.9            | 84.2            | 8.3          | 8.0        |
| Poverty index, middle tercile          | 93.1            | 91.8            | 9.4          | 8.6        |
| Poverty index, highest tercile         | 100.4           | 89.7            | 10.7         | 9.5        |

| Education level of parent or caregiver | WRAT-R variable | WISC-R variable |
|----------------------------------------|-----------------|-----------------|
| Less than HS graduation                 | 85.1            | 83.4            | 8.2          | 7.7        |
| HS graduation                          | 92.0            | 90.9            | 9.2          | 8.5        |
| More than HS graduation                 | 89.7            | 97.6            | 10.5         | 9.4        |

Note. WRAT-R variables are set to a mean of 100 and S.D. of 15 and WISC-R variables are set to a mean of 10 and S.D. of 3 in their respective standardization samples.

of 2 months and 16 years of age. If cell sizes are large enough, a nonrepresentative sample can be weighted to mimic the target population characteristics and weights were derived by the original investigators for the NHANES III data that were intended to ensure it would be representative of the target population. Samples that are nonrepresentative produce results that lack external validity, that is; the results cannot be generalized outside of the sample in which they were obtained. Several of the findings reported by Lanphear et al. (2000) suggest problems either with the collection of the cognitive data, with the sampling design, the MEC weights, or some combination of these problems.

In Table 2 of their paper Lanphear et al. report MEC weighted mean scores for their total sample and for various subgroups. Table 1 provides a summary of some of these data and is abstracted from Lanphear et al. (2000) (but were independently replicated by our own analyses). These values reflect a series of oddities.

The WRAT-R standardization data were collected in the early 1980s, and the WISC-R standardization data were collected between about 1970 and 1972. The mean score levels on the tests gradually crept upward, one of the factors leading to the development of the third edition of each test, the WISC-III (Wechsler, 1991) and the WRAT-III (Wilkinson, 1993). Gradual increases in cognitive test scores are a well-documented phenomenon (e.g., Kamphaus, 2001) commonly referred to as the Flynn Effect (e.g., see Flynn, 1998, for a review and discussion) and has been observed over the last 100 years of research on individual differences. For example, by 1990, the WISC-R Block Design mean score had risen from 10 to 11.3, nearly 0.5 S.D. above the original estimate of the population mean of 10. Digit Span, as is characteristic of verbal tests, showed a much smaller increase with a mean of 10.4 (Wechsler, 1991) versus 10. Parent educational level and socioeconomic status (SES) also are associated with scores on such tests, with higher educational levels and higher SES associated with scores considerably above the population mean in many instances (e.g., see Kaufman, 1979; Kaufman, McLean, & Reynolds, 1988; Reynolds, Chastain, Kaufman, & McLean, 1987).
The mean scores on the cognitive variables in the NHANES III sample do not conform to known patterns of performance on these variables. For the total sample, the discrepancies from the original standardization sample means are as follows: Arithmetic, $-0.46$ S.D.; Reading, $-0.54$ S.D.; Block Design, $-0.17$ S.D.; and, Digit Span, $-0.43$ S.D. The mean discrepancy in standard deviation units is $-0.4$ S.D. Although this effect size is many times larger than that reported by Lanphear et al. (2000) in their analyses of the association between BLLs and these scores, the NCHS, in reply to a query about these values, characterized them as being “...slightly below the mean...” (NCHSED, personal communication of 1-31-02). Of note, the average effect size here of $-0.4$ S.D. across these four cognitive variables is almost certainly an underestimate due to the Flynn Effect. If one uses the 1990 estimate of mean levels of performance on Block Design of 11.3 (Wechsler, 1991), the discrepancy becomes $-0.60$ S.D., nearly two thirds of a standard deviation.

However, even more striking disparities from well-known patterns appear. According to the NHANES III MEC weighted sample statistics, the mean arithmetic performance of children whose parents or caregivers have a level of education that includes post-secondary education is 0.69 S.D. below the population mean. Their reading and Digit Span scores also fall below the population mean, even without considering the Flynn Effect, which would widen this gap. The arithmetic scores of the less educated group are higher, not lower, than the group with the highest educational level. If one reviews the means for the poverty index ratio classes, similar problems appear. The highest tercile scores 0.69 S.D. below the population mean in reading and below the scores of the middle tercile. Within-group discrepancies also appear that are unexplainable. For example, on the WRAT-R Arithmetic scale, African Americans score nearly 0.40 S.D. below their performance on the reading scale, a pattern contrary to other literature (e.g., see Reynolds, 1982). Such patterns suggest an odd sample or errors in administration or scoring of the tests. Other results suggest similar problems with the data.

Each of the four cognitive measures used in NHANES III is reported in age-corrected deviation scaled scores, which in the general population have a zero correlation with age (e.g., see Reynolds, 1998). Using the MEC sample weights, the Pearson correlation between age and each of the scaled scores for the four cognitive variables was:

| Test         | Correlation | $P$  |
|--------------|-------------|------|
| Arithmetic   | $-0.102$    | .0001|
| Reading      | $0.127$     | .0001|
| Block Design | $-0.064$    | .0001|
| Digit Span   | $-0.065$    | .0001|

These correlations are indeed small in magnitude but are highly significant when their value is tested against a predicted value of zero. The correlations between age and raw scores should be quite high for these variables. Raw scores are reported only for Arithmetic and Reading and the age-raw score $r$ values are .566 and .184 respectively. Note that the correlation with age and Reading raw score is near the correlation with the age-corrected scaled scores. The inconsistency of the sign (direction) of the correlation for reading is also of concern.
In response to a question on this issue (to nchsquery@cdc.gov), NCHS suggested five potential reasons for these findings. The NCHS suggested explanations (NCHSED, personal communication of 1-31-02) are given below in quotes with our comments.

2.1. NCHS1

“Temporal changes in performance on standardized tests.” As we have noted, the Flynn Effect moves the means upwards, thus, biasing against the effects seen in this sample. The known direction and magnitude of temporal changes in performance on standardized tests indicate the sample differences we have noted above actually underestimate the problem.

2.2. NCHS2

“Differences in testing environment (perhaps the MECs were a more stressful or not as serious a testing situation as the one done by the WISC and WRAT companies [sic]).” If this is the reason, it could indeed be a serious problem but the effect appears not to have been random but systematically biasing downward but also not in equal increments across relevant demographic variables (e.g., see Table 1). The latter is important because BLLs are also associated with some of the demographic variables. Examiners were likely blind to BLLs but not to the social or related demographic status of the examinees and examiner biases may be present as an additional confound, especially given that neither professional psychologists nor educational diagnosticians were used to conduct the cognitive testing.

2.3. NCHS3

“... The types of people doing the testing could also have had an effect (NCHS did not use a psychologist, whereas the other companies probably did).” Kaufman (2001a, 2001b) has criticized the use of poorly trained or unsupervised individuals in the collection of intelligence test and related data although Needleman and Bellinger (2001) defend the practice as acceptable. Performance on intelligence tests is highly sensitive to changes in administration of the tests and can introduce biasing effects (e.g., see Lee, Reynolds, & Willson, 2002). The substantial discrepancies present here in obtained data versus population estimates argue against such practices and support Kaufman’s (2001a, 2001b) view of this issue.

2.4. NCHS4

“Bias/error in how representative [the original WISC-R and WRAT-R] the samples were.” The standardization sample of the WISC-R essentially set the modern standard in the psychological testing industry and is lauded by many as exemplary (e.g., Cronbach, 1990; Kamphaus, 2001; Kaufman, 1979; Sattler, 2001). If these samples are faulty as NCHS suggests and are in fact nonrepresentative of the population of children in the United States at the time they were gathered, hundreds of thousands of children have been misdiagnosed and misplaced. Such suggestions/allegations should not be made lightly. In the absence of evidence and analysis to the contrary, the WISC-R and WRAT-R samples remain persuasive to us.
2.5. NCHS5

“Differences in the race/ethnicity of the samples.” This may be partly an issue, but the weights as well as another biasing factor should have mitigated against this variable producing lower scores. While 20% of the NHANES III Youth sample was African American, only 15% of the sample receiving cognitive testing was African American. Since the African American children earn lower scores, this should not have produced the bias seen in the total, weighted NHANES III Youth sample.

It is worthy of note that NCHS does share our views that a nonrepresentative sample or errors in the data collection are important to the ability to draw valid inferences about the relationships seen within the NHANES III sample that generalize to the population of the United States. According to NCHS, “Regardless of the reason (emphasis added), the fact that the mean test scores are not 100 should not be of concern... What should be more important and more interesting is how they vary within the NHANES sample and how they are modified by other variables.” We disagree and view this position as poor science. To argue that regardless of the reason one may have error-laden data or a nonrepresentative sample, one should go forth and analyze the data and draw conclusions that affect public health policy and the lives of our youth seems irresponsible at best. If the NHANES III Youth dataset produces results that cannot be generalized outside of sample to the target population, then it is useful solely to generate research hypotheses and not for drawing conclusions about public policy or for resolving any scientific issues.

We cannot determine totally whether the NHANES III cognitive test score data were inap- propriately gathered or whether the sample is just biased in some peculiar way or whether the weighting schema are fatally flawed—or whether the effects seen are due to some interaction of these variables. What is apparent, however, is that conclusions drawn from the NHANES data lack external validity (generalizability beyond the sample) and may have internal validity problems as well, especially if the cognitive data were collected in ways that introduce systematic error through such matters as examiner bias or use of incorrect administration procedures.

The sampling plan and weighting schema along with missing values for variables assessed and missing values create other analytical and logical problems with these data, particularly as they relate to estimating effects of BLLs on cognitive performance. The largest effect sizes (which are on Reading and Digit Span) reported by Lanphear et al. (2000) occur on the two cognitive tasks whose means deviate the most from overall population values contributing to our conclusion that the reported associations may be due to biased data.

3. Missing values

As noted by even those who disagree vehemently about interpretations of the association of low BLLs with cognitive function (e.g., Kaufman, 2001a, 2001b; Needleman & Bellinger, 2001), when looking at the association of BLLs to other variables, a number of other variables must be controlled. For example, SES correlates with both BLLs and cognitive test scores, as does ethnicity and many other demographic variables. Nonstatus variables such as cotinine
level and nutritional variables may correlate with one or more variables related to the association of BLL and cognitive test scores. Such relationships become increasingly complex since nutritional status (particularly calcium, iron, and zinc deficiencies that are more common among low SES groups) affect lead absorption and possibly potency (e.g., Juberg, 2000).

It is thus, necessary to control such variables to determine any independent association of BLLs to cognitive test scores. When data are incomplete or missing on such control variables, complications in the analyses and their interpretations arise.

Lanphear et al. (2000) presented several sets of coefficients representing the association of BLLs to the variation in cognitive test scores of the NHANES III Youth sample (math, reading, Digit Span, and Block Design). The number of observations was reported for each of the regressions run to produce the BLL associations. To attain the number of observations reported required the replacement of missing values for one or more of the explanatory variables, as well as the dependent variables (Table 4). For example, the regression run on the total sample reported 4,853 observations. The number of subjects with valid scores for the math test was 4,542. Thus, 311 missing math test scores had to be replaced, possibly with the mean math score value for the sample, in order to maintain the 4,853 total observations reported by Lanphear et al. Similar replacement of missing values was required for several explanatory variables, such as, education level of the reference adult, poverty index ratio, serum ferritin level, and serum cotinine level. In fact, over 2,628 missing values had to be replaced for the serum cotinine level values, over 54% of the total sample. Table 2 summarizes the missing data rates for the variables extracted by Lanphear et al.

In addition to the issue of missing values and how the replacement of the missing values for dependent and independent variables affects the outcome of the regression analyses reported by Lanphear et al. (2000) violation of the sampling design is also a concern. None of the Lanphear et al. regressions can be produced with fully ranked data (using only valid values provided by the subjects) without violating the sampling design weighting by PSUs and stratum. The replacement of missing values is necessary in order to maintain the integrity of the sampling design in the regression analyses, which produce the regression coefficients, given

| Variable                        | Number of missing values | Percent of total sample |
|---------------------------------|--------------------------|-------------------------|
| Math score                      | 311                      | 6.41                    |
| Reading score                   | 441                      | 9.09                    |
| Digit Span score                | 281                      | 5.79                    |
| Block Design score              | 278                      | 5.73                    |
| Gender                          | 0                        | 0                       |
| Race/ethnicity                  | 0                        | 0                       |
| Poverty index ratio             | 381                      | 7.85                    |
| Education level of reference adult | 34                     | 0.70                    |
| Serum ferritin level            | 73                       | 1.50                    |
| Serum cotinine level            | 2,628                    | 54.15                   |
| Total observations in sample    | 4,853                    |                         |
Table 3
Regression coefficients for standardized math score

| Variable                        | Coefficient full rank | t value | Coefficient back-filled | t value |
|---------------------------------|-----------------------|---------|-------------------------|---------|
| Gender, male                    | −0.94                 | −1.00   | −0.67                   | −0.86   |
| African American                | −5.32                 | −5.51*  | −5.03                   | −5.25*  |
| Mexican American                | −2.95                 | −2.90*  | −2.59                   | −2.63   |
| Other races                     | −1.67                 | −0.99   | −2.31                   | −1.30   |
| Poverty index ratio             | 2.157                 | 5.49*   | 1.99                    | 5.36*   |
| Adult non-high school graduate  | −8.37                 | −5.71*  | −8.72                   | −6.89*  |
| Adult high school graduate      | −4.51                 | −4.29*  | −4.73                   | −4.98*  |
| Serum ferritin level            | 0.00                  | 0.60    | 0.00                    | 0.63    |
| Serum cotinine levela           | −0.07                 | −3.91*  | −0.06                   | −4.11*  |
| Blood lead level                | −0.76                 | −4.58*  | −0.70                   | −5.33*  |
| Constant                        | 96.38                 | 57.78*  | 96.30                   | 61.73*  |

Number of observations 4,102 4,853

a Serum cotinine level was back-filled with the sample mean to make the number of observations closer for comparison and to avoid violating the sample design, which means the equation would not compute because of strata with only one PSU.

* Statistically significant at \( P < .01 \).

the explanatory variables proposed by Lanphear et al. However, given the complex sampling design of the NHANES III dataset and the need to weight by PSU and strata to obtain supposedly representative scores, using total sample means to replace missing values for all PSUs across strata seems questionable.

Use of the sample means for replacing missing data can also make a significant difference in the size and statistical significance of the coefficients. If the Lanphear et al. (2000) equation is used, one can compare the coefficients and statistical significance between an equation estimated using back-filled data and one using fully ranked data (total sample, 4,853), maintaining the integrity of the sampling design. Tables 3 and 4 present the results of such a comparison for the standardized math and Digit Span scores. As the coefficients for lead indicate, the use of the sample mean for missing data can cause the estimated size of the effect to increase (Table 3) or decrease (Table 4). If the imputation of data was for a random sample of the subjects, the problem would be less serious but the missing data are not randomly distributed across relevant demographic groups suggesting a biasing effect of using back-filled or imputed values in Lanphear et al’s equations.

As indicated by the footnotes in Tables 3 and 4, estimation of the Lanphear et al. (2000) equations without violating the integrity of the sample design is impossible without replacing missing values for dependent and independent variables. When evaluating equations for BLLs less than 2.5 (Lanphear et al., 2000, Table 4), even the replacement of missing values is not sufficient to prevent the violation of the NHANES III sample design. As Table 5 demonstrates, the sample for BLLs less than 2.5 leaves stratum number 37 with only one PSU, even though all missing values for the equation in Tables 3 and 4 have been replaced with sample means.

It should be noted that the NHANES sample design affects the standard error and the significance levels of the coefficients in a regression analysis. For example, Table 6 provides the
Table 4
Regression coefficients for Digit Span scaled score

| Variable                        | Coefficient full rank | t value | Coefficient back-filled | t value |
|---------------------------------|-----------------------|---------|-------------------------|---------|
| Gender, male                    | −0.35                 | −2.55   | −0.28                   | −2.40*  |
| African American                | −0.62                 | −4.07*  | −0.60                   | −4.29*  |
| Mexican American                | −1.1                  | −6.01*  | −1.19                   | −7.10*  |
| Other races                     | −0.73                 | −3.19*  | −0.77                   | −3.54*  |
| Poverty index ratio             | 0.23                  | 3.69*   | 0.20                    | 3.26*   |
| Adult non-high school graduate  | −0.88                 | −4.87*  | −0.91                   | −5.60*  |
| Adult high school graduate      | −0.58                 | −3.26*  | −0.60                   | −3.70*  |
| Serum ferritin level            | −0.00                 | −0.77   | −0.00                   | −0.82   |
| Serum cotinine levela           | −0.00                 | −1.11   | −0.00                   | −1.21   |
| Blood lead level                | −0.05                 | −2.05   | −0.06                   | −2.84*  |
| Constant                        | 9.22                  | 39.42*  | 9.25                    | 42.01*  |

Number of observations 4,128 4,853

* Serum cotinine level was back-filled with the sample mean to make the number of observations closer for comparison and to avoid violating the sample design, which means the equation would not compute because of stratums with only one PSU.

* Statistically significant at P < .01.

coefficients, standard errors, and t values for three separate equations with the same specifications using components for all of the aspects of the sample design. Stratification only affects the standard error and thus, the t value, but estimation without weighting the data results in a significant difference in the coefficient, as well as the standard error and t value.

4. Other explanatory factors

The NHANES III Youth dataset contains a number of variables that may be associated with cognitive test scores, some of which show colinearity with BLLs and some of which do not. In order to isolate the association of BLLs below 10 μg/dl to cognitive test scores, these variables must be evaluated. Lanphear et al. (2000) used the following as control variables: gender, race/ethnicity, poverty index ratio, region of residence, parent or caregiver educational level and marital status, serum ferritin, and serum cotinine. Given the small effects seen, examination of other variables in the data seems appropriate. Of course, these analyses are subject to the sampling problems and related issues (e.g., data imputation) noted previously. However, they do serve to illustrate how the results can be affected by ignoring relevant variables in the dataset.

4.1. Have you ever repeated a grade?

In the NHANES Youth sample used by Lanphear et al. (2000), 21.20% of the survey respondents repeated a grade, more than one of every five respondents (this also suggests sampling problems as this value greatly exceeds the population values across the United States). When the responses are weighted using the NHANES sampling design, the percentage
repeating a grade declines to 17.13% of the representative population. (This is double the rate reported on the National Center for Educational Statistics website and seems to stack the sample with lower scoring individuals who also are more likely to be from urban centers, be minority, and be exposed to more lead in the environment.) This confounds greatly our ability to draw inferences about the relationship of BLLs to cognitive development in this sample.

Whether respondents repeated a grade or not significantly affected their test scores. Table 7 presents the survey means for each of the four test scores by whether the respondents repeated

| Stratum | PSU 1 | PSU 2 | Stratum | PSU 1 | PSU 2 |
|---------|-------|-------|---------|-------|-------|
| 1       | 15    | 33    | 2       | 26    | 68    |
| 2       | 5     | 16    | 3       | 27    | 23    |
| 3       | 14    | 22    | 4       | 28    | 7     |
| 4       | 5     | 16    | 5       | 29    | 84    |
| 5       | 6     | 19    | 6       | 30    | 58    |
| 6       | 16    | 21    | 7       | 31    | 26    |
| 7       | 5     | 12    | 8       | 32    | 20    |
| 8       | 27    | 20    | 9       | 33    | 33    |
| 9       | 18    | 16    | 10      | 34    | 32    |
| 10      | 29    | 28    | 11      | 35    | 7     |
| 11      | 34    | 43    | 12      | 36    | 4     |
| 12      | 17    | 37    | 13      | 37*   | 2     |
| 13      | 34    | 22    | 14      | 38    | 18    |
| 14      | 13    | 26    | 15      | 39    | 10    |
| 15      | 11    | 50    | 16      | 40    | 39    |
| 16      | 23    | 32    | 17      | 41    | 8     |
| 17      | 34    | 48    | 18      | 42    | 4     |
| 18      | 20    | 39    | 19      | 43    | 32    |
| 20      | 38    | 40    | 21      | 44    | 23    |
| 21      | 26    | 53    | 22      | 45    | 11    |
| 22      | 35    | 29    | 23      | 47    | 7     |
| 23      | 37    | 45    | 24      | 48    | 19    |
| 24      | 26    | 46    | 25      | 49    | 21    |

* Stratum with only one PSU.

Table 5

Frequency distribution of strata and PSUs for sample of blood lead levels less than 2.5

| Stratum | PSU 1 | PSU 2 | Stratum | PSU 1 | PSU 2 |
|---------|-------|-------|---------|-------|-------|
| 1       | 15    | 33    | 2       | 26    | 68    |
| 2       | 5     | 16    | 3       | 27    | 23    |
| 3       | 14    | 22    | 4       | 28    | 7     |
| 4       | 5     | 16    | 5       | 29    | 84    |
| 5       | 6     | 19    | 6       | 30    | 58    |
| 6       | 16    | 21    | 7       | 31    | 26    |
| 7       | 5     | 12    | 8       | 32    | 20    |
| 8       | 27    | 20    | 9       | 33    | 33    |
| 9       | 18    | 16    | 10      | 34    | 32    |
| 10      | 29    | 28    | 11      | 35    | 7     |
| 11      | 34    | 43    | 12      | 36    | 4     |
| 12      | 17    | 37    | 13      | 37*   | 2     |
| 13      | 34    | 22    | 14      | 38    | 18    |
| 14      | 13    | 26    | 15      | 39    | 10    |
| 15      | 11    | 50    | 16      | 40    | 39    |
| 16      | 23    | 32    | 17      | 41    | 8     |
| 17      | 34    | 48    | 18      | 42    | 4     |
| 18      | 20    | 39    | 19      | 43    | 32    |
| 20      | 38    | 40    | 21      | 44    | 23    |
| 21      | 26    | 53    | 22      | 45    | 11    |
| 22      | 35    | 29    | 23      | 47    | 7     |
| 23      | 37    | 45    | 24      | 48    | 19    |
| 24      | 26    | 46    | 25      | 49    | 21    |

Blood lead level coefficients based on different types of weighting schemes

| Coefficient               | Standard error | t value |
|--------------------------|----------------|---------|
| Full sample specifications| -0.76          | 0.17    | -4.58  |
| No PSU specification     | -0.76          | 0.14    | -5.55  |
| No strata or PSU specification | -0.76      | 0.14    | -5.55  |
| No weighting             | -0.38          | 0.09    | -4.36  |
a grade or not. Consistently across all four tests, weighted test score means were statistically lower (at the 99% level of confidence) for respondents who repeated a grade. Thus, an equation which is trying to explain the variation in test scores must include an explanatory variable for “repeat a grade” to be properly specified particularly when it occurs more frequently in ethnic minority groups where exposure to lead also is more likely.

African Americans and Mexican Americans clearly were more likely to have “repeated a grade” (see Table 8) and to score lower on the various cognitive tests.

4.2. Spanish language used in the interview

Another variable that affects the variation in the values for test scores is taken from the question, “the language of the interview for Family questionnaire.” Over 17% of the respondents had their family interview conducted in Spanish (6.26% weighted) as indicated in Table 9. The issue is whether Spanish being spoken in the household is related to the MEC test scores.
Table 10
Weighted mean test scores by “Spanish language used in interview”

| Test type         | Spanish interview | No Spanish interview | t-test value |
|-------------------|-------------------|----------------------|--------------|
| Math, standardized| 82.74             | 94.33                | 12.02        |
| Reading, standardized| 78.60         | 92.76                | 14.75        |
| Digit Span, scaled| 6.44              | 8.82                 | 19.23        |
| Block Design, scaled| 7.85            | 9.62                 | 7.35         |

*a All mean values are weighted.

Table 11
Distribution of “Spanish language used in the MEC tests”

| Respondent used Spanish in MEC tests (question MXPLANG) | Frequency (unweighted) | Percent (unweighted) |
|---------------------------------------------------------|------------------------|----------------------|
| No                                                      | 4,156                  | 88.41                |
| Yes                                                     | 545                    | 11.59                |
| Total                                                   | 4,701                  | 100.00               |

As Table 10 demonstrates, when the family interview was conducted in Spanish, test scores were significantly lower on average.

4.3. Spanish language used in the MEC cognitive tests

In addition, the language used by the respondent when taking the MEC tests also affects the variation in the test score values. This information is taken from the question, “language used by sample person in MEC.” Over 11% of the respondents used Spanish when taking their MEC tests as indicated in Table 11. Proper weighting of the responses was not possible due to small samples, which eliminates one or more of the PSUs within a stratum. This leaves one or more strata with only one PSU. The issue is whether Spanish being spoken in the MEC tests is related to the MEC test scores. As Table 12 demonstrates, when the respondent used Spanish in the MEC tests, test scores were, once again, significantly lower on average.

Table 13 presents the simple Pearson pairwise correlations between the test scores and whether Spanish was used in the interview or in the MEC tests. All simple correlations are statistically significant ($P \leq .01$) and negative, indicating the lower than average test scores for Spanish speaking respondents or households.

Table 12
Unweighted mean test scores by “Spanish language used in the MEC tests”

| Test type         | Spanish* | No Spanish* | t-test value |
|-------------------|----------|-------------|--------------|
| Math, standardized| 84.43    | 89.69       | 6.77         |
| Reading, standardized| 76.82    | 87.87       | 12.72        |
| Digit Span, scaled| 6.22     | 8.20        | 14.58        |
| Block Design, scaled| 7.79     | 8.52        | 4.63         |

* All mean values are unweighted.
Table 13
Unweighted pairwise correlations

|                  | Spanish language | Math | Reading | Digit Span | Block Design |
|------------------|------------------|------|---------|------------|--------------|
| Interviewa       | −0.14            | −0.22| −0.26   | −0.07      |              |
| MEC testsa       | −0.10            | −0.19| −0.21   | −0.07      |              |

a All pairwise correlations are statistically significant at \( P \leq .01 \).

This is extremely important since the validity of the test scores as reflecting cognitive development is highly questionable when the tests are administered in Spanish. In fact, the normative data used to derive the scaled scores cannot be applied accurately and will result in lowered scores. If these minority populations have higher BLLs (which they do) and have artificially deflated test scores (by virtue of using inappropriate tests, e.g., see Cronbach, 1990; Reynolds, Lowe, & Saenz, 1999), this sample will inflate any BLL-cognitive test correlation or regression coefficient spuriously and may cause associations to appear that are entirely artifactual.

4.4. Blood lead level variations by age at interview

Another factor that affects the BLLs is age. Table 14 provides the weighted and unweighted BLLs by age. In both cases, BLLs indicate a clear tendency to decline as the age of the respondent at the interview increases. In addition, the unweighted BLLs are clearly larger than their weighted counterparts at each age, though the difference between weighted and unweighted BLLs tends to decline as the age of the respondent at the interview increases. The actual number of respondents from each age group tends to be more represented by the younger ages, 6 through 11.

Estimating the Lanphear et al. (2000) equations by age group should provide a consistently significant blood lead level coefficient for each and every age group. In addition, by estimating an equation for each age group, the respondents used in the estimation of each equation have

Table 14
Weighted and unweighted mean blood lead levels

| Age | Number of observations | Percent of total | Weighted mean | Unweighted mean | Difference in means |
|-----|------------------------|-------------------|---------------|-----------------|---------------------|
| 6   | 468                    | 9.64              | 3.44          | 4.15            | −0.70               |
| 7   | 480                    | 9.89              | 3.05          | 3.66            | −0.61               |
| 8   | 458                    | 9.44              | 2.87          | 3.80            | −0.93               |
| 9   | 503                    | 10.36             | 2.62          | 3.52            | −0.89               |
| 10  | 497                    | 10.24             | 2.61          | 3.42            | −0.81               |
| 11  | 523                    | 10.78             | 2.48          | 3.11            | −0.63               |
| 12  | 396                    | 8.16              | 2.66          | 3.02            | −0.36               |
| 13  | 388                    | 8.00              | 2.20          | 2.78            | −0.58               |
| 14  | 388                    | 8.00              | 2.26          | 2.60            | −0.34               |
| 15  | 357                    | 7.36              | 2.08          | 2.45            | −0.37               |
| 16  | 395                    | 8.14              | 1.92          | 2.32            | −0.39               |
Table 15
Blood lead level coefficients by age at interview

| Age    | Math Coefficient | Math t value | Reading Coefficient | Reading t value |
|--------|------------------|--------------|---------------------|-----------------|
| All ages | −0.75            | −4.59*       | −0.85              | −5.25*          |
| 6      | −0.52            | −1.64        | −0.78              | −2.48           |
| 7      | −2.11            | −3.09*       | −1.09              | −2.05           |
| 8      | −0.35            | −1.41        | −0.72              | −1.76           |
| 9      | −0.66            | −1.95        | −1.84              | −3.54*          |
| 10     | −1.13            | −2.07        | −0.90              | −2.71*          |
| 11     | −0.18            | −0.33        | 0.14               | 0.29            |
| 12     | −0.96            | −1.40        | −0.43              | −0.54           |
| 13     | 0.03             | 0.08         | −0.30              | −0.68           |
| 14     | −1.43            | −3.87*       | −0.95              | −2.79*          |
| 15     | −0.20            | −0.43        | 0.48               | 0.93            |
| 16     | −0.64            | −0.66        | −2.17              | −2.88*          |

| Digit Span Coefficient | Digit Span t value | Block Design Coefficient | Block Design t value |
|------------------------|--------------------|--------------------------|----------------------|
| All ages               | −0.03              | −2.06                    | −0.12                | −3.64*              |
| 6                      | −0.11              | −1.89                    | −0.22                | −4.09*              |
| 7                      | −0.08              | −1.48                    | −0.124               | −1.58               |
| 8                      | −0.09              | −0.67                    | −0.14                | −1.35               |
| 9                      | −0.11              | −1.48                    | −0.22                | −2.71*              |
| 10                     | −0.06              | −0.92                    | −0.08                | −1.16               |
| 11                     | 0.07               | 0.94                     | −0.07                | −0.62               |
| 12                     | −0.15              | −1.04                    | −0.11                | −0.86               |
| 13                     | 0.01               | 0.12                     | 0.02                 | 0.16                |
| 14                     | −0.02              | −0.15                    | −0.25                | −3.96*              |
| 15                     | −0.16              | −3.07*                   | −0.02                | −0.16               |
| 16                     | −0.27              | −1.79                    | 0.05                 | 0.46                |

* Statistically significant at $P \leq 0.01$.

had the same “time based” opportunity for exposure to lead. Table 15 presents the results of the estimation of the Lanphear et al. equation specification for each age group for each test. The math test coefficient for BLL is statistically significant ($P \leq 0.01$) for only two of the eleven age groups, 7 and 14. The reading test coefficient for BLL is statistically significant ($P \leq 0.01$) for four of the eleven age groups, 9, 10, 14, and 16. The Digit Span test coefficient for BLL is statistically significant at .01 for only one of the eleven age groups, 15. The Block Design test coefficient for BLL is statistically significant at .01 for three of the eleven age groups, 6, 9, and 14. Note that in each of the equation results for the age groups and tests, that at least one low age and one high age group is statistically significant. In fact, the age group 14 was statistically significant at .01 for three of the four tests, even though it has one of the smaller sample sizes as indicated in Table 14. Another curious sampling issue occurs here. The age-corrected cognitive test scores (based on their respective standardization samples) are in fact correlated with age—the standardized Reading scores increase with age and all other scores decline with age—and BLLs also decline with age!
Table 16 presents the BLL coefficients using an expanded equation. The expanded equation includes repeat a grade, Spanish spoken at the interview and test, parents/guardian served in the military, time of day tested, and existence of an impairment/health problem. When adding these relevant explanatory variables to the equation, the amount of the variation being explained by the BLLs decreases.

The math test coefficients for BLLs are still statistically significant at .01 for the two age groups 7 and 14 (compared to Table 15). The reading test coefficients for BLLs are statistically significant (.01) only once in the eleven age groups, age 9. This is a large reduction for the reading test from the four statistically significant coefficients exhibited in Table 15. The Digit Span test coefficients for BLLs exhibit no statistical significance across all eleven age groups (.01). The Block Design test coefficients for BLLs are statistically significant at .01 for two of the eleven age groups, 6 and 14. Thus, improved specification of the equation, that is, including

| Age | Math Coefficient | Math t value | Reading Coefficient | Reading t value |
|-----|-----------------|--------------|---------------------|-----------------|
| All ages | -0.53 | -3.49* | -0.69 | -5.08* |
| 6 | -0.14 | -0.42 | | |
| 7 | -1.94 | -3.68* | -0.93 | -2.61 |
| 8 | -0.29 | -1.02 | -0.63 | -1.67 |
| 9 | -0.42 | -1.69 | -1.43 | -3.72* |
| 10 | -0.58 | -1.22 | -0.41 | -1.32 |
| 11 | 0.02 | 0.03 | 0.25 | 0.58 |
| 12 | -0.81 | -1.68 | -0.51 | -0.73 |
| 13 | 0.45 | 1.10 | 0.23 | 0.62 |
| 14 | -1.26 | -3.75* | -0.81 | -2.19 |
| 15 | 0.41 | 0.92 | 0.67 | 1.40 |
| 16 | 0.39 | 0.47 | | |

| Age | Digit Span Coefficient | Digit Span t value | Block Design Coefficient | Block Design t value |
|-----|------------------------|-------------------|--------------------------|---------------------|
| All ages | -0.03 | -1.26 | -0.08 | -2.47** |
| 6 | -0.06 | -1.42 | -0.19 | -3.13* |
| 7 | -0.05 | -1.12 | -0.07 | -1.22 |
| 8 | 0.02 | 0.38 | -0.01 | -0.13 |
| 9 | -0.05 | -1.12 | -0.15 | -1.85 |
| 10 | -0.01 | -0.22 | -0.00 | -0.06 |
| 11 | 0.09 | 1.29 | -0.06 | -0.53 |
| 12 | -0.17 | -1.46 | -0.09 | -0.94 |
| 13 | 0.01 | 0.19 | 0.06 | 0.62 |
| 14 | 0.00 | 0.05 | -0.26 | -4.00* |
| 15 | -0.12 | -1.87 | 0.11 | 1.48 |
| 16 | -0.17 | -1.16 | 0.14 | 1.34 |

* Statistically significant at $P \leq .01$.
** Statistically significant $P \leq .02$. 
explanatory factors that are important to the explanation of the variation in test scores, reduces
the magnitude and statistical significance of the coefficients of the BLLs across all age groups
and cognitive tests.

4.5. Other specifications

As earlier analyses have suggested, other information collected by the NHANES survey,
such as repeating a grade, language used for interview, and so forth, are correlated with the
variation in test scores. The specification by Lanphear et al. (2000) was very limited, potentially
biasing the results (Judge, Griffiths, Hill, Lutkepohl, & Lee, 1980). In addition, Lanphear et al.
used a linear specification of the relation between test scores and BLLs, attempting to test for
nonlinearity or significance through restriction of the sample by BLLs (less than 10, less than
7.5, less than 5, etc.). In essence, Lanphear et al. were searching for the point at which the
BLL begins to affect the variation in the test scores, that is, decreasing aptitude. To achieve
these goals, other specifications may be more appropriate, as well as more accurate.

Figure 1 presents a graph of the NHANES III Youth mean test scores for math and reading
by BLL. The most dramatic change occurs at BLLs of less than 8 μg/dl. At BLLs beyond
8 μg/dl no consistent downward trend exists. This is contrary to the hypothesis that increased
BLLs are expected to continue to decrease test scores. Figure 2 presents the same graphs for
Digit Span and Block Design with the same pattern.
Table 17

| Variable    | Coefficient | Standard error | t value | 95% confidence interval |
|-------------|-------------|----------------|---------|-------------------------|
| PBP ≤ 2.5   | 15.92       | 2.01           | 7.90*   | 11.87 – 19.96           |
| 2.5 < PBP ≤ 5 | 10.35       | 2.28           | 4.53*   | 5.76 – 14.93            |
| 5 < PBP ≤ 7.5 | 5.22        | 2.28           | 2.28**  | 0.62 – 9.81             |
| 7.5 < PBP ≤ 10 | 3.91        | 3.32           | 1.17    | −2.75 – 10.57           |
| Constant    | 80.21       | 2.12           | 37.85*  | 75.94 – 84.46           |

* Statistically significant at \( P \leq 0.01 \).
** Statistically significant at \( P \leq 0.03 \).

One of the contributing factors behind this phenomenon is the rapidly shrinking cell size. Of the 4,853 NHANES subjects aged 6–16 in the Lanphear et al. (2000) study, only 166 (3.42%) exhibit BLLs of more than 10 \( \mu \text{g/dL} \). In fact, 4,075 (83.97%) subjects have BLLs less than or equal to 5 \( \mu \text{g/dL} \) and 2,023 (41.69%) have BLLs less than or equal to 2 \( \mu \text{g/dL} \). Thus, most of the Lanphear et al. sample is concentrated in the lower end of the BLLs with little or no consistent trend (upward or downward) detectable beyond 8 \( \mu \text{g/dL} \).

To explore this relationship further, Table 17 presents the results of a linear specified regression using categorical variables for BLLs as explanatory variables. As the Table indicates, BLLs between 7.5 and 10 \( \mu \text{g/dL} \) are not significantly different from BLLs above 10 \( \mu \text{g/dL} \). BLLs between 5.0 and 7.5 \( \mu \text{g/dL} \) are significantly different at \( P \leq 0.03 \). This implies that the decreasing aptitude level as measured by math scores becomes approximately constant above 7.5 \( \mu \text{g/dL} \), which is inconsistent with existing hypotheses. In fact, the 95% confidence intervals around the estimated coefficients overlap significantly across all four categorical variables. The overlaps are sufficiently large as to cause the low estimate of each of the intervals for BLLs to nearly encompass the estimated coefficient for the next interval (exception, interval 2.5–5 \( \mu \text{g/dL} \) and 5–7.5 \( \mu \text{g/dL} \), misses by only 0.54). Thus, though the first three BLL intervals are statistically significant, their confidence intervals substantially overlap, suggesting that the relationship portrayed by the regression could be random noise or variation around the mean.

Figure 3 presents the weighted mean values for test scores, math and reading, by age of the subject at the time of the interview. The mean values of the test scores are nearly constant across age. Conversely, Figure 4 clearly indicates that the weighted mean BLLs decline as age at the
interview rises. Thus, some factors directly affect the BLLs, such as age, while having little or no affect on the variation in test scores. One could easily develop an explanatory equation for the variation in BLLs using such factors as age, demographics, poverty index ratio, etc. This suggests the variation in both BLLs and test scores are affected by some of the same factors.

5. Missing variables

Additionally, there are numerous variables that are known to be associated with lower cognitive test scores that are also more prevalent in populations with higher BLLs. Some of these variables include substance abuse during pregnancy (especially crack cocaine) and very low birth weights (any below 1500 g and especially those under 1000 g). These variables are not included in the NHANES III data and thus, make it impossible to draw firm conclusions regarding the association of BLLs to cognitive test scores. Their absence results in a spurious correlation between BLLs and test scores if any of these variables are correlated both with BLL and test scores, and several certainly are.

6. Endogeneity of BLLs and cognitive test scores with demographic characteristics of the NHANES III Youth dataset

Lanphear et al. (2000) estimated four equations as variants of multiple regression using a limited set of control variables. If the BLLs are part of a system of equations in which aptitude (represented by the four test scores) and BLLs are jointly determined (Greene, 1993; Pedhazur & Schmelkin, 1991), then the BLLs are endogenous to the simultaneous system of equations. The specification of the BLLs as an explanatory variable to aptitude can be tested within the specification of the simultaneous system. A simultaneous system can be estimated using a simultaneous systems estimator (such as seemingly unrelated regression, Zellner, 1962; or two-stage least squares regression, Kelejian, 1971). In either case, BLLs, which are endogenous to the system of equations, could be specified as a dependent variable in the equation specification of the aptitude measures (four test scores) for hypothesis testing.

Table 18 presents the estimated coefficients of a linear regression with BLLs as the dependent variable that is predicting BLL from basic demographic data. Seven of the 13 explanatory variables are statistically significant at \( P \leq .01 \), male, African American, poverty index ratio,
Table 18
Regression coefficients for blood lead levels

| Variable                                | Coefficient | t value |
|-----------------------------------------|-------------|---------|
| Male                                    | 0.72        | 4.82*   |
| African American                        | 1.19        | 5.81*   |
| Mexican American                        | 0.45        | 2.26    |
| Other races                             | −0.04       | −0.16   |
| Poverty index ratio                     | −0.35       | −7.97*  |
| Adult non-high school graduate          | 0.43        | 2.17    |
| Adult high school graduate              | −0.00       | −0.04   |
| Age at time of interview                | −0.11       | −7.65*  |
| Non-United States birth                 | 0.35        | 1.57    |
| Northeast                               | 1.18        | 3.15*   |
| Midwest                                 | 0.72        | 2.78*   |
| South                                   | 0.13        | 0.62    |
| Constant                                | 3.45        | 10.46*  |

* Statistically significant at $P \leq .01$.

Table 19
Regression coefficients for math test scores

| Variable                                | Predicted$^a$ | Actual$^b$ |
|-----------------------------------------|---------------|------------|
|                                         | Coefficient   | t value    | Coefficient   | t value    |
| Blood lead level                        | −0.23         | −0.24      | −0.61         | −3.59*     |
| Male                                    | −0.25         | −0.24      | −0.05         | −0.05      |
| African American                        | −5.04         | −3.10*     | −4.68         | −4.73*     |
| Mexican American                        | −1.66         | −1.59      | −1.71         | −1.65      |
| Other races                             | −1.99         | −1.33      | −2.02         | −1.29      |
| Poverty index ratio                     | 2.00          | 4.15*      | 1.88          | 4.96*      |
| Adult non-high school graduate          | −7.78         | −5.59*     | −7.70         | −6.04*     |
| Adult high school graduate              | −4.08         | −3.98*     | −4.14         | −4.07*     |
| Serum ferritin level                    | 0.00          | 0.96       | 0.00          | 0.91       |
| Serum cotinine level$^b$                | −0.05         | −3.05*     | −0.05         | −3.27*     |
| Military service                        | −1.28         | −1.32      | −1.34         | −1.39      |
| Adult 17 years of age                   | 17.83         | 9.25*      | 17.02         | 9.46*      |
| Interview provided in Spanish           | −3.74         | −2.73*     | −3.34         | −2.47      |
| Morning test session                    | 2.10          | 2.02       | 2.09          | 2.02       |
| Afternoon test session                  | 1.17          | 1.20       | 1.12          | 1.16       |
| Impairment/health problem               | −3.39         | −1.33      | −3.40         | −1.29      |
| Repeat a grade                          | −12.32        | −10.85*    | −12.15        | −10.96*    |
| Constant                                | 95.79         | 38.05*     | 96.83         | 60.04*     |
| $R$-square                              | .26           |  .26       |

$^a$ Predicted blood lead levels used in estimation.

$^b$ Actual blood lead levels used in estimation.

* Statistically significant at $P \leq .01$. 
age at time of interview, Northeast and Midwest regions and the constant term. Two others, Mexican American and adult non-high school graduate, are statistically significant at $P \leq .05$. The $R$-square for the BLL equation is .18, an effect size considerably larger than effects associated with low BLLs on neurodevelopmental phenomena.

If one uses the “predicted” BLLs based on the equation in Table 18, as an explanatory variable in an equation to explain the variation in test scores, then the BLLs are considered to be endogenous to the test scores system. Table 19 presents the results of the math test score equations using the “predicted” BLLs and the actual BLL. The difference between the two equations lies primarily with the coefficient of the BLL. The coefficient produced with the predicted BLLs is statistically insignificant while the coefficient resulting from the actual BLLs is statistically significant.

If the primary focus is to determine whether blood lead content has a significant effect on the variation in test scores, then all factors that affect the variation in test scores should be properly accounted within the scope of the available data. This does not imply that all factors, which affect the variation in the test scores, will be included since some of those factors may not be available in the data. But, to ignore relevant information, which may enhance the explanatory power of a mathematical specification, produces inadequate or misleading models, especially when endogeniety has been demonstrated (e.g., see Pedhazur & Schmelkin, Table 20

| Variable                  | Math         |          | Reading       |          |
|---------------------------|--------------|----------|---------------|----------|
|                           | Coefficient  | $t$ value| Coefficient   | $t$ value|
| Blood lead level          | $-0.32$      | $-0.34$  | $-3.57$       | $-3.35^*$|
| Male                      | $-0.21$      | $-0.20$  | $2.16$        | $1.81$   |
| African American          | $-4.97$      | $-3.21^*$| $-1.49$       | $-0.85$  |
| Mexican American          | $-1.70$      | $-1.66$  | $-2.38$       | $-1.94$  |
| Other races               | $-2.00$      | $-1.31$  | $-2.8$        | $-1.56$  |
| Poverty index ratio       | $1.98$       | $4.32^*$ | $1.0613$      | $2.10$   |
| Adult non-high school graduate | $-7.79$        | $-5.80^*$| $-6.39$       | $-5.96^*$|
| Adult high school graduate| $-4.11$      | $-4.00^*$| $-3.47$       | $-2.74^*$|
| Serum ferritin level      | $0.01$       | $0.97$   | $0.01$        | $1.12$   |
| Serum cotinine levelb     | $-0.06$      | $-3.14^*$| $-0.02$       | $-1.47$  |
| Military service          | $-1.29$      | $-1.33$  | $-0.43$       | $-0.37$  |
| Adult 17 years of age     | $17.45$      | $7.44^*$ | $16.09$       | $9.08^*$ |
| Interview provided in Spanish | $-3.54$        | $-2.22$  | $-5.88$       | $-3.40^*$|
| Morning test session      | $2.09$       | $2.03$   | $0.38$        | $0.38$   |
| Afternoon test session    | $1.14$       | $1.18$   | $1.13$        | $1.15$   |
| Impairment/health problem | $-3.41$      | $-1.32$  | $-5.95$       | $-1.94$  |
| Repeat a grade            | $-12.23$     | $-11.07^*$| $-8.94$       | $-8.68^*$|
| Constant                  | $96.00$      | $39.88^*$| $102.19$      | $34.89^*$|

$^a$ Actual blood lead levels used in estimation.

$^b$ Serum cotinine level was back-filled with the sample mean to avoid violating the sample design, which means that the equation would not compute because of strata with only one PSU.

$^*$ Statistically significant at $P \leq .01$. 

Downloaded from https://academic.oup.com/acn/article-abstract/18/3/219/2185 
by guest 
on 28 July 2018
The following discussions highlight factors included in the NHANES data related to the variation in test scores that were ignored in the analysis performed by Lanphear et al. (2000). As we have noted above, several of the factors that affect the variation in test scores also affect the variation in BLLs. According to Saving and DeVany (1982), among others, this commonality of explanatory variables for the dependent and independent variable suggests a system which exhibits endogeniety (Saving, DeVany, Shughart, & May, 1978; Saving, Stone, Looper, & Taylor, 1985; Stone, Saving, Turner, Looper, & Enquist, 1991). Systems comprised of endogenous explanatory variables often use a two-stage analysis to account for the effect of the endogenous explanatory variables (Judge et al., 1980; Pedhazur & Schmelkin, 1991).

Tables 20 and 21 present the results from the two-stage regression analysis of the test scores. The result presented in Table 19 holds true for three of the four test scores; reading remains statistically significant when using the predicted BLLs. The largest $t$ values across all four equations occur for the “adult under 17 years of age” variable, followed by the “repeat a grade” variable. The “adult under 17 years of age” variable only represents a few subjects, but they consistently score above average across all four tests. Variables which are statistically significant across all four equations are “adult under 17 years of age (positive),” “repeat a grade...

Table 21
Two-stage regression coefficients for Digit Span and Block Design test scores

| Variable                        | Digit Span Coefficient | $t$ value | Block Design Coefficient | $t$ value |
|---------------------------------|------------------------|-----------|--------------------------|-----------|
| Blood lead level*               | 0.07                   | 0.53      | 0.13                     | 0.95      |
| Male                            | −0.27                  | −1.87     | 0.66                     | 3.17*     |
| African American                | −0.62                  | −3.37*    | −2.37                    | −10.35*   |
| Mexican American                | −0.91                  | −5.08*    | 0.03                     | 0.13      |
| Other races                     | −0.62                  | −2.33     | −0.47                    | −1.35     |
| Poverty index ratio             | 0.22                   | 3.21*     | 0.39                     | 4.62*     |
| Adult non-high school graduate  | −0.77                  | −4.41*    | −1.40                    | −6.02*    |
| Adult high school graduate      | −0.53                  | −2.92*    | −0.71                    | −3.50*    |
| Serum ferritin level            | −0.00                  | −0.06     | −0.00                    | −0.41     |
| Serum cotinine level*           | −0.00                  | −0.55     | −0.00                    | −2.01     |
| Military service                | −0.26                  | −1.41     | −0.04                    | −0.18     |
| Adult 17 years of age           | 3.44                   | 9.71*     | 6.13                     | 11.40*    |
| Interview provided in Spanish   | −1.28                  | −5.27*    | −1.21                    | −3.70*    |
| Morning test session            | 0.13                   | 1.01      | 0.30                     | 1.43      |
| Afternoon test session          | 0.11                   | 0.68      | 0.48                     | 2.37      |
| Impairment/health problem       | −2.06                  | −2.91*    | −2.35                    | −2.94*    |
| Repeat a grade                  | −1.45                  | −9.22*    | −0.89                    | −4.42*    |
| Constant                        | 9.01                   | 20.08*    | 8.95                     | 18.70*    |
| $R$-square                      | .134                   |           | .201                     |           |

a Actual blood lead levels used in estimation.
b Serum cotinine level was back-filled with the sample mean to avoid violating the sample design, which means that the equation would not compute because of stratums with only one PSU.

* Statistically significant at $P \leq .01$. 
Variables for poverty index ratio (positive) and “interview provided in Spanish (negative)” were statistically significant for three of the four equations. The important difference between using the two-stage approach to modeling the association of BLLs to test scores versus simple multivariate regression is the quantitative change in the significance of the coefficients of BLLs.

7. Conclusions

Attempts to study the association of BLLs to cognitive test scores and development of children are a complex undertaking. To use the NHANES III Youth dataset to study this issue is geometrically more difficult and likely inappropriate if not impossible. While one can certainly generate equations that purport to show relationships and “explain” variance, the sampling and data collection issues seem insurmountable. Analyses such as those by Lanphear et al. (2000) simply lack external validity.

As we have shown (admittedly in a laborious manner), changes in the specifications of the equations reduce the already small effect sizes seen in the original Lanphear et al. (2000) equations. Given the potential sampling problems or weighting problems (or both) of NHANES III, these small effect sizes must be viewed with a healthy dose of skepticism. Even more importantly, even if the sampling were correct, key variables such as substance abuse during pregnancy and currently in the home are missing that could cause the observed relationships of BLL to cognitive test scores to vanish. Variables such as birthweight, which may affect neuropsychological development (and is correlated significantly with all cognitive test scores in the NHANES III sample), often have missing values for half or more of the cases in the sample. The magnitude of the missing data problem is substantial in light of the complex weights used and the resulting cell sizes (in some cases all of the subjects in a cell will end up with imputed data). The distributions of the test scores in NHANES III are also nonnormal, being skewed to the right, and the distribution of BLLs is also substantially skewed to the right (which is a desirable public health outcome). However, the grossly disproportionate cell sizes at various BLLs enhance the difficulty of seeking a threshold effect for BLLs. The rather odd findings in Figures 1–4 illustrate this problem and are contradictory to other findings in the lead literature (e.g., Juberg, 2000; Needleman & Bellinger, 2001). We would not attempt to generalize these findings outside of the NHANES III sample at present.

We do not believe the NHANES III Youth data can be used to study the issue of BLL associations with cognitive test scores. Too many peculiarities occur and the sampling is greatly suspect as is the accuracy of the collection of the cognitive data. There are too many missing values to breed confidence in the effect sizes seen when such large-scale imputation is undertaken and too many entire variables are missing by design. Inferences of causality between BLLs and cognitive loss are also inappropriate from this data base; too many variables that may affect the relationships are either absent from the data-collection effort or contain too many missing values when they are present and the research design itself is inadequate to the task of establishing cause (e.g., see Reynolds, 1999, for a more detailed discussion). Neither
Policy nor scientific problems related to cognitive and other neurodevelopmental problems should be considered using the NHANES III Youth dataset.

References

Brown, R. T. (2001). Behavioral teratology/toxicology: How do we know what we know? *Archives of Clinical Neuropsychology, 6*(4), 389–402.

Cronbach, L. J. (1990). Essentials of psychological testing (5th ed.). New York: Harper & Row.

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

Greene, W. H. (1993). *Econometric analysis.* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Hartman, D. (1995). *Neuropsychological toxicology.* New York: Plenum.

Hebben, N. (2001). Low lead levels and neuropsychological assessment: Let us not be mislead. *Archives of Clinical Neuropsychology, 6*(4), 353–358.

Jastak, S., & Wilkinson, G. S. (1984). *The Wide Range Achievement Test—Revised: Administration manual.* Wilmington, DE: Jastak Associates, Inc.

Juberg, D. R. (2000). *Lead and human health: An update.* New York: American Council on Science and Health.

Judge, G. G., Griffiths, W. E., Hill, R. C., Lutkepohl, H., & Lee, T. (1980). *The theory and practice of econometrics.* New York, NY: Wiley

Kamphaus, R. W. (2001). *Clinical assessment of child and adolescent intelligence.* Boston: Allyn & Bacon.

Kaufman, A. S. (1979). *Intelligent testing with the WISC-R.* New York, NY: Wiley

Kaufman, A. S. (2001a). Do low levels of lead produce IQ loss in children? A careful examination of the literature. *Archives of Clinical Neuropsychology, 6*(4), 303–342.

Kaufman, A. S. (2001b). How dangerous are low (not moderate or high) doses of lead for children’s intellectual development? *Archives of Clinical Neuropsychology, 6*(4), 403–431.

Kaufman, A. S., McLean, J. E., & Reynolds, C. R. (1988). Sex, race, residence, region and educational differences on the 11 WAIS-R subtests. *Journal of Clinical Psychology, 44*(2), 231–248.

Kelejian, H. (1971). Two stage least squares and econometric systems: Linear in parameters but nonlinear in the endogenous variables. *Journal of the American Statistical Association, 66*, 373–374.

Kramer, R. A., Allen, L., & Gergen, P. J. (1995). Health and social characteristics and children’s cognitive functioning: Results from a national cohort. *American Journal of Public Health, 85*, 312–318.

Lanphear, B. P., Dietrich, K., Auinger, P., & Cox, C. (2000, November). Cognitive deficits associated with blood lead concentrations <10 μg/dl in United States’ children and adolescents. U.S. Department of Health and Human Services; Public Health Reports. Public Health Report 2000 (Vol. 115, pp. 521–529).

Nation, J. R., & Gleaves, D. H. (2001). Low-level lead exposure and intelligence in children. *Archives of Clinical Neuropsychology, 6*(4), 375–388.

Needleman, H. L., & Bellinger, D. (2001). Studies of lead exposure and the developing central nervous system: A reply to Kaufman. *Archives of Clinical Neuropsychology, 6*(4), 359–374.

Pedhazur, E., & Schmelkin, L. (1991) *Measurement, design, and analysis.* Hillsdale, NJ: Erlbaum.

Plan and Operation of the Third National Health and Nutrition Examination Survey, 1988–1994. (1994, July). U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Center for Health Statistics.

Reynolds, C. R. (1982). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 178–208), NY: Wiley.

Reynolds, C. R. (1998). Fundamentals of measurement and assessment in psychology. In C. R. Reynolds (Ed.), *Assessment* (pp. 33–56). In A. Bellack & M. Hersen (Eds.), *Comprehensive clinical psychology* (Vol. 4). Oxford: Elsevier Science.

Reynolds, C. R. (1999). Inferring causality from relational data and designs: Historical and contemporary lessons for research and clinical practice. *The Clinical Neuropsychologist, 13*, 386–395.
Reynolds, C. R., Chastain, R., Kaufman, A. S., & McLean, J. E. (1987). Demographic influences on adult intelligence at ages 16 to 74 years. *Journal of School Psychology, 25*(4), 323–342.

Reynolds, C. R., Lowe, P. A., & Saenz, A. (1999). The problem of bias in psychological assessment. In C.R. Reynolds & T.B. Gutkin (Eds.), *The handbook of school psychology* (3rd ed., pp. 549–596). New York: Wiley.

Sattler, J. M. (2001). *Assessment of children*, (4th ed.). San Diego: Jerome M. Sattler.

Saving, T. R. & DeVany, A. S. (1982, September). Life-cycle job choice and the demand and supply of entry level jobs: Some evidence from the air force. *Review of economics and statistics.*

Saving, T. R., DeVany, A. S., & Shughart, W. (1978, May). *The supply rate and equilibrium inventory of air force enlisted personnel: A simultaneous model for the accession and retention markets incorporating force level constraints.* Air Force Human Resources Laboratories (Tech. Rep.). Brooks Air Force Base, Texas: Air Force Systems Command.

Saving, T. R., Stone, B. M., Looper, L., & Taylor, J. N. (1985). *Retention of air force enlisted personnel: An empirical examination.* Air Force Human Resources Laboratory (Tech. Paper 85-6). Brooks Air Force Base, Texas: Air Force Systems Command.

Stone, B. M., Saving, T. R., Turner, K. L., Looper, L. T., & Enquist, S. K. (1991). *A simultaneous estimation model of air force accession and retention.* Human Resources Directorate, Manpower and Personnel Research Division, Brooks Air Force Base, Texas: Air Force Systems Command.

Wasserman, G. A., & Factor-Litvak, P. (2001). Methodology, inference and causation: Environmental lead exposure and childhood intelligence. *Archives of Clinical Neuropsychology, 6*(4), 343–352.

Wechsler, D. (1974). *Wechsler Intelligence Scale for Children—Revised.* San Antonio, TX: Psychological Corporation.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children-III.* San Antonio, TX: The Psychological Corporation.

Wilkinson, G. (1993). *Administration and scoring manual for the WRAT-3.* Wilmington, DE: Jastak Associates.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias. *Journal of American Statistical Association, 57*, 500–509.