COMMENTARY

Inconsistencies in the Lead-Effects Literature Exist and Cannot be Explained by “Effect Modification”

CLAIRE B. ERNHART

Department of Psychiatry, Case Western Reserve University and MetroHealth Medical Center, Cleveland, OH 44109-1998

THE inconsistency of findings within and between studies of the relationship of low level lead exposure and child development is perplexing. Bellinger (4) tried to circumvent the problem (a) by asserting that meta-analyses demonstrate consistency, (b) by evading recognition of some of the methodological problems that produce inconsistency, and (c) by arguing that the inconsistencies are due to effect modification, or interaction. I believe that he is mistaken in each of these endeavors.

INCONSISTENCY

Intelligence Test Scores

Table 1 documents inconsistency. The estimates listed are based on regression coefficients from prospective studies that have included at least a lead measure at 2 years and an intelligence measure at age 4 or 5 years. Mean lead levels at age 2 years ranged from 6.5 μg/dL (10) to 35.5 μg/dL (54, in one of two towns). The entries are predicted losses in intelligence related to an increase in blood lead level from 10 to 50 μg/dL. The range of 10 to 50 μg/dL was chosen to increase equivalence among studies with log transformed and nontransformed lead measures. A selection of 10 to 20 μg/dL used elsewhere (48,49) exaggerates the estimate for studies that used log transformation.

The lead levels chosen were either a cumulative preschool index (which would be the most stable index), or if this was not available, the age 2 years index. At the onset of the prospective studies, it was expected that larger effects would be seen for prenatal exposure because this is the period of most rapid development of the nervous system. When most studies found no effect beyond infancy, the emphasis shifted to exposure at age 2 years. For the Boston study, this was the only postnatal lead measure related significantly to any outcome measure.

Multiple analyses are reported for several studies to illustrate within-study inconsistencies. The Full Scale IQ, or equivalent thereof, was selected for maximal reliability. All coefficients were adjusted for confounding except for the Costa Rica study. (Because estimates without confounder control were very small, control of confounding was deemed unnecessary.) The Costa Rica and Sydney investigators did not report regression coefficients because they found no evidence of effect. The effect estimate for these studies was set at 0.0.

These variations in effect estimates are striking in a field in which it is sometimes argued that the “loss” of two IQ points is crucial. The wide range of estimates would have been greater yet if all analyses from each study were included. Among the listed studies, there are over 600 statistical tests of the hypothesis that lead has adverse effects on cognitive development. Since approximately 8% of these reached statistical significance (p < 0.05), the contribution of chance cannot be ignored. (Those within a study are, of course, not independent.) The figure of 8% would be smaller, approaching 5%, if all tests actually conducted were available. The issue of multiple comparisons was not raised by most investigators.

Bellinger (4) asserted that intra-study agreement should be given higher weight than concordance between studies. Intra-study consistency, however, reflects repeated use of the identical measures (as, say, 2 year lead level) from one analysis to another as well as the to-be-expected coherence among measures of the same subjects. Even so, within-study consistency is not strong, as in the discrepancy between the 5 and 10 year Boston effect estimates calculated using the same age 2 year lead measure. Notable within-study inconsistencies also include failure to replicate findings at a later stage that were reported at an earlier date. This is particularly so for interactions or other results of exploratory analyses.

Effect estimates are not very precise. The estimate for the Boston, age 10 year, testing could range from -6.8 to -39.6 points for the 10 to 50 μg/dL range (95% confidence interval). The linear projection for this study is problematic since further extrapolation would predict impossible values for many
adults who have had extremely high lead levels in childhood and who show no evidence of impairment (46).

**Measures Other Than Full Scale IQ**

If Table 1 included measures of Verbal and Performance IQ scores, further inconsistency would be seen. While some reports describe a greater effect for Verbal than Performance IQ (51), some studies describe significant effects for Performance (8,21) but not Verbal scores, and in one study (3,34) the discrepancy between Performance and Verbal scores was reversed with age of testing. Although other prospective studies measured early language development, only the Cleveland group (25) has reported its findings. Because the results from the other studies are probably also small, their dissemination would further weaken the hypothesis of lead-related adverse effects. Inconsistency is also seen for subscore or specific end-point measures (52).

**META-ANALYSIS**

Meta-analysis can be an effective way of integrating findings from randomized clinical trials; it cannot compensate for problems other than sample size (44). A major problem for the analysis of nonexperimental studies is the selection from multiple measures of both risk factors and outcome variables. Furthermore, biases, such as those due to confounding by other factors, compound rather than being canceled out as the studies are combined. Because bias due to confounding is important, but often ignored in lead effects research, it will be considered in more detail below. Meta-analysis requires the collection of all studies. Nevertheless, there exists a strong bias against publication of negative findings, particularly in politically charged fields of study, such as research on the effects of cocaine and lead (15,17,31). The pool of studies available for meta-analysis on these topics thus has a potential for being inherently biased.

Meta-analyses in the lead effects area have not included all studies; the criteria for inclusion thus is an important issue. For example, an analyst might require publication of regression coefficients for inclusion. Regression coefficients are less likely to be reported for studies with negative findings because there is no need to report units of change in the outcome linked to the risk factor when there is no apparent change. This criterion was used in the selection of prospective studies for the meta-analysis conducted under the auspices of the World Health Organization (58).

Selection criteria are likely to be set by individuals who know the outcomes of the studies. Two recent overlapping meta-analyses (48,49) included only seven studies each. The choice in each included (a) the largest single effect estimate from the analyses of the Boston study, (b) studies with little or no control of confounding, and (c) one only of the eight studies in the European Multicenter study (56). The latter study (29) was the only one from that group with a significant lead effect estimate. Even that was misrepresented since the plotted data showed that lower intelligence test scores were found for lower, not higher lead levels in the range of exposure (10 to 20 pg/dL) studied in the meta-analysis.

For the second version, two studies, one with a positive effect estimate, were replaced by studies with effect estimates in the desired direction. For both versions, the main selection criteria appeared to be the statistician’s interests.

**CONFOUNDING**

It is two decades since publication of the first lead-effects study (41) to enter confounders into the analytic model, yet results without control of confounding are still used in support of the lead-effects hypothesis. This is one of the most serious problems in the lead effects literature (22,51), yet Bellinger tends to downplay it — “My biostatistician can lick your biosstatistician.” Figure 1 of his commentary was of data that were apparently the unadjusted data presented with adjustment for confounding in the original publication (10). A previous article (11) featured four graphs that included unadjusted data from other studies. No caveats were given. The consequences of the use of confounded findings have been profound; two of the latter graphs were copied into a document (14) which has been critical for the setting of costly governmental policies and two have been copied into an authoritative review (38).

Covariate control adjusts for bias and increases precision

<table>
<thead>
<tr>
<th>Study</th>
<th>Lead Measure</th>
<th>IQ Measure</th>
<th>Estimate 10-50 µg/dL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston</td>
<td>24 month</td>
<td>GCI 57 month</td>
<td>-4.7*</td>
</tr>
<tr>
<td>Boston</td>
<td>24 month</td>
<td>FS IQ 10 year</td>
<td>-23.2*</td>
</tr>
<tr>
<td>Cincinnati</td>
<td>Lifetime</td>
<td>MPC 5 year</td>
<td>-2.8</td>
</tr>
<tr>
<td>Cincinnati</td>
<td>Lifetime</td>
<td>FS IQ 6.5 year</td>
<td>-5.2</td>
</tr>
<tr>
<td>Cleveland</td>
<td>Average to 3 year</td>
<td>FS IQ 4 year, 10 month</td>
<td>+10.0</td>
</tr>
<tr>
<td>Costa Rica</td>
<td>12-24 month</td>
<td>GCI 5 year</td>
<td>0.0</td>
</tr>
<tr>
<td>Pt. Pirie (M)</td>
<td>0 to 3 year</td>
<td>GCI 4 year</td>
<td>-1.2</td>
</tr>
<tr>
<td>Pt. Pirie (F)</td>
<td>0 to 3 year</td>
<td>GCI 4 year</td>
<td>-12.1*</td>
</tr>
<tr>
<td>Pt. Pirie (M)</td>
<td>0 to 3 year</td>
<td>FS IQ 7 year</td>
<td>-3.8</td>
</tr>
<tr>
<td>Pt. Pirie (F)</td>
<td>0 to 3 year</td>
<td>FS IQ 7 year</td>
<td>-11.4*</td>
</tr>
<tr>
<td>Sydney</td>
<td>Birth to 7 year</td>
<td>FS IQ 7 year</td>
<td>0.0</td>
</tr>
<tr>
<td>Yugoslavia</td>
<td>24 month</td>
<td>MDI 24 month</td>
<td>-3.7</td>
</tr>
<tr>
<td>Yugoslavia</td>
<td>24 month</td>
<td>GCI 4 year</td>
<td>-7.3*</td>
</tr>
</tbody>
</table>

*p < 0.05. Abbreviations: (M): male, (F): female, GCI: General Cognitive Index (McCarthv Scales), MPC: Mental Processing Composite (Kaufman ABC), MDI: Mental Development Index (Bayley Scales).
The influence of confounding in most lead effects research is substantial. In the Cleveland study (24), for example, the set of covariates accounted for over 50% of the variance for analyses of IQ at age 4 years, 10 months. In none of these analyses did lead account for more than 2% of the variance; in most it was less than 1%. The effects of lead, if any, are minor indeed when compared with other conditions that affect child development.

Confounding can also be demonstrated using regression coefficients. In the Cleveland data the unadjusted regression coefficient is a highly significant -8.9 IQ points for each increase of 10 μg/dL in blood lead at age two years. (The change for 10 to 50 μg/dL would be 35.6 points.) With control of confounding, it was small and nonsignificant, -1.8 points. This shift in effect estimates is not unique to the Cleveland study.

Logistic difficulties prevent measurement of some likely confounders. Some measured co-factors are limited surrogates for nutrition, paternal intelligence, poverty, childhood illness, abuse and neglect, parental use of drugs, including tobacco and alcohol, and so forth. In the Cleveland study, history of medical problems, including otitis media, was a confounder; these illnesses are not unusual. Methods of sampling or conducting a study sometimes result in study-specific confounders. Unmeasured confounders have been called phantom variables (39). They are not phantoms; they exist and they can bias the inferences we draw.

Lead effects regression coefficients are likely to be more dependent on the quality of readily available covariates (if any), than on the true relationship of lead and the outcome measure.

Misuse of spurious findings with little or no control of confounding undermines the credibility of this field of study. It is well past the time to reach a consensus on the use of confounded findings in policy documents, meta-analyses, commentaries, and other published statements.

ERRORS IN MEASUREMENT

Although errors-in-measurement is receiving more attention in epidemiological research (42) most analyses of estimated lead effects are conducted under the assumption of no measurement error. In one exception (26), lead and cognitive ability scores, but not the covariates, were "disattenuated" for unreliability. Two other analyses of a given dataset reached different conclusions. In one instance (1) the effect of covariate measurement error was dismissed. The second report (30) concluded that the true regression coefficient for lead might be reversed with consideration of error in measurement.

The most comprehensive approach to errors-in-measurement in lead-effects research (28) used sensitivity analysis to test a range of plausible assumptions about measurement error in the lead variable. Both the estimated lead effect and the confidence intervals of that estimate were shown to increase with increasing estimates of measurement error. Thus, while adjustments for error measurement in the lead variable will increase the size of a regression coefficient, they can also decrease the precision of the estimate.

The effect on the lead estimate of assumptions about errors in measurement in covariates was then evaluated. When estimates of measurement error in several covariates were included in the analysis, the estimate of the lead effect decreased, but the precision of the estimate was unchanged. Changes with correction for errors-in-measurement were not large, but because lead effect estimates are small and of low precision, the consequences may be considerable.

EFFECT MODIFICATION

Because some of the inconsistencies among lead-effects studies are hard to ignore, Bellinger (4) suggested that failure of replication across studies is due to specific characteristics of the samples rather than to internal validity factors. The most concrete of these effect modifiers can be tested within studies as interactions of a lead measure with a purported modifier. Some potential effect modifiers have not been tested within studies and can be evaluated only across studies.

A well known instance of effect modification is a finding that the effects of low birth weight are attenuated in children raised in middle class families in comparison to children of more disadvantaged families (2,33,55), but even this finding is not replicated consistently. For example, the development of preterm infants with severe intraventricular hemorrhage (IVH) was found not to be enhanced by family conditions that do improve the outcomes in a noninteractive way for mild IVH and nonIVH neonates (12). In still another study (13), both preterm birth and SES level were related to IQ without evidence of interaction.

Schwartz (49) hypothesized a different effect moderation paradigm for disadvantaged lead exposed children, "... that the effects of lead are muted in children who have already had their higher order functioning disturbed by other factors," hence middle class children are more likely to be impaired by lead. Perhaps we can "explain" modification by SES, or family risk factors, whether they either diminish or enhance lead effects estimates.

It is easy to add product terms into one's model. Obviously, working through numerous interaction analyses increases the risk of chance associations. This may explain some unusual findings, such as the report (7) from the Boston study that "recovery" from the effects of high cord blood lead levels was facilitated if contemporary blood lead level was low, if the child was female, or if the family was of higher SES. Another danger that emerges in the search for effect modification is that investigators might sift their samples into various subgroups to identify a "vulnerable" status (37) and treat the findings as demonstrated "facts" rather than hypothesis-generating observations.

As with any other scientific observation the impact of an effect modifier should be generalizable, or replicated, across studies. We consider some possible effect modifiers.

Species

Taxonomic classification is the most likely of all effect modifiers. This is the case for a number of disease conditions as well as for susceptibility to toxins or teratogens (as thalidomide) which are specific to a given species.

Insofar as I know, there is no systematic research on various subhuman species with the same blood lead level, the same experimental paradigm, and the same endpoint measure. If consistent between-species findings at the nonhuman level cannot be obtained, generalization of effect to humans is tenuous.

Dose

Between studies. Exposure level seems to be a reasonable candidate for effect modification. Bellinger quoted Schwartz (49) as reporting a steeper lead-related rate of decline in IQ for cohorts with lower blood lead levels, but the conclusion depended on study selection. (This plot was based on seven studies selected for one of the meta-analyses summarized above.) If one study, a study with very limited control of...
confounding, were excluded, only the 10 year estimate of the Boston study would be inconsistent with a trend toward a higher slope at higher mean blood lead levels. Among the prospective studies, there is no apparent ordering of effect estimates by blood lead level.

**Within studies.** Bellinger described a graph from the Pt. Pirie study (3) as indicating that the slope of estimate is greater for lower lead levels, but, once again, the choice is of data not adjusted for confounding. There have been two within-study analyses of data of the Boston study with results that are not consistent. Using the 2 year lead level, the association for age 10 years IQ was as strong below the mean as for those at the higher end (49), but for the age 5 years intelligence index the relation was positively accelerated, i.e., the effect estimate was greater at higher lead levels (48).

In the Lavrion study (29) still a different pattern is seen in that the relationship was positive to about 70 μg/dL (median = 21.5) before becoming negative. The rate of decrease at the higher end of the range was greater than that for any other covariate adjusted study aside from the Boston study (10). The Lavrion study thus does not support the hypothesis of a greater effect estimate at low than at high levels. This was not replicated in the other 7 studies in the European Multicenter study. Another study (45) of children with lead levels above 20 μg/dL found no relation between lead level and cognitive index. Inconsistencies are seen at all levels of lead exposure.

At present, only the Boston study supports an interpretation of an effect at very low levels and that inference is based on a few analyses from a large number. The evidence that lead level is an effect modifier is weak and inconsistent within as well as between studies.

**Genetic Differences**

The methods of behavioral genetics do not seem to be readily applied to the trait, vulnerability to lead, which is difficult to operationalize in groups and essentially impossible in individuals. Bellinger et al. (9), however, identified 5 individuals who were heterozygotic with respect to ALA-D isozyme phenotype and noted that these individuals tended to have lower dentine lead levels and also tended to have more optimal scores on various psychological tests. These tendencies are obviously very tenuous. Some findings in the area of behavioral genetics that have been thought to be firm have failed in replication (37).

**Within studies.** Because members of different ethnic groups vary genetically, we can test whether lead effect estimates differ by ethnic group. In the Cincinnati study (18) blood lead measured at 3 months was related to Bayley MDI scores at 6 months for white but not black infants. It was suggested that early development among black infants is more canalized and more resistant to adverse conditions. In the Cleveland study (23) no interaction of race and lead for 21 combinations of lead measures and test scores in the infancy period was found. Neither group has reported significant race by lead interactions at later ages.

**Sex Differences**

**Within studies.** A greater vulnerability to lead for males was suggested by a slightly steeper lead effect estimate for boys than for girls in one study (43). Early results (18) from the Cincinnati study included a relationship of maternal lead level and 6 month Bayley MDI scores for males but not females. The interaction apparently did not persist; the possibility was tested, but the authors did not report a sex by lead interaction at age 4 years (19). The findings of greater effects for males are the reverse of the dramatic differences by sex in the Pt. Pirie study (35), shown in Table 1. There were no sex × lead interactions in the Cleveland study and no other reports substantiating sex as an effect modifier were found. The evidence for sex-related differences in vulnerability is not consistent.

**Other Exposures**

**Within studies.** The design of the Cleveland study permitted tests of the interaction of lead level with maternal use of alcohol and tobacco in pregnancy and also with seropositivity for Toxocara canis (40). None of the interactions were significant. Maternal use of alcohol and tobacco were also related to cord blood lead level in the Boston study (5), but there was no mention of interaction.

**Socioeconomic Status**

SES is a complex and rather untidy variable. First, it may be causally related to biological risk factors, as low birth weight, poor nutrition, iron deficiency anemia, and an increased likelihood of illness and injury. To the extent that this is so, biological factors may increase vulnerability to the adverse effects of another risk factor. Alternatively, if we accept Schwartz's interpretation (49), these children have already been reduced to a more basic functioning that is more resilient to lead effects. SES is also a distal surrogate for family risk factors, including parental intelligence, parental teaching styles, child rearing capabilities, crowding, single parent family, family stress, etc. (12). Many of these more proximal risk conditions have been measured more directly in child development research. In general, the most plausible descriptions are of additive models of effect (2,12,47). A compilation of family risk factors may act more effectively as an effect modifier for lead effects than does the more distal traditional SES index. Interactions of lead with such variables were not significant in the Cleveland study. Because these interactions have not been reported for most studies with positive findings, the discussion is limited to the reports that utilized traditional measures of SES.

**Between studies.** It has been suggested (4,49) that a nondisadvantaged population provides a “cleaner” sample to test the effect of lead. If so, a ranking of studies by SES should be related to estimates of effect. The prospective study that is second to the Boston study in SES is the Sydney study (16) of upper middle class ethnically homogeneous children. No significant lead effects were seen in this study through age 7 years. No evidence of a lead effect was seen in another homogeneous sample (57) that showed little evidence of confounding. A clean field does not seem to be the critical factor.

A wide range of SES was represented in the European Multicenter study (56). Of the 8 centers in that study, the only significant estimate was for the Lavrion study of low SES smelletown children. Effect estimates do not change systematically among studies on the basis of SES.

**Within studies.** Among the studies that have enough variation in SES to test interactions, one well conducted study (50) of the effects of dentine lead found no evidence of a greater estimate in the lower social classes. Trends toward a greater within-study lead effect estimate for lower SES children were reported by a few investigators. The Cincinnati group, in particular, found larger effect estimates for the lowest portion of its SES range (19), but this finding did not persist (21). There
were no significant interactions of lead with SES or home environment measures in the Pt. Pirie study (35).

A finding of greater effect for upper SES children appears to be limited to the Boston study.

**The Boston Study**

Much of the evidence cited by Bellinger (4) is from the Boston study. Since the study is atypical, it is possible that effect at low level and modifications of effect are unique to the cohort. In the spirit of identifying the specific experimental system, some of what is known about the study is reviewed.

The study originated when cord blood lead data were related to numerous items from medical records. From this large pool of 11,837 mother-infant pairs, 1207 were deemed eligible on the basis of cord blood lead levels in the lower, middle and upper deciles as well as other criteria.

Of the 1207 eligible infants, 249 were enrolled. The three lead groups differed in the proportions excluded for different reasons. More than 30 variables showed statistically significant differential enrollment or attrition to age 2 years (36). Risk factors that were inexplicably related to enrollment and attrition included maternal drinking, smoking, and maternal hemoglobin. The incidence of respiratory distress was higher in the high lead than the low lead group, although lead level was related negatively to respiratory distress in the perinatal study. Being Jewish and being on welfare were each related to cord blood lead group and may have been related to mental development. Infants in the more socioeconomically advantaged group tended to have higher prenatal lead levels. Each variable in the covariate list (except maternal IQ) for age 2 years was involved in differential enrollment, differential retention, a curvilinear relation to cord blood level and the cognitive measures, or all three.

The 169 children retained to age 5 years had significantly more favorable home environment scores and to have significantly higher lead levels than those not retained (8). This resulted in overrepresentation of high SES, high lead level children. Higher SES was also related to retention at age 10 years (10). (A very strange attrition factor was having donated a tooth; donors were much less likely than nondonors to be included in the 10 year examination.) The cohort of 148 children at age 10 years comprised only 12% of the original group of provisionally eligible children.

Significant findings for postnatal lead exposure were limited to the age 2 year lead sampling although both lead had been sampled at 6 months, 12 months, 18 months, 57 months and 10 years. The mean lead level at age 24 months was 6.8 µg/dL. Capillary sampling was used. The analyses of these samples had an average difference between assays of 1 µg/dL and a difference of 3.5 µg/dL or more in 10% of the pairs. The effect of error in measurement for this critical index has not been discussed.

Variables related to lead exposure (as maternal IQ) in most studies were not related to lead level in the Boston study at age 2 years (6). The only statistically significant covariate (maternal involvement with the child) was related positively, not negatively, to lead level. Measures used as covariates in other research were thus not relevant to the findings of the study, except insofar as they may have increased the precision of the analyses. Statistical control of covariates, however, did decrease regression coefficients for 2 year lead level and the ages 5 and 10 year outcome measures. In contrast to other studies, covariate control increased the association of cord blood lead level and infancy outcome measures.

As shown in Table 1, for age 5 years, a small but statistically significant relationship of 24 month blood lead and intelligence was reported; for age 10 years the effect estimate was very large. The effect was found at a low lead level with no evidence of a threshold (49), but the small number of cases and the notable imprecision of the lead measure may have clouded detection of a threshold. Dentine lead level was not related to outcome measures. The data of the study have been well explored, which may explain some of the nonreplicated aspects of the findings.

The Boston study is unique. It is difficult to know exactly how selection and retention biases have affected its results, but given these eccentricities, one may ask whether the findings can be generalized. If evidence of effect under given conditions exists in only one study, it is little more than a curiosity without meaning for practical or theoretical science.

**CONCLUSION**

"If you torture your data long enough, they will tell you whatever you want them to hear." (37) We can ignore data that is telling us what we don't want to hear. We can conduct numerous statistical analyses, the not so nice term is dredging. We can also misuse the data of others, both by convenient selection of studies for meta-analyses and by drawing conclusions from findings not adjusted for confounding. Even with use of these strategies there is considerable inconsistency in the lead effects literature.

Some problems that can contribute to inconsistency have received little attention. Error-in-measurement, for example, can be critically important in studies with small and imprecise effect estimates. Biases in selection and attrition can act like loose cannons on a deck, with unpredictable effects. Little attention has been paid to the risk of over emphasizing occasional positive findings in multiple analyses.

Bellinger has proposed the idea that different groups, each reflecting a different "experimental system," are differentially vulnerable to the effects of lead. Whatever contextual factors are found to differentiate groups with and without notable lead effect estimates may thus become effect modifiers. Given the large number of analyses of interaction that can be done, numerous candidates should emerge. The problem lies in replication. Bellinger's speculations were based primarily on research with animal models. At present, at least, the notion does not apply to research with children. The literature shows little evidence of consistent findings regarding candidate effect modifiers.

In view of the methodological limitations, we may never be able to say with certainty that low level lead exposure has adverse effects. What we can say is that, if there is an effect when other conditions are considered, it is small.

**REFERENCES**


