

## **APPENDIX G**

### **Multi-Media Model (Empirical Model) for Use in the Section 403 Risk Assessment**

## APPENDIX G

### MULTI-MEDIA MODEL (EMPIRICAL MODEL) FOR USE IN THE SECTION 403 RISK ASSESSMENT

#### EXECUTIVE SUMMARY

##### Purpose Of The Appendix

This appendix documents development and evaluation of an empirical regression model relating measures of lead in a residential environment to geometric mean children's blood-lead concentrations. The model is used as one tool in the Section 403 risk assessment to estimate blood-lead concentrations of children exposed to lead in paint, dust and soil as measured in the HUD National Survey. This model is also employed to evaluate various options for risk management for the Section 403 standards. In this analysis, EPA estimated a national distribution of blood-lead levels (and, ultimately, estimated health effects) before enactment of the Section 403 standards, and then employed models to relate environmental levels of lead to children's blood-lead levels to estimate a national distribution of blood-lead levels (and health effects) after enactment of specific 403 standards. Environmental measures of lead from the HUD National Survey are used as inputs to the empirical model to predict the national distribution of blood-lead concentrations. Therefore, the model development was constrained to variables in the HUD National Survey data set. The goal was to develop a model that could be used to give an approximation of expected blood-lead concentrations related to residential environmental lead based on a single source of data.

In this appendix the empirical model is presented and its prediction of a national distribution of blood-lead concentrations is compared to the results of Phase 2 of the Third National Health and Nutrition Examination Survey (NHANES III).

##### Model Development Issues

The choice and construction of variables, the mathematical form of the empirical model, assessment of goodness-of-fit and influential points, and the treatment of measurement error in predictor variables were all given consideration during the development of the empirical model.

One particular difficulty was that the empirical model was constructed using dust lead results collected from wipe sampling in the Rochester study, whereas dust lead results in the HUD National Survey were collected from blue nozzle vacuum sampling. Similarly, the empirical model was constructed using soil lead concentrations observed from drip-line sample locations in the Rochester study, whereas soil lead results in the HUD National Survey were based on an average concentration of lead in soil from drip-line, entryway and remote locations. A statistical method was developed to account for both systematic differences as well as differences in error structures between the sampling methods employed in the Rochester study and the HUD National Survey.

## The Empirical Model

The form of the empirical model is:

$$\ln(\text{PbB}) = \beta_0 + \beta_1 \cdot \ln(\text{PbF}_{\text{BN}}) + \beta_2 \cdot \ln(\text{PbW}_{\text{BN}}) + \beta_3 \cdot \ln(\text{PbS}) + \beta_4 \cdot \text{PbP} + e$$

where PbB represents the blood-lead concentration,  $\text{PbF}_{\text{BN}}$  and  $\text{PbW}_{\text{BN}}$  correspond to dust-lead loading from interior floors and window sills respectively (on a Blue Nozzle Vacuum Scale), PbS represents soil-lead concentration, PbP represents paint/pica hazard, and e represents the residual error left unexplained by the model.

## Results Of The Comparison With NHANES III

The predicted distribution of blood-lead concentrations for children aged 1-2 years obtained by applying the empirical model to the HUD National Survey Data was compared to Phase 2 of NHANES III. Results of this comparison indicate:

- ! The national geometric mean blood-lead concentration (pre-intervention) was properly calibrated to the geometric mean reported in NHANES III.
- ! The variability in the national distribution of blood-lead concentrations predicted by the empirical model using the HUD National Survey is approximately 1.71 (GSD), in contrast to a GSD of 2.09 for Phase 2 of NHANES III.
- ! The estimated proportions of blood-lead concentrations exceeding 10, 20 or 30  $\mu\text{g}/\text{dL}$  using the empirical model predictions are much lower than the corresponding proportions estimated by NHANES III. For example, the percentage of children aged 1-2 years estimated to have blood-lead concentrations above 10  $\mu\text{g}/\text{dL}$  using the empirical model was 1.54% in comparison to 5.75% estimated in Phase 2 of NHANES III.

Differences between the Rochester study population and the national population represent the primary limitation when using the empirical model based on data from the Rochester Study to predict a national distribution of blood-lead concentrations.

## Use Of The Empirical Model

The empirical model is used in the Section 403 risk assessment and economic analyses to predict a distribution of childhood blood-lead concentrations based on measures of lead in paint, dust and soil at the child's primary residence. This information is used to evaluate various options for risk management for the proposed Section 403 Standards. In these analyses, the model is used to predict national distributions of children's blood-lead concentrations both before and after the rule is proposed. Estimates of environmental levels of lead before and after enactment of the Section 403 standards and after interventions resulting from the standards will be used as inputs to the model. The empirical model should only be used to predict a distribution of blood-lead

levels when environmental levels for all media are known or estimated. It is not intended as a general dose-response model, but rather as a predictive model developed specifically for use in the Section 403 Risk Assessment and specifically to predict blood-lead concentrations from estimates of environmental lead as measured in the HUD National Survey or as measured by a standard Section 402 risk assessment.

## G1.0 INTRODUCTION

In order to better inform risk managers as they consider various options for the Section 403 standards, EPA estimated the range of risk reductions that are expected to result from a variety of potential standards. In order to do this, EPA estimated a national distribution of blood-lead levels (and, ultimately, potential health effects) before enactment of the Section 403 standards, and then relied on models relating environmental levels of lead to children's blood-lead levels to estimate a national distribution of blood-lead levels (and potential health effects) after enactment of specific 403 standards. The empirical model is used in the Section 403 risk assessment and economic analysis to predict a distribution of blood-lead concentrations related (jointly) to measures of lead in three media at the child's primary residence: paint, dust, and soil. Environmental measures of lead from the HUD National Survey were used as inputs to the empirical model to predict the national distribution of blood-lead concentrations. Therefore, the model development was constrained to variables in the HUD National Survey data set. Given time and budget constraints the goal for the empirical model development could not include construction of the best possible model based on multiple data sources. Rather, the goal was to develop a model that could be used to give an approximation of expected blood-lead concentrations related to residential environmental lead based on a single source of data. This model has not undergone formal validation and is based on only one data set. It is not intended as a general dose-response model, but rather as a predictive model developed specifically for use in the Section 403 Risk Assessment and specifically to predict blood-lead concentrations from estimates of environmental lead as measured in the HUD National Survey or as measured by a standard Section 402 risk assessment. The model was used to estimate the benefits of the 403 rule in the post-403 situation by estimating the reduction in children's blood lead concentrations resulting from application of various options for the 403 standards via risk assessments in residential housing.

In this appendix the empirical model is presented and its prediction of a national distribution of blood-lead concentrations is compared to the results of the NHANES III Survey as follows:

A national distribution of housing and population characteristics was estimated using the HUD National Survey of environmental levels of lead in paint, dust, and soil in residential housing along with pertinent Census information. The Census information and the HUD National Survey measurements of environmental lead (after appropriate conversions) were used as inputs to the model to predict a national distribution of children's blood-lead levels before enactment of the Section 403 standards. This pre-rulemaking distribution was compared to the national distribution of children's blood-lead concentrations estimated by the NHANES III survey to assess the adequacy of the model and its applicability on a national level.

The empirical model is also used to predict the national distribution of children's blood-lead levels after enactment of the Section 403 standards. Estimates of environmental levels of lead after the conduct of interventions performed in response to various options for the Section

403 standards are used as inputs to the model. Comparison of the pre- and post-rulemaking distributions allow estimation of the benefits associated with the rulemaking.

The empirical model is not intended to be used to estimate the effect of a single media on blood-lead levels. The model should only be used to predict a distribution of blood-lead levels when environmental levels for all media are known or estimated. Individual parameter estimates should not be interpreted in isolation.

The choice and construction of variables, the mathematical form of the empirical model, assessment of goodness of fit and influential points, and the treatment of measurement error in predictor variables were all given consideration during the development of the empirical model, and are described in detail in this document.

## **G2.0 DESCRIPTION OF DATA**

The purpose of the statistical analysis was to provide a predictive model which relates childhood blood-lead concentration to measures of lead exposure from paint, dust and soil. Variables which represent lead exposure in environmental media were based on data that was available in both the HUD National Survey, and in the Rochester Lead-in-Dust study. Data from the HUD National Survey and from NHANES III are based on surveys that were designed to be nationally representative of the housing stock and the population of children, respectively. The HUD National Survey was a survey of pre-1980 housing that was adjusted using data from the 1993 American Housing Survey to represent the 1997 housing stock as described in the Section 403 Risk Assessment document. The Rochester Study was based on a targeted sample limited to a single geographic area as were other candidate epidemiological (epi) studies. It is unclear as to whether inferences drawn from any particular epi-study can be generalized to the national population of children and/or housing. Following is a brief discussion of each individual source of data, as well as a rationale and description of the variables that were included in the statistical analyses.

### **G2.1 SOURCES OF DATA**

#### **G2.1.1 Rochester Lead-in-Dust Study**

The Rochester Lead-in-Dust Study is a cross sectional study which recruited 205 children from live births at three local hospitals using a stratified sampling scheme. The sampling scheme was designed to recruit a high proportion of low income families living in older (pre 1940) housing. Blood-lead and hand-lead sample collection from recruited children occurred between August 31 and November 20, 1993. A detailed questionnaire was also completed at the time of blood sample collection. Environmental assessment of the primary residence of each recruited child was generally completed within three weeks of the date of blood sample collection, and included samples of dust from floors, window sills and wells, samples of soil from the dripline adjacent to the foundation and the child's play area, and measurements of painted interior and exterior surfaces (condition of paint and XRF paint lead loading).

#### **G2.1.2 HUD National Survey**

The HUD National Survey collected environmental samples of paint, dust, and soil from 284 private homes between 1989 and 1990. The objective of the study was to obtain data for estimating the prevalence of lead-based paint and lead-contaminated dust and soil in the nation. The presence or absence of children with elevated blood-lead was not part of the sampling design. One floor-dust sample was collected from each of three rooms, and one window sill and window well sample was collected from each of two rooms using a blue nozzle vacuum sampler. Three soil samples were collected from the dripline, entryway and remote locations. Paint sampling included XRF measures of paint-lead loading and condition of paint from generally two interior rooms and one side on the exterior of each residential unit.

In the HUD National Survey, each unit was assigned a sampling weight equal to the number of pre-1980 privately-owned, occupied units in the national housing stock that were represented by the given unit in the survey. The total of all 284 sampling weights equaled the number of pre-1980 privately-owned, occupied units in the national housing stock at the time of the survey. Sampling weights in the National Survey were determined according to four demographic variables associated with the units:

- ! Age category of unit
- ! Number of units in the building
- ! Census region
- ! Presence of a child under age 7 years

Since EPA's Risk Assessment uses 1997 as a base year for Section 403 activities, it was desirable to use the environmental-lead levels from the National Survey to characterize environmental-lead levels in the 1997 national housing stock. Therefore the sampling weights of National Survey units were revised to represent the 1997 occupied housing stock. The revised weights indicate the number of units in the 1997 national housing stock that are associated with the given National Survey unit, and therefore, with its distribution of environmental-lead levels.

### **G2.1.3 National Health and Nutritional Educational Survey (NHANES) III**

The Third National Health and Nutrition Examination Survey (NHANES III), conducted from 1988 to 1994, was the seventh in a series of national examination studies conducted by CDC's National Center for Health Statistics (NCHS) to trace the health and nutritional status of the non-institutionalized, civilian U. S. population. The target population for NHANES III included the civilian non-institutionalized population 2 months of age and older.

To provide for a nationally representative sample and sufficient precision in characterizing key subgroups, a complex survey design was employed in NHANES III. Approximately 40,000 persons were sampled in NHANES III, including approximately 3,000 children aged 1 to 2 years. Although estimates of national population health and nutrition parameters were the primary objectives of the survey, suitably precise estimates for certain age and race groups were obtained through over sampling. As a result, the NHANES III provides a solid basis for obtaining national estimates of the distribution of childhood blood-lead concentrations. Details on the study design and how the survey was conducted are available from CDC, 1992 and CDC, 1994.

### **G2.1.4 Other Candidate Epi Studies Considered**

There are various other epi studies that were potential data sources on which to base the empirical model. Given time and budget constraints the goal for the empirical model development could not include construction of the best possible model based on multiple data sources. Rather the goal was to develop a model that could be used to give an approximation of expected blood-lead concentrations related to residential lead based on a single source of data. The Rochester Study was chosen because of the following advantages:

1. All media, locations, and surfaces that are being considered for Section 403 standards were measured for lead in the Rochester Study.
2. The Rochester Study includes dust-lead loadings from wipe sampling and the Section 403 dust standard is expected to be based on dust-lead loading from wipe sampling.
3. The selection of homes and children in the Rochester Study, although targeted, was more random and more representative of a general population than is the case with most recent epidemiological studies of lead exposure in non-smelter communities.
4. The Rochester Study is recent.

The primary limitation associated with the Rochester Study is concern over the degree to which the Rochester Study may be considered representative of the nation as a whole. The limitations of the Rochester Study are discussed in more detail in Section G8.

Other data sets considered for use in constructing the empirical model included:

1. Pre-intervention data from the Baltimore Repair and Maintenance (R&M) Study. The R&M Study is a prospective longitudinal study which was designed to investigate the potential health and environmental benefits associated with performing R&M interventions on urban housing with lead-paint hazards. The pre-intervention sample included 115 children living in 87 homes. Samples of blood were collected from each participating child, and samples of dust, soil and water were collected from each house during the pre-intervention campaign. Due to the fact that the housing stock in this study consisted primarily of Baltimore City rowhouses, only 42 children living in 29 homes had soil samples. The absence of measures of lead in soil would have limited the use of this data in the development of an empirical model focused on all three media: paint, dust and soil.
2. Pre-intervention data from the Boston Soil Lead Abatement Demonstration Project. The Boston 3-City Study recruited 152 children living in 101 houses from four different urban neighborhoods during the pre-intervention campaign. The main restrictions for recruitment into the study were that the children had to be under the age of 5 and have an initial blood-lead concentration between 7 and 24  $\mu\text{g}/\text{dL}$ . For each household recruited into the study, a detailed environmental assessment was conducted concurrently with the blood-sampling. This environmental assessment included the collection of samples from paint, dust, soil and water. All dust samples from the Boston 3-City Study were collected using the Sirchee-Spitler Method. This method entails the use of a modified Black & Decker Dustbuster vacuum, and its properties with respect to other sampling methods are not well understood at the current time. Collection of a handwipe sample from each participating child and the completion of a questionnaire was also conducted with each blood sample.

- The restricted range of blood-lead concentrations recruited into this study was likely to have a large impact on parameter estimates of the relationships under investigation, and therefore, this source of data was not considered optimal for use in developing the empirical model.
3. Pre-intervention data from the Baltimore Soil Lead Abatement Demonstration Project. The Baltimore 3-City Study recruited 402 children living in 204 houses from two different urban neighborhoods during three rounds of pre-intervention sampling. There were no restrictions on the blood-lead concentration of children recruited into the study, however children had to be under the age of seven. For each household recruited into the study, a detailed environmental assessment was conducted once during the pre-intervention campaign. This environmental assessment included the collection of samples from dust, soil, exterior paint, and water. The Baltimore 3-City Study did not include samples of lead in paint or dust from window sills or window wells. Samples of interior paint were collected after the soil abatement intervention took place. In addition, all dust samples from the Baltimore 3-City Study were collected using the Sirchee-Spitler Method, and its properties with respect to other sampling methods are not well understood at the current time. Therefore, this source of data was not considered optimal for use in developing the empirical model.
  4. Pre-intervention data from the Cincinnati Soil Lead Abatement Demonstration Project. The Cincinnati 3-City Study included 201 children living in 129 houses from six different urban neighborhoods in the first (pre-intervention) phase of the study. The households recruited into the study were mostly single family residential units within multi-unit apartment buildings. It was believed that lead-based paint was removed from participating residential units in the early 1970's as part of a housing rehabilitation project. The pre-intervention environmental assessment consisted of the collection of interior and exterior dust and paint from each participating residential unit, and samples of soil from neighborhood recreation areas such as parks and playgrounds. Dust samples were collected using the DVM sampling method. Soil abatement was performed on a neighborhood scale, in parks, play areas, and other common grounds. Exterior dust was also removed from the neighborhood streets, alleys, and sidewalks as part of the intervention. Since soil samples could not be related to individual residences, this source of data was not considered optimal for use in developing the empirical model.
  5. Data from the Cincinnati Longitudinal Study. The Cincinnati Longitudinal Study is a prospective study which followed a cohort of several hundred children from birth to five years of age. It was designed to assess the impact of urban lead exposure on children's blood-lead concentrations. Once a year, blood-lead and hand lead samples were collected from each participating child. Progress in social, behavioral and cognitive development for each child was also measured over the course of the study. Environmental samples which included interior surface dust, XRF paint and exterior surface scrapings were collected from the residences of each participating child at approximately the same time as blood sample collection. There was also a qualitative

housing evaluation that was conducted for each residence included in the study. The Cincinnati Longitudinal Study provides data on the relationship between blood-lead and environmental lead over time. Although it is uncertain as to whether the exterior surface scrapings are representative of exterior dust or soil (or both), it appeared as though the Cincinnati Longitudinal Study was a good potential source of data for the empirical model; however these data have not yet been publicly released by the University of Cincinnati.

6. Data from the HUD Lead-Based Paint Hazard Control Grant Program in Private Housing (HUD Grantee data). HUD has provided grants to states and units of general local government (Grantees) for environmental interventions in privately owned low- and moderate-income housing. HUD requires Grantees to conduct dust-wipe testing and blood testing prior to environmental intervention. Paint and soil sampling are optional. Data from this program was not available for analysis at the time of preparation of the empirical model.

## **G2.2 VARIABLES UNDER CONSIDERATION**

Following is a rationale and description of the variables that were most closely examined for inclusion in the empirical model. These variables represent a subset of all the variables originally considered. They were selected based on several properties, including strength of association with blood-lead concentration in bivariate models, predictive power when included into a model with competing sources of lead exposure, interpretation, ability to construct the variable across different sources of data, and applicability to data collected by a standard Section 402 risk assessment.

The criteria used for the selection of variables in the empirical model emphasized use of measures of environmental lead and other factors observed in both the Rochester Lead-in-Dust Study and the HUD National Survey. Variables whose definition provided a convenient translation when applied to the National Survey, whose predictive power in Rochester were high, and whose spread in the National Survey populations covered a wide enough range of values, were used in the empirical model.

The first group of variables are subject specific, constructed from measurements on each child recruited into the empirical studies. The second group of variables are property specific, representing observations from the primary residences of each of the subjects. Because the Rochester Study included only one child per household, all of the variables measured in this statistical analysis can be organized using an identifier for household, represented by the subscript, *i*, throughout this document.

### **G2.2.1 Subject Specific Variables**

Table G-1 gives descriptions of the subject-specific variables: blood-lead concentration, age, pica and race.

**Table G-1. Subject-Specific Variable Descriptions**

Variable	Description
Blood-Lead	<p>Blood lead concentration on a venous sample is reported in units of micrograms of lead per deciliter (<math>\mu\text{g}/\text{dL}</math>). Because the distribution of blood-lead concentration is usually skewed, a natural log transformation was applied to blood lead concentration for use as a response variable in the statistical models. The natural log transformation helps the distribution of observed blood-lead levels meet normality assumptions required by the statistical models.</p> <p><math>\text{LPb}_i</math> = Natural log of the blood lead concentration measured from the <math>i</math>th child.</p>
Pica	<p>It has been hypothesized that sources of lead exposure in environmental media influence blood-lead concentration as a function of the hand-to-mouth activity or mouthing behavior of the child. A child who exhibits "strong" mouthing behavior or pica may be at higher risk for attaining an elevated blood-lead concentration. The following two questions were included in the Rochester Lead-in-Dust study as part of the questionnaire, and were designed to measure mouthing behavior or pica tendencies in children: (1) How often does the child put paint chips in his/her mouth?, and (2) How often does the child put dirt or sand into his/her mouth? The following choices were given as a possible response to these questions.</p> <p>0 Never                      3 Often 1 Rarely                      4 Always 2 Sometimes</p> <p>The following Pica variables were constructed based on the parental/guardian responses to the above two questions:</p> <p>Paint <math>\text{Pica}_i</math> = Tendency of the <math>i</math>th child to put paint chips in the mouth (on a scale of 0 to 4). Soil <math>\text{Pica}_i</math> = Tendency of the <math>i</math>th child to put dirt or sand in the mouth (on a scale of 0 to 4).</p>
Age	<p>Age has been documented as having a nonlinear effect on blood lead concentration when children are young (CDC, 1991). Therefore the age of each subject (in years) measured at the time of blood sampling was considered as a potential covariate in the statistical analysis.</p> <p><math>\text{Age}_i</math> = Age (continuous measure in years) of the <math>i</math>th child.</p>
Race	<p>It is quite possible that there are biological, cultural and/or behavioral differences among children recruited into the Rochester study that cannot be explained by any of the other measured variables barring race. Indicator variables representing race were therefore explored as covariates for the statistical analyses:</p> <p>White<math>_i</math> = 1 if the <math>i</math>th child is Caucasian.           = 0 Otherwise Black<math>_i</math> = 1 if the <math>i</math>th child is of African American descent.           = 0 Otherwise Other<math>_i</math> = 1 if the <math>i</math>th child is not Caucasian or not African American.           = 0 Otherwise</p>

### G2.2.2 Property Specific Variables

The property specific variables that were investigated in this statistical analysis correspond to measures of lead exposure from paint, dust and soil. There are many different ways of constructing lead exposure variables from the various different samples that were collected from

each environmental media. The variables discussed below represent one way of characterizing lead levels in environmental media.

**Table G-2. Property-Specific (Dust and Soil) Variable Descriptions**

Exposure	Description
<p>Paint (75th Percentile)</p>	<p>Interior and exterior paint lead loading was measured on multiple different painted surfaces within each residential unit using portable XRF instruments. Usually the condition of the paint was also measured for each painted surface that was sampled. Several variables were constructed using a combination of observed paint lead loadings and condition of the paint from both the interior and exterior of each residential unit. Two variables were chosen for the statistical analyses, which represent the presence and severity of deteriorated interior and exterior lead-based paint. The following formula describes the construction of the paint-lead variables, and was applied separately for interior and exterior paint samples within each residential unit:</p> <p>Let <math>XRF_{ij}</math> represent the observed paint lead loading (<math>mg/cm^2</math>) from the <math>j</math>th component within the <math>i</math>th residential unit, if the XRF value was greater than or equal to <math>1 mg/cm^2</math>. An observed XRF paint-lead loading greater than or equal to one is considered lead-based paint. If the observed paint lead loading was less than <math>1 mg/cm^2</math>, <math>XRF_{ij}</math> is equal to zero.</p> <p>Condition of the paint is characterized as Good whenever less than 5% of the surface is deteriorated; Fair whenever 5% to 15% of the surface is deteriorated; and Poor whenever more than 15% of the surface is deteriorated. By combining categories, let <math>Cond_{ij}</math> represent the condition of the paint on the <math>j</math>th component within the <math>i</math>th residential unit; <math>Cond_{ij}</math> is equal to one if the surface was rated Fair or Poor, and is equal to zero if it was rated Good. Then we have a measure of deteriorated LBP, which is given by <math>DETLBP_{ij} = XRF_{ij} \cdot Cond_{ij}</math></p> <p><math>Paint_i</math> is defined as the 75th percentile of the <math>j</math> observed levels of <math>DETLBP_{ij}</math>. It is a variable which represents the presence and severity of deteriorated lead-based paint within a residential unit. Residential units in which less than 25% of the sampled painted surfaces had deteriorated lead-based paint result in a <math>DETLBP_{ij}</math> value that is equal to zero. Residential units with 25% or more of the sampled painted surfaces having deteriorated lead-based paint result in <math>DETLBP_{ij}</math> values that are greater than or equal to one.</p> <p><math>Int\_pnt_i = Paint_i</math> based on interior painted surfaces.  <math>Ext\_pnt_i = Paint_i</math> based on exterior painted surfaces.</p>
<p>Paint/Pica Hazard</p>	<p>An additional paint variable combined paint condition, lead-based paint and pica. An indicator variable which was nonzero whenever each of the following conditions existed in a residential unit: presence of deteriorated or damaged interior paint in the household; and presence of interior lead-based paint in the household; and presence of a child with paint pica in the household.</p> <p>The paint variable had values of:</p> <ul style="list-style-type: none"> <li>0 No LBP (XRF reading <math>&lt; 1</math>), or condition is Good, or child does not exhibit paint pica;</li> <li>1 LBP (XRF reading <math>\geq 1</math>), condition is Fair or Poor, and child exhibits paint pica rarely;</li> <li>2 LBP (XRF reading <math>\geq 1</math>), condition is Fair or Poor, and child exhibits paint pica at least sometimes.</li> </ul> <p>In the Rochester Study, a child's tendency towards paint pica was characterized as:  0 = Never, 1 = Barely, 2 = Sometimes, 3 = Often and 4 = Always.</p> <p>Because of limited sample size in each category, Paint pica was collapsed for this modeling to have values: 0 = No paint pica, 1 = Child exhibits paint pica rarely, and 2 = Child exhibits paint pica at least sometimes.</p> <p>A value of 1.5 was chosen as the input value for those children exhibiting pica at least rarely in applying the empirical model to the HUD National Survey. The average value of this pica variable for children who exhibited any pica in the Rochester Study was 1.25</p>

**Table G-2. Property-Specific (Dust and Soil) Variable Descriptions (Continued)**

Exposure	Description
<p>Floor Dust Combined With Proportion of Carpeted/ Uncarpeted Surfaces</p>	<p>There were residential units in which all floor surfaces that were sampled were either carpeted or uncarpeted, resulting in missing values for the variables Floor_C<sub>i</sub> or Floor_U<sub>i</sub>. A second set of floor-dust exposure variables were therefore pursued in an effort to recapture residential units with missing values.</p> <p>Let PC<sub>i</sub> represent the proportion of floor dust samples collected from carpeted surfaces within the i<sup>th</sup> house: <math>PC_i = \frac{[\text{Number of carpeted floor surfaces}]_i}{[\text{Total number of floor surfaces sampled}]_i}</math></p> <p>Then Carp_flr<sub>i</sub> = Floor_C<sub>i</sub> * PC<sub>i</sub>, and            Bare_flr<sub>i</sub> = Floor_U<sub>i</sub> * (1-PC<sub>i</sub>) where            Carp_flr<sub>i</sub> represents the area weighted arithmetic average dust-lead loading from carpeted floors multiplied by the proportion of floor dust samples that were collected from carpeted surfaces in the i<sup>th</sup> residential unit. Note that Carp_flr<sub>i</sub> is equal to zero for residential units that had no carpeted surfaces sampled.            Bare_flr<sub>i</sub> represents the area weighted arithmetic average dust-lead loading from uncarpeted floors multiplied by the proportion of floor dust samples that were collected from uncarpeted surfaces in the i<sup>th</sup> residential unit. Note that Bare_flr<sub>i</sub> is equal to zero for residential units that had no uncarpeted surfaces sampled.</p>
<p>Dust (Window Trough, Window Sill and Floor)</p>	<p>Samples of interior household dust were collected from floors, window sills and window wells from residential units in the Rochester Study. Dust samples were collected using both wipe and vacuum samples, thus measures of dust-lead loading were available for all dust samples, and measures of dust-lead concentration are available for those dust samples that were collected using vacuum samples. Variables were constructed which represent the area weighted arithmetic average dust-lead loading and the mass weighted arithmetic average dust-lead concentration for each component type tested within each residential unit. Due to a lack of understanding of potential differences between the exposure mechanism between carpeted and uncarpeted surfaces, floor dust samples collected from carpeted and uncarpeted surfaces were treated as separate component types in the construction of variables. An initial assessment comparing dust-lead loading variables to dust-lead concentration variables (for samples collected using vacuum sampling) in the Rochester Lead-in-Dust Study demonstrated that the lead-loading variables were consistently stronger predictors of blood-lead concentrations. In addition, it is expected that dust standards will be specified in terms of dust-lead loading from wipe samples. Therefore, the following measures of wipe dust-lead loading were considered as potential variables in the predictive model:</p> <p>Floor_A<sub>i</sub> represents the area weighted arithmetic average dust-lead loading from all surface (carpeted or uncarpeted) floors in the i<sup>th</sup> residential unit.            Floor_C<sub>i</sub> represents the area weighted arithmetic average dust-lead loading from carpeted floors in the i<sup>th</sup> residential unit.            Floor_U<sub>i</sub> represents the area weighted arithmetic average dust-lead loading from uncarpeted floors in the i<sup>th</sup> residential unit.            W_Sill<sub>i</sub> represents the area weighted arithmetic average dust-lead loading from window sills in the i<sup>th</sup> residential unit.            W_Well<sub>i</sub> represents the area weighted arithmetic average dust-lead loading from window wells in the i<sup>th</sup> residential.</p>

**Table G-2. Property-Specific (Dust and Soil) Variable Descriptions (Continued)**

Exposure	Description
Soil	<p>Composite samples of soil were collected using a coring tool from several different locations within the yard of each residential unit. In the Rochester Lead-in-Dust Study, the laboratory analysis of the composite soil samples resulted in measures of soil-lead concentration (<math>\mu\text{g/g}</math>) for a fine soil fraction and a coarse soil fraction. An initial assessment of the soil-lead data from the Rochester Lead-in-Dust Study Data showed no statistically significant difference in predictive power between the fine and coarse soil fractions. Soil samples usually undergo some degree of sieving (note the HUD Guidelines protocol for soil sampling, Appendix 13.3, page App 13.3-3). Historically, the fine soil fraction has been used as a predictor variable in lead exposure studies, because it was thought that the fine-soil fraction is more bioavailable to children. We therefore considered only the fine-soil fraction in the statistical analyses. The following soil-lead exposure variables were considered as potential predictor variables in the statistical models:</p> <p>Drip_Soil<sub>i</sub> represents the observed lead concentration in a composite soil sample collected from the dripline (adjacent to the foundation) of the <i>i</i>th home.</p> <p>Play_Soil<sub>i</sub> represents the observed lead concentration in a composite soil sample collected from the play area of the <i>i</i>th home. Note that Play_Soil<sub>i</sub> could be considered a subject specific variable.</p>

### G3.0 FORMS OF THE STATISTICAL MODELS

This section contains a discussion of the different forms of mathematical models considered for characterizing the relationship between blood-lead and measures of lead exposure that were considered as part of the modeling effort. The following five mathematical model forms were investigated for the development of a multi-exposure predictive model for childhood blood-lead concentrations. Each model is individually discussed in terms of statistical assumptions, biological/physical assumptions, and mathematical ease of use. Although biological/physical plausibility is an important issue, the objective of the empirical Model was to predict a rational distribution of blood-lead concentrations. Thus, the primary basis for choosing a model was based on predictive ability. It should be noted that there is currently no definitive model accepted by the scientific community for the relationship between childhood blood-lead and environmental-lead. The final form of the empirical model is presented in Section G6.

#### G3.1 LOG-LINEAR MODEL

The log-linear model expresses natural-log transformed blood-lead concentration as a linear combination of natural-log transformed exposure variables and select covariates. A multimedia exposure log-linear model for blood-lead concentrations (in generic form) would appear as follows:

$$\ln(\text{PbB}_i) = \beta_0 + \beta_1 \cdot \ln(\text{Dust}_i) + \beta_2 \cdot \ln(\text{Soil}_i) + \beta_3 \cdot \ln(\text{Paint}_i) + \gamma \cdot \text{Covariate}_i + e_i$$

where  $e_i$  (the residual error) is assumed to follow a normal distribution with mean zero and variance  $\sigma_{\text{Error}}^2$ .

One main advantage of the log-linear model is its mathematical convenience. The log-linear model is easily fitted using standard linear regression methods (although in the development of a multiple-exposure model it may be necessary to fit the log-linear model using a numerical approximation method while constraining parameter estimates for exposure variables to positive values; i.e.  $\beta_1, \beta_2,$  and  $\beta_3 \geq 0$ ). Another mathematical convenience of the log-linear model is the fact that calculation of tolerance intervals and exceedance proportions, and adjusting for the effects of measurement error in predictor variables is relatively straight-forward.

With respect to biological/physical assumptions, the log-linear model when translated back into the original scale of observed blood-lead concentrations, results in a multiplicative relationship for environmental-lead:

$$\text{PbB}_i = \exp(\beta_0) \cdot \text{Dust}_i^{\beta_1} \cdot \text{Soil}_i^{\beta_2} \cdot \text{Paint}_i^{\beta_3} \cdot \text{Covariate}_i^{\gamma} \cdot \exp(e_i)$$

Thus, the effect of dust-lead on blood-lead is dependant on the combined effects of all of the other variables included in the model. Furthermore, the difference in predicted blood-lead

concentration for children exposed to dust-lead loadings of 5 and 50  $\mu\text{g}/\text{ft}^2$  is the same as the difference in predicted blood-lead concentration for children exposed to dust-lead loadings of 500 and 5000  $\mu\text{g}/\text{ft}^2$ . Although the multiplicative interpretation of the log-linear model is not considered biologically/physically plausible, it often fits the data better than statistical models with a more plausible, biological/physical basis for data with low to moderately exposed children (Rust, et al., 1996).

### G3.2 LOG-ADDITIVE MODEL

Whereas the log-linear model when translated back to the original scale of measurement results in an assumed multiplicative relationship, the log-additive model results in an assumption of additivity among the exposure variables. The log-additive model expresses natural-log transformed blood-lead concentration as the natural-log of a linear combination of exposure variables and select covariates. A multimedia exposure log-additive model for blood-lead concentrations (in generic form) would appear as follows:

$$\ln(\text{PbB}_i) = \ln(\beta_0 + \beta_1 \cdot \text{Dust}_i + \beta_2 \cdot \text{Soil}_i + \beta_3 \cdot \text{Paint}_i + \gamma \cdot \text{Covariate}_i) + e_i$$

where  $e_i$  (the residual error) is assumed to follow a normal distribution with mean zero and variance  $\sigma_{\text{Error}}^2$ .

Since the response variable in the log-additive model is expressed as a non-linear function of the exposure variables, it must be fitted using a non-linear regression algorithm. Thus, the mathematical conveniences of the log-linear model do not apply to the log-additive model.

With respect to biological/physical assumptions, the log-linear model when translated back into the original scale of observed blood-lead concentrations, results in an additive relationship for environmental-lead:

$$\text{PbB}_i = (\beta_0 + \beta_1 \cdot \text{Dust}_i + \beta_2 \cdot \text{Soil}_i + \beta_3 \cdot \text{Paint}_i + \gamma \cdot \text{Covariate}_i) \cdot \exp(e_i)$$

Thus, the effect of each measure of environmental lead on blood-lead is not dependant on the combined effects of all of the other variables that were included in the model. The model is attractive in that it is reasonable and biologically plausible that the relationship between blood-lead and environmental lead would be additive at low levels of environmental exposure. However, there is also evidence that saturation of the effect of environmental lead on blood-lead concentration occurs at higher levels of lead exposure, in which case additivity may no longer hold.

### **G3.3 ALTERNATE LOG-ADDITIVE MODEL**

Although the additive interpretation of the log-additive model is more biologically plausible than the multiplicative interpretation of the log-linear model, the tendency of the log-additive model to over predict blood-lead at higher levels of environmental lead exposure may present a problem. One method for solving the problem is to use mathematically transformed measures of environmental-lead (such as the natural-log transformation) in the log-additive model. This “Alternate Log-Additive Model” would preserve the additivity property associated with the log-additive model, while also accounting for a saturation of the effect of environmental lead on blood-lead concentration at higher levels of lead exposure. A multimedia exposure version of an alternate log-additive model for blood-lead concentrations (in generic form) would appear as follows:

$$\ln(\text{PbB}_i) = \ln[\beta_0 + \beta_1 \cdot \ln(\text{Dust}_i) + \beta_2 \cdot \ln(\text{Soil}_i) + \beta_3 \cdot \ln(\text{Paint}_i) + \gamma \cdot \text{Covariate}_i] + e_i$$

where  $e_i$  (the residual error) is assumed to follow a normal distribution with mean zero and variance  $\sigma^2_{\text{Error}}$ .

The alternate log-additive model must also be fitted using non-linear regression, and therefore the alternate log-additive model does not have the same mathematical conveniences that are associated with the log-linear model. When using the alternate log-additive model, particular attention should be paid to the mathematical transformation that is applied to the environmental lead exposure variables. A transformation that is too strong may result in a model in which the effect of saturation at high environmental-lead levels is over-predicted, resulting in a model which under-predicts blood lead.

### **G3.4 ACTIVE/PASSIVE UPTAKE MODEL**

Another method of adjusting the log-additive model to compensate for saturation of the response at high levels of environmental lead is to parameterize the saturation effect itself. The following “Active/Passive Uptake” Model demonstrates one method for parameterizing the saturation effect:

Let  $\text{Exposure}_i$  represent a linear combination of the exposure variables (on the original scale) similar to the linear combination that appears inside the natural-log function in the log-additive model;

$$\text{Exposure}_i = \beta_0 + \beta_1 \cdot \text{Dust}_i + \beta_2 \cdot \text{Soil}_i + \beta_3 \cdot \text{Paint}_i + \gamma \cdot \text{Covariate}_i$$

The Active/Passive Uptake Model is then expressed as:

$$\ln(\text{PbB}_i) = \ln(\text{Exposure}_i) + \ln\left( F_{\text{Passive}} + \frac{1 - F_{\text{Passive}}}{1 + \frac{\text{Exposure}_i}{\theta}} \right) + e_i$$

where  $0 \leq F_{\text{Passive}} \leq 1$  and  $0 < \theta$

Figure G-1 provides a plot of blood-lead concentration as a function of  $\text{Exposure}_i$ , assuming that  $\theta = 10 \mu\text{g/dL}$  and that  $F_{\text{Passive}}$  takes on values of 0, 0.05, 0.1, 0.5 and 1. The plot shows that when  $F_{\text{Passive}}$  is equal to zero,  $\theta = 10 \mu\text{g/dL}$  provides an asymptote for the maximum blood-lead concentration that is predicted as a function of  $\text{Exposure}_i$ . In the Active/Passive Uptake Model,  $F_{\text{Passive}}$  represents the portion of  $\text{Exposure}_i$  which has a linear effect on blood-lead concentration beyond the saturation point of  $\theta(1 - F_{\text{Passive}})$ . When  $F_{\text{Passive}}$  equals 1, the Active/Passive Uptake Model is identical to the log-additive model, and therefore does not compensate for saturation of the response at high levels of exposure.

Advantages include biological/physical plausibility, goodness of fit relative to other candidate models (as is seen in the tables of Section G13) and the fact that this model is similar in nature to the relationship modeled within the IEUBK model. Disadvantages include the fact that this model may overparameterize the relationship between blood-lead and environmental lead in these data. Also, the active/passive uptake model does not have the same mathematical conveniences associated with the log-linear model.

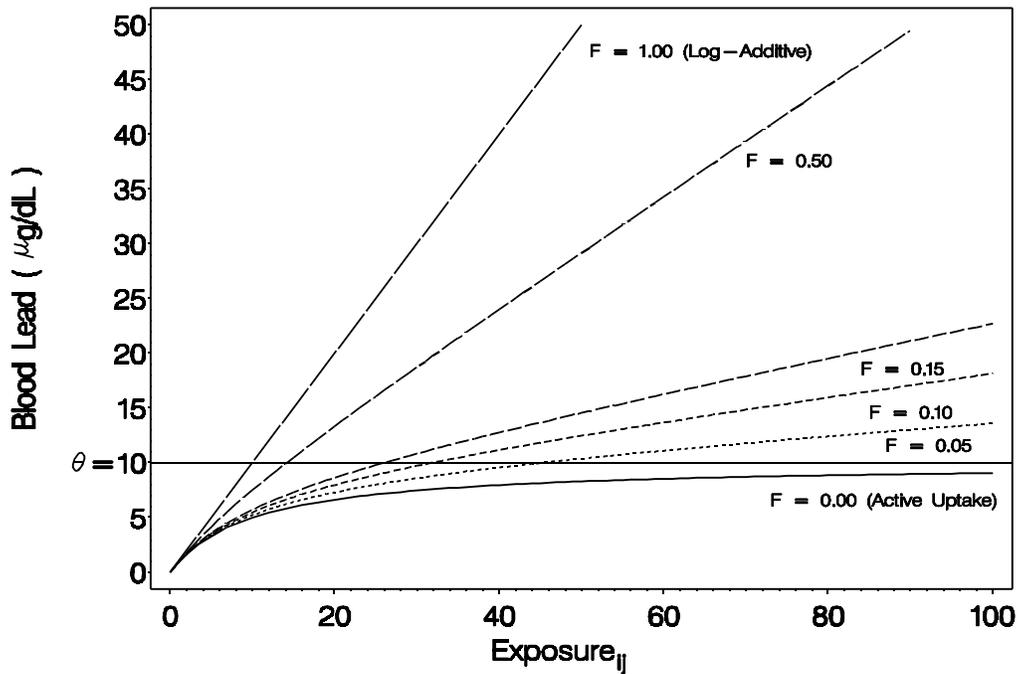


Figure G-1. Plot of Blood-Lead Concentration as a Function of Lead Exposure Using the Active/Passive Uptake Model with  $\theta = 10 \mu\text{g/dL}$  and  $F_{\text{Passive}}$  Ranging from Zero to One.

### G3.5 ACTIVE UPTAKE MODEL

The Active Uptake Model is simply a reduced form of the Active/Passive Uptake model in which the parameter  $F_{\text{Passive}}$  is held fixed at zero. This model includes properties similar to the Active/Passive Uptake model, and may in some cases provide more interpretable parameter estimates for situations in which the Active/Passive Uptake Model is overparameterized.

## **G4.0 MEASUREMENT ERROR**

The fact that the lead predictor variables for paint, dust and soil are subject to measurement error raises issues about the need to account for this measurement error in the model building process. In addition, the fact that different sampling methods were used in the Rochester Study and the HUD National Survey also raises issues about similar adjustments for different sampling methods when applying the empirical model to the HUD National Survey Data. Choosing an appropriate statistical methodology for adjustment is dependant on several factors including use of the model, interpretation of the predictor variables, the definition of the components of measurement error in the predictor variables, and the mathematical form of the model relating blood-lead to environmental lead.

Sections G4.1, G4.2, and G4.3 of this appendix discuss, respectively, the questions of:

1. what is being measured (and modeled) in the empirical model;
2. what adjustment for the effects of measurement error in predictor variables or differences in sampling methods is appropriate (with respect to Section 403 rulemaking activities);
3. the definition and characterization of measurement error associated with dust predictor variable.

### **G4.1 WHAT IS BEING MEASURED (AND MODELED)**

The purpose of the empirical model is to assess the changes in the distribution of blood-lead levels of children one to two years old that are likely to result from the application of the 403 Rule standards. The vehicle for application of the 403 Rule standards is a risk assessment conducted in accordance with the work practice standards in the 402 Rule and following the detailed approach for risk assessments in the 1995 HUD Guidelines. Accordingly, the multi-media model defined in this document seeks to establish a relationship between children's blood-lead levels and environmental-lead levels as would be measured in a risk assessment. Environmental and blood-lead data from the Rochester Lead-in-Dust Study provided the means to develop the multi-media model. The relationship between blood-lead and environmental-lead observed in the Rochester Study was to be applied to environmental inputs from the HUD National Survey (with weights adjusted to represent housing in 1997). In most cases, environmental variables included in the multi-media model based on the Rochester data were constructed similarly using environmental-lead levels observed in the HUD National Survey. However, dust and soil measures were sufficiently different between these two studies, and a statistical adjustment procedure had to be developed to allow dust-lead and soil-lead measures from the HUD National Survey to be properly used as inputs to the model. This adjusted relationship between blood-lead and environmental-lead as observed in the HUD National Survey results in what this document refers to as the empirical model. For development of 403 Rule standards, the empirical model will be used to assess different options for the standards, and the resulting changes in the children's blood-lead distribution will be assessed to estimate the benefits

of the various options. Proper development of the empirical model requires attention to and balancing of three key features: 1) how environmental lead measurements are made in a risk assessment, 2) how environmental measurements were made in the Rochester Study and in the HUD National Survey, and 3) in the Rochester Study, what environmental measures are strong predictors of children's blood-lead levels.

The lead exposure variables used in the statistical model(s) were constructed from measured levels of lead in various different samples of paint, dust, or soil from the primary residence of each subject child. Protocols for environmental sampling were used in each study to assure that the measures of lead in environmental media were consistent across the various different houses recruited into that particular study. These protocols required detailed sampling in an effort to characterize the levels of lead in paint, dust, and/or soil at the time of sampling. The collection of environmental samples from a child's primary residence usually occurred within a few weeks of the collection of that child's blood-sample.

In the variable selection phase of the statistical analysis, various different ways of combining the lead-loading (or lead concentration) of same media samples within a residence into a lead exposure variable were investigated in terms of (1) their association with blood-lead in bivariate model(s), (2) their association with blood-lead in multimedia exposure model(s), and (3) ease of interpretation. In each case, the resulting variable was designed to characterize the child's exposure to lead in paint, dust, or soil *from the primary residence*.

Although a child's blood-lead concentration is a product of cumulative exposure to lead, most of the available data from the lead exposure studies only provide information on the lead levels in environmental media at one point in time. Thus, the lead exposure variables that were constructed for use in the statistical models represent an estimate of the child's exposure to lead from paint, dust or soil from the primary residence at the time of sampling. The exposure variables (environmental lead) characterize current exposure to lead, rather than cumulative exposure to lead, whereas the response variable (blood-lead) is a measure of cumulative exposure. These exposure variables, including dust-wipe lead loadings, are similar to the measures that would be collected in a standard Section 402 risk assessment.

Therefore, the empirical model provides an estimate of the relationship between childhood blood-lead concentrations (indicative of a child's cumulative exposure to lead) and sampled measurements of lead from paint, dust, or soil *from the primary residence at the time of sampling*. Further discussion of the decision to focus on exposure from the primary residence at the time of sampling is provided in Section G4.3 below in the sections on spatial and temporal variability.

## **G4.2 WHAT ADJUSTMENT FOR MEASUREMENT ERROR IS APPROPRIATE**

The first question to be asked when addressing measurement error is: Is an adjustment for measurement error necessary? The appropriateness of an adjustment for measurement error is dependent on the use of the statistical model.

### **G4.2.1 Errors-In-Variables Adjustment To Model "True" Lead Exposure**

A primary differentiation in model use concerns whether the model is being used to characterize the relationship between observed blood-lead levels in children and “true” lead exposures or whether the model is being used to predict blood-lead concentrations based on measured levels of environmental lead. The former case is the classic measurement error problem (Carroll, et al., 1995). Although this case may be of interest to EPA in documenting the extent of the lead problem, the primary use of the empirical model in the Section 403 rulemaking is for the latter case, prediction.

*Therefore, because the empirical model is not intended to be used as a dose-response model, but rather is intended to be used to predict blood-lead levels based on measured levels of environmental lead, a classic errors-in-variables approach that would model the relationship between "true" lead exposure and children's blood lead concentrations was considered inappropriate for this analysis.*

### **G4.2.2 Adjustment To Account For Differences In Measurement Error Between Dust Sampling Methods Used In The Rochester Study And Those Used In The HUD National Survey**

In order to predict the national distribution of childhood blood-lead concentrations (prior to, and following implementation of Section 403 rules), the empirical model based on the Rochester Study must be combined with environmental data observed in a nationally representative sample (the HUD National Survey). As mentioned earlier, the dust and soil sampling methods were different between these two studies and therefore an adjustment for both systematic differences and differences in measurement error between the Rochester dust-lead and soil-lead predictor variables and the HUD National Survey dust-lead and soil-lead predictor variables must be considered.

An empirical model unadjusted for the effects of differences in the lead exposure predictor variables would be appropriate for prediction of the national distribution of blood-lead concentrations (prior to, and following Section 403 interventions) if the following four assumptions are met:

1. The sampling scheme for environmental lead implemented in the Rochester Lead-in-Dust Study (or other studies used for model building) is similar to the sampling scheme implemented in the HUD National Survey.

2. The sampling collection devices and instruments used to measure lead have similar properties with respect to measurement error between the Rochester Study and the HUD National Survey.
3. The distribution of observed environmental lead levels is similar between the Rochester Lead-in-Dust Study and the HUD National Survey.
4. The characteristics of the relationship between blood-lead and environmental lead in Rochester is the same as in the U.S. as a whole.

If either of the first two assumptions are not met, it would be necessary to adjust the model for differences in measurement error between variables constructed using the Rochester data and variables constructed using the HUD National Survey data. Although this can be considered an adjustment for “measurement error,” the resulting model would not be interpreted as the relationship between blood-lead and “true” environmental lead levels (measured without error). Rather, this adjustment will account for differences in variability related to the different sampling methods to facilitate a more accurate prediction of the national distribution of childhood blood-lead concentrations.

If the third or fourth assumptions are not met, it raises the question as to whether the data from the Rochester Lead-in-Dust Study is an appropriate source of data for informing decisions concerning lead exposures nationwide.

Initial investigation of the data suggested that the first two assumptions were not met by the observed data in the two studies; and therefore, *an adjustment for the differences between dust-lead and soil-lead predictor variables used in the model building process and dust-lead and soil-lead input variables from the HUD National Survey used in the prediction process is warranted.*

A related issue concerns the degree to which equation error (or an incorrect mathematical form of the model) can affect the accuracy and precision of model predictions. Measurement error and the form of the model are directly related in that the specific methodology for a measurement error adjustment is dependent on the form of the model.

#### **G4.3 DEFINITION AND CHARACTERIZATION OF MEASUREMENT ERROR ASSOCIATED WITH EACH PREDICTOR VARIABLE**

While it was determined that a classic adjustment for measurement error (Carroll, et al., 1995) was not appropriate for this particular use of the model, the statistical adjustment to the model for differences in sampling methods requires estimates of the variability associated with measuring the dust-lead and soil-lead exposure predictor variables. The following equation represents the three sources of variability that contribute to an estimate of measurement error in a dust-lead (or soil-lead) sample from the primary residence at the time of sampling and that are taken into account in the statistical adjustment to the model for differences in sampling methods:

$$\sigma^2_{\text{Measurement Error}} = \sigma^2_{\text{Spatial}} + \sigma^2_{\text{Sampling}} + \sigma^2_{\text{Laboratory}}$$

Another potential component of variability was temporal variability,  $\sigma^2_{\text{Temporal}}$ , but this component of variability was not included in any measurement error adjustments for reasons that are discussed below. The question of whether or not it is appropriate to consider any particular component of variation as part of the estimated measurement error for an exposure variable is dependant on the interpretation of the exposure variable and the way it is being used in the statistical model. Each component, including temporal variability, is discussed in the following subsections with respect to characterizing measurement error in the lead exposure predictor variables.

Details concerning the estimation of variability associated with measurement error in the dust-lead predictor variables are provided in Section G10.

#### **G4.3.1 Spatial Variability**

Spatial variability ( $\sigma^2_{\text{Spatial}}$ ) represents variability in environmental lead levels among all possible locations on the surface(s) being tested as part of the sampling scheme. Although an ideal lead exposure variable would characterize lead-levels from all the surfaces which are related to a child's lead exposure (both inside and outside of the primary residence), the environmental data corresponding to a subject's lead exposure is usually limited to the sampling schemes implemented during a study. (For residential risk assessments, it is limited to the sampling schemes specified by the Section 402 rule.) It is an assumption that the sampling schemes that were implemented in these studies provide a sample of environmental lead as would be obtained in a risk assessment.

Lead measures outside the primary residence are unlikely to be taken in a risk assessment. There appear to be two ways of viewing lead exposures that occur outside the primary residence (such as in a day care center):

1. Lead exposure that occurs outside the primary residence is not captured by the observed lead exposure variables. Outside exposure represents a group of covariates that are not included in the statistical models, and therefore,  $\sigma^2_{\text{Spatial}}$  would be limited to the variability of environmental lead that occurs among all possible locations within the primary residence.
2. Lead exposure that occurs outside the primary residence is captured by the observed lead exposure variables (measured within the primary residence), based on an assumption that levels of environmental lead inside the primary residence are similar to levels of lead found outside the primary residence. Under this assumption, the definition of  $\sigma^2_{\text{Spatial}}$  would be expanded to include the variability of environmental lead that occurs among all possible locations to which a child has been exposed (both inside and outside the primary residence).

We accepted the first viewpoint of spatial variability ( $\sigma^2_{\text{Spatial}}$ ) based on the following three facts:

1. There is no known information that can be used to verify the assumption that lead-levels in paint, dust, or soil within the primary residence are representative of lead-levels that occur outside the home.
2. There is no known information that can be used to estimate  $\sigma^2_{\text{Spatial}}$  under an expanded definition which includes all surfaces to which a child is exposed (both inside and outside of the primary residence). However, there is information that can be used to estimate spatial variability in environmental lead levels that occur within a primary residence.
3. Environmental interventions that will occur under Section 403 will likely be focussed on reducing residential exposure to lead. It may therefore be inappropriate to develop a model in which the predictor variables are interpreted in a way which represents exposure that occurs outside of the primary residence.

Spatial variability was taken into account in the statistical adjustments to the model for differences in dust and soil sampling methods.

#### **G4.3.2 Sampling Variability**

Sampling variability  $\sigma^2_{\text{Sampling}}$  represents variability introduced during the physical collection of environmental samples, and is a typical source of measurement error associated with the lead exposure predictor variables. Examples of variability that may be classified as sampling variability when collecting dust samples include:

- ! variability associated with sampling methods, e.g. wipe versus vacuum sampling
- ! variability associated with sampled surfaces, e.g. carpeted versus uncarpeted floors
- ! variability associated with properties of the given sample, e.g. particle size and dust-loading.

Examples of variability that may be classified as sampling variability when collecting soil samples include:

- ! variability associated with sampling methods, e.g. coring tool versus grab sample
- ! variability associated with sampled surfaces, e.g. bare soil versus covered soil
- ! variability associated with properties of the given sample, e.g. fraction of soil sample that is fine (versus coarse).

Sampling variability was taken into account in the statistical adjustments to the model for differences in dust and soil sampling methods.

### **G4.3.3 Laboratory Variability**

Laboratory variability ( $\sigma^2_{\text{Laboratory}}$ ) represents variability in the laboratory analysis of an environmental sample, and includes error in sample preparation and analytical error. It is often the case that laboratory error is a very small component of the total measurement error associated with a sample result.

Laboratory variability was taken into account in the statistical adjustments to the model for differences in laboratory methods for measuring lead in dust and soil samples.

### **G4.3.4 Temporal Variability**

Temporal variability ( $\sigma^2_{\text{Temporal}}$ ) represents the variability over time in environmental lead levels on the location(s) selected to be part of the sample. Although lead levels in paint may not be subject to substantial temporal variability, it is documented that lead levels in dust and soil vary over time.

Since we are interpreting the lead exposure variables as being representative of current lead exposure (as would be measured in a Section 402 Risk Assessment) rather than cumulative lead exposure, temporal variability in environmental lead levels was not taken into account in the statistical adjustments to the model for differences in dust and soil sampling methods.

## **G5.0 MODEL BUILDING BASED ON DATA FROM THE ROCHESTER STUDY**

This chapter describes the steps involved in the development of a multi-media predictive model based on data observed in the Rochester Lead-in-Dust Study. First, single media models of the Rochester data were investigated, then the variables identified from them were used to explore joint media models. Diagnostic analyses are described which were used to validate assumptions made during model development. Finally, information from these efforts was used to develop a multi-media predictive model based on data observed in the Rochester Study.

### **G5.1 USE OF SINGLE MEDIA MODELS (Bivariate Relationships Between Blood-Lead and Each Potential Variable)**

Statistical modeling of the data from the Rochester Lead-in-Dust Study began with an initial evaluation of the bivariate relationship between blood-lead concentration and each individual exposure variable or select covariate. This evaluation included an assessment of all five candidate statistical models discussed in Section G3.

Section G11 contains for each potential exposure variable constructed from the Rochester Lead-in-Dust Study Data, a figure which displays the estimated regression curve for each candidate statistical model plotted along with the observed data, as well as a table which summarizes parameter estimates and associated standard errors for each candidate model. Note that parameter estimates and associated standard errors for the active/passive uptake model are not included in the tables in Section G11, because in most cases, the  $F_{\text{Passive}}$  parameter was estimated as zero in the bivariate models, and thus, the active/passive uptake model reduces in form to the active uptake model. Candidate models and the strength of the relationship between blood-lead and each variable were compared using measures of  $R^2$  and estimated likelihood ratios.  $R^2$  (also called the coefficient of determination) is a measure of the proportion of the variability in childhood blood-lead concentrations that is explained by a model. Estimated likelihood ratios were calculated using parameter estimates from each model and the observed data. Use of the likelihood ratio as a diagnostic tool is discussed in Section G5.3 on regression diagnostics.

Results of the bivariate statistical analysis of the relationship between blood-lead concentration and each potential exposure variable from the Rochester Lead-in-Dust Study Data demonstrated the following:

1. The variables representing the presence and severity of interior deteriorating lead-based paint were significant predictors of blood-lead. The variables representing the presence and severity of exterior deteriorating lead-based paint were only borderline significant at the 0.05 level.
2. Measures of floor dust-lead loading from uncarpeted surfaces were better predictors of blood-lead than measures of floor dust-lead loading from carpeted surfaces.

3. Measures of dust-lead loading from window wells were better predictors of blood-lead than measures of dust-lead loading from window sills.
4. Both measures of soil-lead concentration (Dripline & Play-Area) were strong predictors of children's blood-lead concentration. Using Dripline soil Pb concentration (n=186) allowed more children/houses to enter the model versus Play-area (n=87).
5. Pica for paint chips was a significant predictor of blood-lead. Pica for soil was borderline significant.
6. The indicator variable representing race (black) was the strongest single predictor of blood-lead concentrations.
7. Age was not significantly associated with blood-lead in the Rochester data.

## **G5.2 DESCRIPTION OF JOINT MEDIA MODELS** **(Development of a Multimedia Exposure Statistical Model)**

After assessing the bivariate relationships with each variable under consideration, the variables were systematically evaluated in an effort to develop a parsimonious multimedia exposure model for each source of data. There were a number of technical issues involved in the fitting of these models, including variable selection, collinearity among environmental exposure variables, and details concerning the use of non-linear regression:

### **G5.2.1 Variable Selection and Collinearity**

Variable selection for the multimedia exposure model was based on several properties, including strength of relationship with blood-lead concentration as estimated using the bivariate statistical models, predictive power of each variable when included into a model with competing sources of lead exposure, and interpretability of the parameter estimates. Another goal related to variable selection was to develop a predictive model that was based on lead exposure from the three environmental media; paint, dust and soil. Thus, measures of lead exposure from paint, dust, and soil were considered as primary variables in the statistical analyses, and all other variables were considered as secondary variables. If a secondary variable was competing with a primary exposure variable in the multimedia exposure model (in terms of explaining variability in childhood blood-lead concentration), the secondary variable was excluded from the model in its final form.

Another issue in variable selection is the fact that the multimedia exposure models included variables which represent lead-levels in paint, dust, and soil from each residential unit. These measures tend to be correlated, and may result in meaningless parameter estimates when jointly added to the same statistical model (i.e. the association between blood-lead and environmental-lead might be estimated as negative for one or more sources of exposure in the

joint model). To avoid negative parameter estimates for lead exposure predictor variables, all five candidate models were originally fitted using non-linear regression models with constraints on the parameter estimates associated with exposure variables (the parameter estimates for these variables were constrained to be greater than or equal to zero). Log-linear models with positive parameter estimates for lead exposure predictor variables were later fitted using standard linear regression models. The models occasionally converged to local maximums rather than the global maximum likelihood solution, however, this problem was resolved by identifying improved starting values for each model. Further discussion of collinearity diagnostics is presented in Section G5.3 and Section G12.

### **G5.2.2 Multimedia Exposure Model Development**

As discussed above, many combinations of variables were considered for the multi-media exposure model. Section G13 presents details of statistical model fittings for four sets of variables which met the variable selection criteria discussed above. The variable selection and model development work resulted in the following general conclusions:

1. Measures of soil-lead concentration from the dripline, dust-lead loading from floors, dust-lead loading from window sills, interior deteriorated lead based paint, pica for paint, and race were consistent predictors of blood-lead concentrations. Window sill lead loading appeared to compete with interior deteriorated lead-based paint as a predictor of blood-lead concentration.
2. A reduced set of variables (including measures of lead in paint, dust and soil, race and pica for paint) resulted in statistical models which were able to explain roughly 40% of the variability in children's blood-lead concentrations.
3. The log-additive model was outperformed by the other candidate models, as indicated by log likelihood statistics presented in Section G13, largely due to a saturation of the response at higher levels of environmental lead.
4. The  $F_{\text{Passive}}$  parameter in the Passive/Active Uptake model was consistently estimated at or very close to zero. The Active Uptake model may therefore be a more appropriate model (since it won't be over-parameterized).
5. The log-linear model consistently outperformed all other candidate models (with the same variables) based on an evaluation of log likelihoods, as can be seen in Section G13.

Parameter estimates and associated standard errors of a series of four different multi-media exposure models (each of which included a different set of predictor variables) are provided in Section G13. Each table in Section G13 contains the results of fitting all five candidate statistical model forms to data from the Rochester Lead-in-Dust Study.

### **G5.3 REGRESSION DIAGNOSTICS**

This section describes the diagnostic analyses performed as part of development of the multi-media predictive model using data from the Rochester Lead-In-Dust Study. Through the use of regression diagnostics the adequacy of fit of the various candidate models developed to the data observed can be determined, and model assumptions can be verified. For these models, the following regression diagnostic procedures were performed:

1. A normal quantile plot of the residuals was created. The normal quantile plot approximated a straight line indicating that residuals (errors) were approximately normally distributed, as assumed.
2. Residual values were plotted versus predicted values. This scatterplot did not indicate signs of nonconstant variance (if points spread out or tighten up as you move from left to right) or nonlinearity (if points look quadratic or bow-shaped). The scatterplot exhibited no pattern, indicating no such problems. Similarly, plots of residuals versus predictors indicated no discernible pattern.
3. Cook's distance and DFFITS (both measures of influence) were plotted versus studentized residuals (a measure of how far an observation deviates from the modeled relationship) to indicate potential outliers - points with undue influence and points lying far outside the model's prediction. These plots of Cook's distance and DFFITS were produced only for the log-linear models, which were implemented using standard linear regression, and identified no obvious outliers or influential points.
4. For a closer examination of how points influence model parameter estimates, the models were fit while excluding a single point at a time. Analysis of the coefficients adjusted for their standard error (intercept, and coefficients of PbS, PbF, PbW and PbP), including plots, again identified no major problems with influential data points.
5. Partial regression leverage plots were created for the environmental measures of lead exposure: dripline soil, floor dust from carpeted and uncarpeted floors, paint/pica hazard, and window sill dust. A partial regression leverage plot that exhibits a strong linear relationship between blood-lead and the variable under consideration is indicative of a strong linear relationship between blood lead and the environmental measure of lead exposure while controlling for all the other variables in the model. Partial regression leverage plots were produced only for the log-linear models, which were implemented using standard linear regression, and indicated an adjusted positive relationship for each lead exposure variable included in the multi-media predictive model.
6. Partial  $R^2$  comparisons between predictor variables included in the model were calculated. A high partial  $R^2$  indicates greater importance in predicting blood-lead concentration.

7. Estimated log-likelihoods were calculated using parameter estimates from each model and the observed Rochester data, and the likelihood ratios between different models were then assessed. The likelihood ratio (LR) is equivalent to the ratio of the data's probability under one model compared to its probability under a second model. The likelihood ratio evaluation consistently indicated that the log-linear model provided the best fit to the data.
8. An analysis into the effects of collinearity using several methods was conducted during the development of the multi-media predictive model. Estimates of the tolerance statistic and the variance inflation factor associated with each predictor variable in the model were calculated, along with a single value decomposition for the design matrix of observed predictor variables in the model. These analyses suggested that the model did not suffer from a problem with collinearity.

The above regression diagnostics and tests of collinearity among explanatory variables for the multi-media predictive model are provided in detail in Section G12. Based on the regression diagnostics on the multi-media predictive model it was concluded that:

- ! no influential or outlying points should be deleted from the analysis,
- ! the model developed fits the data observed,
- ! model assumptions are verified, and
- ! the model does not appear to suffer from a severe problem with collinearity.

#### **G5.4 THE MULTI-MEDIA PREDICTIVE MODEL BASED ON ROCHESTER DATA**

The criteria used for the selection of variables in the multi-media predictive model emphasized use of measures of environmental lead and other factors observed in both the Rochester Lead-in-Dust Study and the HUD National Survey. Variables whose definition provided a convenient translation when applied to the National Survey, whose predictive power in Rochester were high, and whose spread in the National Survey populations covered a wide enough range of values, were used in the empirical model. For example, the paint/pica variable was chosen for use in the multi-media predictive model because it was a better predictor and because application of the paint (75th percentile) variable in the HUD National Survey data resulted in a variable that provided very little discrimination between houses in the survey. Another example is that although the variable Bare\_flr was a stronger predictor of blood-lead than the variable Floor\_A in the Rochester Study, Floor\_A was a more appropriate choice for construction in the HUD National Survey, and was therefore selected for use in the multi-media predictive model. Therefore, measures of lead in soil, floor dust, window sill dust and the paint/pica variable were chosen for use in the multi-media predictive model. The final mathematical form of this model was:

$$\ln(\text{PbB}) = \beta_0 + \beta_1 \cdot \ln(\text{PbF}) + \beta_2 \cdot \ln(\text{PbW}) + \beta_3 \cdot \ln(\text{PbS}) + \beta_4 \cdot \text{PbP} + e$$

where PbB represents the blood-lead concentration, PbF corresponds to measurements from interior floor dust, PbW represents environmental lead from window sills, PbS represents soil-lead, PbP represents paint hazard, and e represents the residual error left unexplained by the model. Parameter estimates and associated standard errors, and measures of R-squared and the residual standard deviation for the empirical model are provided in Table G-3. Note that the parameter estimate associated with floor dust-lead loading was only borderline statistically significant when considered jointly with the effect of window sill dust-lead loading (and other exposure variables) in the multi-media predictive model.

**Table G-3. Parameter Estimates and (Associated Standard Errors) for the Multi-Media Predictive Model Based on Data from the Rochester Lead-in-Dust Study**

Parameter	Variable Description	Estimate
0	Intercept	0.418 (0.240)
1	log (PbF): Area-Weighted Arithmetic Mean (Wipe) Dust-Lead Loading from Any Floor (Carpeted or Uncarpeted)	0.066 (0.040)
2	log (PbW): Area-weighted Arithmetic Mean (Wipe) Dust-Lead Loading from Window Sills	0.087 (0.036)
3	log (PbS): Dripline Soil-Lead Concentration (fine soil fraction)	0.114 (0.035)
4	PbP: Indicator of Interior Paint/Pica Hazard	0.248 (0.100)
R <sup>2</sup>	Coefficient of Determination	21.67%
	Root Mean-Square Error (Residual Error)	0.56188

The above multi-media predictive model is used in the Section 403 Risk Assessment to determine the probability that a child in the Rochester Study exposed to specific levels of lead in paint, dust and soil will have a blood-lead concentration exceeding 10 µg/dL.

## G6.0 THE EMPIRICAL MODEL

The goal of the empirical model is to provide a relationship between blood-lead concentration and various environmental lead exposures as measured in the HUD National Survey for use in the Section 403 risk assessment. Unfortunately, the HUD National Survey contains no information about blood-lead concentration. However, data from the Rochester Lead-in-Dust Study (i.e. the multi-media predictive model) can provide a basis for the empirical model. At issue is how to use the multi-media predictive model based on the Rochester data set to develop an empirical model applicable to the data observed in the HUD National Survey.

Matters are complicated by the fact that the sampling methodology used to measure lead exposures in HUD is different from that used in Rochester. Thus, some variables have a different interpretation in each of these two studies. Specifically, two of the lead exposure measurements in HUD are blue nozzle floor dust lead loading and blue nozzle window sill dust lead loading, compared to floor wipe dust lead loading and window sill wipe dust lead loading in Rochester. Another example is that the soil variable in Rochester was based on a composite sample from the dripline area adjacent to the house, whereas in the HUD National Survey, the soil variable was based on a weighted average of samples collected from dripline, entryway and remote locations (with weights of 25%, 25% and 50%, respectively). Also the paint/pica hazard predictor variable was constructed differently between the Rochester Study and the HUD National Survey data. The primary difference was that the paint/pica hazard input variable from the HUD National Survey data was based on the measures of paint on both interior and exterior surfaces, whereas the variable used in Rochester for estimation of the effect of paint/pica hazard was based on measure of paint on only interior surfaces. Lead based paint on deteriorated exterior surfaces was not considered in the estimation of the paint/pica model parameter based on Rochester data because approximately 84 percent of houses in the Rochester Study were built prior to 1940 and as a result virtually every home surveyed in the Rochester Study had lead based paint on exterior surfaces. Therefore, a paint/pica hazard variable which included presence of exterior lead based paint in Rochester lost its statistical significance and its predictive power. The differences in paint/pica variable construction between the Rochester and HUD National Survey is considered minor in comparison to the differences in dust and soil sampling methodologies. Table G-4 provides details comparing the construction and interpretation of variables in both the Rochester Lead-in-Dust Study and the HUD National Survey.

The following statistical method was used to account for differences in dust and soil sample collection methods between the Rochester Study and the HUD National Survey when assessing the impact of 403 rulemaking on children's blood-lead levels. The method involves establishing a relationship between blood-lead and environmental variables as measured by methods used in the Rochester Study (i.e. the multi-media predictive model based on Rochester Data), and then adjusting this relationship to use dust-lead and soil-lead variables as measured in the HUD National Survey. The adjustment takes into account both systematic differences and differences in error structures between the Rochester wipe dust-lead and drip-line soil-lead predictor variables versus the HUD National Survey Blue Nozzle dust-lead and averaged soil-lead predictor variables. The method provides a relationship between blood-lead concentration,

**Table G-4. Variable Construction in the Rochester Lead-in-Dust Study and the HUD National Survey**

Predictor Variable	Rochester Study	HUD National Survey Input Variables
Soil	Natural log transformation of dripline soil-lead concentration (fine soil fraction).	The natural log transformation of the weighted average of dripline, entryway and remote soil-lead concentrations, with weights of 25%, 25% and 50% respectively when all three soil samples were collected. If these values were missing, an imputed value <sup>1</sup> was used.
Floor Dust	The natural logarithm of the area weighted arithmetic average (wipe) dust-lead loading from carpeted and uncarpeted floors.	The natural logarithm of the area-weighted arithmetic average dust-lead loading (Blue Nozzle Vacuum) from 3 sample locations (wet, dry and entry rooms) was used as the measure of lead in dust. If the dust-lead loadings from all of the 3 sample locations were missing, an imputed value <sup>1</sup> was used.
Window Sill Dust	The natural logarithm of the area-weighted arithmetic average (wipe) dust-lead loading from window sills.	The natural logarithm of the area-weighted arithmetic average dust-lead loadings (Blue Nozzle Vacuum) from window sills from 2 sample locations (wet and dry rooms). If the window sill dust-lead loadings from both sample locations were missing, an imputed value <sup>1</sup> was used.
Interior Pica/Paint	<p>An indicator variable which was nonzero when the following conditions each existed in a residential unit: presence of deteriorated or damaged interior paint; presence of interior lead-based paint; and presence of a child with paint pica. The paint variable had values of:</p> <ul style="list-style-type: none"> <li>0 No LBP (XRF reading &lt; 1), or condition<sup>a</sup> is Good, or child does not exhibit pica;</li> <li>1 LBP (XRF reading ≥ 1), condition is Fair or Poor, and child exhibits pica rarely;</li> <li>2 LBP (XRF reading ≥ 1), condition is Fair or Poor, and child exhibits pica at least sometimes.</li> </ul>	<p>HUD National Survey homes were determined to have deteriorated LBP whenever there is any deterioration in interior or exterior lead-based paint, as measured by square footage (that is, square footage of deteriorated LBP surface &gt; 0). That is, the LBP indicator was defined as</p> <ul style="list-style-type: none"> <li>1 Whenever square footage of surface exhibiting deteriorated LBP (interior and exterior) &gt; 0</li> <li>0 Otherwise</li> </ul> <p>The pica factor was only considered for houses with deteriorated LBP. In these houses, it was assumed that 9% of U.S. children aged 1-2 years have pica for paint. For the children with pica for paint, the pica value was defined to be 1.5<sup>b</sup>.</p>

<sup>1</sup> Imputed values for dust and soil were based on a presence of LBP indicator variable and on a house age-specific indicator. The presence of LBP indicator was defined as:

- 0 Predicted maximum XRF < 1 for both interior and exterior samples
- 1 Predicted maximum XRF ≥ 1 for either interior or exterior samples.

The house age-specific indicator had categories: Pre-1940, 1940-1960, 1960-1979, Post-1979. The imputed values for dust and soil were constructed by taking the means for the associated subsets formed by crossing the paint and age of house categories.

<sup>a</sup> Condition of the paint in the Rochester Lead-in-Dust Study is described in Table G-2.

<sup>b</sup> The Paint/Pica Hazard Variable was described in Table G-2. A value of 1.5 was chosen as the input value for those children exhibiting pica in applying the empirical model to the HUD National Survey.

floor and window sill dust-lead loadings, soil-lead concentrations, and other covariates as observed in the HUD National Survey. An errors in variables measurement error adjustment is applied as an intermediate step in reaching this goal. The method for adjusting the multi-media predictive model may be described as follows, and is provided with complete detail in Appendix G1.

The first step involves fitting an errors in variables measurement error adjusted multi-media exposure model that assumes blood-lead concentration is a function of true unobserved floor and window sill dust-lead loadings and dripline soil-lead concentrations along with other covariates (paint/pica hazard) used in the model. While the dependence of blood-lead concentration on true dust-lead loadings, dripline soil-lead concentrations, and other covariates can not be observed, they can be estimated via equations (2.2.12) and (2.2.16) in Fuller, 1987. In order to use these equations for estimating this relationship, the measurement error associated with each particular dust-lead loading and soil-lead concentration must be obtained. This is achieved by taking individual measurements of dust-lead loadings and soil-lead concentrations within households and calculating their variability. The average of all within household variances is then used as an estimate of the true measurement error associated with each particular dust-lead loading and soil-lead concentration. The estimated measurement errors are then used to calculate parameter estimates for a model based on Rochester data that relates blood-lead concentration to true dust-lead loadings, dripline soil-lead concentrations, and other covariates (paint/pica hazard). Keep in mind that the model must be developed using Rochester data because there is no blood-lead concentration variable in the HUD data set.

If the goal had been to identify the nature of the dependence of blood-lead concentration on true floor and window sill dust-lead loadings, dripline soil-lead concentrations, and the other covariates, then the adjustment described above would have been all that was required. However, the relationship of interest is blood-lead concentration as a function of floor and window sill dust-lead loadings, average soil-lead concentrations, and other covariates (paint/pica hazard) as observed in the HUD National Survey. Therefore, adjusting for measurement error is only the first step toward a final solution to this problem.

The next step in this process is to define the relationship between blood-lead concentrations, observed dust-lead and soil-lead predictor variables as measured in both Rochester and HUD, dust-lead and soil-lead predictor variables measured without error on the scale of measure used in Rochester, and any other covariates (paint/pica hazard) in the multimedia exposure model. It is assumed that these random variables jointly follow a multivariate normal distribution. Standard statistical theory then allows for deriving the distribution of blood-lead concentration conditioned on floor and window sill dust lead loadings, average soil lead concentrations, and other covariates as measured in HUD. Estimates of the parameters for a multimedia exposure model that relates blood-lead concentration to lead exposures as measured in the HUD National Survey are obtained from this conditional distribution.

The final step in developing the empirical model was to derive an estimate for the intercept. The empirical model intercept was designed to calibrate the model so that the predicted national (pre-403) geometric mean blood-lead concentration obtained from applying the empirical

model to data observed in the HUD National Survey equals the geometric mean blood-lead concentration estimated in Phase 2 of NHANES III.

The empirical model involves an adjustment to the multi-media predictive model based on the Rochester Study to allow use of Blue-Nozzle dust-lead loadings rather than wipe dust-lead loadings and average soil-lead concentration rather than dripline soil-lead concentration. The final mathematical form of this model is:

$$\ln(\text{PbB}) = \beta_0 + \beta_1 \cdot \ln(\text{PbF}_{\text{BN}}) + \beta_2 \cdot \ln(\text{PbW}_{\text{BN}}) + \beta_3 \cdot \ln(\text{PbS}) + \beta_4 \cdot \text{PbP} + e$$

where PbB represents the blood-lead concentration,  $\text{PbF}_{\text{BN}}$  and  $\text{PbW}_{\text{BN}}$  correspond to dust-lead loading from interior floors and window sills respectively (for samples collected in the HUD National Survey with the blue nozzle vacuum), PbS represents average soil-lead concentration, PbP represents paint/pica hazard, and e represents the residual error left unexplained by the model. Table G-5 provides parameter estimates and associated standard errors for the empirical Model developed to predict the national distribution of children’s blood-lead concentrations using data as observed in the HUD National Survey. The standard errors provided in Table G-5 were estimated using a Bootstrap Algorithm which is detailed in Section G10.4.

**Table G-5. Parameter Estimates and Associated Standard Errors for the Empirical Model used to Predict the National Distribution of Children’s Blood-Lead Concentration Based on Data from the HUD National Survey**

Variable	Parameter	Estimate (Standard Error)
Intercept	0	0.650 (0.154)
Floor Dust-Lead Loading (Blue Nozzle Vacuum)	1	0.032 (0.044)
Window Sill Dust-Lead Loading (Blue Nozzle Vacuum)	2	0.050 (0.031)
Average Soil-Lead Concentration	3	0.094 (0.043)
Paint/Pica Hazard	4	0.256 (0.098)
Error	<sup>2</sup> Error	0.313

## G7.0 ESTIMATING THE NATIONAL DISTRIBUTION OF BLOOD-LEAD USING THE EMPIRICAL MODEL

As stated previously, the empirical model will be used in the Risk Assessment to predict a national distribution of children's blood-lead concentrations both before and after interventions resulting from the Section 403 standards. A nationally representative sample of environmental conditions in housing is required as input to the empirical model to predict a national distribution of children's blood-lead concentrations. The HUD National Survey is a nationally representative study which assessed environmental lead-levels in paint, dust and soil in residential housing. Environmental conditions observed in the HUD National Survey were used as input to the EPI model for predicting blood-lead levels in children 1-2 years old. A population of children aged 1-2 years is both the target age group for EPA's Risk Assessment, and the age group that was recruited in the Rochester Lead-in-Dust Study (thus the empirical model is representative of children in this age group). The empirical model is used to estimate an average log-transformed childhood blood-lead concentration associated with each home in the HUD National Survey.

As noted in Table G-5, the variables used for prediction are average soil-lead concentration, blue-nozzle vacuum dust-lead loading on floors (carpeted or uncarpeted), blue-nozzle vacuum dust-lead loading on window sills, and an indicator of paint/pica hazard. These variables, constructed from observed levels of lead in each HUD National Survey residential unit, are used as input to the empirical model for predicting the pre-403 national distribution of children's blood-lead concentrations.

To predict a post-403 national distribution of children's blood-lead concentrations, the following method was used to prepare the HUD National Survey Data for input into the empirical model:

- [1] Observed levels of lead in environmental variables in the HUD National Survey were compared to proposed section 403 standards. Blue-nozzle vacuum floor and window sill dust-lead loadings were converted to wipe dust-lead loadings before comparison to the 403 standards.
- [2] Section 403 interventions were triggered in HUD National Survey residential units that had levels of lead in environmental variables that were above the proposed standard. If an intervention was triggered, assumed post-intervention lead levels in environmental variables were substituted for observed levels according to the Section 403 risk assessment assumptions. Post intervention dust-lead levels that were specified in terms of wipe dust-lead loadings were converted to a blue nozzle vacuum scale for use in the prediction.

The distribution of blood-lead concentrations associated with each home was characterized by assigning a geometric mean (predicted by the empirical model) and a geometric standard deviation. A geometric standard deviation of 1.6 was assumed for the distribution of blood-lead concentrations associated with each home. The default geometric standard deviation of blood-lead concentrations for children at similar environmental-lead levels for the IEUBK

model is 1.6 and the estimated variability from the multi-media predictive model based on the Rochester Data was 1.76 as measured by the exponentiation of the root mean square error. Thus, a population of children (aged 1-2 years) associated with environmental lead levels found at each home in the HUD National Survey was constructed using the geometric mean blood-lead concentration predicted by the empirical model, an assumed geometric standard deviation of 1.6, and population weights based on the 1993 American Housing Survey adjusted to 1997.

The predicted national distribution of blood-lead concentrations can be characterized using a geometric mean and a geometric standard deviation. The predicted national geometric mean is calculated by taking a weighted geometric mean of the empirical Model predicted blood-lead concentration associated with each home in the HUD National Survey, using the adjusted weights for 1997. The predicted national geometric standard deviation is calculated by taking the square root of the sum of the predicted between-house variability and the assumed within-house variability. The predicted between-house variability is estimated as a weighted geometric variance among the empirical Model predicted blood-lead concentration associated with each home in the HUD National Survey, using the adjusted weights for 1997. Thus, the between-house variability represents the variability among the predicted blood-lead concentrations associated with the environmental conditions observed in each home in the HUD National Survey. The assumed within-house variability was  $(1.6)^2$ , and represents the expected variability among children who are exposed to similar environmental conditions. The predicted national geometric standard deviation relies on an assumption that the between-homes distribution of blood-lead concentration is log-normally distributed.

The predicted national distribution of children's blood-lead concentrations can also be characterized using exceedance percentiles (i.e. the percentage of children estimated to have blood-lead concentrations above a specified level, such as 10, 20 and 30  $\mu\text{g}/\text{dL}$ ). These exceedance proportions were calculated in two ways, first by using normal probability theory combined with the estimated national geometric mean and standard deviation, and second by empirical evaluation of a national population built by summing discretized populations of children associated with each home.

The second approach is robust to deviations from the assumed log-normal distribution of blood-lead concentrations between homes, and can be described as follows:

A distribution of blood-lead concentrations is constructed for each home using the empirical Model predicted geometric mean and the assumed within house geometric standard deviation of 1.6. Each of these distributions are then partitioned into seven discrete blood-lead intervals. Table G-6 provides the specific method for partitioning a distribution of log blood-lead concentrations into the seven intervals about the log of the geometric mean (predicted from the empirical model). Figure G-2 graphically illustrates this partitioning. The two tails of the distribution represent log blood-lead concentrations below or above 2.5 standard deviations from the mean, respectively. The percentage of the distribution assigned to each of these intervals, 0.62%, is based on the area under a standard normal curve for z-values less than -2.5 in the lower tail or greater than 2.5 in the upper tail. The assigned log blood-lead concentration for the lower tail is the expected value of a standard normal random deviate lying in the interval from  $-\infty$  to -

2.5; the assigned log blood-lead concentration was similarly chosen for the upper tail, and mid-points were used for the finite-length intervals. The assigned blood-lead concentration for each interval was obtained by exponentiating the assigned log blood-lead for the interval. For example, for the lower tail,

$$e^{\mu - 2.82 \times \sigma} = e^{\mu} \times e^{-2.82 \times \sigma} = GM \times GSD^{-2.82} = \frac{GM}{GSD^{2.82}}$$

**Table G-6. Allocation of Blood-Lead Distribution to Seven Intervals**

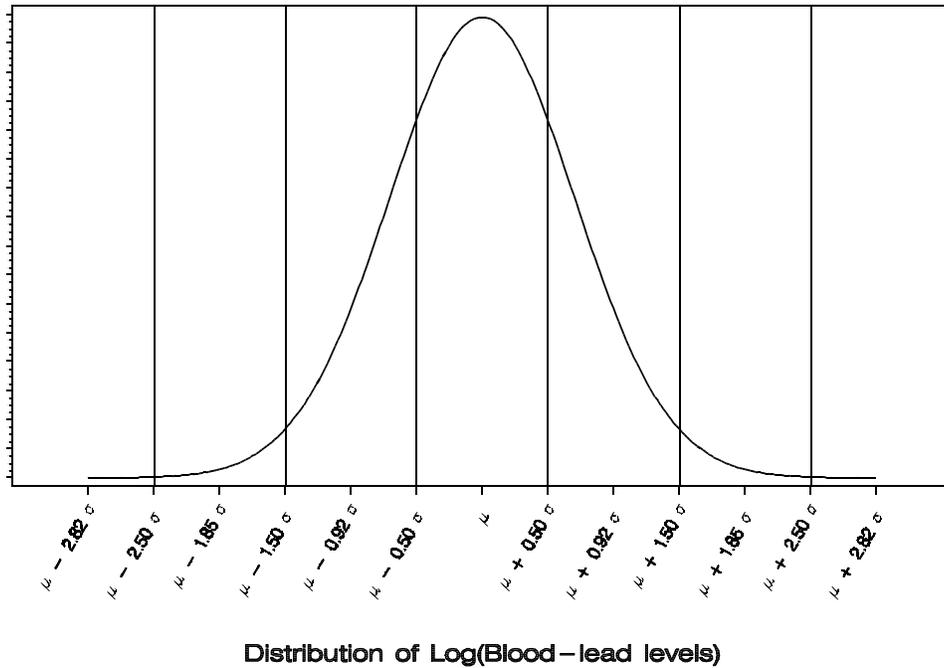
Log Blood-Lead Concentrations			Assigned Blood-Lead Concentration for Interval
Interval for Log Blood Lead <sup>a</sup>	Percentage of Distribution in Interval	Assigned Log Blood Lead for Interval	
$[-\infty, \mu - 2.5^* ]$	0.0062	$\mu - 2.82^* \text{ }^b$	$GM/[GSD^{2.82}]$
$[\mu - 2.5^* , \mu - 1.5^* ]$	0.0606	$\mu - 2.00^*$	$GM/[GSD^{2.00}]$
$[\mu - 1.5^* , \mu - 0.5^* ]$	0.2417	$\mu - 1.00^*$	$GM/[GSD^{1.00}]$
$[\mu - 0.5^* , \mu + 0.5^* ]$	0.3830	$\mu$	GM
$[\mu + 0.5^* , \mu + 1.5^* ]$	0.2417	$\mu + 1.00^*$	$GM*[GSD^{1.00}]$
$[\mu + 1.5^* , \mu + 2.5^* ]$	0.0606	$\mu + 2.00^*$	$GM*[GSD^{2.00}]$
$[\mu + 2.5^* , + \infty]$	0.0062	$\mu + 2.82^* \text{ }^c$	$GM*[GSD^{2.82}]$

<sup>a</sup> Blood-lead concentrations were assumed to have a log-normal distribution with the geometric mean (GM) predicted by the empirical model and a geometric standard deviation (GSD) of 1.6 (the default geometric standard deviation for the IEUBK model). The distribution of log blood-lead concentrations was assumed to be normal with mean  $\mu$  given by  $\log(GM)$  and standard deviation given by  $\log(GSD = 1.6)$ .

<sup>b</sup> The expected value of a normal random deviate known to lie in the interval  $[-\infty, -2.5]$  is  $-2.82$ .

<sup>c</sup> The expected value of a normal random deviate known to lie in the interval  $[2.5, + \infty]$  is  $+2.82$ .

For this lower tail, if N children were associated with the specific housing condition (according to weights in the 1993 American Housing Survey adjusted to 1997) then 0.62 percent of the N children were assigned a blood-lead concentration of  $GM/GSD^{2.82}$ . The remaining 99.28 percent were similarly assigned to the other blood-lead concentrations presented in Table G-5 using the percentages given in the second column of the table. In this manner, the distribution of blood-lead concentrations of the N children were allocated to a distribution of blood-lead concentrations centered around the GM predicted by the empirical model with a GSD of 1.6. The predicted distributions at each housing condition were then combined to generate a distribution of childhood blood-lead levels over all of the housing conditions present in the HUD National Survey.



**Figure G-2. Distribution of Blood-Lead Levels About Geometric Mean on Logarithmic Scale.**

The exceedance percentiles can then be assessed by empirically tabulating the proportion of children in this constructed distribution who are above the target blood-lead concentrations of 10, 20 and 30  $\mu\text{g}/\text{dL}$ .

### **G7.1 RESULTS OF THE COMPARISON WITH NHANES III**

The predicted distribution of blood-lead concentrations obtained by applying the empirical model to the HUD National Survey Data was compared to NHANES III as a check on how well the empirical model performed. Table G-7 contains characteristics of the predicted blood-lead distribution for the empirical model, including estimates of exceedance proportions (the estimated proportion of blood-lead concentration exceeding 10, 20 or 30  $\mu\text{g}/\text{dL}$ ), the geometric mean, and the geometric standard deviation. Results in Table G-7 for the NHANES III distribution, the distribution of children recruited into the Rochester Lead-in-Dust Study, and the predicted national distribution based on applying the empirical model to data from the HUD National Survey (both before and after Section 403 interventions take place) are presented first with exceedance proportions calculated from the discretized distribution and second for exceedance proportions calculated assuming a log-normal distribution with the calculated geometric mean and geometric standard deviation.

**Table G-7. Predicted National Distribution Characteristics for Empirical Model Compared to Rochester and NHANES III**

Predicted Model Results	Parameter	Pre-Intervention Blood-lead Levels			Post-Intervention Blood-lead Levels <sup>1</sup>
		NHANES III	Rochester Study	Empirical Model	Empirical Model
National Geometric Mean	$\mu_p$	3.14	6.36	3.14	3.03
National Geometric Standard Deviation	$\rho$	2.09	1.85	1.71	1.67
Discretized Distribution Exceedance Percentiles (% of Population $\geq$ 10, 20 & 30 $\mu\text{g}/\text{dL}$ )	% $\geq 10 \mu\text{g}/\text{dL}$	5.88%	22.90%	0.00%	0.00%
	% $\geq 20 \mu\text{g}/\text{dL}$	0.43%	2.90%	0.00%	0.00%
	% $\geq 30 \mu\text{g}/\text{dL}$	0.07%	1.00%	0.00%	0.00%
Log-Normal Distribution Exceedance Percentiles (% of Population $\geq$ 10, 20 & 30 $\mu\text{g}/\text{dL}$ )	% $\geq 10 \mu\text{g}/\text{dL}$	5.75%	23.10%	1.54%	1.00%
	% $\geq 20 \mu\text{g}/\text{dL}$	0.59%	3.13%	0.03%	0.01%
	% $\geq 30 \mu\text{g}/\text{dL}$	0.11%	0.01%	0.0013%	0.0004%

<sup>1</sup> For illustration of a calculation of a post-intervention blood-lead distribution, standards were set at: 100  $\mu\text{g}/\text{ft}^2$  for floor dust-lead loading (wipe), 500  $\mu\text{g}/\text{ft}^2$  for window sill dust-lead loading (wipe), 2000  $\mu\text{g}/\text{g}$  for soil removal, 5  $\text{ft}^2$  damaged LBP for paint repair, and 20  $\text{ft}^2$  damaged LBP for paint abatement. Post-403 lead levels for homes that were above the standard were adjusted to 40  $\mu\text{g}/\text{ft}^2$  for floor dust-lead loading (wipe), 100  $\mu\text{g}/\text{ft}^2$  for window sill dust-lead loading (wipe), 150  $\mu\text{g}/\text{g}$  for soil removal, and 0  $\text{ft}^2$  damaged LBP for paint repair or abatement.

The results of the comparison with NHANES III for the revised empirical model indicate:

- ! The national geometric mean blood-lead concentration (pre-intervention) was calibrated to the geometric mean reported in NHANES III.
- ! The variability in the national distribution of blood-lead concentration predicted by the empirical model using the HUD National Survey (pre-403) is estimated at 1.71 (GSD), in contrast to a GSD of 2.09 for NHANES III.
- ! The estimated proportions of blood-lead concentrations of at least 10, 20, or 30  $\mu\text{g}/\text{dL}$  using the empirical model predictions are much lower than the corresponding proportions estimated by NHANES III.

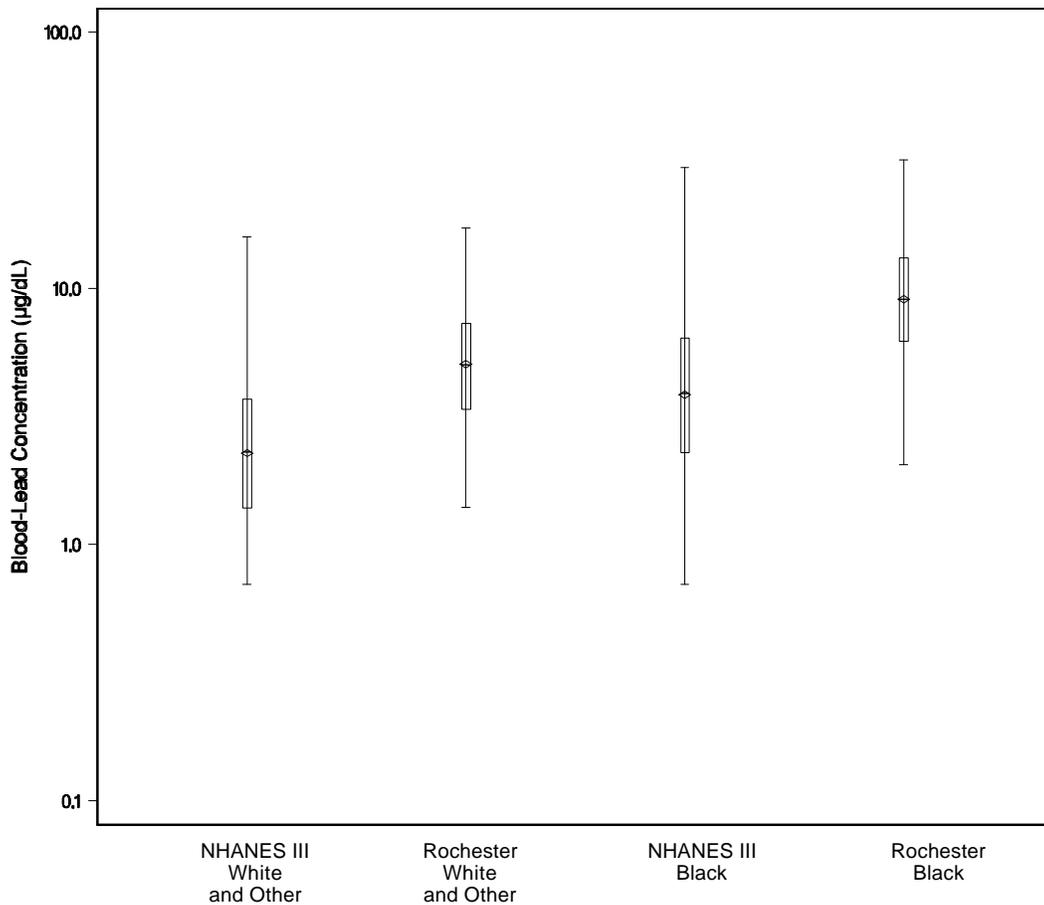
It should also be noted that NHANES III itself is only an estimate of the true national distribution of blood-lead concentrations (pre-403), and that an "exact" match of NHANES III does not mean an exact match of the true national distribution, nor does it guarantee that the model is appropriate for predicting a post-403 national distribution.

## G8.0 DISCUSSION

The primary limitation associated with the Rochester Study is concern over the degree to which the Rochester Study may be considered representative of the nation as a whole. Differences between the Rochester Study population and the national population include the following:

- a. Almost one-quarter (22.9%) of the Rochester children had observed blood-lead concentrations above 10  $\mu\text{g}/\text{dL}$ , whereas only 5.9% of children aged 1-2 years nationwide were estimated to have blood-lead concentrations above 10  $\mu\text{g}/\text{dL}$  by Phase 2 of NHANES III.
- b. The geometric mean blood-lead concentration in Rochester is 6.4, whereas the geometric mean blood-lead concentration nationwide as estimated by NHANES III is 3.1. The GSD for Rochester is 1.9, compared to 2.1 for NHANES III.
- c. Approximately 84 percent of the housing included in the Rochester Study was built prior to 1940, and there is a well documented relationship between age of housing and presence of lead-based paint. Only approximately 20% of housing nationwide was built prior to 1940.
- d. Approximately 40% of the sample of children in the Rochester Study were African Americans, compared to an estimated 13% of the population of children nationwide (from 1997 US Census Projections), and compared to approximately 7% in the HUD National Survey.
- e. Environmental levels of lead in soil in the Rochester Study were higher than would be expected in the HUD National Survey. For example, the geometric mean dripline soil-lead concentration in the HUD National Survey was approximately 75 ppm whereas the Rochester geometric mean was approximately 730 ppm.
- f. Subjects recruited into the Rochester Study represent children whose primary exposure to lead was from dust, soil and paint at the primary residence. Children whose parents had lead exposure, who spent time away from the home, or whose homes underwent renovation or remodeling were excluded from the study. Only 376 of 1,536 families were eligible to participate in the study after the initial telephone screening. The selection criteria utilized in the Rochester Study may have resulted in a biased sample of children, since children who had potential lead exposure outside of the primary residence were excluded.

The difference in the observed blood-lead distributions between the Rochester Study and NHANES III is illustrated in Figure G-3. Although there are limitations associated with the Rochester Study, there are also positive aspects of the study that recommend its use:



**Figure G-3. Box Plot of Blood-lead Concentrations for Children Aged 1-2 Years for Phase II of NHANES III versus Rochester Data Sets.**

- a. all media, locations, and surfaces that are being considered for Section 403 standards were measured for lead in the Rochester Study.
- b. the Rochester Study includes dust-lead loadings from wipe sampling and the Section 403 dust standard is expected to be based on dust-lead loading from wipe sampling.
- c. the selection of homes and children in the Rochester Study, although targeted, was more random and more representative of a general population than is the case with most recent epidemiological studies of lead exposure in non-smelter communities.

The ability of an empirical model to predict the national distribution of blood-lead concentrations following Section 403 lead hazard reduction activities may be most severely limited by factors that are not included in the model. Reflecting its use in the Section 403 Risk Assessment, the empirical model accounts only for factors related to environmental lead

exposures at the residence, and does not account for other factors that might affect childhood blood lead. Such factors that may affect children's blood-lead concentration but may not be able to be controlled by the Section 403 rule include:

- (1) home and personal cleaning habits,
- (2) diet and nutritional status,
- (3) bio-availability of the lead found in residential environmental media,
- (4) non-residential exposures,
- (5) inhalation exposure,
- (6) children's behavior,
- (7) socio-economic factors,
- (8) renovation and remodeling (R&R) activity,
- (9) hobbies,
- (10) occupation.

Finally, it should be noted that the empirical model contains variables that differ from variables created for a best-fit of the Rochester data, because the goal of the empirical model was to provide a basis for using measures of lead from the HUD National Survey to predict a national distribution of childhood blood-lead concentrations. In particular, the empirical model differs from the multimedia regression model used to characterize the dose-response relationship between environmental-lead and blood-lead.

## **G9.0 REFERENCES**

See Chapter 7 of Volume I for references cited within this appendix.

**G10: Appendix on Methodology for Adjusting for  
Different Sampling Methods**

## Statistical Details

Section G10 of Appendix G is comprised of four sections that describe the statistical details associated with the Empirical Model. Section G10.1 explains statistical methodology used to account for differences in sample collection methods used in the Rochester Lead-in-Dust Study and the HUD National Survey. Section G10.2. describes the classic errors in variables regression model. Section G10.3 provides details on the estimation of variance components used as input to the above two statistical models. Finally, Section G10.4 explains the bootstrap algorithm used for approximating the standard errors associated with parameter estimates of the model that accounts for differences in sampling methods.

### G10.1 STATISTICAL ADJUSTMENTS TO ACCOUNT FOR DIFFERENCES IN SAMPLE COLLECTION METHODS USED IN THE ROCHESTER LEAD-IN-DUST STUDY AND THE HUD NATIONAL SURVEY

The goal of this section is to provide a statistical methodology for adjusting the multi-media predictive model to appropriately use environmental lead levels observed in the HUD National Survey as inputs to the model. The adjustment takes into account both systematic differences and differences in error structures between the Rochester predictor variables and the HUD National Survey predictor variables. The method provides a relationship between blood-lead concentration and a set of lead exposure variables and other covariates as they were measured in the HUD National Survey. As an initial overview the method may be described as follows. Assume:

- Y represents children's blood-lead levels,
- R represents wipe dust lead loading observed in the Rochester Study,
- H represents blue nozzle dust lead loading observed in the HUD National Survey, and
- C represents covariates of interest which appear both in the Rochester Study and in the HUD National Survey.

The density of interest is children's blood lead levels as a function of lead exposures measured in the HUD National Survey, namely

$$F_{Y|H,C}(y|h,C) = \int F_{Y|R,H,C}(y|r,h,C) \cdot F_{R|H,C}(r|h,C) \cdot dr$$

Given that we do not have a source of data with Y,R,H and C observed simultaneously, the method used for estimating  $F_{Y|H,C}(y|h,C)$  is:

$$F_{Y|H,C}(y|h,C) = \int F_{Y|X,C}(y|x,C) \cdot F_{X|H,C}(x|h,C) \cdot dx$$

where X is a latent variable that represents dust lead loading measured without error.

This method assumes that Y can be modeled as a function of X using an errors-in-variables approach.

Details of the method are provided in the following subsections. Section G10.1.1 presents the methodology for the specific case of an errors-in-variables adjustment of a single covariate. This section is provided to aid in the understanding of the theoretical development of the model parameters. Section G10.1.2 presents the methodology for the general case of an errors-in-variables adjustment of one or more covariates. The Empirical model involves an errors-in-variables adjustment of three covariates: floor wipe dust lead loading, window sill wipe dust lead loading, and drip-line soil lead concentration. Thus, the Empirical model parameter development follows the methodology detailed in Section G10.1.2.

### **G10.1.1 MODELING BLOOD-LEAD AS A FUNCTION OF ONE VARIABLE MEASURED WITH ERROR AND OTHER SELECT COVARIATES.**

The following theoretical development of the Empirical model parameters is specific to an errors-in-variables adjustment of a single covariate. Details are given for this specific case for two reasons:

1. The original theory was developed in this context.
2. The theoretical development is easiest to follow for a single variable adjustment.

In general, the theory applies to errors-in-variables adjustments for any number of covariates in the model. Section G10.1.2 below uses matrix notation to present the general theoretical details, which includes as a special case the errors-in-variables adjustment of a single covariate.

#### **Definitions and Assumptions**

Define the following variables:

- Y = The response variable, log of blood-lead concentration.
- R = Log of the area weighted arithmetic mean of the floor wipe dust lead loading as observed and measured in the Rochester Lead-in-Dust Study.
- H = Log of the area weighted arithmetic mean of the blue nozzle floor dust lead loading as observed and measured in the HUD National Survey.
- X = Log of the “true” unobserved area weighted arithmetic mean floor wipe dust lead loading (measured without error).
- C = A vector (or scalar) of remaining covariates used as independent variables. For the model detailed in this section, which adjusts for the measurement error in floor wipe dust lead loading only, C is a vector consisting of the variables drip-line soil lead concentration and paint/pica hazard. These covariates are assumed to be measured using identical methods in the Rochester Lead-in-Dust Study and the HUD National Survey.

The model assumes

(A)

$$\begin{bmatrix} Y \\ R \\ H \\ X \\ C \end{bmatrix} \sim N \left\{ \begin{bmatrix} \mu_Y \\ \mu_R \\ \mu_H \\ \mu_X \\ \mu_C \end{bmatrix}, \begin{bmatrix} \sigma_Y^2 & \sigma_{YR}^2 & \sigma_{YH}^2 & \sigma_{YX}^2 & \sigma_{YC}^2 \\ & \sigma_R^2 & \sigma_{RH}^2 & \sigma_{RX}^2 & \sigma_{RC}^2 \\ & & \sigma_H^2 & \sigma_{HX}^2 & \sigma_{HC}^2 \\ & & & \sigma_X^2 & \sigma_{XC}^2 \\ & & & & \sigma_C^2 \end{bmatrix} \right\},$$

(B)

$$\begin{aligned} Y &= \alpha_{Y|X,C} + \beta_{Y|X(C)}X + \beta_{Y|C(X)}C + e_{Y|X,C} & e_{Y|X,C} &\sim N(0, \sigma_{Y|X,C}^2) \\ R &= X + e_{R|X} & e_{R|X} &\sim N(0, \sigma_{R|X}^2) \\ H &= \alpha_{H|X} + \beta_{H|X}X + e_{H|X} & e_{H|X} &\sim N(0, \sigma_{H|X}^2) \\ X &= \alpha_{X|C} + \beta_{X|C}C + e_{X|C} & e_{X|C} &\sim N(0, \sigma_{X|C}^2) \\ C &= \mu_C + e_C & e_C &\sim N(0, \sigma_C^2) \end{aligned},$$

and all errors are independent of one another.

The parameters  $\alpha_{Y|X,C}$ ,  $\beta_{Y|X(C)}$ , and  $\beta_{Y|C(X)}$  represent the intercept and slopes, respectively, associated with a regression of Y on X(unobserved) and C; and  $\sigma_{Y|X,C}^2$  is the variability in Y unexplained by X and C.  $\sigma_{R|X}^2$  is the measurement error associated with wipe floor dust lead loading in the Rochester Study.  $\sigma_{H|X}^2$  is the measurement error associated with blue nozzle vacuum floor dust lead loading in the HUD National Survey.  $\alpha_{H|X}$  represents a location shift in the distribution of H relative to the distribution of X. Similarly,  $\beta_{H|X}$  represents a scale shift in the distribution of H relative to the distribution of X.  $\alpha_{X|C}$  and  $\beta_{X|C}$  are the intercepts and slopes, respectively, associated with a regression of X on the covariates in C.  $\sigma_{X|C}^2$  represents the variability in X unexplained by the covariates in C.

In addition, the calculations that follow rely heavily on the assumption that the conditional distribution of X given C ( $X/C$ ) is the same in the Rochester Lead-in-Dust Study and the HUD National Survey. This assumption will be referred to as an assumption of transportability.

### Parameter Development

Using assumption (A) of Section G10.1.1, normal distribution theory implies that Y conditioned on H and C is normally distributed with the following parameters:

$$\mu_{YH,C} = \mu_Y + [\sigma_{YH}^2 \quad \sigma_{YC}^2] \begin{bmatrix} \sigma_H^2 & \sigma_{HC}^2 \\ & \sigma_C^2 \end{bmatrix}^{-1} \begin{bmatrix} H - \mu_H \\ C - \mu_C \end{bmatrix}$$

$$\sigma_{YH,C}^2 = \sigma_Y^2 - [\sigma_{YH}^2 \quad \sigma_{YC}^2] \begin{bmatrix} \sigma_H^2 & \sigma_{HC}^2 \\ & \sigma_C^2 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{YH}^2 \\ \sigma_{YC}^2 \end{bmatrix}$$

Solving the inverse matrix above and using assumption (B) of Section G10.1.1 for substitutions yields:

$$\mu_{YH,C} = \mu_Y + \begin{bmatrix} \frac{\beta_{HX} \beta_{YX(C)} \sigma_{XC}^2}{\beta_{HX}^2 \sigma_{XC}^2 + \sigma_{HX}^2} \\ \frac{\beta_{XC} \beta_{YX(C)} \sigma_{HX}^2}{\beta_{HX}^2 \sigma_{XC}^2 + \sigma_{HX}^2} + \beta_{YC(X)} \end{bmatrix}^T \begin{bmatrix} H - \mu_H \\ C - \mu_C \end{bmatrix}$$

$$\sigma_{YH,C}^2 = \sigma_{YX,C}^2 + \frac{\beta_{YX(C)}^2 \sigma_{HX}^2 \sigma_{XC}^2}{\beta_{HX}^2 \sigma_{XC}^2 + \sigma_{HX}^2}$$

Using (B), observe that

$$(C) \quad \beta_{HX}^2 \sigma_{XC}^2 + \sigma_{HX}^2 = \sigma_{HC}^2 \quad ,$$

where the left-hand side of (C) represents the portion of  $\sigma_H^2$  that remains after conditioning on C.

From (C),

$$\beta_{YH(C)} = \frac{\beta_{HX} \beta_{YX(C)} \sigma_{XC}^2}{\beta_{HX}^2 \sigma_{XC}^2 + \sigma_{HX}^2} = \frac{\beta_{YX(C)}}{\beta_{HX}} \left( \frac{\beta_{HX}^2 \sigma_{XC}^2}{\beta_{HX}^2 \sigma_{XC}^2 + \sigma_{HX}^2} \right) = \frac{\beta_{YX(C)}}{\beta_{HX}} \left( 1 - \frac{\sigma_{HX}^2}{\sigma_{HC}^2} \right) \quad ,$$

$$\begin{aligned}\beta_{Y|C(H)} &= \beta_{Y|C(X)} + \frac{\beta_{X|C} \beta_{Y|X(C)} \sigma_{HX}^2}{\beta_{HX}^2 \sigma_{X|C}^2 + \sigma_{HX}^2} \\ &= \beta_{Y|C(X)} + \left\{ \beta_{X|C} \beta_{Y|X(C)} \left( \frac{\sigma_{HX}^2}{\sigma_{H|C}^2} \right) \right\},\end{aligned}$$

and

$$\begin{aligned}\sigma_{YH,C}^2 &= \sigma_{YX,C}^2 + \frac{\beta_{Y|X(C)}^2 \sigma_{HX}^2 \sigma_{X|C}^2}{\beta_{HX}^2 \sigma_{X|C}^2 + \sigma_{HX}^2} \\ &= \sigma_{YX,C}^2 + \beta_{Y|X(C)}^2 \left( \frac{\sigma_{HX}^2 \sigma_{X|C}^2}{\sigma_{H|C}^2} \right).\end{aligned}$$

The equations above provide formulas for the slope parameters and the variance of the model. The remaining model parameter to be considered is the intercept,  $\alpha_{Y|H,C}$ , which can be expressed as a function of the slope parameters derived above and the mean of the variables Y, H, and C. The formula for the model's intercept is as follows:

$$\alpha_{YH,C} = \mu_Y - \left[ \left( \beta_{Y|H(C)} \mu_H \right) + \left( \beta_{Y|C(H)} \mu_C \right) \right].$$

### **G10.1.2 MODELING BLOOD-LEAD AS A FUNCTION OF ONE OR MORE VARIABLES MEASURED WITH ERROR AND OTHER SELECT COVARIATES.**

#### **Definitions and Assumptions**

In the notation that follows, matrices are indicated by bold capital letters and vectors are indicated by underlined letters. Also, squares and square roots of the elements of diagonal matrices are written as the matrix raised to a power (e.g.,  $^2$  or  $^{1/2}$ ).

Define the following variables:

- Y** = The response variable, log of blood-lead concentration.
- R** = A vector (or scalar) of observed Rochester Lead-in-Dust Study covariates measured with error (for the Empirical model this vector consists of the log of the area weighted arithmetic mean of floor wipe dust lead loading, the log of the area weighted arithmetic mean of window sill wipe dust lead loading, and the log of the drip-line soil lead concentration).

- $\underline{H}$  = A vector (or scalar) of observed HUD National Survey covariates measured with error (for the Empirical model this vector consists of the log of the area weighted arithmetic mean of floor blue nozzle vacuum dust lead loading, the log of the area weighted arithmetic mean of window sill blue nozzle vacuum dust lead loading, and the log of the average soil lead concentration).
- $\underline{X}$  = A vector (or scalar) of unobserved covariates measured without error (for the Empirical model this vector consists of the “true” unobserved log of the area weighted arithmetic mean of floor wipe dust lead loading, the “true” unobserved log of the area weighted arithmetic mean of window sill wipe dust lead loading, and the “true” unobserved log of the drip-line soil lead concentration).
- $\underline{C}$  = A vector (or scalar) of remaining covariates (for the Empirical model this variable is the scalar paint/pica hazard). These covariates are assumed to be measured using identical methods in the Rochester Lead-in-Dust Study and the HUD National Survey.

The model assumes

(A)

$$\begin{bmatrix} Y \\ \underline{R}_{p_1 \times 1} \\ \underline{H}_{p_1 \times 1} \\ \underline{X}_{p_1 \times 1} \\ \underline{C}_{p_2 \times 1} \end{bmatrix} \sim N \left\{ \begin{bmatrix} \mu_Y \\ \underline{\mu}_R \\ \underline{\mu}_H \\ \underline{\mu}_X \\ \underline{\mu}_C \end{bmatrix}, \begin{bmatrix} \sigma_Y^2 & & & & \\ -\gamma_{YR} & RR & & & \\ -\gamma_{YH} & RH & HH & & \\ -\gamma_{YX} & RX & HX & XX & \\ -\gamma_{YC} & RC & HC & XC & CC \end{bmatrix} \right\},$$

(B)

For a random sample of size N generated from the distribution in (A):

$$\begin{aligned} \underline{Y}_{N \times 1} &= \underline{1}_N \alpha_{YX,C} + \underline{X} \underline{\beta}_{YX(C)} + \underline{C} \underline{\beta}_{YC(X)} + \underline{e}_{YX,C} & \underline{e}_{YX,C} &\sim N\left(\underline{0}, \sigma_{YX,C}^2 \underline{I}_N\right) \\ \underline{R}_{N \times p_1} &= \underline{X} + \underline{E}_{RX} & \underline{E}_{RX} &\sim N\left(\underline{0}, \underline{I}_N \otimes \underline{I}_{RX}\right) \\ \underline{H}_{N \times p_1} &= \underline{1}_N \underline{\alpha}_{HX}^T + \underline{X} \underline{B}_{HX} + \underline{E}_{HX} & \underline{E}_{HX} &\sim N\left(\underline{0}, \underline{I}_N \otimes \underline{I}_{HX}\right) \\ \underline{X}_{N \times p_1} &= \underline{1}_N \underline{\alpha}_{XC}^T + \underline{C} \underline{B}_{XC} + \underline{E}_{XC} & \underline{E}_{XC} &\sim N\left(\underline{0}, \underline{I}_N \otimes \underline{I}_{XC}\right), \end{aligned}$$

and all errors are independent of one another.

The parameters  $\alpha_{Y|X,C}$ ,  $\beta_{Y|X(C)}$ , and  $\beta_{Y|C(X)}$  represent the intercept and slopes, respectively, associated with a regression of  $Y$  on  $\underline{X}$ (unobserved) and  $\underline{C}$ ; and  $\sigma_{Y|X,C}^2$  is the variability in  $Y$  unexplained by  $\underline{X}$  and  $\underline{C}$ .  $\Sigma_{Ri|Xi}$  is a  $p_1$  by  $p_1$  diagonal matrix with  $i$ th diagonal element equal to  $\sigma_{Ri|Xi}^2$ , the measurement error associated with the  $i$ th covariate in Rochester measured with error.  $\Sigma_{Hi|Xi}$  is defined analogously for HUD. The  $i$ th element of the  $p_1$  by 1 vector  $\alpha_{Hi|X}$  represents a location shift in the distribution of the  $i$ th variable in  $\underline{H}$  relative to the distribution of the  $i$ th variable in  $\underline{X}$ . Similarly,  $\mathbf{B}_{Hi|X}$  is a  $p_1$  by  $p_1$  diagonal matrix with  $i$ th diagonal element representing a scale shift in the distribution of the  $i$ th variable in  $\underline{H}$  relative to the distribution of the  $i$ th variable in  $\underline{X}$ . The  $p_1$  by 1 vector  $\alpha_{X|C}$  and the  $p_2$  by  $p_1$  matrix  $\mathbf{B}_{X|C}$  are the intercepts and slopes, respectively, associated with a regression of  $\underline{X}$  on the covariates in  $\underline{C}$ .  $\Sigma_{X|C}$  is a diagonal matrix with  $i$ th diagonal element equal to the variability in the  $i$ th element of  $\underline{X}$  unexplained by the covariates in  $\underline{C}$ .

In addition, the calculations that follow rely heavily on the assumption that the conditional distribution of  $\underline{X}$  given  $\underline{C}$  ( $\underline{X}/\underline{C}$ ) is the same in the Rochester Lead-in-Dust Study and the HUD National Survey. This assumption will be referred to as an assumption of transportability.

### Parameter Development

Using assumption (A) of Section G10.2.1, normal distribution theory gives the following result for the conditional distribution of  $Y$  given  $\underline{H}$  and  $\underline{C}$ :

$$Y|\underline{H}, \underline{C} \sim N(\mu_{YH,C}, \sigma_{YH,C}^2); \text{ where,}$$

$$\mu_{YH,C} = \mu_Y + \begin{bmatrix} T & T \\ -YH & -YC \end{bmatrix} \begin{bmatrix} HH & \\ HC & CC \end{bmatrix}^{-1} \begin{bmatrix} \underline{H} - \underline{\mu}_H \\ \underline{C} - \underline{\mu}_C \end{bmatrix}$$

and

$$\sigma_{YH,C}^2 = \sigma_Y^2 - \begin{bmatrix} T & T \\ -YH & -YC \end{bmatrix} \begin{bmatrix} HH & \\ HC & CC \end{bmatrix}^{-1} \begin{bmatrix} -YH \\ -YC \end{bmatrix}.$$

Solving for the inverse above and using (B) for substitutions gives:

$$\mu_{YH,C} = \mu_Y + \begin{bmatrix} (\mathbf{B}_{HX}^2 \quad X|C + \quad H|X)^{-1} (\mathbf{B}_{HX} \quad X|C \quad \beta_{Y|X(C)}) \\ \mathbf{B}_{X|C} (\mathbf{B}_{HX}^2 \quad X|C + \quad H|X)^{-1} (\quad H|X \quad \beta_{Y|X(C)}) + \beta_{Y|C(X)} \end{bmatrix}^T \begin{bmatrix} \underline{H} - \underline{\mu}_H \\ \underline{C} - \underline{\mu}_C \end{bmatrix}$$

and

$$\sigma_{YH,C}^2 = \sigma_{Y|X,C}^2 + \beta_{Y|X(C)}^T (\mathbf{B}_{HX}^2 \quad X|C + \quad H|X)^{-1} \quad H|X \quad X|C \quad \beta_{Y|X(C)}.$$

Upon substituting the equality  $\beta_{H|C} = \mathbf{B}_{H|X}^2 \beta_{X|C} + \beta_{H|X}$  into the equation above, assumptions (A) and (B) yield the following slope and variance estimates for the Empirical model:

$$\beta_{YH(C)} = (\mathbf{I}_{p_1} - \beta_{H|X} \beta_{H|C}^{-1}) \mathbf{B}_{H|X}^{-1} \beta_{YX(C)},$$

$$\beta_{YC(H)} = \beta_{YC(X)} + \mathbf{B}_{X|C} \beta_{H|X} \beta_{H|C}^{-1} \beta_{YX(C)},$$

and

$$\sigma_{YH,C}^2 = \sigma_{YX,C}^2 + \beta_{YX(C)}^T \beta_{H|X} \beta_{H|C}^{-1} \beta_{X|C} \beta_{YX(C)}.$$

Finally, the formula used to estimate the Empirical model's intercept is given by:

$$\alpha_{YH,C} = \mu_Y - \beta_{YH(C)}^T \mu_H - \beta_{YC(H)}^T \mu_C.$$

### G10.1.3 PARAMETER ESTIMATION

Sections G10.1 and G10.2 above provide equations for the model parameters after adjusting for differences between sample collection methods in the Rochester Lead-in-Dust Study and the HUD National Survey. Each variable appearing in the above equations first must be estimated in order to obtain the final estimates of the Empirical model parameters. The following text describes the methodology used to estimate the variables that appear in the final Empirical model formulas of Section G10.2.2. Note that all the variance components described below are provided in Table G10.1.

In the discussion that follows, all estimates for parameters from the HUD National Survey are weighted estimates. The weights correspond to the 1993 American Housing Survey adjusted to 1997. Weights are used because the HUD National Survey is designed to be nationally representative and each observation in HUD is weighted with respect to the population it represents.

#### Estimation of Parameters Used in Deriving the Empirical Model Slopes and Variance

For estimating the parameters  $\beta_{Y|X(C)}$ ,  $\beta_{Y|C(X)}$ , and  $\sigma_{Y|X,C}^2$ , a classic errors-in-variables model is applied to the Rochester data. The application of this model requires an estimate of the true measurement errors associated with the elements of  $\mathbf{R}$  (i.e.,  $\mathbf{R}_{iX}$ ). For further detail on the errors-in-variables model and the estimation of measurement errors associated with both  $\mathbf{R}$  and  $\mathbf{H}$  (i.e.,  $\mathbf{R}_{iX}$  and  $\mathbf{H}_{iX}$ ), see Sections 2 and 3 below.

The  $i$ th diagonal element of  $\mathbf{R}_{iC}$  and  $\mathbf{H}_{iC}$  is estimated by the mean squared error from a least squares regression of the  $i$ th element of  $\mathbf{R}$  on the covariate vector  $\mathbf{C}$  in Rochester and the

mean squared error from a weighted least squares regression of the  $i$ th element of  $\underline{H}$  on the covariate vector  $\underline{C}$  in HUD, respectively. For example, in the Empirical model, the first diagonal element of  $\Sigma_{R|C}$  is estimated by the mean squared error from the least squares regression of log floor wipe dust lead loading on paint/pica hazard in the Rochester data set.

Using assumption (B) from Section 1.2.1 along with the assumption that all errors are independently distributed yields

$$\Sigma_{H|C} = \mathbf{B}_{HX}^2 \Sigma_{X|C} + \Sigma_{HX} \quad \text{and} \quad \Sigma_{R|C} = \Sigma_{X|C} + \Sigma_{RX}.$$

The estimate of  $\Sigma_{X|C}$  is derived easily from the second equality given above. Using both equalities above,  $\mathbf{B}_{HX}$  is estimated as

$$\mathbf{B}_{HX} = [(\Sigma_{H|C} - \Sigma_{HX}) (\Sigma_{R|C} - \Sigma_{RX})^{-1}]^{1/2}.$$

Since  $\underline{X}$  is a latent variable, the parameter  $\mathbf{B}_{X|C}$  cannot be observed. However, from assumption (A) in Section 1.2.1,

$$E(\mathbf{R}) = E(\mathbf{X}) = \mathbf{1} \alpha_{X|C}^T + \mathbf{C} \mathbf{B}_{X|C}.$$

So,  $\mathbf{B}_{X|C}$  is estimated by  $\mathbf{B}_{R|C}$ , which is obtained from a least squares regression of  $\underline{R}$  on the covariate vector  $\underline{C}$  in Rochester.

### Estimation of Parameters Used in Deriving the Empirical Model Intercept

Estimates of the slope parameters,  $\beta_{Y|H(C)}$  and  $\beta_{Y|C(H)}$ , follow from Section G1-1.3.1 above. The mean parameters,  $\mu_H$  and  $\mu_C$ , are estimated by weighted means of  $\underline{H}$  and  $\underline{C}$ , respectively, as observed in the HUD National Survey. Unfortunately,  $Y$  is not measured in the HUD National Survey; therefore,  $\mu_Y$  can not be estimated directly from HUD data. As a result, using the intercept formula given in Section G10.1.2 requires an alternative estimate of  $\mu_Y$ .

Given the intent of the Empirical model, the alternative estimate that is used for  $\mu_Y$  is the mean of the log of blood-lead concentration in the NHANES III data set. This decision was arrived at for the following reasons:

- (1) NHANES III data provide a perfectly legitimate estimate of  $\mu_Y$ , the national mean of log(blood-lead concentration), with the added appeal of guaranteeing the model's predicted national mean equals the targeted national mean.
- (2) The only other sensible estimate of  $\mu_Y$ , the sample mean from the Rochester study, may be a poor estimator since the distribution of covariates in Rochester is different from the distribution of covariates in HUD. Subsequently, mean blood-lead

concentrations (being a function of the covariates) can be expected to differ across studies as well.

## G10.2 REGRESSION PARAMETER ESTIMATION IN THE PRESENCE OF MEASUREMENT ERROR

Let

$$Y = X\beta + \epsilon \quad (1)$$

where,

$Y$  = an  $n \times 1$  vector containing the  $n$  values of the dependent variable,

$X$  = an  $n \times p$  matrix where each column contains the  $n$  values of one independent variable in the regression model (in a model with an intercept term, one of the columns would be a column of ones),

$\beta$  = a  $p \times 1$  vector of regression coefficients, and

$\epsilon$  = an  $n \times 1$  vector of random error terms.

In a standard regression model it is assumed that  $X$  is a matrix of fixed and known constants,  $\beta$  is a vector of fixed and unknown constants, and  $\epsilon$  is distributed as  $MVN(0, \sigma^2 I)$  where  $MVN(\mu, \Sigma)$  represents a multivariate normal distribution with mean vector  $\mu$  and covariance matrix  $\Sigma$ . Estimates of regression parameters for this standard regression model are obtained as follows:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$\hat{\sigma}^2 = Y^T [I - X(X^T X)^{-1} X^T] Y / (n-p) \quad (2)$$

$$\text{Cov}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1}$$

In the presence of measurement error, it is assumed that

$$Y = R\beta + \epsilon \quad (3)$$

where,

$X$  = an  $n \times p$  matrix of fixed but unknown constants representing the values of the independent variables if measured without error;

$R$  =  $X + \epsilon$  is an  $n \times p$  matrix representing the values of the independent variables observed with measurement error, and

= an  $n \times p$  matrix of the random measurement errors associated with each of the observed values of the independent variables.

$Y$  and  $\epsilon$  are as defined above. It is assumed that  $\epsilon$  is distributed as  $MVN(0, I_{\otimes})$  where  $I_{\otimes}$  is known and  $\epsilon$  is stochastically independent of  $X$ . Under this measurement error model, estimates of regression parameters are obtained as follows:

$$\hat{\beta} = (R^T R - n^{-1})^{-1} R^T Y$$

$$MSE_{Y|R} = Y^T [I - R(R^T R - n^{-1})^{-1} R^T] Y / (n-p) \quad (4)$$

$$C\hat{o}v(\hat{\beta}) = MSE_{Y|R} (R^T R - n^{-1})^{-1} R^T R (R^T R - n^{-1})^{-1}$$

These estimators are equivalent to those recommended in Equations (2.2.11) and (2.2.12) by Fuller (Measurement Error Models, 1987).

It can be shown that

- 1a. The difference between  $[(R^T R - n^{-1}) / n]$  and  $[X^T X / n]$  converges in probability to zero as  $n \rightarrow \infty$ ;
- 1b. The difference between  $[(R^T R - (n-p)^{-1}) / (n-p)]$  and  $[X^T X / (n-p)]$  converges in probability to zero as  $n \rightarrow \infty$ ; and
2. The difference between  $[R^T Y / n]$  and  $[X^T Y / n]$  converges in probability to zero as  $n \rightarrow \infty$ .

Additionally, it is assumed that

3.  $X$  is distributed as  $MVN(\underline{1} \mu_x^T, I_{\otimes x})$  and is stochastically independent of both  $\epsilon$  and  $Y$ ,

and hence all inferences are based on the conditional distribution of  $Y$  given  $X$ .

### G10.3 DETAILS ON MEASUREMENT ERROR ESTIMATION

The statistical models that account for differences in sample collection methods used in the Rochester Lead-in-Dust Study and the HUD National Survey require estimates of variance components associated with the dust-lead and soil-lead predictor variables in each study. In the notation that follows, a subscript of “f” represents floor dust lead loadings, a subscript of “w” represents window sill dust lead loadings, and a subscript of “s” represents soil lead concentrations. Specifically, we need to obtain the following estimates:

1. The “between homes” variance of observed values of the dust-lead and soil-lead predictor variables for Rochester ( $\sigma^2_{Rf}$ ,  $\sigma^2_{Rw}$ , and  $\sigma^2_{Rs}$  corresponding to the diagonal elements of  $\Sigma_{RR}$  from Section G10.1.2 above) and for HUD ( $\sigma^2_{Hf}$ ,  $\sigma^2_{Hw}$ , and  $\sigma^2_{Hs}$  corresponding to the diagonal elements of  $\Sigma_{HH}$  from Section G10.1.2 above),
2. After adjusting for the effects of covariates included in the Empirical model, the “between homes” variance of observed values of the dust-lead and soil-lead predictor variables for Rochester ( $\sigma^2_{Rf/C}$ ,  $\sigma^2_{Rw/C}$ , and  $\sigma^2_{Rs/C}$  corresponding to the diagonal elements of  $\Sigma_{R/C}$  from Section G10.1.3 above) and for HUD ( $\sigma^2_{Hf/C}$ ,  $\sigma^2_{Hw/C}$ , and  $\sigma^2_{Hs/C}$  corresponding to the diagonal elements of  $\Sigma_{H/C}$  from Section G10.1.3 above), and
3. The “within homes” variance attributable to measurement error associated with the dust-lead and soil-lead predictor variables for Rochester ( $\sigma^2_{Rf/Xf}$ ,  $\sigma^2_{Rw/Xw}$ , and  $\sigma^2_{Rs/Xs}$  corresponding to the diagonal elements of  $\Sigma_{R/X}$  from Section G10.1.2 above) and for HUD ( $\sigma^2_{Hf/Xf}$ ,  $\sigma^2_{Hw/Xw}$ , and  $\sigma^2_{Hs/Xs}$  corresponding to the diagonal elements of  $\Sigma_{H/X}$  from Section G10.1.2 above).

The following four sections provide details on the methods used to estimate each of the above variance components.

### **G10.3.1 “BETWEEN HOMES” VARIANCE OF DUST-LEAD AND SOIL-LEAD PREDICTOR VARIABLES**

Between home variances of log(floor wipe dust-lead loading), log(window sill wipe dust-lead loading), and log(drip-line soil lead concentration) from the Rochester Lead-in-Dust Study are represented by  $\sigma^2_{Rf}$ ,  $\sigma^2_{Rw}$ , and  $\sigma^2_{Rs}$ , respectively. Each variance is estimated by the sample variance of the respective variable as observed in the Rochester dataset. Specifically,

$$\sigma^2_{Rf} = \frac{1}{N-1} \sum_{i=1}^N (Rf_i - \bar{Rf})^2, \quad \sigma^2_{Rw} = \frac{1}{N-1} \sum_{i=1}^N (Rw_i - \bar{Rw})^2, \quad \text{and} \quad \sigma^2_{Rs} = \frac{1}{N-1} \sum_{i=1}^N (Rs_i - \bar{Rs})^2,$$

where  $Rf_i$  and  $Rw_i$  represent the floor and window sill dust-lead predictor variables and  $Rs_i$  represents the soil-lead predictor variable associated with each home in the Rochester Study.

$\bar{Rf}$  represents the sample mean of log(floor dust lead loading) among all homes from the Rochester Study,  $\bar{Rw}$  represents the sample mean of log(window sill dust lead loading) among all homes from the Rochester Study, and  $\bar{Rs}$  represents the sample mean of log(drip-line soil lead concentration) among all homes from the Rochester Study.

Between home variances of log(blue nozzle floor dust lead loading), log(blue nozzle window sill dust lead loading), and log(average soil lead concentration) from the HUD Survey are represented by  $\sigma^2_{Hf}$ ,  $\sigma^2_{Hw}$ , and  $\sigma^2_{Hs}$ , respectively. In contrast to the Rochester between home variance estimates described above, HUD Survey between home variance estimates are weighted.

Each observation is weighted using weights from the 1993 American Housing Survey adjusted to 1997. Weights are used because the HUD Survey is designed to be nationally representative and each observation in HUD is weighted with respect to the population it represents. Specifically,

$$\sigma_{Hf}^2 = \frac{\sum_{i=1}^N w_i (Hf_i - \bar{Hf})^2}{\sum_{i=1}^N w_i - 1}, \quad \sigma_{Hw}^2 = \frac{\sum_{i=1}^N w_i (Hw_i - \bar{Hw})^2}{\sum_{i=1}^N w_i - 1}, \quad \text{and} \quad \sigma_{Hs}^2 = \frac{\sum_{i=1}^N w_i (Hs_i - \bar{Hs})^2}{\sum_{i=1}^N w_i - 1},$$

where  $w_i$  represents the population weight for the  $i$ th home in the HUD National Survey,  $Hf_i$ ,  $Hw_i$ , and  $Hs_i$  represent the floor and window sill dust-lead predictor variables and soil-lead predictor variable associated with each house in the HUD National Survey,  $\sum_{i=1}^N w_i = N$ , and  $\bar{Hf}$ ,  $\bar{Hw}$ , and  $\bar{Hs}$  are weighted means calculated as follows:

$$\bar{Hf} = \frac{\sum_{i=1}^N w_i Hf_i}{\sum_{i=1}^N w_i}, \quad \bar{Hw} = \frac{\sum_{i=1}^N w_i Hw_i}{\sum_{i=1}^N w_i}, \quad \text{and} \quad \bar{Hs} = \frac{\sum_{i=1}^N w_i Hs_i}{\sum_{i=1}^N w_i}.$$

### **G10.3.2 COVARIATE ADJUSTED "BETWEEN HOMES" VARIANCE OF DUST-LEAD AND SOIL-LEAD PREDICTOR VARIABLES**

$\sigma_{Rf|C}^2$ ,  $\sigma_{Rw|C}^2$ ,  $\sigma_{Rs|C}^2$ ,  $\sigma_{Hf|C}^2$ ,  $\sigma_{Hw|C}^2$ , and  $\sigma_{Hs|C}^2$  represent the portion of between home variance ( $\sigma_{Rf}^2$ ,  $\sigma_{Rw}^2$ ,  $\sigma_{Rs}^2$ ,  $\sigma_{Hf}^2$ ,  $\sigma_{Hw}^2$ , and  $\sigma_{Hs}^2$ , respectively) that remains after adjusting for the other covariates included in the Empirical model. An estimate of these quantities can be obtained from the mean squared error of a least squares regression of the variables (Rf, Rw, Rs, Hf, Hw, or Hs) on the other covariates in the Empirical model. The least squares regression model treats the covariates as fixed; and the resulting mean squared error estimates the remaining variability of the variable in the presence of the fixed covariates.

The covariate adjusted between home variances from the Rochester Lead-in-Dust Study,  $\sigma_{Rf|C}^2$ ,  $\sigma_{Rw|C}^2$ , and  $\sigma_{Rs|C}^2$ , are estimated using mean squared errors obtained from ordinary least squares regressions of log(floor wipe dust lead loading), log(window sill wipe dust lead loading), and log(drip-line soil lead concentration) on the remaining model covariates, respectively. Similarly, the covariate adjusted between home variances from the HUD Survey,  $\sigma_{Hf|C}^2$ ,  $\sigma_{Hw|C}^2$ , and  $\sigma_{Hs|C}^2$ , are estimated using mean squared errors obtained from weighted least squares regressions of log(floor wipe dust lead loading), log(window sill wipe dust lead loading), and log(average soil lead concentration) on the remaining model covariates, respectively. Again, least squares regressions involving HUD data are weighted because the HUD Survey is designed to be nationally representative.

### G10.3.3 MEASUREMENT ERROR ASSOCIATED WITH PREDICTOR VARIABLES

The dust-lead predictor variables in the statistical models represent area-weighted arithmetic average individual sample dust-lead loadings from floors and window sills. The following equation represents the three sources of variability that must be accounted for in an estimate of measurement error for these dust-lead predictor variables:

$$\sigma^2_{\text{Measurement Error}} = \sigma^2_{\text{Spatial}} + \sigma^2_{\text{Sampling}} + \sigma^2_{\text{Laboratory}} ,$$

where  $\sigma^2_{\text{Spatial}}$  represents the variability in dust-lead levels among all possible locations on the surface being tested,  $\sigma^2_{\text{Sampling}}$  represents variability in the collection of dust from the surface, and  $\sigma^2_{\text{Laboratory}}$  represents variability in the chemical analysis of the sample. This definition of measurement error is consistent with the interpretation of each predictor variable as exposure to lead from floor or window sill dust found at the primary residence at the time of sampling. Thus there was no attempt to estimate a component of variation associated with temporal variability. The following two subsections contain details on estimating the measurement associated with dust-lead and soil-lead predictor variables.

#### G10.3.3.1 Measurement Error Associated with Dust-Lead Predictor Variables

Several sources of data were considered for providing information about the variability in dust sample results due to measurement error, including field duplicate data and data that included multiple dust samples (of a given component type) collected from within the same house. Since the predictor variables included in the statistical models represented area weighted averages of multiple dust sample results collected within a house, the individual sample lead loading results from the Rochester Lead-in-Dust Study and the HUD National Survey were used to assess the measurement error. Specifically, let

$\text{Dust}_{ijk}$  represent the dust-lead loading from the  $k$ th component type (floor or window sill) from the  $j$ th location within the  $i$ th residential unit,

$\text{Area}_{ijk}$  represent the area of the sample from the  $k$ th component type from the  $j$ th location within the  $i$ th residential unit, and

The following model was then fitted separately for floors and window sills from each study to estimate the within house variability in dust-lead loadings between individual dust samples:

$$\ln(\text{Dust}_{ijk}) = \ln(\mu_k) + H_{ik} + E_{ijk} ,$$

where  $\mu_k$  is the geometric mean of  $\text{Dust}_{ijk}$  among all samples of component  $k$ ,  $H_{ik}$  is the random effect associated with the  $i$ th House, and  $E_{ijk}$  is the random within-house error term associated with  $\text{Dust}_{ijk}$ .  $H_{ik}$  is assumed to follow a normal distribution with mean zero and variance  $\sigma^2_{\text{Between Houses}}$ , and  $E_{ijk}$  is assumed to follow a normal distribution with mean zero and variance  $\sigma^2_{\text{Within Houses}}$ .

$\sigma^2_{\text{Between Houses}}$  characterizes the variability between houses.  $\sigma^2_{\text{Within Houses}}$  characterizes the variability within a house; attributed to a combination of spatial, sampling, and laboratory variability. The following two subsections describe how weights were used with the above model to calculate the measurement error variance components  $\sigma^2_{\text{Rf|Xf}}$  and  $\sigma^2_{\text{Rw|Xw}}$  corresponding to the Rochester Lead-in-Dust Study, and  $\sigma^2_{\text{Hf|Xf}}$  and  $\sigma^2_{\text{Hw|Xw}}$  corresponding to the HUD National Survey.

### Rochester Lead-in-Dust Study

Since area weighted (arithmetic) mean floor and window sill dust-lead loadings were used to characterize the dust-lead levels in each house in the Rochester Lead-in-Dust Study, the above model was fitted using weights corresponding to the percent of total area that was associated with each sample:

$$Weight_{ijk} = \frac{Area_{ijk}}{\sum_{j=1}^{n_{ik}} Area_{ijk}}$$

where  $n_{ik}$  is the number of samples collected from component k within the ith house.

Values of  $\sigma^2_{\text{Within Houses}}$  calculated in this weighted analysis are used as estimates of  $\sigma^2_{\text{Rf|Xf}}$  and  $\sigma^2_{\text{Rw|Xw}}$  in the statistical models described in Sections 1 and 2 of this appendix. In actuality, these estimates of  $\sigma^2_{\text{Rf|Xf}}$  and  $\sigma^2_{\text{Rw|Xw}}$  correspond more closely to measurement error in area weighted geometric mean dust-lead loadings from floors and window sills within each house. Table G10-1 provides estimates of  $\sigma^2_{\text{Rf|Xf}}$  and  $\sigma^2_{\text{Rw|Xw}}$  as calculated from the Rochester Lead-in-Dust Study data.

### HUD National Survey

Since area weighted (arithmetic) mean floor and window sill dust-lead loadings were used to characterize the dust-lead levels in each house in the HUD National Survey, the above model was fitted using a combination of weights corresponding to the percent of total area that was associated with each sample, and the survey weight associated with each home sampled:

$$Weight_{ijk} = \frac{Area_{ijk}}{\sum_{j=1}^{n_{ik}} Area_{ijk}} \cdot \frac{HSW_i}{\frac{1}{n} \sum_{i=1}^n HSW_i}$$

where  $n_{ik}$  is the number of samples collected from component k within the ith house, n is the number of homes included in the HUD National Survey, and  $HSW_i$  is the survey weight associated with the ith home in the HUD National Survey.

**Table G10-1. Components of Variation Used to Implement an Adjustment of the Rochester Multi-Media Predictive Model for Use with Environmental Lead Levels as Measured in the HUD National Survey.**

Study	Parameter	Final Empirical Model
-------	-----------	-----------------------

Rochester Lead-in-Dust	$\sigma^2_{Rf Xf}$	0.2082
	$\sigma^2_{Rw Xw}$	0.5708
	$\sigma^2_{Rs Xs}$	0.3898
	$\sigma^2_{Rf C}$	1.3323
	$\sigma^2_{Rw C}$	1.8505
	$\sigma^2_{Rs C}$	1.6497
	$\sigma^2_{Rf}$	1.3410
	$\sigma^2_{Rw}$	1.8592
	$\sigma^2_{Rs}$	1.6640
HUD National Survey	$\sigma^2_{Hf Xf}$	0.6125
	$\sigma^2_{Hw Xw}$	1.6937
	$\sigma^2_{Hs Xs}$	0.3016
	$\sigma^2_{Hf C}$	2.3589
	$\sigma^2_{Hw C}$	5.2881
	$\sigma^2_{Hs C}$	2.2125
	$\sigma^2_{Hf}$	2.3767
	$\sigma^2_{Hw}$	5.3225
	$\sigma^2_{Hs}$	2.2434

Values of  $\sigma^2_{\text{Within Houses}}$  calculated in this weighted analysis are used as estimates of  $\sigma^2_{Hf|Xf}$  and  $\sigma^2_{Hw|Xw}$  in the statistical models described in Sections 1 and 2 of this appendix. In actuality, these estimates of  $\sigma^2_{Hf|Xf}$  and  $\sigma^2_{Hw|Xw}$  correspond more closely to measurement error in area weighted geometric mean dust-lead loadings from floors and window sills within each house. Table G10-1 provides estimates of  $\sigma^2_{Hf|Xf}$  and  $\sigma^2_{Hw|Xw}$  as calculated from the HUD National Survey data.

### G10.3.3.2 Measurement Error Associated with Soil-Lead Predictor Variables

Due to the fact that there was analytical information available from only one composite drip-line soil sample collected from each home in Rochester, we were unable to derive an estimate of measurement error ( $\sigma^2_{Rs|Xs}$ ) using data observed in the Rochester Lead-in-Dust Study. We therefore derived estimates of measurement error associated with soil-lead predictor variables in both the Rochester Study ( $\sigma^2_{Rs|Xs}$ ) and the HUD National Survey ( $\sigma^2_{Hs|Xs}$ ) using data collected in the HUD National Survey.

Up to three different soil samples were collected from each home in the HUD National Survey: an entryway soil sample, a drip-line soil sample, and a remote soil sample. The soil-lead predictor variables used in the Empirical Model can be regarded as weighted averages of these multiple soil sample results collected within each HUD National Survey home. Specifically, we considered the Rochester soil-lead predictor variable to be representative of the average between entryway and drip-line soil samples collected in the HUD National Survey (each sample receiving weight of 0.5). The HUD National Survey predictor variable was constructed as the average between the remote soil sample and the average between entryway and drip-line soil samples (remote sample receiving weight of 0.5, and drip-line and entryway samples each receiving weight of 0.25). The individual soil-lead concentration results from the HUD National Survey were used to assess the measurement error variance components as follows:

Let  $Soil_{ij}$  represent the soil-lead concentration from the  $j$ th location within the  $i$ th residential unit. The following model was then fitted to estimate the within house variability in soil-lead concentration between individual soil samples:

$$\ln(Soil_{ij}) = \ln(\mu) + H_i + E_{ij} ,$$

where  $\mu_k$  is the geometric mean of  $Soil_{ij}$  among all samples,  $H_i$  is the random effect associated with the  $i$ th House, and  $E_{ij}$  is the random within-house error term associated with  $Soil_{ij}$ .  $H_i$  is assumed to follow a normal distribution with mean zero and variance  $\sigma^2_{\text{Between Houses}}$ , and  $E_{ij}$  is assumed to follow a normal distribution with mean zero and variance  $\sigma^2_{\text{Within Houses}}$ .

$\sigma^2_{\text{Between Houses}}$  characterizes the variability between houses.  $\sigma^2_{\text{Within Houses}}$  characterizes the variability within a house; attributed to a combination of spatial, sampling, and laboratory variability. Weights were used with the above model to calculate the measurement error variance components  $\sigma^2_{Rs|Xs}$  corresponding to the Rochester Lead-in-Dust Study, and  $\sigma^2_{Hs|Xs}$  corresponding to the HUD National Survey as follows:

$$Weight_{ij} = \frac{W_{ij} \cdot HSW_i}{\frac{1}{N} \sum_{i=1}^n HSW_i}$$

where  $n$  is the number of homes included in the HUD National Survey,  $HSW_i$  is the survey weight associated with the  $i$ th home in the HUD National Survey, and  $W_{ij}$  is the weight corresponding to each individual sample being averaged:

Soil Sample Location	Value of $W_{ij}$ when Estimating	
	$\sigma^2_{Rs Xs}$	$\sigma^2_{Rs Xs}$

Drip-line	0.5	0.25
Entryway	0.5	0.25
Remote	0.0	0.5

Values of  $\sigma^2_{\text{Within Houses}}$  calculated in this weighted analysis are used as estimates of  $\sigma^2_{\text{Rs|Xs}}$  and  $\sigma^2_{\text{Hs|Xs}}$  in the statistical models described in Sections 1 and 2 of this appendix. Table G10-1 provides estimates of  $\sigma^2_{\text{Rs|Xs}}$  and  $\sigma^2_{\text{Hs|Xs}}$  as calculated from the HUD National Survey data.

#### **G10.3.4 EFFECT OF IMPUTING BLUE NOZZLE WINDOW SILL DUST LEAD LOADINGS IN THE HUD DATASET ON ESTIMATED VARIANCE COMPONENTS**

The floor and window sill dust-lead loading predictor variable was imputed for several of the homes in the HUD National Survey (for homes that did not include any dust samples from window sills) in an effort to keep as many homes in the analysis as possible, and thus maintain its property of being nationally representative (with appropriate survey weights).

The HUD sample used for calculating  $\sigma^2_{\text{Hf}}$ ,  $\sigma^2_{\text{Hw}}$ , and  $\sigma^2_{\text{Hs}}$  includes imputed values, and is the same for the preliminary and final Empirical models; therefore the estimate of  $\sigma^2_{\text{Hf}}$  is consistent across the rows in Table G10-1. The HUD sample used for calculating  $\sigma^2_{\text{Hf|C}}$ ,  $\sigma^2_{\text{Hw|C}}$ , and  $\sigma^2_{\text{Hs|C}}$  also includes imputed values.

In contrast,  $\sigma^2_{\text{Hf|Xf}}$ ,  $\sigma^2_{\text{Hw|Xw}}$ , and  $\sigma^2_{\text{Hs|Xs}}$  can only be estimated using those houses in which floor and window sill dust samples and soil samples were collected. Values for  $\sigma^2_{\text{Hf|Xf}}$  were therefore calculated separately for each version of the empirical model.

#### **G10.3.5 ESTIMATED COMPONENTS OF VARIATION**

The following table provides the components of variation used to implement an adjustment of the Rochester multi-media predictive model for use with environmental lead levels as measured in the HUD National Survey.

### **G10.4 BOOTSTRAP ESTIMATION OF STANDARD ERRORS**

In ordinary least squares regression, formulas are readily available for calculating standard errors associated with the model's parameter estimates. For the parameter estimates of the model that accounts for differences in sample collection methods used in the Rochester Lead-in-Dust Study and the HUD National Survey, no such simple formulas exist. As a result, 48 standard errors can only be approximated. The method of approximation used for estimating the standard errors corresponding to the parameters of the adjusted model is a basic bootstrap algorithm, which is described below. Note that the following definitions and algorithm are taken directly from Efron and Tibshirani, "An Introduction to the Bootstrap," 1993 pp. 45-47.

Let  $\vec{x} = (x_1, x_2, \dots, x_n)$  represent a sample dataset.

Let  $\hat{\theta}(\vec{x})$  be an estimator of a parameter of interest  $\theta$ , where  $\hat{\theta}(\vec{x})$  is such that its standard error is not easily obtained.

Define  $\hat{F}$  to be the empirical distribution that assigns probability  $1/n$  to each of the  $n$  observations in the sample dataset.

Define a bootstrap sample,  $\vec{x}^{(b)}$ , as a random sample of size  $n$  drawn with replacement from  $\hat{F}$ .

A bootstrap estimate of the standard error of  $\hat{\theta}(\vec{x})$  is obtained as follows:

1. Collect  $B$  independent bootstrap samples,  $\vec{x}_1^{(b)}, \vec{x}_2^{(b)}, \dots, \vec{x}_B^{(b)}$ .
2. For each bootstrap sample, calculate  $\hat{\theta}(\vec{x}_i^{(b)})$ ,  $i=1, 2, \dots, B$ .
3. Estimate the standard error of  $\hat{\theta}(\vec{x})$  as:

$$\hat{se}_B = \left\{ \sum_{i=1}^B \left[ \hat{\theta}(\vec{x}_i^{(b)}) - \bar{\theta}^{(b)} \right]^2 / (B-1) \right\}^{1/2}, \text{ where } \bar{\theta}^{(b)} = \sum_{i=1}^B \hat{\theta}(\vec{x}_i^{(b)}) / B.$$

The above algorithm is used to estimate the standard errors of the estimators described in Section 1 (  $\beta_{Y|Hf(Hw, Hs, C)}$ ,  $\beta_{Y|Hw(Hf, Hs, C)}$ ,  $\beta_{Y|Hs(Hf, Hw, C)}$ ,  $\beta_{Y|C(Hf, Hw, Hs)}$ , and  $\alpha_{Y|Hf, Hw, Hs, C}$ ). Because the adjustment procedure is based on data from the Rochester Lead-in-Dust Study, the Rochester dataset is treated as the sample dataset,  $\vec{x}$ . Data from the HUD National Survey are held fixed in the implementation of the algorithm. In essence, the adjusted model parameters are viewed as functions of sample data (Rochester dataset) that are calibrated to correspond to population values (HUD dataset). Thus, their variability is assumed, in this preliminary assessment, to stem from the Rochester dataset only.

Finally, observe that,

$$\lim_{B \rightarrow \infty} \hat{se}_B = se_{\hat{F}}.$$

That is, as the number of bootstrap replications increases, the estimated standard error approaches the population standard error; where the population distribution is estimated by  $\hat{F}$ . Efron and Tibshirani (1993) recommend between 25 and 200 bootstrap replications for adequate approximations. 200 bootstrap replications were used in the application of the bootstrap algorithm to approximate the standard errors of parameters in the adjusted model.