Comments from Dr. Sherry Dixon, National Center for Healthy Housing (NCHH), on EPA's Approach for Developing Lead Dust Hazard Standards for Residences (November 5, 2010 Draft)

I have some concerns about the statistical methodology employed as described in my comments (in particular see comments 9 , 12 and 13).

1. **Detection Limits**

Pg 10 "In addition,14 percent of floor dust loading samples and 36 percent of window sill loading samples are below the detection limit (0.16 μg for floors and 2 μg for sills). Dixon et al (2009) chose the common method of substitution with LOD/√2 (where LOD is level of detection). Although recommended by NHANES for use with its data, this approach, however, can skew the distribution and introduce bias in the regression estimates."

Comment: You should point out that because sill areas sampled varied from home to home, the sill loadings in the homes with lead below the detection limits have different loading values and hence impact analysis less than for floors, where one square foot was always sampled.

2. **Main Effect of Log Floor Dust Lead Loading**

Pg 11 "Floor-dust lead loading enters into the log-log model fit by Dixon et al (2009) only in the form of interaction terms. For the Agency's reanalysis, models were explored where ln(floor-dust loading) was included both as a main effect and in the interaction terms."

Pg 13 "The reason for not including floor dust as a main effect in the model (e.g., whether in the presence of the other covariates it became nonsignificant) was not explained by Dixon et al (2009)."

Comment: In the model construction, the interaction of log floor dust lead loading and surface/condition was found to be significant (p=0.064) when the main effect for log floor dust lead loading was in the model. When the model was run for presentation of the parameter estimates, the main effects for log floor dust lead loading was excluded from the model so that the parameter estimates would be easier to interpret. Because floor surface/condition is a nominal variable, the two models are equivalent. Table 1 gives the parameter estimates related to floor dust lead loading from Table 4 of Dixon et al (2009). Table 2 gives the parameters related to floor dust lead loading from the model with the main effect for floor dust lead loading. As you see, the parameter estimates are identical.

Note that there is no main effect for floor surface/condition. This effect was purposely and meaningfully excluded because it doesn't make sense that blood leads would vary between children in homes with and without "smooth & cleanable or carpeted floors" if there is no floor dust lead.

3. **Page 13. Imputation of Missing Sill Dust Lead Loadings**

Pg 13. "In their blood- lead regression, Dixon et al (2009) chose to impute the missing values based on an unweighted regression that included only floor dust as an explanatory variable: ln(sill PbD) = 2.654 + 0.524 × ln(floor PbD)… These approaches of dealing with missing values introduced an element of collinearity into the model that could have biased the regression coefficients and standard error estimates. That is, the sill-dust values in the model are actually just transformed floor-dust values, and thus are perfectly correlated with them. For the Agency reanalysis, a regression model for sill dust was developed that included not only floor dust but also other significant covariates, assuming that such a model would explain more of the variance in the window-sill dust levels and thus provide a more reliable imputation of missing sill dust values"

Comment: Collinearity is present in the model to predict log blood lead even when missing sill dust lead loadings are not imputed. The imputation may add to the collinearity but it is unclear why EPA believes their method of prediction with a full complement of covariates would lead to less collinearity because the same covariates are predictors in the blood lead model.

4. Missing values

Pg 17 "The models were fitted to 2,055 records that had blood lead, floor dust lead loading, and sill dust lead loading measurements. Of these, sill dust measurements were imputed using the regression model described in Section 3.2.2."

Comment: The document does not explain how missing values were handled. The dataset contained missing values of year of construction, PIR, smoker in house, and floor dust lead that had to be accounted for.

5. Interpreting regression coefficients

Pg. 18 "When the empirical model was used, the coefficient for floor dust concentration was slightly more than 100 times that for window sill lead concentration. In the regression that used dust concentrations from the mechanistic model as its inputs, the coefficient for floor dust was approximately 700 times that for sill dust concentrations."

Comment: Comparison of the regression coefficients isn't relevant unless both floor and sill dust are standardized before inclusion into the model

6. Pg18. Table 3-4 QL models

Comments: It is unclear whether carpeted floors were combined with smooth & cleanable floors. Footnote c and d are identical - it appears d is in error

7. High dust lead levels (pg 46)

Pg 46 "As noted in Section 3.3, the best fitting log-log model of these data predicts supra-linearity at low exposures and decreasing blood-lead levels with exposure at high dust levels."

Comment: The slope between log blood lead and log floor dust lead is non-significant at higher floor dust lead levels. Relevant results based on the Dixon et al (2009) model:

- *For smooth & cleanable or carpeted floors:*
  - *The slope between blood lead and floor dust is significantly positive at 17 µg/ft^2 or lower ($p<0.1$).*
  - *The slope between blood lead and floor dust is is non-significant ($p>0.1$) at 18 µg/ft^2 or higher. The slope is negative above floor dust loading of 46 µg/f^{t2} but the slope is not significantly different from zero (e.g., $p=0.992$ at 46 µg/ft^2, $p=0.399$ at the maximum of 126.5 µg/ft^2)*
- *For non-smooth & cleanable and non-carpeted floors:*
  - *The slope between blood lead and floor dust is significantly positive at 10 µg/ft^2 or lower ($p<0.1$).*
  - *The slope between blood lead and floor dust is non-significant ($p>0.1$) at 11 µg/ft^2 or higher. The slope is negative above floor dust loading of 20 µg/ft^2 but the slope is not significantly different from zero (e.g., $p=0.932$ at 20 µg/ft^2, $p=0.618$ at 31.2 µg/ft^2)*

8. Pg 32- 39: Figures 5-2 to 5-5:

Comment: It would be useful to change the line formats so that you can see the difference when printed in black & white.

9. Pg 36 Section 5.2

Comment: Dixon et al (2009) predicted the probability that blood lead is above 5 and 10 µg/dL based on logistic models presented in that article. In linear models without complex survey designs (i.e., clustering and unequal survey weights), a tolerance bound approach can be used to predict the probabilities that the dependent variable is above a threshold. At the time the Dixon et al (2009) model was published, the authors did a thorough literature search to see if the tolerance bound approach could be generalized to the NHANES models with clustering and unequal survey weight but we did not find any reference to such an approach. The methods used for estimating these probabilities must be clarified because they are not standard methods. The methodology used for the QL models is not specified and I'm not sure the appropriate method exists.

Also, it would be useful to present confidence intervals for the predicted probabilities. The confidence intervals would be very large at the upper ends of the dust leads so they should be considered.

10. Pg 41. Table 6-1

Comment: Additional explanation of Table 6-1 is needed. It's unclear why this information is valuable.

11. Pg D-31

Comment: The variable "floor condition missing" should be "floor dust missing" as was presented in Table 4 of Dixon et al 2009

12. Page D-3 Scenario Definitions

Comments:

- *CT Scenario for Dixon 2009: For ethnicity, I believe that EPA used "non-Hispanic white" for ethnicity but "not nhblack" is specified*
- *Remove sill dust assumptions because all scenarios are evaluated at various floor and sill loadings*
- *Please add some motivation for the selected values of the variables used. Why did EPA deviate from the least-squares means approach for the central tendency scenario? That approach is not equivalent to evaluating the model at the most common responses. The assumptions given do not give predictions for the average child.*
- *For the upper-end scenario, shouldn't EPA use the oldest year of construction?*
- *EPA wants to make the scenario assumptions for the Dixon model and the NHANES QL model as close as possible but it doesn't appear this was done (e.g., year of construction for CT and upper, smokers in home for upper)*

13. Pages 34 and D-4 to D-6 for the Dixon et al (2009)

Comment: With the original models from Dixon et al (2009) I could not replicate the predicted GM blood lead or percent above the thresholds (5, 2.5 and 1 µg/dL) on pages 34 and D-4 to D-6 for the Dixon et al (2009) model with the model assumptions on page D-3. It is not clear why EPA's calculations using the Dixon 2009 log-log model could not be replicated; perhaps there is a problem with the presentation of the model assumptions.

14. Pg 7. The median PIR is 1.2, not 1.1

Table 1: Excerpt from Table 4 (Dixon 2009). Linear model results for log children's PbB

Variables	Levels	Estimate (SE)	P-value
Floor surface/condition* log floor PbD	Intercept for missing	0.178(0.094)	0.065
	Not smooth & cleanable	0.386(0.089)	<0.001
	Smooth & cleanable or carpeted	0.205(0.032)	<0.001
Floor surface/condition* $(\text{log floor PbD})^2$	Not smooth & cleanable	0.023(0.015)	0.124
	Smooth & cleanable or carpeted	0.027(0.008)	0.001
Floor surface/condition* $(\text{log floor PbD})^3$	Uncarpeted not smooth & cleanable	-0.020(0.014)	0.159
	Smooth & cleanable or carpeted	-0.009(0.004)	0.012

Table 2: Equivalent model with Main Effects for log floor dust lead loading

Variable	**Level**	**Estimate (SE)**	**p-value**
Floor DPb	Intercept for Missing	0.178(0.094)	0.065
Log floor DPb	-	0.386(0.089)	<.001
$(\text{Log floor DPb})^2$	-	0.023(0.015)	0.124
$(\text{Log floor DPb})^3$	-	-0.020(0.014)	0.159
Floor surface/condition*log floor PbD	Smooth & Cleanable or Carpet	-0.181(0.089)	0.048
Floor surface/condition*log floor PbD	Not (Smooth & Cleanable or Carpet)	0.000(0.000)	
Floor surface/condition* $(\text{log floor PbD})^2$	Smooth & Cleanable or Carpet	0.004(0.017)	0.813
Floor surface/condition* $(\text{log floor PbD})^2$	Not (Smooth & Cleanable or Carpet)	0.000(0.000)	
Floor surface/condition* $(\text{log floor PbD})^3$	Smooth & Cleanable or Carpet	0.011(0.015)	0.470
Floor surface/condition* $(\text{log floor PbD})^3$	Not (Smooth & Cleanable or Carpet)	0.000(0.000)	

Prediction equations for different floor surface/conditions

- Missing Floor dust: log(blood)=A+ 0.178
- Smooth & Cleanable or carpeted : $\log(\text{blood})=A+(0.386-0.181)*\log(\text{floor})+(0.023+0.004)*(\log(\text{floor}))^2+(-.020+0.011)*(\log(\text{floor}))^3 = A+0.205*\log(\text{floor})+0.027*(\log(\text{floor}))^2+-.009*(\log(\text{floor}))^3$
- Not Smooth & Cleanable and Not carpeted: $\log(\text{blood})=A+0.386*\log(\text{floor})+0.023*(\log(\text{floor}))^2 - 0.020(\log(\text{floor}))^3+$

where A depended on other covariates