RALPH A. BRADLEY
325 Hickory Hill Drive
Watkinsville, Georgia 30677
Telephone: (404) 769-5250

Mr. Ben F. Fort
Engineering and Mathematical Sciences
Research and Development Department
Ethyl Corporation
P.O. Box 341
Baton Rouge, LA 70821

Dear Ben:

I am responding to your letter of December 21, 1982 with my comments on the two papers that you enclosed:

(i) Comments on the DuPont Analysis of the NHANES II Blood Lead Data by James L. Pirkle, December 1, 1982 and

(ii) The Relationship between Gasoline Lead Usage and Blood Lead Levels in Americans: A Statistical Analysis of the NHANES II Data prepared by ICF Incorporated, December, 1982.

I shall comment on the two in order. My general thesis of my earlier report, "A Discussion of Issues and Conclusions on Gasoline Lead Use and Human Blood Lead Levels," still stands, namely,

> "No causal effect of use of lead in gasoline on human blood lead levels has been established nor can be established from the NHANES II data. At best such data may suggest relationships, real or spurious, for careful experimentation or study. The latter has not been done."

The approach in (i) above is to contend that a downward trend in time in human blood lead levels exists after control on demographic variables, urbanization and SMSA. It is interesting that time is retained in all of the regressions with time (Pages 14, 15 of (i)) only as "time-squared" after the described "manual backward elimination"

RALPH A. BRADLEY
325 Hickory Hill Drive
Watkinsville, Georgia 30677
Telephone: (404) 769-5250

of regression variables. Vernon N. Houk, in his April 14 Statement to the House Subcommittee on Environment, Energy and Natural Resources, in his Figure 2, showed a time trend on blood lead levels with a strong linear component. One can only conclude that the linear time trend component in Pirkle's analyses has been accounted for by the demographic, urbanization and SMSA variables if his analyses are correct.

Let us consider the simplest of Pirkle's regression equations, that for the Group: White 18-74 yrs. The equation is

$$y = 2.5452 + 0.1490M - 0.0670R - 1.5944T^2 ,$$

where y is log blood level of lead, M = 1 for male, 0 otherwise, R = 1 for rural and 0 otherwise, while T = time in days from 2/20/76 divided by 2800. A simple calculation shows that blood lead level for a non-rural female should drop to the low level of 1 µg/dl in less than 6 years after the study and was at that level less than 10 years before the beginning of the study. This calculation is an extrapolation and is nonsense, but it does show that at best the regression equation can only approximate the data within the time period of the study and cannot be interpreted as a reasonable explanatory scientific model. The same argument applies to the other regression equations given.

The process of manual backward elimination has arbitrary and subjective aspects. If two of the independent variables in a regression equation are highly correlated, positively or negatively, essentially

RALPH A. BRADLEY
325 Hickory Hill Drive
Watkinsville, Georgia 30677
Telephone: (404) 769-5250

only one of them is needed and either one may be retained and the other eliminated - the additional value of either one, given choice of the other, may be non-significant. This was the situation in Table 1 of my discussion paper; there was no effect of local gasoline lead use after time was included as an independent variable. In Pirkle's analyses with gasoline lead use, GAS 1, GAS 2 and GAS 3, on pages 16 - 21 of (i), time seems not to have been considered as an independent variable along with a GAS variable. If time had been included, it seems almost certain that the effect of GAS would have been non-significant after adjustment for the demographic, urbanization and SMSA variables and time. With time and GAS highly correlated, it is seen that variables retained in the regressions are similar with time-squared and GAS being somewhat interchangeable; one wonders why time-squared is not replaced by GAS-squared but this may again reflect a subjective choice of order of variables considered for elimination in the manual backward elimination process.

Many questions arise as to how Pirkle did his analyses. It is noted on page 4 of (i) that 9937 blood lead values were available. The mystique of SURREGR as the required regression program because of the complex survey design is brought in again; we have already commented on this in our earlier discussion paper. Use of complicated weights would not come in unless one were attempting to make subgroup population estimates for each demographic group for each time period. We are not told if log blood lead level, the dependent regression variable, is used for each of 9937 individuals or for some aggregate

RALPH A. BRADLEY
325 Hickory Hill Drive
Watkinsville, Georgia 30677
Telephone: (404) 769-5250

of individuals with the latter seemingly the choice as error degrees of freedom are reduced to an incredibly small number, 32. We are not told why the number of error degrees of freedom remains constant, independent of the number of regression variables in the equation. No information is given on how F-statistics in the various appendix tables were calculated; the sequencing of these tests is most important in the usual analysis of variance table in order to preserve orthogonality. Are they orthogonal? We suspect not. We suspect also that the error variance with 32 degrees of freedom is somehow estimated separately from the residual variation in the regressions and that unspecified assumptions were made in order for it to be used.

It is interesting that the Pirkle regression analyses move much nearer to the Ethyl Corporation's approach using indicator variables to code demographic, urbanization and SMSA variables. One would expect urbanization and SMSA variables to have some of the effect of the population density variable used by Ethyl Corporation. However, individual data seem not to be used by Pirkle, perhaps under fascination with SURREGR and we have stated in our earlier report that use of individual data avoids the need for unbiased population estimates. It is surprising that data on lead use in gasoline were used by quarters or six-month periods coded into GAS 1, GAS 2 and GAS3 when monthly data were available - see the October 1982 version of (ii).

We do not see that the regressions of (i) add any new insights. The argument is essentially that time-squared may be replaced by one of

RALPH A. BRADLEY
325 Hickory Hill Drive
Watkinsville, Georgia 30677
Telephone: (404) 769-5250

the GAS variables. It isthen argued that blood lead level decreases cannot be explained by lead in paint or lead in the diet and, being unable to think of other lead sources or another variable that might correlate with gasoline lead use and blood lead levels, it is concluded that the decrease in blood lead levels is caused by the decrease in gasoline lead use. The logic is dubious and certainly does not constitute proof.

We have not seen the DuPont Analysis but the argument about analytical error in estimating the percentage of the population with lead toxicity is peculiar. Variation seems to be confused with bias in blood lead determinations. Measurement error correction procedures are discussed at the bottom of page 10 of (i). This discussion suggests a possible correction for bias - all readings too low or too high. But the coefficient of variation is also discussed on the same page and the conclusion drawn that "the NHANES II data either correctly estimate or underestimate the number of persons with 'lead toxicity'". The coefficient of variation is a function of standard deviation s relative to the mean $\bar{x}$, CV = $s/\bar{x}$ or CV = 100 $s/\bar{x}$ %. Measurement error as variation should be reasonably symmetric about $\bar{x}$ and the possibility of underestimation seems no more likely than overestimation for an unbiased estimation procedure.

We have no information permitting comment on the quality control data discussed in (i) on page 11. The problem of quality control seems a minor one if it exists relative to other problems in the use of

RALPH A. BRADLEY
325 Hickory Hill Drive
Watkinsville, Georgia 30677
Telephone: (404) 769-5250

the NHANES II data to attempt to show a causal relationship between blood lead and gasoline lead use.

The second report by ICF incorporated is a revision and extension of their earlier report of October 1982 that I reviewed earlier. Many of the introductory sections of (ii) have only minor revisions. However, the main analyses of Section II have been redone and merit discussion. Also (ii) provides background on some of the statements in (i) and the analyses in (i) seem to be very similar to those in (ii) including the use of SURREGR, the 32 degrees of freedom for error independent of the number of regression variables used, the failure to specify the source of the dependent variable, and the apparent non-orthogonality of the F-tests in regression analysis tables.

The report (ii) does provide an explanation of the Pirkle statement on underestimation discussed above that seems plausible - see the discussion on Misclassification Error on pages 26 - 30.

The regression analyses of Section II of (ii) are more elaborate than before and represent a further move in the direction of use of the independent variables used by the Ethyl Corporation. Income and season are brought in as variables and both time and gasoline lead use appear for the first time in the same regression equation along with a one-month lagged gasoline lead use. It is asserted on page 31 of (ii) that "throughout this preliminary analysis current and one-month lagged gasoline lead remained highly significant and both had very stable coefficients". It should be noted that time was also significant and, while one-month lagged gasoline lead use was significant, current

RALPH A. BRADLEY
325 Hickory Hill Drive
Watkinsville, Georgia 30677
Telephone: (404) 769-5250

gasoline lead use was not always significant.

Our main criticism of these regression analyses is that the significance tests seem to be tests of regression coefficients and not the tests of the appropriate ordered entries in the proper analysis of variance table for the regression analysis. It would again be our contention that, if one were to claim an effect of gasoline lead use on blood lead level even after adjustment for time as is done in this report, the proper method is to obtain the sum of squares for gasoline lead use, lagged or otherwise, as the reduction achieved in the residual sum of squares due to the extra term(s) in the model after fitting the other independent variables. This has not been done and, if it were done, it is likely that gasoline lead use would be insignificant. We still believe that the procedure of Table 1 of our discussion report is the proper one.

Similar comments apply to the logistic regressions of (ii). Time is not considered as an independent variable, only one-month lagged gasoline lead use is included and income is introduced. The tests again seem to be tests of regression coefficients.

In (ii), the causality argument is explicitly stated on page 11 : "Second, one would be hard pressed to suggest an omitted variable which could explain seasonality and trends in blood lead levels and show a lagged effect". In addition to my earlier comments above, a lagged effect is clearly not relevent and one could probably find many variables that would do so long as the variable is strongly

RALPH A. BRADLEY
325 Hickory Hill Drive
Watkinsville, Georgia 30677
Telephone: (404) 769-5250

correlated with time. It is interesting to note that in (i) it is " time-squared" that is used while "time" alone is used in (ii). This seems only to suggest that there are a number of models that will fit the NHANES II data and that none can be claimed to describe the functional dependence of blood lead on the chosen independent variables. The proliferation of models now being presented suggests some insecurity in dependence upon the primitive models initially used.

I hope that these comments are helpful to you.

Yours sincerely,