International Journal of Occupational Medicine and Environmental Health, 2004; 17(1): 111—114

VALUES, ERRORS, AND PRECAUTIONS

HERBERT L. NEEDLEMAN

Department of Psychiatry
University of Pittsburgh School of Medicine
Pittsburgh, PA, USA

Abstract. In the environmental health literature, errors in interpreting studies or data are not infrequent. Many are of the Type II variety. Common solecisms of this type are: treating the criterion of $p < 0.05$ as a sacrament; demanding complete confounder control; arguing for the existence of phantom confounders; arguing that the effect size is trivial; building nonveridical models; arguing for no effect from inadequate sample size; demanding causal proof; arguing that causality is reversed; conducting a ballot of published studies. These are examined in this paper.

Key words:
Type I and Type II errors, Epidemiology, Methodology

Explicating or exercising a precautionary principle inescapably involves value judgements and choices. The principle itself resides in the tension between Type I and Type II errors. Avoiding Type I errors, that is, accepting a false finding as true, is considered rigorous scientific behavior, and is generally approved of by the scientific community. Avoiding Type II errors, rejecting a true finding as false, receives consideraby less attention and approbation.

Type I errors are also known as "producers" errors. If a Type I error is committed, e.g., a harmless product is declared to be toxic, the producer suffers. Money is lost. Type II errors are known to as "consumers" errors. If such an error is committed, e.g., a toxic pesticide is reported to be harmless, and its distribution permitted, consumers will suffer. People will be sickened.

A general bias in favor of avoiding Type I errors has permeated the literature. This is shown in Table 1, taken from the work of Rubin and Rosenthal [1], which presents the ratio of β to α errors in studies of frequently encountered sample and effect sizes. It can be seen that for studies with 200 subjects and fewer, at small effect sizes (r =.10) the β/α ratio ranges between 16 to 18. Only at larger effect size (r =.30) and large sample size (n = 200) does the α risk exceed the β. This bias was largely ignored until the publication of Jacob Cohen's book on statistical power in 1969.

Some Type I and Type II errors are the product of lack of understanding or innocent methodological oversights. Some errors, particularly Type II errors in the study of industrial toxicants, have more unprincipled roots. It is this phenomenon that is examined here.

Table 1. The ratio of β to α risks at various sample and effect sizes

No. subjects	R = 0.10		No. subjects	R = 0.30	
	α = 0.05	α = 0.01		α = 0.05	α = 0.01
50	18	97	50	9	67
80	18	96	80	4	61
100	17	94	100	3	44
200	16	88	200	0.2	31

Received: January 19, 2004. Accepted: January 30. 2004.
Address reprint requests to Prof. H.L. Needleman, Keystone Building, 3520 Fifth Ave., Suite 310, Pittsburgh, PA 15213, USA (e-mail: hlnlead@pitt.edu).

One place where type II errors of disingenuous character have been frequently committed is in the study of lead toxicity. The errors take two forms. The first is biased reviews of the literature. The second is rarer: dismissing or minimizing the finding of one's own data in order to reach a null conclusion.

In many reviews of the childhood lead poisoning literature, one finds a consistent pattern of minimizing the size of any lead effect, or dismissing it entirely. Listed below are some of the common Type II errors encountered in lead study criticisms:

■ **Worship of the sacrament of $p < 0.05$.** Studies that present p values greater than 0.05 have been dismissed as demonstrating no effect, or even more egregiously, as evidence that no association exists in nature. This of course is not what a p value means. What is the origin of the $p = 0.05$ criterion? Fisher [2] is credited with establishing this in 1925 as the conventional level for rejection of the null hypothesis in analysis of variance data:

> It is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant.

Fisher's operative word is " convenient." There is no reason other than convenience to select this criterion. Only time and mediocre thinking have fixed 0.05 and endowed it with sacramental properties.

■ **Arguing that because complete confounder control has not been accomplished, no causal inference can be proffered.** This criticism is like many others, disingenuous. Multivariate space is infinite, and sample sizes (and budgets) are finite. Complete confounder control is impossible to achieve. Another covariate can always be nominated for control.

■ **Arguing for the existence of phantom confounders**. One critic of the lead-IQ hypothesis has argued that since the incorporation of confounders into the model has reduced the effect size, if one could find the relevant confounder, unidentified, of course, the effect size would go to zero. No attempt is made, in this criticism, to name the missing variate.

■ **Arguing that the effect size, if there is one, is small, and therefore of no consequence**. The mean difference in IQ scores between high and low lead groups is about 4–6 points. Critics of the lead – IQ hypothesis claim that this is less than the standard deviation of the distribution, and could be due to the error in measurement. Fig. 1, taken from our 1979 study, shows that a 6-point shift in means between high and low lead groups is associated with a four-fold increase in severe deficits: IQ scores below 80. Also shown is an overlooked effect, the number of children with superior function, IQs greater than 125, is reduced by lead exposure. The meaning of "small" in this context is not "inconsequential;" rather it is "difficult to see in a crowded multivariate field."

■ **Building nonveridical models**. Some studies have controlled in the analysis for school placement or temperament. These, because they are affected by lead exposure, are in the causal pathway, and controlling for them constitutes overcontrol.

■ **Arguing for null effect on the basis of small samples.** The effect size in many lead studies is small. Studies of 80 or 150 subjects generally do not have the statistical power to detect such an effect. To find an effect in this range requires upwards of 400 subjects. To argue for a null ef-

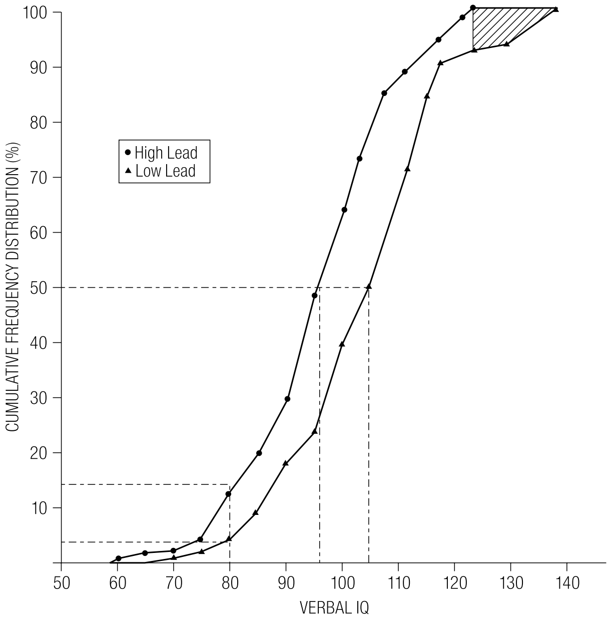

Fig. 1. Cumulative frequency distribution of verbal IQ scores in high and low lead subjects. A shift in the median score of 6 points is associated with a four-fold increase in the risk of IQ below 80.

fect from samples inadequate to find an effect is a classic Type II error.

■ **Demanding proof of causality.** This is also disingenuous. David Hume in the mid-18th century showed that causality is a set of relations between mental events, not physical events, and as such is not subject to empirical demonstration. Yet the demand for proof of causality continues to be raised by lead industry spokespersons.

■ **Arguing that causality is reversed.** Seizing on the observation that mentally retarded children have much more mouthing behaviour, critics argue that the lead-cognition association proceeds from poor cognition to low IQ. Ignored repeatedly are the observations that most subjects were not mentally delayed; the mean IQ score in our high lead group was 108 [3]. Also ignored are studies of the association of IQ at 24 months or later in relation to lead levels in umbilical cord blood taken at birth, and finally, studies of psychological performance in animals systematically dosed with lead [4].

■ **Examining studies in isolation and arguing that not all studies have found a significant effect.** Not all published studies have reported a statistically significant association between lead and IQ. The ballot approach (counting and comparing positive and negative studies) to drawing a summary conclusion from the available data ignores fundamental principles of probability. Instead of simply counting the number of positive and negative studies, one should ask: "What is the probability under the null hypothesis of N positive studies out of the set of studies that attempted to find an effect at the p = 0.05 level? The answer is readily obtained through binomial expansion. Listed below the probabilities of finding 4, 5, and 6 positive studies if 10 studies were conducted.

Table 2. Probability of finding positive studies under the null hypotesis

Number of studies conducted	Number positive at p = 0.05	Probability
10	4	0.0009
10	5	0.00006
10	6	0.0000026

Anyone who referees papers will encounter authors who strain to conclude an effect is present in their data when there is none, or only minimal evidence for one. In the lead literature, however, one occasionally encounters a singular behavior: authors whose data seem to demonstrate a lead effect, but whose reports tend to dismiss or deprecate it. In 1974, Perino and Ernhart [5] reported that children's blood lead levels were significantly related to cognitive, verbal and perceptual abilities after covariate adjustment.

> While the effects of subclinical lead intoxication may not be noted in the individual cases seen in a pediatric clinic, analysis of group data indicate quite clearly that performance on an intelligence test is impaired. It seems that the criteria set for lead poisoning need re-examination and stepped-up efforts for the prevention of lead ingestion in preschool children should be emphasized.

Five years later, Ernhart et al. [6] followed 63 of these subjects, and although they found that blood lead levels were significantly related to general cognitive, verbal and motor IQ scores, they drew a startling conclusion:

> The few statistically significant findings of this study are due to methodological difficulties inherent in this area of research…If there are in fact behavioral and intellectual sequellae of low levels of lead burden… these effects of lead are minimal [6].

The misstatements summarized above have served to cloud understanding of lead toxicity at low exposures, and have resulted in a misperception in some quarters that the question is still controversial. These techniques are in the industrial armory and will be used again to obscure understanding of other toxicants. It is in the hope of making them available for examination and refutation that they are presented here.

REFERENCES

1. Rosenthal R, Rubin DB. *Statistical analysis: summarizing evidence versus establishing facts. Psychol Bull* 1985; 97: 521–9.
2. Fisher RA. *Statitsical Methods for Research Workers*. Edinburgh: Oliver & Boyd; 1925.
3. Bellinger D, Leviton A, Waternaux C, Needleman H, Rabinowitz M. *Longitudinal analysis of prenatal and postnatal lead exposure and early cognitive development*. New Engl J Med 1987; 316: 1037–43.

4. Rice D, Karpinsky K. *Lifetime low-level lead exposure produces deficits in delayed alternation in adult monkeys*. Neurotox Teratol 1988; 10: 207–13.

5. Perino J, Ernhart C. *The relation of subclininical lead level to cognitive and sensorimotor impairment in black preschoolers.* J Learning Dis 1974; 7: 26–30.

6. Ernhart C, Landa B, Schell N. *Subclinical levels of lead and developmental deficit – a multivariate follow-up reassessment.* Pediatrics 1981; 67: 911–9.