Behavioral teratology/toxicology: How do we know what we know?

Robert T. Brown

Department of Psychology, University of North Carolina at Wilmington, Wilmington, NC 28403, USA

Accepted 28 August 2000

Abstract

Knowledge about human behavioral teratology/toxicology must necessarily be limited by the methods that can ethically be used. Research in the field is an enterprise calling for tolerance for uncertainty if not chaos. Among the main points of this paper are: (a) Kaufman’s criticisms of research on low blood-lead level (BLL) and children’s IQ are generally valid and apply to virtually all human natural-groups research; (b) Relative to some contexts, research on low BLL on children’s IQ is exemplary; (c) Relative to other contexts, the conclusion that low BLL have linear effects is well supported; (d) Owing to necessary design limitations, all natural-groups studies have shortcomings; (e) Inference of causality is uncertain under any conditions; and (f) Some authors readily leap from molehills of data to mountains of causal conclusions.

Keywords: Behavioral teratology/toxicology; Lead poisoning; Low BLL; Children’s intelligence

Under the best of circumstances, human behavioral teratology/toxicology research is an uncertain enterprise, one in which knowledge is necessarily limited by the methods that must necessarily be used. A major point of the present paper is that research in one area cannot adequately be evaluated independently of several contexts. Kaufman’s thoughtful and detailed analysis of the literature on low blood-lead level (BLLs) and children’s IQ provokes much thought about those contexts. In the case of the effects of low BLLs on children’s IQ, those contexts include other behavioral teratology/toxicology research, relevant experimental non-human research, and developmental research in general. In particular, I became interested in how well developmental research in general would stand up to Kaufman’s criticisms. Among
the main points of this paper are: (a) Kaufman’s criticisms are generally valid and apply to virtually all human natural-groups research; (b) Relative to some contexts, research on low BLL on children’s IQ is exemplary; (c) Relative to other contexts, the conclusion that low BLL have linear effects is well supported; (d) Owing to necessary design limitations, all natural-groups studies have shortcomings; (e) Inference of causality is uncertain under any conditions; and (f) Some authors readily leap from molehills of data to mountains of causal conclusions.

Many of Kaufman’s criticisms apply to other current issues. Does childhood physical and sexual abuse invariably cause long-term adverse effects? Does prenatal maternal smoking cause low birth weight and other adverse consequences in children? Does a high-fiber diet protect against colon cancer? Does a high-salt diet cause high blood pressure? Answers to these questions change frequently and are in some cases a matter of considerable scientific, political, and legal controversy. They also have major methodological difficulties and inconsistent research bases. Questions over acceptable design, control, data, analyses, and interpretation have reached from the laboratory through the scientific literature and into the courtroom. The effects of low BLLs have implications for prevention, intervention, and, unfortunately, litigation that may at least equal those of other issues.

1. Some fairly obvious contexts of behavioral teratology and toxicology

1.1. The causal context

True experimental teratology and toxicology research on humans involving randomization and control to eliminate potentially confounding variables is not possible. My line to experimental psychology students was that the heroes of randomization and control slay the devious and evil villain of confounding. Obvious ethical constraints preclude random assignment of subjects, whether children or pregnant women, to varying amounts of a potential toxin as an independent variable while controlling other variables. Studies with humans necessarily use natural groups that differ in degree of exposure not only to the toxin of interest but to other variables as well. Potential confounding variables are inevitable, and researchers can only try to minimize them through actual or statistical control. Data from such studies are correlational, and causal inferences must be guarded (see Reynolds, 1999 for a recent discussion).

But causal inferences must always be guarded. Inference is “the act of passing from one proposition, statement, or judgment considered as true to another whose truth is believed to follow from that of the former” (Webster, 1972, p. 432). Causality involves belief and is subject to the vagaries of human reasoning and inference. From Aristotle’s unmoved mover through the certainty of Francis Bacon to the skepticism of Hume, Kant’s innate categories of human understanding, John Stuart Mill’s Canons of scientific proof (or induction), and C.S. Pierce’s ringing defense of the experimental method, philosophers have argued whether and how causality can be determined. As Aiken (1957, p. 141) stated:

Mill understood the fallacy of post hoc, ergo propter hoc [After this, therefore because of this], and he was quite aware of the crucial importance of experimentation for the scientific testing of theories. Real continuities in nature ... do not always lie on the surface of
experience. Valid causal connections can be established only through meticulous sorting and independent variation of the complex antecedent conditions and consequences that are concealed from casual observation. Unwashed appearances, including appearances of correlation, are notoriously deceptive. It should be the function of inductive logic, among other things, to provide reliable technique for cleaning them.

Unfortunately, “inductive logic” is an oxymoron, as Hume demonstrated. Also unfortunately, I have lost the citation for the quotation that went something like, “Logic textbooks spend the first half of the book on deductive logic describing logical fallacies and the second half on inductive logic committing those fallacies.”

The problem of causality under natural, as opposed to laboratory, conditions underlies Campbell and Stanley’s (1966) pioneering work on quasi-experimental designs, and their discussion of threats to internal validity should be regularly reviewed by all conducting natural-groups research. In a recent important article, Reynolds (2000) has not only argued cogently against the Surgeon General’s criteria for inferring causality from an association between two variables but has also demonstrated empirically that the criteria simply do not work. Reynolds (2000) aptly quotes Baumrind, who stated that her purpose in critiquing a particular theory of heroin usage was to “discourage the fantasy that any statistical system can justify drawing causal inferences from correlational data.”

Further, causal inferences are uncertain under the best of conditions. True experiments establish independent variable–dependent variable relationships only within limits of probability; replication reduces but does not eliminate those limits. Further, correlations may have implications, although limited, for causal interpretations. Surely, our confidence in Patterson, Reid, and Dishion’s (1992) theory that preschool boys’ temper tantrums in the home are causally related to later antisocial behavior is increased by knowledge that such tantrums are indeed correlated with later antisocial behavior. Causality can never be determined with certainty, and we must cope with a continuum of uncertainty over causality rather than a dichotomy between certain causal and noncausal relationships. In a science farm as opposed to an animal farm, Orwell might have suggested that all causal relationships are uncertain, but some are more uncertain than others.

1.2. The natural environment context

In studies of humans in natural environments (if our current environments can be described as natural), the variables listed above and many others are by necessity uncontrolled. A limited number can be held constant through subject selection or statistical manipulation, but many more go uncontrolled and are, as Kaufman states, potential confounding variables. They also, however, as Kaufman also states, potentially increase within-group (error) variability. A basic principle of behavioral teratology is that individual differences in genetic background influence susceptibility to teratogens (e.g., Vorhees, 1986), further increasing within-group variability. Given the small effects expected with low doses of a potential toxin, relatively high within-group variability may mask relatively low between-group variability. Further, some uncontrolled variables may interact with the toxin, either as provocative or protective factors, increasing or decreasing effects, respectively. Expense in locating and testing subjects and measuring potentially confounding variables may result in small sample sizes, further decreasing sensitivity of designs.
The factors described above increase the possibility of both Types I and II errors, as Kaufman indicates. Since different studies differ in characteristics of subjects and environmental variables, inconsistent results among studies are likely. Well-conducted meta-analyses should help to determine the best overall inferences about a body of research literature.

1.3. The converging evidence context

Owing to uncertainties associated with human research, converging evidence is particularly important. Results of natural-groups studies on humans should be interpreted only relative to those of true experiments on appropriate nonhuman models. Only experimental nonhuman research can isolate lead from potentially confounding variables, specify the degree of lead exposure, and establish probable causal relationships. In some areas, for example fetal alcohol syndrome, results on animal models have been critical to an understanding of the phenomenon in humans. In laboratory behavioral teratology and toxicology experiments with nonhumans, subjects are randomly assigned to condition and extraneous variables such as temperature, rearing and feeding conditions, and handling are held constant and thus controlled. Owing to breeding practices, genetic variability among subjects may also be relatively low. Control over variables not only eliminates them as potential confounding variables but also reduces within-group variability, making easier demonstration of an effect of a potential toxin.

Appropriateness of animal models is not always easily determined. Not only are equivalence of dosage and measures difficult to determine but also apparently minor methodological differences between human and nonhuman studies may be quite important. Consider whether or not subjects have control over intake of a substance. Dworkin, Mirkes, and Smith (1995) yoked rats that controlled their intake of cocaine through lever pressing with those who received equal but uncontrolled intake. Rats that did not control intake died at a much higher frequency (5 of 13) than did those that controlled intake (2 of 25). Across a number of replications, approximately 60% of uncontrolled intake subjects, but only 5% of controlled intake subjects, died (Dworkin, personal communication, April 2000). This finding suggests, for example, a potential limitation to the generality of nonhuman models of prenatal nicotine/smoking on postnatal development since humans smoke voluntarily, whereas nonhumans are forcibly exposed.

However, with comparable presentation, a range of doses, and standardized behavioral measures, experimental nonhuman research can provide findings critical to causal inferences from human research. In the case of lead, exposure in both humans and nonhumans is involuntary. Further, some research has tested monkeys with varying experimentally induced BLLs on several complex learning and memory tasks that are similar to items on human IQ test. Monkeys’ nervous systems and developmental trends are similar to ours, increasing validity of cross-species generalizations. Similar results have been obtained using rats as subjects, further demonstrating generality of findings. The results provide strong experimental support for causal interpretations of similar findings in human natural-groups studies:

It is well established that monkeys exhibit behavioral impairment as a result of developmental exposure to lead. This has been demonstrated in two different species of
monkeys, in different laboratories, in a number of different groups of monkeys, and on a
variety of behavioral tasks.  

(Rice & Silbergeld, 1996, p. 671)

1.4. The developmental research context

Many critiques of a body of literature consider only research in that area, evaluating it relative to absolute standards. That is, of course, one-cell data, which are not very useful for scientific purposes. Claims that masturbation causes blindness and insanity first made by the famous psychiatrist Tissot in the eighteenth century and that XYY males were unusually prone to violence made in the 1960s were based on one-cell data. Later studies that involved comparison groups obviously showed the claims to be unjustified.

Are studies of low BLL in children below the standards of other developmental research? In order to obtain information on the base rate of occurrence of some procedures, I used the Brown Rapid Ocular Wandering Scan Estimator (BROWSE) on a sample of current well-regarded developmental journals. I attempted to identify all primary research articles that measured children’s IQ, made causal inferences on the basis of natural-groups designs, or used natural-groups designs where parental IQ would be a relevant correlate. I recorded (a) information on qualifications of testers, (b) IQ tests used on children, (c) tests of maternal/paternal intelligence, and (d) measures of socioeconomic status and home conditions. Too few articles used multiple measures or had more than three groups of subjects to enable comparison with Kaufman’s criticism nos. 3 and 4, Failure to control for multiple comparisons and Comparison of IQs of two extreme “lead-level” groups, respectively. I scanned titles, abstracts, and method sections of primary research articles in late 1999 to early 2000 issues of journals available at the UNCW library. The journals (and number of issues scanned) were Child development (4), Developmental medicine and child neurology (3), Developmental psychology (1), Journal of autism and developmental disabilities (6), Journal of clinical child psychology (1), Journal of child psychology and psychiatry (3), and Journal of developmental medicine and behavioral pediatrics (7). Since the goal was not to critique those articles but to provide an overview of currently accepted methods, individual studies are not cited in order to protect the guilty. The results of the BROWSE suggest that studies of low BLL are overall better controlled than many other recently published natural-groups studies; details are in the evaluation of Kaufman’s criticisms that follows. Kaufman’s criticisms, then, are in many cases justified but should be seen as applying to the whole field.

2. Evaluation of Kaufman’s five shortcomings +2 in low BLL research

In his critique, Kaufman criticizes five areas where he claims that low BLL studies are poorly designed or conducted and at least two interpretations of the research. This section deals with each of his criticisms relative to the above contexts.
2.1. Shortcoming no. 1 — uncontrolled variables cloud conclusions drawn from even the best studies

Of course, uncontrolled variables inevitably are potential confounds in human natural-groups studies. Researchers cannot control all potential confounds and must try to deal with those deemed most important. Kaufman rightfully points out avoidable shortcomings in the weaker BLL research. Researchers should at least accurately measure SES, conditions in the home, and overall health. To expect them to go beyond the better available measures is unrealistic. Thus, many of Kaufman’s criticisms of HOME would apply to other available and potentially available instruments — How could any control for “all of the pertinent variance?” (p. 308). Lack of control over possible confounding variables and inappropriate inference of causality cloud developmental research in general. Some studies are particularly outrageous. Consider Russell (1990, p. 293): “Over 90% of teenagers who smoke three to four cigarettes are trapped into a career of regular smoking that lasts for some 30–40 years.” Quite a claim, particularly since it was based on second-hand self-report one-cell data with no control for potential confounding variables and published in an obscure journal.

Kaufman’s suggestion that lead studies should specifically control for otitis media has characteristics of a red herring. I know of no teratological/toxicological studies that have controlled for otitis media. Otitis media would be a confound only if it was correlated with, but not a consequence of, low BLL. If otitis media or other hearing disorders were consequences, then control would reduce deficits in IQ attributable to lead, hardly an appropriate procedure. Prenatal alcohol has been implicated in otitis media and other hearing disorders (Church & Gerkin, 1988), and lead itself may impair both hearing and processing of speech, providing one possible pathway by which lead could then impair IQ and learning (Bellinger, 1996).

Results from the BROWSE: Only the eight natural-groups studies that made clear causal inferences were included this evaluation. One suggested that early neuropsychological skills were causal factors in later development, but controlled for no external variables. Of the remaining, three had no measures of parental intelligence, two had brief WAIS scores on the mother, and one had PPVT scores on the caretakers. One study compared infants of mothers who either did or did not use cocaine prenatally. Information on use of other licit or illicit drugs was based on maternal self-reports, which to this author qualifies as an uncontrolled variable. One study’s abstract claimed to have measured SES and maternal education, but the procedure section gave no information on these measures. Except for one study that used the HOME and one that used mothers’ education as an index of SES(!), none measured SES, home conditions, or prenatal care. Since such measures were requisite for even the most cautious interpretation of causality in these studies, poor control was the rule. Not surprisingly, no studies controlled for otitis media.

2.2. Shortcoming no. 2 — parental IQ is typically measured poorly or not at all

No arguments from me here, particularly since I have criticized studies of prenatal maternal smoking just on this basis. From the literature, one could well conclude that those
who study environmental influences on intelligence are ignorant of both basic psychometrics and behavior genetics. Given the parent–child correlation in IQ of about 0.50 and the general care with which researchers measure the children’s intelligence, their often cavalier attitude towards measurement of parental intelligence is puzzling. Kaufman cannot be bettered when he states that “despite the rigorous attempts that experimenters made to use state-of-the-art IQ tests for the children in the studies, a similar rigor was not followed when assessing their parents” (p. 313, emphasis in original).

Maternal PPVT score as the only measure of parental intelligence, as is often the case in developmental research, is simply inexcusable. One need only consult an undergraduate test and measurement textbook, let alone the Buros Mental Measurement Yearbook, for information that the PPVT is at best a measure of receptive vocabulary. As Kaufman indicates, acceptable if hardly ideal, shorter forms of standard intelligence tests are available and would be a clear improvement. Sadly, to the uninformed, use of the PPVT may provide research with an unjustified facade of control. More sadly, it appears to provide the same facade to presumably informed reviewers and editors.

BLL studies again appear to be of higher quality than many of those concerned with other substances. Studies of the effects of maternal smoking on children’s IQ are notably cavalier. Even the most cited studies have no measure of paternal IQ, and most measures of the mothers’ IQ were through the now infamous PPVT. Little, if any, mention of the qualifications of the test administrators is provided. In one study, children’s IQs were taken from whatever scores were in school records and used interchangeably! Studies of FAS are surprisingly deficient in control over extraneous variables, particularly parental IQ. Since IQ is related to SES, one would expect that studies of the relationship between SES and FAS would use intelligence as a covariate, so that any effects could be attributed to prenatal alcohol. Not so. In an oft-cited study, Bingol et al. (1987) reported that incidence of FAS, including mental retardation, was significantly higher in offspring of low SES mothers than in offspring of middle-class mothers, even when maternal drinking was controlled. The study had no measure of parental IQ. Further, maternal education and race differed greatly across groups but were not statistically equated. Abel and Sokol (1999) summarized the situation:

Alcohol is but one of many possible risk factors such as social class, maternal illness, genetic susceptibility, smoking, diet, past health history, pregnancy complications, use of drugs, and exposure to environmental pollutants. In epidemiology, statistical tests are used to try to “even out” as many of these cofactors as possible. The goal is to match people as closely as possible except for their alcohol use. If you are going to do this carefully, you especially need to control for heredity … Consider IQ scores [sic]. If you are studying the causes of subnormal IQ, isn’t it reasonable to start with parental IQ? Claire Ernhart and her coworkers did this when they examined the effects of lead in children. They found lead was associated with decreased scores in cognitive and verbal tests, but when parental IQ was incorporated into the analysis, the relationship was no longer significant. No studies in fetal alcohol research have bothered to include parental IQ as a factor (p. 4, emphasis added).

Note that the epidemiological study Abel and Sokol describe as an example of the importance of controlling for parental intelligence is one that Kaufman criticizes, rightly so, for using maternal PPVT scores as the sole measure of parental intelligence. One might argue that in studying FAS, measures of parental intelligence are unimportant owing to the
classic physical signs of FAS. However, those physical signs are not so obvious, as only those with considerable training may be able to recognize the Gestalt of FAS, and a record of maternal alcohol abuse is still considered necessary for a diagnosis of FAS, and fetal alcohol effects may be diagnosed in the absence of clear physical features (Stratton, Howe, & Battaglia, 1996).

Results of the BROWSE: Of 10 studies for which measures of parental intelligence were potentially relevant correlates, seven (70%) did not directly measure parental intelligence at all, and only two of those reported parental education level as a substitute. Of the remaining three, one measured both maternal and paternal IQ with a brief WAIS, one measured maternal IQ with a brief WAIS, and one used the apparently near ubiquitous PPVT. Again, studies of low levels of lead appear to be at least in the mainstream of developmental research.

2.3. Shortcoming nos. 3 — failure to control for multiple comparisons, and 4 — comparison of IQs of two extreme “lead-level” groups

Statistical flaws of this type are all too common in behavioral research in general, so Kaufman’s point about low BLL studies is well taken. Either selecting a small number of significant paired comparisons or selecting extreme groups from multiple ones risks a Type I error. In some cases, the data appear to have been inappropriately analyzed at the outset. The sophistication of researchers does not always match that of computer analysis packages. Even if paired comparisons are based on an overall analysis, correction for multiple paired comparisons should be used. Fortunately, concern about multiple comparisons is apparently increasing in research in general. However, the Bonferroni procedure to which Kaufman refers is indeed conservative and increases likelihood of Type II errors.

2.4. Shortcoming no. 5 — lack of quality control in measuring children’s IQ

Administrators of individual intelligence tests need extensive training. Graduate students may spend an entire course learning to administer Wechsler tests. The old joke applies: Tourist in NYC to old man carrying a violin case, “How do you get to Carnegie Hall?” Answer, of course, “Practice, my boy, practice.” Some authors have not described the training or experience of test administrators, justifying Kaufman’s conclusion that they did not demonstrate awareness of the need for quality control. However, his conclusion that the studies actually lacked such control is unwarranted and turns an error of omission (Studies did not describe test administrators’ experience) into an error of commission (Studies used inexperienced test administrators). Kaufman provides no evidence of such lack of control other than in an early study by Gregory, Lehman, and Mohan (1976, cited in Kaufman). His claim that in “large-scale research projects, it is not uncommon for inexperienced examiners . . . to be used to collect IQ data . . .” which makes it “feasible that examiner errors . . . occurred in some of the 26 ‘best’ lead–IQ studies” (p. 318) is reasonable, but hardly justifies a conclusion that errors actually occurred. Thus, his heading, “Shortcoming no. 5 — Lack of Quality Control in Measuring Children’s IQ” (p. 318) is misleading.

Results of the BROWSE: Of the 36 articles purporting to measure children’s intelligence, 34 (94.4%) reported no information on tester qualifications. The two that
gave information simply stated that the tests were administered (1) by trained graduate students or (2) by a graduate student or the author. Many studies suffered from a more serious quality control issue, the test used to measure children’s intelligence. A total of 17 studies (47%) used full versions of WISC-R, WISC-III, or McCarthy Scales. However, 11 studies (31%) used the PPVT or BPVT, labeled variously as a test of intelligence, mental age, or verbal mental age. Of those 11, one used a brief version of the PPVT, and another, by a respected behavior geneticist, used PPVT to estimate genetic effects on intelligence! Three studies used brief versions, generally vocabulary and block design tests, of the WISC. One study each used Raven’s Progressive Matrices, brief Stanford-Binet, Mullen Scales of Early learning (with no information on validity as a test of intelligence), WISC-R, or Stanford Achievement Test scores converted to IQ and treated as interchangeable, and raw scores on a Canadian sample compared directly with raw scores on an unstandardized Hebrew translation of the PPVT on an Israeli sample. In two additional studies where child IQ was a relevant correlate, no measure was obtained. Lack of information on tester qualifications appears to be the norm, and use of unsuitable tests to measure intelligence is common practice in developmental research.

2.5. No documented linear relationship between lead level and IQ

Arguments over linear vs. threshold effects of toxic/teratological substances are of more than academic importance. A linear relationship implies that exposure to even low levels of the substance is dangerous, whereas a threshold relationship implies that exposure to low levels is harmless. Currently, the issue of linear vs. threshold effects of prenatal exposure to alcohol is highly controversial. Citing well-known literature that fully expressed FAS appears only in the offspring of heavy-drinking women, Abel (e.g., Abel, 1998; Abel & Sokol, 1999) has argued for a threshold effect and that the term “fetal alcohol syndrome” should be replaced with “fetal alcohol abuse syndrome.” The Washington group (e.g., Bookstein, 1999) has vigorously argued for a linear effect. As might be expected, both can cite supporting evidence. Given high variability common to human studies particularly at low levels of a toxin, one should not be surprised at small and inconsistent findings.

A study on low BLL published after Kaufman prepared his review (Mendelsohn et al., 1999) supports a linear relationship between BLL and development in young children. Apparently well-controlled in terms of demographic characteristics, measures of SES, cognitive stimulation in the home, and chronic disease, the study suffered mainly from use of the PPVT-R to measure maternal intelligence. With confounding variables controlled, young children with BLL of less than 10 μg/dl showed significantly higher Bayley Scale scores at ages 12–36 months than did children with BLL of 10–24.9 μg/dl. Further, the semipartial correlation between Bayley scores and BLL for all children was highly significant (−.24). Whether these findings will hold up as the children develop is, of course, unknown.

Kaufman’s restriction of his evaluation to studies of IQ in children presents problems in regard to the issue of linear vs. threshold relationship, since it again eliminates the
experimental nonhuman literature. Consider the continuation of the quotation from Rice and Silbergeld (1996, p. 671) cited above:

Moreover, behavioral impairment has been observed in every group studied, with no evidence of a threshold effect. Specifically, research has demonstrated impairment in a group of monkeys with steady-state blood lead levels of 11 μg/dl and a peak of 15 μg/dl early in life, compared to a “control” group with blood lead levels of 3–5 μg/dl [the lowest dose group in the colony]. Moreover, deficits persisted into adulthood.

2.6. Interpreting fractions of an IQ point has no scientific meaning

Kaufman uses the fact that IQs are only accurate within limits to argue that they cannot be divided into fractions of a point. Certainly for individuals, that is true, and in any case, claims about 0.678 of a point, one of Kaufman’s examples, is absurdly false precision. But IQ has from the outset been treated as an interval measure — We specify mean and standard deviation. Fractionation is both reasonable and accepted practice in presenting group means. After all, if two individuals have IQs of 95 and 96, is not the best estimate of their mean IQ 95.5 rather than 95 or 96? For a particularly graphic example of IQ fractionation, one need only go to a commonly cited study, Zajonc and Markus’ (1975) confluence model of birth-order effects and IQ. Having transformed scores on the Raven’s Progressive Matrix Test reported by other authors in six categories to a scale that had the mean Raven’s IQ for an only child at 100, Zajonc and Markus report decline in scores across birth order to the third decimal point! Since virtually all authors present group IQ means at least to the first and frequently the second decimal point, I may have missed Kaufman’s point.

3. Evaluation of Kaufman’s five shortcomings +2: valid criticisms, impossible standards

In closely evaluating aspects of the literature on low levels of lead on IQ, Kaufman has revealed shortcomings that commonly mar research on other toxicological/teratological agents including alcohol and maternal smoking. Indeed, most of the BLL studies Kaufman evaluated are models of design rectitude when compared with frequently cited research claiming to show adverse effects of maternal smoking on children’s development. Researchers should long have been aware of these shortcomings and taken steps to avoid them. Consultation of undergraduate textbooks on research design and tests and measurements would have avoided many problems.

Kaufman has provided an ideal model for human teratological/toxicological research. Unfortunately, his model is almost Platonic — immaculate but impossible in reality. Apparently for Kaufman, we not only do not know whether low levels of lead adversely affect children’s IQ, but we cannot: “Technically, a confound in a lead–IQ study correlates significantly with both the dependent variable of children’s IQ and with the independent variable of BLL [blood-lead level]. But it is important to control for variables ... that are believed or known to vary alongside an outcome variable (such as IQ), even if their relationship to BLL is non-significant or unknown ... Even more of a threat to the validity
of the lead–IQ studies, however, are variables associated with intelligence that are either unknown or unmeasurable” (p. 310). By ms. p. 17, the latter have become “what are undoubtedly a plethora of unknown but potentially potent variables.” Kaufman makes a terminological error in describing BLL as an independent variable; it is a subject variable on which individuals are classified, not one to which they are randomly assigned. Internal validity can always be questioned since confounding cannot be completely avoided. Further, no study, not even a true experiment, can meet the demand that unknown or unmeasurable variables be randomized or controlled. Dossey (1995) similarly reasons that experimental evaluations of alternative medical therapies are inappropriate, since groups could differ even if subjects had been randomly assigned. He provides a virtually impossible hypothetical example to support his point while missing the point that random assignment is designed to equate groups as well as possible. (To digress briefly, Dossey also suggests that since the meta-experimenter in double-blind studies knows which subjects are in which experimental condition, s/he can influence them through mental telepathy and thus alter the experimental results! Some go to inordinate lengths to discredit research.) In suggesting that variables that are unknown or unmeasurable are even more of a threat, Kaufman leads himself into an untestable position. No matter how well-controlled a study might be, its results may be clouded by “a plethora of unknown but potentially potent variables.” Strictly speaking, he is right, but the position is Humean skepticism to the point of nihilism.

Since much human research is necessarily correlational, it can under the best of circumstances only be tentatively and cautiously interpreted causally. Reynolds’ (1999) recent discussion of the problems should be “must-reading” for researchers and policy makers. Owing to the limitations of human research, causal relationships should be made on in the context of supporting nonhuman experimental research. Kaufman’s omission of that body of research weakens his conclusions about the effects of low BLL.

4. Two additional concerns about Kaufman’s critique

4.1. Population homogeneity and the extent of genetic influence

In his overview, Kaufman (2001, p. 322) states that “… the relative contributions of genetics and environment to a child’s IQ differ, sometimes markedly, from country to country. The genetic component tends to be population specific and may be larger in countries that are homogeneous in their population …” Determining relative contributions of genetics and environment to an individual’s development cannot be done. As D.O. Hebb stated decades ago, “Trying to determine how much of an individual’s development is due to heredity and how much to environment is like trying to determine how much of a rectangle’s area is due to its length and how much to its height. The area is due 100% to length and 100% to height.” We can only estimate the amount of variance in a measure that owes to variance in genetic background, variance in environmental background, and the interaction between the two for groups of individuals. This estimate is the well-known population genetics statistic, heritability, which is indeed population specific. However, since it is a proportion and ranges between 0 and 1, as genetic variance decreases as in a homogeneous population, environ-
mental variance must increase. For example, if we have a completely inbred strain with identical genetic background, any trait variance must owe to environmental variance, and heritability will be 0. Thus, in a relatively homogeneous population, variance in a trait attributed to genetic variance will be relatively low.

4.2. A surprising shift in Kaufman’s position

In his introduction, Kaufman limits his analysis to studies of low BLL on IQ: “The current paper . . . addresses the question of what impact — if any — do low BLLs (10–20 μg/dl) have on children’s performance on conventional measures of intellectual functioning” (p. 304). He limits his discussion to IQ regardless of literature on other adverse effects of lead exposure and evidence that in syndromes/disorders owing to other substances, behavioral deficits may be relatively greater than IQ deficits. In FAS, for example, serious deficits in attention and conduct may accompany relatively small deficits in IQ (e.g., Streissguth & Kanter, 1997). He then argues that even the best studies are so flawed that a conclusion that low BLLs affect IQ is unwarranted and that evidence suggests that BLLs have a threshold rather than linear negative relationship with IQ.

Late in the paper, Kaufman makes a surprising turn in his argument. Having limited his analysis to IQ, Kaufman claims that it is not particularly useful:

IQ tests deliberately measure a limited aspect of human functioning. They are not intended to be used as the sole criterion for making any decisions that have educational, vocational, neurological, or societal implications. They are too narrow in scope and design. They are not intended to measure interpersonal skills (social intelligence), creativity, special talents, or any of a number of qualities that are commonly associated with intelligent people (p. 328).

Kaufman proposes that Sternberg’s triarchic theory is a “more pertinent theory of intelligence for evaluating societal impact” since it “encompasses diverse aspects of intellectual functioning” (p. 329) and concludes:

As currently measured, IQ is too narrow a concept to have societal implications even if exposure to very low levels of lead should be shown to lower IQ by a few points . . . IQ tasks are not sufficiently real-world-oriented and do not tap the kinds of activities that are necessary to maintain and advance a society. When other, more comprehensive, theories are applied (such as Sternberg’s), and when future, broader based intelligence tests are perfected [such as individually administered versions of the Sternberg Triarchic Abilities Test (STAT)], then perhaps the results of the lead–IQ investigations will prove to have societal implications (p. 329).

In the first place, adherents of “g” or “g+s” theory have much data to counter Kaufman’s sweeping condemnation of the importance of IQ (e.g., Carroll, 1993; Eysenck, 1998; Humphreys, 1992; Jensen, 1998). Kaufman appears to imply that tests such as Sternberg’s will become available and be more valid measures of intelligence in general. But to become available, they will presumably have to pass APA’s criteria for measurement and be published. Kaufman’s citation for Sternberg’s test and supporting evidence is a 1993 unpublished paper. Since Sternberg is hardly known to be anal retentive about publishing,
one might well wonder if his test remains unpublished because Sternberg has yet to establish his test’s reliability and validity.

Kaufman’s reasoning appears to be: “I am limiting my analysis to IQ because it has been most studied. Deficits in IQ associated with low BLLs are inconsistent and small. Even the best studies have known or unknown confoundings that severely limit their interpretation. But even if the effects were larger and the confoundings not a problem, the results wouldn’t matter because IQ is too narrow to have societal implications.” A defense attorney might say: “We should consider only my choice of evidence. My evidence indicates that any number of other perpetrators may be guilty. Even if my evidence indicates that my client is guilty, my evidence should be ignored because it is too narrow to indicate that my client is guilty of a crime that has meaningful consequences to society. Maybe in the future sufficient evidence will be available, but not now. Therefore, although he may not be innocent, we cannot find him guilty.”

5. Are we maintaining appropriate standards for published research?

Regardless of my criticisms, many of Kaufman’s concerns appear valid generally, not just in the context of effects of BLL, as Reynolds’ (1999) critique of causal inferences from correlational research indicates. Although human research necessarily will be subject to confounding and multiple interpretations, some basic research principles, well known by undergraduate psychology majors, should be observed in the professional literature. Research that violates these principles routinely appears in respected journals, indicating that not only the researchers but also reviewers and editors have blind spots. The possibility that the blind spots are neither random nor consistently applied is discomfiting. Political/social sensitivities may inappropriately intrude into what should be objective scientific decisions.

Once in the literature, findings tend to remain accepted regardless of flawed design or interpretation, contrary findings, or failures to replicate. The same studies then become fair game for citation in reviews, popular literature, legislative hearings, and litigation, contributing to the spread of junk science in the classroom and congress as well as the courtroom (e.g., Crossen, 1994; Huber, 1991). Note however that Crossen suffers from serious bias. We would be well advised to keep junk science out of science. Problems of quality control appear widespread in many scientific areas including, as Reynolds (1999) and Kaufman have indicated, psychology. If the question, “Will psychology be regarded as science or pseudoscience?” concerns us, we should address these problems.

References


