Psychology in the Schools, Vol. 41(6), 2004
Published online in Wiley InterScience (www.interscience.wiley.com).

DOI: 10.1002/pits.20004

STRONG OPINIONS ARE NO SUBSTITUTE FOR BALANCED ARGUMENTS: COMMENTS ON CICCHETTI, KAUFMAN, AND SPARROW'S CRITICAL APPRAISAL OF PCB COHORT STUDIES

GERHARD WINNEKE, JENS WALKOWIAK, AND URSULA KRÄMER

Heinrich-Heine-Universität Düsseldorf

This paper comments on a critical review of cohort studies on PCB-related neurodevelopmental deficit in young children by D.V. Cicchetti, A.S. Kaufman, and S.S. Sparrow (CKS). Major points of criticism of CKS, namely alleged violation of statistical principles, presumed lack of clinical significance of findings, and alleged insufficient control of confounding are dealt with in appropriate detail. It is argued that much of this criticism is inappropriate and biased, and that, in particular, arguments dealing with basic rules of statistical inference rely too heavily on the narrow principles of Neyman-Pearson testing which are discussed controversially in modern epidemiology. Other critical arguments concerning the presumed lack of validity of assessment instruments, the apparent lack of reliability checks in some studies, and the presumed inappropriate treatment of longitudinal data are also discussed as being poorly founded. It is finally concluded that, although there are inconsistencies and weaknesses both within and between individual PCB-studies, the almost unconditional rejection of the full set of cohort studies by CKS is in no way supported by good reasoning. © 2004 Wiley Periodicals, Inc.

To an important degree, progress in science has always been pushed by elements of doubt. The principle of unlimited doubt (*de omnibus est dubitandum* = "everything must be in doubt"), formulated by Descartes (1596–1650) and certainly not suitable for everyday life, has therefore always been an intrinsic corrective element of scientific endeavors. Doubt-based replication or nonreplication of published observations are the necessary prerequisites of either continuing or discontinuing the further elaboration of scientific theories.

Thus, in principle, the motivation behind Cicchetti, Kaufman, and Sparrow's (CKS) critical appraisal of PCB cohort studies might be considered to be fully in line with good scientific tradition. However, to be helpful, scientific doubt must be based on sound reasoning, must be valid and objective, and must be as systematic and unbiased as possible. These fundamental requirements are met by our critics only to a very limited degree. Instead, this critical appraisal differs widely from established standards of how a systematic evaluation of the literature, e.g., for the purpose of a meta-analysis, should be performed. For example, of the 16 criteria listed in recognized guidelines such as SIGN50 (2001) for evaluating the internal validity of studies, only three (namely reliability/validity of measuring instruments and reliability checks, appropriate control for confounding, and control for chance findings, where appropriate) were chosen by our critics, apparently based on their own personal preferences. The two additional criteria selected, namely statistical versus clinical significance and longitudinal design/analysis are not even regarded as being elements of checking for the internal validity of a study. The statement, "these are objective criteria that are uniformly used by both behavioral and medical scientists" (p. 591) is an assertion but no argument. In addition, CKS did not even try to describe the cohort studies in any systematic manner along their simplified set of criteria; instead, the descriptions are erratic, haphazard, and mostly difficult to follow.

Despite the above-mentioned shortcomings, we will, in the subsequent discussion, follow the sequence of the six criteria given in the critical appraisal. We will, furthermore, focus our comments largely on those aspects directly related to critique raised against our own work (Walkowiak

Correspondence to: Gerhard Winneke, Division of Neurobehavioral Toxicology, Medical Institute of Environmental Hygiene, Heinrich-Heine-Universität Düsseldorf, Auf'm Hennkamp 50, D-40225 Düsseldorf, Germany. E-mail: gerhard.winneke@uni-duesseldorf.de

et al., 2001; Winneke et al., 1998), although selected issues of more general interest will also receive comment.

Reliability and Validity of Assessment Instruments

Objectivity, reliability, and validity are key criteria in evaluating the psychometric quality of psychodiagnostic or psychodevelopmental tests. We agree with our critics that in some PCB studies suboptimal tests have been used. In our case (Winneke et al., 1998), following the pioneering work of the Michigan group (Jacobson, Fein, Jacobson, Schwartz, & Dowler, 1985), we administered the Fagan Test of Infant Intelligence (FTII; Fagan & Shepherd, 1987) in its 10-item version at 7 months of age, and were unable to come up with negative PCB effects. We did however, as pointed out by CKS, check for re-test reliability and found it to be zero or even slightly negative. Unlike our critics, however, we were more careful in our conclusions in pointing out that this disappointing outcome might have resulted from using the mobile test-version in the child's home which, in contrast to the more standardized hospital setting used by the Michigan group, might well have been more distracting.

Although the psychometric properties of the FTII are clearly suboptimal, it remains true that the Oswego study was able to replicate the negative FTII/PCB associations of the Michigan group at 7 and also at 12 months of age (Darvill et al., 2000). If in doubt, however, established psychometric tools like the Wechsler-Scales (WISC), the McCarthy-Scales (MSCA), the Bayley-Scales (BSID), and the Kaufman-Scales (K-ABC) should be and indeed have been used in most of the cohort studies (Dutch, Faroe Islands, German, North Carolina, Michigan), and therefore the lengthy criticism of our critics concerning more experimental instruments like the FTII may perhaps serve the intended purpose of raising doubts as to the overall adequacy of the cohort studies, but are less helpful for the reader to develop a balanced opinion.

Appropriateness of Reliability Assessments

Another issue raised by our critics relates to the argument that, even if good instruments like the BSID II (Bayley, 1993) were administered, inter-examiner reliabilities were not consistently checked. This may be true for some studies. As for the European studies, it is true that only one well-trained examiner administered the BSID within the individual cohorts at a particular age or even across ages, such that checking for inter-observer reliability would seem as a redundant exercise. Yet to be able to pool data from the different European cohorts at some later point in time, we nevertheless did run such a check by exchanging videotapes from BSID sessions between the examiner from the Faroe Island cohort (Steuerwald et al., 2000) and from the German Düsseldorf cohort (Winneke et al., 1998). The outcome of this check was highly satisfactory, namely $r = 0.85$ for the BSID II mental scale and $r = 0.98$ for the psychomotor scale. We presented it in our paper (Winneke et al., 1998, p. 425), but our critics failed to mention it. Perhaps because they take the view that the standard Pearson Product Moment Correlation (PPMC) is an inadequate measure of association because if "scores for two examiners co-vary in the same order, PPMC will be high, falsely signaling high reliability, even when the pairs of scores are widely discrepant" (p. 596).

This is true in a very restricted sense. However, high-level textbooks in psychodiagnostics in children (e.g., Sattler, 1992) recommend the PPMC as a satisfactory measure of association, if the pattern of agreement rather than the level of agreement between the observer's ratings on an interval scale are of interest. Only if pattern plus level of agreement are important for interval data the intraclass correlation coefficient may be used (Sattler, 1992, p. 511). Of course, these are details but, because our critics never miss an opportunity to create doubts as to the scientific validity of existing cohort studies, they have to be mentioned, even more so since "the answer to Question 2, whether appropriate reliability assessments were made on study variables, is "No"

(p. 597). The correct answer could have been "Sometimes," but the adjective "appropriate" allows CKS to apply their own narrow definition of appropriateness to arrive at an unconditional "No"!

Control for Chance Findings

This is an important issue over and above the PCB cohort studies, and also over and above—but related to—our own studies (Walkowiak et al., 2001; Winneke et al., 1998). The main subtopics here are *statistical significance*, *directional versus nondirectional hypotheses*, and *correction for multiple testing*. What our critics think and write about these topics is extremely mechanical and in no way compatible with more up-to-date conceptual thinking about statistical issues. We recommend a thoughtful recent textbook, *Modern Epidemiology* by Rothman and Greenland (1998), in particular, chapter 12 on "Approaches to Statistical Analysis" (pp. 183–199), to gain a deeper insight into what inferential statistics in research is or should be.

CKS use statistical significance as given by some conventional inequality threshold (e.g., $p < .05$) as the one and only decision rule to qualify an observation as being "real" or "factual." Observations not passing this threshold simply do not exist. This is compatible with the fundamentals of (dichotomous) Neyman-Pearson Hypothesis Testing but is no longer acceptable today. Let's hear what thoughtful public health-oriented epidemiologists have to say about this (Rothman & Greenland, 1998):

> A *P*-value is a continuous measure of the compatibility between an hypothesis and data. Although its utility as such a measure can be disputed . . . a worse problem is that it is often used to force a qualitative decision about rejection of an hypothesis (p. 186).

The authors then go on discussing Type I and Type II errors and the trade-off between them and continue as follows (pp. 186, 187):

> The concepts of alpha-levels, Type I error, Type II error, and power stem from a paradigm in which data are used to decide whether to reject or not reject the test hypothesis. The extent to which decision-making dominates research thinking is reflected in the frequency with which the *P*-value, a continuous measure, is reported only as an inequality (such as $p < .05$ or $p > .05$) or else not at all, with the evaluation focusing instead on 'statistical significance or its absence.'

To illustrate the absurdity of such "unsound practice as Neyman-Pearson (dichotomous) hypothesis testing" (Rothman & Greenland, 1998, p. 187) these authors give an example in which 71 individual clinical trials did not find a formally significant treatment effect, although the great majority of trials found moderate or even rather strong treatment effects. Since, however, each individual study was negative the clinical researchers did not reject the null hypothesis of no treatment effect, thus inflating the Type II error, i.e., falsely accepting the hypothesis of no treatment effect. This is exactly why, in the public health arena, Jacobson & Jacobson (1996) or Schantz et al. (2002) raise concern about the dangers of falsely not detecting an adverse health effect, if it exists (Type II error). This example is clearly reminiscent of how CKS deal with the published negative PCB effects which, in their decision—rather than thought-oriented reasoning, are just absent had proper adjustment for two-tailed probabilities or control for chance effects by means of Bonferroni adjustment been applied.

This has a direct bearing on how our own findings (Walkowiak et al., 2001; Winneke et al., 1998) are dealt with by our critics. In a previous letter to *The Lancet* (Cicchetti & Kaufman, 2002), the statistical part of which is repeated in the present critical appraisal, they raise two arguments: (a) since our basic PCB-related hypothesis was indisputably nondirectional we should

have used two-tailed rather than one-tailed probabilities as decision criteria; (b) had we applied Bonferroni corrections to our observations none of our "significant" effects would have been statistically significant.

As for the first argument, this is clearly wrong and remains wrong upon repetition (pp. 603–604), as well. We stated (Walkowiak et al., 2001, p. 1602) that there is uncertainty whether environmental exposure to PCBs *adversely* affect mental and motor development in early childhood. This simply says that, in our judgment, developmental PCB effects can either be zero or, if not zero, they can only be negative (adverse). In responding to the Cicchetti and Kaufman letter we pointed out, "We tested for adversity of prenatal and neonatal PCB exposure, because the assumption of a beneficial impact of PCBs on motor and mental development is not biologically plausible" (Winneke et al., 2002a, p. 1438). By this we meant not only the few human cohort studies and the majority of neurobehavioral studies in animals but also, and even more so, the three mechanistic hypotheses presently discussed as possibly underlying PCB-induced developmental neurotoxicity, namely disturbance of dopaminergic, thyroid or gonadal functions, each of which indicating adversity. Thus, directionality of hypothesis testing in the case of PCBs (and for lead as well) is not only justified but in our opinion even mandatory.

As for the second argument, our critics in their *Lancet* letter and in their present text (pp. 597–598) demand correction for multiple comparisons by means of Bonferroni adjustment, to control for chance findings. It is certainly true that chance findings can and will occur if many comparisons on the same data are performed, and it may also be true that in some cases PCB researchers have selected only those few findings from a larger number of comparisons which came out as being significant, perhaps on a chance basis. Such selectivity is not scientifically defensible. However, in our case (Walkowiak et al., 2001, Table 2) we performed nine comparisons involving PCBs and found all of them to exhibit negativity and six of them to be significant at least at $p < .05$ (one-tailed). The likelihood of this combined outcome to be a chance effect is very low ($p < .0001$) as pointed out before (Winneke et al., 2002a, p. 1438). However, it must be emphasized that in the case of correlated observations, as is true for our study with four follow-up observations of the same group of children, a mechanistic application of simple Bonferroni adjustments is totally inadequate anyhow (Perneger, 1998).

However, as mentioned above, we do agree with more thoughtful statisticians that "Declarations of significance or its absence can supplant the need for more refined interpretations of data; the declarations can serve as a mechanical substitute for thought, promulgated by the inertia of training and common practice" (Rothman & Greenland, 1998, p. 187). This is why we did not rely on *p*-values in the first place, but had a closer look at our data to check for effects sizes or dose-response associations as required by toxicological considerations. Such associations were indeed found (Walkowiak et al., 2001, Figs. 1 and 2) and strengthen the validity of our findings. We pointed this out before (Winneke et al., 2002a, p. 1438) but, unfortunately our critics still stick to their rigid Neyman-Pearson frame of reference (pp. 606–607).

One last word concerning adjustment for multiple comparisons. A frequently proposed strategy, and the one also emphasized by CKS, is the Bonferroni correction, i.e., dividing the desired *p*-value by the number of comparisons. For 100 comparison and a desired *p*-value of .05 this means $.05/100 = .0005$ as the alpha-level. This is a valid possibility but a poor one, "because the single intervals it produces are much too wide (conservative) for use in single-association estimation," and for joint estimation purposes "the joint Bonferroni confidence region is unnecessarily imprecise" (Rothman & Greenland, 1998, pp. 227–228). More advanced methods are available (Greenland, 1998) but, unfortunately, CKS emphasize a procedure that is particularly conservative in solving the problem of multiple comparisons. Why they do so we leave to the reader to decide.

Statistical Versus Clinical Significance

This is a topic that is in no way related to the internal validity of the various cohort studies, and is therefore typically not included in systematic reviews of the literature. However, we will deal with it here, because it is of more general interest and also moderately important. Findings from PCB cohort studies such as the 4-point difference of K-ABC results between the upper and the lower quintile of the Dutch study (Patandin et al., 1999) or the 8.5-point difference between the upper 95 and the lower 5 percentile of our study (Walkowiak et al., 2001) are small if applied to an individual child's K-ABC performance on two occasions. However, we do not need CKS' distinction between "statistical" and "clinical" significance to be aware of the problem. In performing epidemiological cohort studies we are interested in group differences and not directly in what these group effects mean for an individual child. This is an important but a totally different issue that, for example, has clearly been emphasized in the WHO-IPCS document on inorganic lead in which it was stated that "estimates of effect size are group averages and only apply to the individual child in a probabilistic manner" (World Health Organization, 1995, p. 31).

It is, furthermore, totally acceptable if concern is raised about the possible public health implications of a 4-point shift for the extremes of a population. This concern has first been raised in the context of pediatric lead studies (Needleman, 1983) and was later translated to observations in PCB-exposed children (Rogan & Gladen, 1991). Our critics do not accept this conclusion as being valid but they only cite a paper of their own in which they apparently deal with the issue (Kaufman, 2001), instead of presenting immediate arguments against it. This is not at all helpful.

Control for Confounding

No one involved in public health related epidemiology denies the crucial importance of confounding. For a variable to qualify as a confounder it is necessary for that variable to correlate both with exposure and outcome. If relevant confounders are not considered associations between exposure and effect may be spurious. Our critics emphasize four main potential confounders or group of confounders, namely parental intelligence (genetic influence), quality of the home environment (HOME), prenatal risk factors, and exposure to organic mercury (MeHg). As for parental intelligence CKS accept that the more recent studies (Dutch, German, Oswego) have done a satisfactory job in at least measuring maternal IQ, although they criticize the administration of the Peabody Picture Vocabulary Test (PPVT) in the Oswego study as being inadequate (p. 607), a criticism which we do not share vis-a-vis the good correlations with children's IQ at 38 and 54 months of age (Stewart, Reihman, Lonky, Darvill, & Pagano, 2003). CKS also do not like that paternal IQ was not measured in addition to maternal IQ (p. 608). Since, however, maternal and paternal IQ are well correlated and since, furthermore, the shared variance of maternal IQ (17.6%) and mid-parent IQ (25%) with child IQ is not widely different (p. 608) this argument is a minor one. They also miss (p. 607) consideration of maternal IQ in an early publication of our cohort at 7 months of age (Winneke et al., 1998). However the Wechsler subtest "Vocabulary" was not administered until at 18 months of age and, above all, maternal IQ in our data had no significant impact on pediatric mental development below 30 months of age.

Besides emphasizing the need for considering genetic and home environment contributions in regression modeling for PCB effects on mental/motor development our critics express particular concern about insufficient coverage of prenatal risk factors, such as smoking, alcohol, drugs, and other lifestyle variables. As for prenatal smoking and alcohol they feel that the "yes/no" coding of these variables in the Dutch and German cohorts was insufficient. What our critics are unaware of is that in the initial questionnaires these variables were treated in a more quantitative

manner on 3- or 5-point scales but were finally collapsed to the yes/no dichotomy after noticing the extremely skewed distributions. It is, furthermore, well known from clinical experience that more quantitative information on smoking and alcohol consumption is extremely unreliable in a medical context. CKS are certainly well aware of that, but in their determined effort to create an atmosphere of doubt they argue against their own professional expertise.

As for other prenatal risk factors such as caffeinated and decaffeinated coffee, tea, herbal tea, caffeinated soda, prescriptions and over-the-counter medications our critics emphasize the assessment efforts of the Oswego group (Darvill et al., 2000) in this respect, and criticize the other cohorts for having paid much less attention here. Thus, they arrive at the following far-reaching apodictic conclusion (p. 610): "Because only the Oswego cohort addressed prenatal variables thoroughly and systematically, it is conceivable that any observed significant differences . . . that were attributed to prenatal PCBs in the five other cohorts . . . may have been due to uncontrolled prenatal variables." Although this is a theoretical possibility it is also a very remote one, because it is extremely unlikely that those prenatal variables exhibit associations with both outcome and PCBs, thus resulting in spurious correlations.

Coexposure to organic mercury (MeHg) is of relevance for cohorts from fish eating populations only, and has been considered in the more recent Oswego (Stewart et al., 2003) and the Faroe Islands studies (Steuerwald et al., 2000), although not in the pioneering early Michigan study (Jacobson et al.,1985). For general population studies such as those from North Carolina (e.g., Rogan et al., 1986; Gladen & Rogan, 1991), from the Netherlands (e.g. Patandin et al., 1999) and from Germany (e.g. Walkowiak et al., 2001) potential coexposure to MeHg can be disregarded, although inorganic lead can be a confounder here. We have measured lead in cord blood, but CKS may have overlooked this; at least it is not mentioned in their critical appraisal.

Since "Biological Plausibility" and "Experimental Evidence" are two out of nine criteria proposed for inferring causality from epidemiological observations (Hill, 1965) it is not surprising that authors presenting findings from PCB cohort studies sometimes also refer to observations from animal studies to support their claim for causality. CKS cover this topic in the context of confounding. Unfortunately, they have no documented expertise in toxicology let alone neurotoxicology. None of their 17 self-citations are original research reports dealing with poisons, and they do not even distinguish between organic and inorganic mercury (p. 612). It is not surprising, therefore, that their arguments here are mostly superficial and nonspecific. All they tell the reader is that animals are not humans, and that following Paracelsus (1493–1541) it is the dose that makes the poison, and that dosage is important for animal–human extrapolation.

This is true. However, our critics fail to distinguish between external and internal (target) dose. Although, for example, it is true that relative to man much higher (two to three orders of magnitude) external lead doses are necessary to induce neurobehavioral toxicity in rats, mainly due to differences in toxicokinetics, such effects occur in young rats and in humans at about the same internal dose, i.e., blood lead levels (PbB; Cory-Slechta, 1998). This is also supported in the WHO document on inorganic lead: "Animal studies provide support for a causal relationship between lead and nervous system effects, reporting deficits of cognitive functions at PbB levels as low as . . . 11–15 μg/dl which can persist well beyond the termination of exposure" (WHO, 1995, p. 32). Similar relations hold true for PCBs. If compared to PCB concentrations in the brains of stillborn babies (Lanting et al., 1998) effective PCB brain levels in newborn rats exposed via their dams are only between 5–30 times higher (Hany et al., 1999; Kaya et al., 2002), which is a small difference. It is, therefore, not at all surprising that Faroon, Jones, and De Rosa (2000) simply state: "Animal studies support human findings" (p. 307). Interestingly, however, CKS use this same review paper to argue against the adequacy of animal-to-child extrapolation for PCBs (p. 616). Why this may be so, and why CKS, for this section, arrive at the overall conclusion that the cohort

studies failed to control appropriately for important confounds, we leave to the reader to judge. Perhaps the adverb "appropriately" helped in this respect.

Design of Longitudinal Studies

CKS criticize that all of the cohort studies were designed as longitudinal or prospective studies, but were not analyzed accordingly. Instead, our critics claim that they were analyzed and reported as cross-sectional studies for each follow-up study period. This is a basic misunderstanding! For each of the six cohorts pre-/perinatal PCBs were taken as the independent variable and developmental effects at various ages were related to this exposure rather than to exposures at follow-up, although with slightly modified regression modeling. These are typical features of an analysis for prospective/longitudinal studies. Probably the only exception is the follow up of our Düsseldorf cohort at 30 and 42 months where, in addition to comparing K-ABC- and BSID-outcome with pre-/perinatal PCB-exposure, we also checked for associations with postnatal PCB-intake via nursing (Walkowiak et al., 2001, p. 1605). We did, indeed, find an additional significant impact of postnatal PCB exposure, which is compatible with the higher PCB transfer via breast milk rightfully noted by CKS (p. 602) and with some animal data, as well (Rice, 1998).

Regrettably, however, CKS have not studied our own cohort report carefully. If they had they would have noticed that in analyzing developmental data from the Düsseldorf cohort we also followed a time-series approach for the BSID follow-up data at 7, 18, and 30 months of age, yielding significant negative PCB associations (Walkowiak et al., 2001, Table 2). Although only children with complete data sets were included, the strength of associations did not change when using simple imputation techniques to account for missing values (p. 1604). Although we have not followed the CKS-proposal of using their preferred method called "individual growth curve analysis (IGCA)" (p. 613), our longitudinal analysis is exactly what our critics are asking for. Therefore, and for the reasons given above, the CKS opinion that none of the cohort studies was appropriately designed and analyzed along longitudinal principles (pp. 613–614) is not valid, unless IGCA is considered the one and only method.

Summary and Conclusions

Our critics have no documented research expertise in neurodevelopmental epidemiology or in developmental neurotoxicology. This need not be a shortcoming. On the contrary, an open unbiased view from outside and a systematic appraisal of the literature from someone not personally involved can be very helpful in elucidating weaknesses and inconsistencies in a particular research area. Unfortunately, however, the CKS appraisal is neither unbiased nor systematic. Its bias lies both in the selection of the six criteria, in the erratic manner of how different studies are treated within these criteria, and in the use of adjectives such as "appropriate," "adequate," or "consistent" which are personal value judgments rather than objective standards.

As for the six criteria chosen, only three of these represent a small subset of criteria typically proposed for systematic reviews of the literature, e.g., for meta-analytic purposes, as pointed out in the beginning of our comments. A systematic effort of how the different studies are treated within each of the six criteria is lacking. Combined with the use of qualifying adjectives, this allows CKS to come up with an almost unconditional rejection of the full set of cohort studies for all of the self-selected criteria. And this, in turn, allows our critics to finally conclude that "These studies have major shortcomings that preclude reaching definitive conclusions about any alleged adverse effects of the ingestion of PCBs . . . on . . . mental functioning or . . . neurobehavioral functioning" (p. 615). In going through the arguments we have shown that this far-reaching conclusion is in no way supported by good reasoning.

This does not mean that PCB cohort studies are without weaknesses and that full consistency of findings exists. Yes, there are shortcomings and inconsistencies between studies and pointing these out in a systematic manner would have been an important service to the field. However, in their determined effort to create nothing but an atmosphere of doubt, CKS have clearly missed an opportunity here. One additional example may help to support this conclusion: In emphasizing the lack of consistency between studies our critics go even as far as pointing at a putative inconsistency within a particular study, namely our Düsseldorf cohort study (p. 615). They notice that whereas significant negative association between PCBs and K-ABC performance occurred at 42 months of age, this was no longer true at age 72 months (Winneke et al., 2002b). This is an interesting observation that only few serious researchers would place into the category of "inconsistency." Rather, alternative and more substantive interpretations like "lack of persistence," "functional recovery" would seem more adequate. Yet, our critics prefer their biased interpretative wording. Why they do so, we leave to the reader to judge.

Where does this generalized bias come from? In their final remarks CKS point out that many of the results from PCB cohort studies have been published in "quite respectable behavioral and medical journals" (p. 619). They go on arguing that papers published in good quality journals are not necessarily good papers. Rather they say or quote someone else saying, "that not only the researchers but also reviewers have blind spots. . . . Political/social sensitivities may inappropriately intrude into what should be objective scientific decisions" (p. 619). Correct! This may happen and happens occasionally. But how about the blind spots and biases of reviewers like CKS who, as indicated on the title page of their article, were paid by the General Electric Company, one of the large PCB polluters worldwide, for a critical appraisal of neurobehavioral PCB cohort studies? We have a proverb in Germany that goes something like this: "He who sits in the greenhouse shall not throw stones." Perhaps CKS have just not been wise enough to take this into account when doing their literature search and writing their review paper.

References

Bayley, N. (1993). Manual for the Bayley Scales of Infant Development (2nd ed.). San Antonio, TX: The Psychological Corporation.

Cicchetti, D.V., & Kaufman, A.S. (2002). Correspondence: Does PCB exposure have any effect on later psychomotor and mental development? Lancet, 359, 1438.

Cicchetti, D.V., Kaufman, A.S., & Sparrow, S.S. (2004). The relationship between prenatal and postnatal exposure to polychlorinated biphenyls (PCBs) and cognitive, neuropsychological, and behavioral deficits: A critical appraisal. Psychology in the Schools, 41, 589–624.

Cory-Slechta, D. (1990). Bridging experimental and human behavioral toxicology studies. In R.W. Russel, P. Ebert-Flattau, & A.M. Pope (Eds.), Behavioral measures of neurotoxicity (pp. 137–158). Washington, DC: National Academy Press.

Darvill, T., Lonky, E., Reihman, J., Stewart, P., & Pagano, J. (2000). Prenatal exposure to PCBs and infant performance on the Fagan Test of Infant Intelligence. Neurotoxicology, 21, 1029–1038.

Fagan, J.F., & Shepherd, P.A. (1987). Fagan Test of Infant Intelligence: Training manual. Cleveland, OH: Infantest Corporation.

Faroon, O., Jones, D., & De Rosa, C. (2000). Effects of polychlorinated biphenyls on the nervous system. Toxicology and Industrial Health, 16, 305–333.

Gladen, B.C., & Rogan, W.J. (1991). Effects of perinatal polychlorinated biphenyls and dichlorodiphenyl dichloroethane on later development. Journal of Pediatrics, 119, 58–63.

Greenland, S. (1998). Introduction to regression modeling. In K.J. Rothman & S. Greenland (Eds.), Modern epidemiology (2nd ed., pp. 401–432). Philadelphia, PA: Lippincott-Raven.

Hany, J., Lilienthal, H., Sarasin, A., Roth-Härer, A., Fastabend, A., Dunemann, L., Lichtensteiger, W., et al. (1999). Developmental exposure of rats to a reconstituted PCB mixture or Arochlor 1254: Effects on organ weights, aromatase activity, sex hormone levels, and sweet preference behavior. Toxicology and Applied Pharmacology, 158, 231–243.

Hill, A.B. (1965). The environment and disease: Association and causation? Proceedings of the Royal Society of Medicine, 58, 295–300.

Jacobson, J.L., & Jacobson, S.W. (1996). Intellectual impairment in children exposed polychlorinated biphenyls in utero. The New England Journal of Medicine, 335, 783–789.

Jacobson, S.W., Fein, G.G., Jacobson, J.L., Schwartz, P.M., & Dowler, J.K. (1985). The effect of intrauterine PCB exposure on visual recognition memory. Child Development, 56, 853–860.

Kaufman, A.S. (2001). Do low levels of lead produce IQ loss in children? A careful examination of the literature. Archives of Clinical Neuropsychology, 16, 303–341.

Kaya, H., Hany, J., Fastabend, A., Roth-Härer, A., Winneke, G., & Lilienthal, H. (2002). Effects of maternal exposure to a reconstituted mixture of polychlorinated biphenyls on sex-dependent behaviors and steroid hormone concentrations in rats: Dose-response relationship. Toxicology and Applied Pharmacology, 178, 71–81

Lanting, C.I., Huisman, M., Muskiet, F.A.J., van der Paauw, C.G., Essed, C.E., & Boersma, E.R. (1998). Polychlorinated biphenyls in adipose tissue, liver, and brain from nine stillborns of various gestational ages. Pediatric Research, 44, 1–4.

Needleman, H.L. (1983). Lead at low doses and behavior of children. NeuroToxicology, 4, 121–133.

Pantandin, S., Lanting, C.I., Mulder, P.G.H., Boersma, E.R., Sauer, P.J.J., & Weisglas-Kuperus, N. (1999). Effects of environmental exposure to polychlorinated biphenyls and dioxins on cognitive abilities in Dutch children at 42 months of age. Journal of Pediatrics, 134, 33–41.

Perneger, T.V. (1998) What is wrong with Bonferroni adjustments? British Medical Journal, 136, 1236–1238.

Rice, D.C. (1998). Effects of postnatal exposure of monkeys to a PCB mixture on spatial discrimination reversal and DRL performance. Neurotoxicology & Teratology, 20, 391–400.

Rogan, W.J., & Gladen, B.C. (1991). PCBs, DDE, and child development at 18 and 24 months. Annual Epidemiology, 1, 401–413.

Rogan, W.J., Gladen, B.C., McKinney, J.D., Carreras, N., Hardy, P., Thullen, J., et al. (1986). Polychlorinated biphenyls (PCBs) and dichlorodiphenyl dichloroethane (DDE) in human milk: Effect of maternal factors and previous lactation. American Journal of Public Health, 76, 172–177.

Rothman, K.J., & Greenland, S. (1998). Modern epidemiology (2nd ed.). Philadelphia, PA: Lippincott-Raven.

Sattler, J.M. (1992). Assessment of children (3rd ed.). San Diego, CA: Jerome M. Sattler.

Schantz, S.L., Gasior, D.M., Polverejan, Humphrey, H.E.B., Gardiner, J.C., McCaffrey, R.J., et al. (2002). Commentary: PCB-induced impairments in older adults. Environmental Health Perspectives, 110, A71–A72.

SIGN50. (2001). Guidelines for evaluating the internal validity of studies. Retrieved February 2003 from http://www.sign.ac.uk/guidelines/fulltext/50/checklist3.html

Steuerwald, U., Weihe, P., Jorgensen, P.J., Bjerve, K., Brock, J., Heinzow, B., Budtz-Jorgensen, E., & Grandjean, P. (2000). Maternal seafood diet, methylmercury exposure, and neonatal neurologic function. Journal of Pediatrics, 136(5), 599–605.

Stewart, P.W., Reihman, J., Lonky, E.I., Darvill, T.J., & Pagano, J. (2003). Cognitive development in preschool children prenatally exposed to PCBs and MeHg. Neurotoxicology and Teratology, 25, 1–12.

Walkowiak, J., Wiener, J.A., Fastabend, A., Heinzow, B., Kramer, U., Schmidt, E., et al. (2001). Environmental exposure to polychlorinated biphenyls and quality of the home environment: Effects on psychodevelopment in early childhood. Lancet, 358, 1602–1607.

Winneke, G., Bucholski, A., Heinzow, B., Kraemer, U., Schmidt, E., Walkowiak, J., et al. (1998). Developmental neurotoxicity of polychlorinated biphenyls (PCBs): Cognitive and psychomotor functions in 7-month old children. Toxicology Letters, 102–103, 423–428.

Winneke, G., Kraemer, U., Walkowiak, J., Fastabend, A., Heinzow, B., Borte, M., et al. (2002b). Delay of neurobehavioral development following pre- and postnatal PCB exposure: Persistent or reversible? Paper presented at The Second PCB Workshop—Recent advances in the environmental toxicology and health effects of PCBs. Brno, Czech Republic. Abstract published in Naunyn-Schmiedeberg's Archives of Pharmacology (2002), Vol. 365 suppl. Heidelberg: Springer-Verlag.

Winneke, G., Walkowiak, J., Kraemer, U., Steingruber, H.-J., & Heinzow, B. (2002a). Correspondence: Authors' reply. Lancet, 359, 1438–1439.

World Health Organization (WHO). (1995). Environmental Health Criteria 165: Inorganic lead. Geneva: International Programme on Chemical Safety.