HERSHEY MEDICAL CENTER TECHNICAL WORKSHOP REPORT: OPTIMIZING THE DESIGN AND INTERPRETATION OF EPIDEMIOLOGIC STUDIES FOR ASSESSING NEURODEVELOPMENTAL EFFECTS FROM IN UTERO CHEMICAL EXPOSURE

Robert W. Amler,
School of Public Health, New York Medical College

Stanley Barone Jr.,
National Center for Environmental Assessment, Office of Research and Development, US Environmental Protection Agency

Aysenil Belger,
Department of Psychiatry, Department of Psychiatry, School of Medicine, University of North Carolina at Chapel Hill

Cheston M. Berlin Jr.,
Department of Pediatrics, Children's Hospital, Milton S. Hershey Medical Center, Pennsylvania State University College of Medicine

Christopher Cox,
Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University

Harry Frank,
Departments of Psychology and Earth & Resource Sciences, The University of Michigan-Flint

Michael Goodman,
Department of Epidemiology, Emory University Rollins School of Public Health

Jean Harry,
Laboratory of Neurobiology, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services

Stephen R. Hooper,
Clinical Center for the Study of Development and Learning, University of North Carolina School of Medicine

Roger Ladda,
Milton S. Hershey Medical Center, Pennsylvania State University College of Medicine

Judy S. LaKind,

Contact: Judy S. LaKind, Ph.D., LaKind Associates, LLC, 106 Oakdale Avenue ,Catonsville, Maryland USA 21228, PH: 410.788.8639, FX: 410.78.1971, lakindassoc@comcast.net

LaKind Associates, LLC, Milton S. Hershey Medical Center, Pennsylvania State College of Medicine, University of Maryland School of Medicine

Paul H. Lipkin,
Division of Neurology and Developmental Medicine, The Kennedy Krieger Institute, The Johns Hopkins University School of Medicine

Lewis P. Lipsitt,
Department of Psychology, Brown University

Matthew N. Lorber,
National Center for Environmental Assessment, Office of Research and Development, US Environmental Protection Agency

Gary Myers,
Division of Pediatric Neurology, University of Rochester Medical Center

Ann M. Mason,
Research Foundation for Health and Environmental Effects

Larry L. Needham,
Division of Environmental Health Laboratory Sciences, Centers for Disease Control and Prevention

Babasaheb Sonawane,
National Center for Environmental Assessment, Office of Research and Development, US Environmental Protection Agency

Theodore D. Wachs, and
Department of Psychological Sciences, Purdue University

Janice W. Yager
Environment Division, Electric Power Research Institute

I. Executive Summary

Neurodevelopmental disabilities affect 3–8% of the four million babies born each year in the U.S. alone, with known etiology for less than 25% of those disabilities. Numerous investigations have sought to determine the role of environmental exposures in the etiology of a variety of human neurodevelopmental disorders (e.g., learning disabilities, attention deficit-hyperactivity disorder, intellectual disabilities) that are manifested in childhood, adolescence, and young adulthood. A comprehensive critical examination and discussion of the various methodologies commonly used in investigations is needed. The *Hershey Medical Center Technical Workshop: Optimizing the Design and Interpretation of Epidemiologic Studies for Assessing Neurodevelopmental Effects from* In Utero *Chemical Exposure* provided such a forum for examining these methodologies. The objective of the Workshop was to develop scientific consensus on the key principles and considerations for optimizing the design and interpretation of epidemiologic studies of *in utero* exposure to environmental chemicals and subsequent neurodevelopmental effects. (The Panel recognized that the nervous system develops post-natally and that critical periods of exposure can span several developmental life stages.) Discussions from the Workshop Panel generated 17 summary points representing key tenets of work in this field. These points stressed the importance of:

- a well-defined, biologically plausible hypothesis as the foundation of *in utero* studies for assessing neurodevelopmental outcomes;

- understanding of the exposure to the environmental chemical(s) of interest, underlying mechanisms of toxicity, and anticipated outcomes;

- the use of a prospective, longitudinal cohort design that, when possible, runs for periods of two to five years, and possibly even longer, in an effort to assess functions at key developmental epochs;

- measuring potentially confounding variables at regular, fixed time intervals;

- including measures of specific cognitive and social-emotional domains along with non-cognitive competence in young children, as well as comprehensive measures of health;

- consistency of research design protocols across studies (i.e., tests, covariates, and analysis styles) in an effort to improve inter-study comparisons;

- emphasis on design features that minimize introduction of systematic error at all stages of investigation: participant selection, data collection and analysis, and interpretation of results; these would include (but not be limited to) reducing selection bias, using double-blind designs, and avoiding post-hoc formulation of hypotheses;

- *a priori* data analysis strategies tied to hypotheses and the overall research design, particularly for methods used to characterize and address confounders in any neurodevelopmental study;

- actual quantitative measurements of exposure, even if indirect, rather than methods based on subject recall;

- careful examination of standard test batteries to ensure that the battery is tailored to the age group as well as what is known about the specific neurotoxic effects on the developing nervous system;

- establishment of a system for neurodevelopmental surveillance for tracking the outcomes from *in utero* exposure across early developmental time periods to determine whether central nervous system injuries may be lying silent until developmentally challenged;

- ongoing exploration of computerized measures that are culturally and linguistically sensitive, and span the age range from birth into the adolescent years;

- routine incorporation of narrative in manuscripts concerning the possibility of spurious (i.e., false positive and false negative) test results in all research reportage (this can be facilitated by detailed, transparent reporting of design, covariates, and analyses so that others can attempt to replicate the study);

- forthright, disciplined, and intellectually honest treatment of the extent to which results of any study are conclusive - that is, how generalizable the results of the study are in terms of the implications for the individual study participants, the community studied, and human health overall;

- confinement of reporting to the actual research questions, how they were tested, and what the study found, and avoiding, or at least keeping to a minimum, any opinions or speculation concerning public health implications;

- education of clinicians and policymakers to critically read scientific reports, and to interpret study findings and conclusions appropriately; and

- recognition by investigators of their ethical duty to report negative as well as positive findings, and the importance of neither minimizing nor exaggerating these findings.

II. Workshop Rationale

Neurodevelopmental disabilities affect 3–8% of 4 million babies born each year in the U.S. alone (Weiss and Landrigan, 2000). Fewer than 25% of these neurodevelopmental disabilities have a known etiology. It is now appreciated that subtle damage that occurs to the nervous system during early development can manifest much later in life. This makes the ability to establish a relationship with events occurring during gestation even more challenging. In an effort to identify the causes of neurodevelopmental disabilities, epidemiologic research is a valuable tool that can be used to identify potential links between disease and genetic and environmental factors. Numerous epidemiologic studies have examined potential links between *in utero* or early postnatal exposure and specific chemicals (e.g., pharmaceuticals, environmental chemicals such as lead, methylmercury, polychlorinated biphenyls [PCBs]) and adverse developmental or behavioral effects in children (see Rice and Barone, 2000, for a recent review). These studies have been invaluable in laying the groundwork for how such investigations should be conducted, and provide an excellent foundation for future studies.

Given the current interest in expanding such studies to address issues related to adverse effects of low-level exposures to environmental factors, an examination of the methodologies commonly used would be of significant value to investigators in the design and analysis of future studies. Such a review would assess the strengths and limitations of methodological approaches used to date, and consider scientific and technical advances in relevant methodologies, such as exposure assessment, neurodevelopmental assessment, interpretation of data, and incorporation of an evidence-based approach to identify health concerns. This review would serve to identify key methodological factors that ultimately determine the value and strength of a study.

The Hershey Medical Center Technical Workshop: Optimizing the Design and Interpretation of Epidemiologic Studies for Assessing Neurodevelopmental Effects From *In Utero* Chemical Exposure was a one-day meeting held in conjunction with the 22nd International Neurotoxicology Conference (Environment and Neurodevelopmental Disorders), Research Triangle Park, North Carolina, September 2005. Within this framework, a multidisciplinary panel was convened to discuss issues as they relate to the design, conduct, interpretation, and dissemination of information of human studies examining the potential adverse effects from gestational exposure to various environmental agents. The Panel was comprised of experts in psychology, medicine, risk/exposure assessment, analytical chemistry, neuroimaging, epidemiology, toxicology, statistics, psychiatry, pediatrics, pediatric neuropsychology, and neurology. Each member had experience and interest in assessing the effects of environmental chemical exposure on human development.

This Workshop was organized to discuss the important principles for detecting the effects of environmental exposures on neurobehavioral development and to make recommendations for the design of future studies evaluating the impact of *in utero* exposures. (The Panel recognized that the nervous system develops post-natally and that critical periods of exposure can span several developmental life stages.) The discussions were initiated by a series of questions related to scientific methodological issues that were posed to the Panel prior to the Workshop (see Table 1). Given that adverse effects may emerge after long latent periods, the Panel discussed how effects that manifest as irreversible damage to the central nervous system, progressive neurodegeneration, or subtle neurological dysfunction first appearing in adolescence and adulthood could be considered and incorporated into study designs. The Panel focused on identifying 'best practices' for such studies which often required revisiting the basic principles underlying current epidemiological studies. The Panel did not evaluate conclusions or findings from previous neurodevelopmental epidemiological studies related to the topic of environmental exposure; however, discussion of focal points from such studies served the basis

for identifying critical points for consideration in any future study designs. The outcome of this Workshop serves to provide input for both the design of future investigations and establishment of standards whereby one can judge the adequacy of reported studies. This report represents a summary of the Panel deliberations, including current basic scientific tenets that have been embraced by this field, as well as considerations for future work.

III. Study Design for Neurodevelopmental Outcomes from *In Utero* Exposure

Developmental epidemiological studies examining perturbations to the nervous system from exposure to environmental factors have provided critical data for determining human heath risks. These studies have set the framework for the design of subsequent studies. Early observations regarding adverse effects on the fetus from maternal exposure to various agents focused primarily on pronounced and obvious adverse outcomes from high exposure levels. Contemporary concerns have shifted and are now focused on more subtle alterations that may occur with low levels of exposure. Such alterations, while not overt, may have long-term impact on human health. The evaluation of such relationships between exposure and outcome presents a significant challenge to the investigator, especially when the effects are subtle or delayed in manifestation. In the discussions, the Panel recognized the importance of revisiting some of the fundamentals of neurodevelopmental epidemiology and basic critical components of study design as they relate to studies of *in utero* exposure. An extensive discussion of epidemiological study design was outside the scope of this Workshop, as was the evaluation of specific published research.

III.A. Types of Designs

The charge to the Panel was to evaluate study design for determining an association between *in utero* environmental chemical exposure and developmental outcomes in exposed children. Given the rationale for this Workshop, the Panel focused part of the deliberations on design strategies for examining various outcomes resulting from such *in utero* exposures. First and foremost, the Panel emphasized the importance of formulating a specific hypothesis or set of hypotheses as part of the study design. Such hypotheses should be informed by and consistent with previous research and, whenever possible, have biological plausibility and be related to an underlying biological mechanism. The degree of focus or specificity of the hypothesis will depend to some extent on the number of different outcome variables. However, hypothesis formulation is important for the design, analysis of data and interpretation of results.

The prospective longitudinal design has proven itself as one of the most useful study designs for examining outcomes associated with *in utero* exposures. Challenges exist when assessing the influence of *in utero* exposures on delayed manifestation of neurotoxicity, and the longitudinal approach may be the primary approach to detect delayed adverse outcomes. The impact of this type of study design comes from the ability to assess exposure at the beginning of the study, with the possibility of obtaining greater accuracy. Most alternative designs require a retrospective assessment of exposure — a useful, but less attractive option. The collection of longitudinal data allows for the possibility of detecting subtle changes in development that may not be apparent in cross-sectional analyses. Longitudinal assessment of both exposure and outcome has required the development of new methods of statistical analysis that can incorporate the resulting correlation among repeated measurements. Such methods permit the assessment of change within individuals over time. This approach has greatly increased the power of the statistical inference in longitudinal studies. The prospective longitudinal design can include developmental testing and assessment of potential confounding variables at regular, fixed time intervals, typically for a period of two to five years, and possibly even longer. The ability to reassess or identify potential confounding variables over the course of the study offers a unique strength to such a study design. However, as in any longitudinal study,

the positive aspects of the extended length of the follow-up period must be balanced with the difficulties of maintaining active subject participation. Special efforts are often required to prevent attrition. Over time, the attrition can become systematic as opposed to random. This, in turn, can affect the results of the data analysis. However, with a longitudinal study design, one can actually examine the various stages of child development. Thus, it may be possible to demonstrate subtle differences due to *in utero* exposure by examining rates of development, as well as specific cognitive, motor, sensory, psychological, or other outcomes.

One design alternative is what is sometimes referred to as an historical cohort study. This involves enrolling a group of children within a particular age window, and retrospectively determining gestational exposure. Such assessment is difficult and will only be possible in very special circumstances, such as when sufficiently detailed, high-quality records, or even biological samples, are routinely available. Alternative useful designs, depending on the questions being asked and the hypotheses being posed, are case-control or case-cohort design (Rothman and Greenland, 1998). These designs are useful if a particular developmental problem is of interest for which standard diagnostic procedures are available. Within a cohort study, one can nest both case-control or case-cohort studies to take advantage of specific diagnostic tests.

This type of design may be especially important in the study of exposure in children, where the developmental stage may be critical in determining the ability to detect effects of neurotoxicants or other environmental risk factors in a given cohort.

III.B. Participant Selection

The design of a particular study, will, in general, dictate the participants selected. For the majority of studies examining in utero exposure, the basic design typically involves assessing exposures that occur during pregnancy and outcomes that are assessed after birth. For such a study, during an initial enrollment period a group of pregnant women is identified early in pregnancy and both the women and children are followed longitudinally for a specified amount of time. The overall goal of participant selection is to select a sample that mirrors the parent population, and thus maximizes the generalizability (external validity) of an observational study (Kalsbeek and Heiss, 2000).

If a cohort is selected completely at random, it is assumed to be representative of the population under study. In a simple random sample, however, the possibility still exists that individuals in certain categories would be missed, over-sampled, or under-sampled. To address this issue, studies may rely on stratified random sampling strategies that involve the creation of mutually exclusive and exhaustive strata (for example, based on age, ethnicity, and residence) and selecting samples from each of those strata (Lemeshow and Stroh, 1988). The main disadvantage of both simple and stratified random sampling methods is that they require a complete list of the population from which a sample is drawn. One solution to this problem is to use a cluster sampling strategy where the original lists consist of clusters, such as villages, wards, or polling districts, rather than individuals (Kalsbeek and Heiss, 2000). Another possibility is a two-stage design in which subjects are selected for recruitment after an initial survey. In many instances, this initial survey would include a preliminary estimate of exposure. Although random samples selected from well-defined sampling frames are uncommon, some thought should be given to the available pool of subjects, and whether this represents the population to which the results of the study should apply. More representative samples can be obtained by using sites with different geographic locations representing different subpopulations, as has been proposed for the National Children's Study (http://nationalchildrensstudy.gov/).

The actual process of selecting a representative sample is an ongoing challenge in both case-control and cohort studies. Technology, privacy laws, and regulations are severely limiting the use of established methods of selection, such as random digit dialing (RDD) (Brogan et al., 2001) and Department of Motor Vehicles (DMV) records (Funkhouser et al., 2000). Hospital-based and neighborhood recruitment strategies can be time-consuming, and are more likely to produce samples that are not representative of the general population (Wacholder et al., 1992). Recently, large comprehensive population directories, merged from multiple sources, have become commercially available. It has been suggested that these directories may be more comprehensive and more cost-effective than the traditional sources of participant selection. Supportive data are not yet available regarding their ultimate utility, and unless very diverse sources of information are used in these directories, they may be subject to stratification bias.

III.B.1. Sources of Bias—The term *selection bias* refers to introduction of systematic differences in characteristics between those who take part in (and provide data for) the study and those of the population, resulting in lack of generalizability and possibly comparability between groups being studied (Grimes and Schulz, 2002). By contrast, *information bias* occurs from errors in obtaining the needed information (Rothman and Greenland, 1998). Examples of bias include: inappropriate selection of controls in case-control studies, Berkson's bias (i.e., when both disease and exposure are likely to be associated with hospitalization, but unrelated to each other in the general population), differential loss to follow up in longitudinal studies, non-response/missing data, volunteer/self-selection, and misclassification of exposure or outcome (Gordis, 2000). To minimize selection bias in the initial recruitment, the researchers may choose to over-sample certain population strata where there is a higher likelihood of refusal or drop-out (Ware and Lee, 1988). Although this approach may help ensure sufficient representation of hard-to-study groups, it does not take into consideration the issue of differential drop-out within the over-sampled categories.

It is important to note that even well-designed studies can become subject to bias in the implementation phase. Participants may refuse to enroll or fail to provide necessary data, or be lost to follow up due to change in residence or lack of motivation. The refusal to participate, failure to provide data, and loss to follow-up may be of particular concern when there is evidence that participation and missing information are not independent of exposure and outcome status. For example, in a large longitudinal study of children in South Africa, Richter et al. (2004) reported that after 10 years the project had retained 70% of the cohort, with different rates of attrition in different subpopulations. In a cohort study evaluating the association between methylmercury exposure and neurobehavioral testing results among Faroese children, the final dataset included approximately 63% of the eligible cohort (Grandjean et al., 1997). This level of participation in and of itself may not introduce bias; however, the hospital with the lowest participation (33%) had the highest median blood mercury concentration (Grandjean and Weihe, 1993).

In those circumstances where a concern about selection bias is justified, an extra effort to recruit a subsample of non-participants may be warranted. This allows researchers to evaluate and quantify the differences between participants and missing subjects—including individuals lost through attrition or non-sampled sub-populations (Caetano et al., 2003). To minimize bias due to subject drop-out, one may: a) obtain information on family or friends as additional contacts; b) maintain regular contact with participants; c) utilize monetary compensation; and/or d) use methods for early identification of attrition problems for correction (Hartman, 2005). Alternatively, one can address the potential impact of missing information during the analysis phase of the study.

During conduct of the study, errors in the collection of the data may also be a concern. Sources of information bias include measurement and misclassification errors, and recall bias in

retrospective assessment. There appears to be growing consensus that bias quantification should become an important part of the collection, analysis, and interpretation of observational data, with an understanding that random variability is not the only important source of error (Caetano et al., 2003;Greenland, 2005;Hernan et al. 2004;Little and Vartivarian, 2003;Maclure and Schneeweiss, 2001). Both selection and information bias issues may be particularly important in cohort studies, especially if existing cohorts are used in the future to test hypotheses that were not considered at the time of the initial recruitment.

III.C. Assessment of Exposure

As noted by Rothman and Greenland (1998), "In large part, the quality of exposure measurement will determine the validity of an environmental epidemiology study." Exposure assessment is clearly a critical component in these types of studies, particularly in studies of low-level toxicant exposures. Before exposure can be assessed, it must be defined. The definition should include an understanding of the exposure setting, the source, nature, and context of the exposure, whether the exposure of interest is a single agent or a mixture, and the possibility of exposure to other compounds not under study (Rothman and Greenland, 1998).

Individual exposures can be relatively constant or highly variable over time. Daily or seasonal variability in exposure can occur as a function of location of primary residence and pattern of daily activity including dietary preferences and cultural practices. This variability determines the sampling interval and may necessitate the need to construct summary measurements for analysis. Such summaries may provide a more accurate measure of total exposure than individual measurements. In addition, the biological kinetics of a compound drive the measures used for analysis. For many toxicants, an average daily exposure can be used; for others, the peak exposure is more relevant. Such decisions depend upon on the nature of the specific toxicant and the actual level and variability of exposure. Studies frequently rely on indirect assessments using biomarkers of exposure rather than on direct exposure measurements. While such markers can be useful, they can have the potential to be misleading and may require additional confirmation.

While accurate exposure measures are often difficult to obtain, they are essential if one is to interpret any statistical associations that might occur. Thus, quantitative measurements are preferred to those relying on subject recall. While occupational or acute exposures can often be determined with a degree of accuracy, chronic exposure to environmental toxicants can be variable and difficult to measure accurately. In this regard, exposure assessment for many agents is best measured prospectively during pregnancy. If applicable, a profile can be generated for postnatal exposure and information obtained retrospectively on pre-pregnancy exposure. The intensity of exposure must be sufficiently great and the length and frequency of follow-up adequate so that the exposure is more likely to affect measures of outcome in the hypothesized manner.

III.C.1. Biomonitoring—Environmental measurements (e.g., diet, soil, air) can be used to assess exposure as part of a neurodevelopmental study. In addition, biomonitoring has been incorporated into epidemiological studies on neurodevelopment. There have, for example, been investigations into the relationship between neurodevelopment and concentrations of selected environmental chemicals, primarily the persistent lipophilic chemicals, in human milk and maternal plasma. Over the past several years, guidance on the selection and collection of biological samples, as well as on sampling time and methods, have been developed. For example, total serum lipids are known to vary with the stage of pregnancy and decrease after delivery (Montes et al., 1984). Consequently, exposure to lipophilic environmental chemicals should be adjusted for the lipid percent at the time of sample acquisition.

Estimates of fetal exposure can be made based on maternal biological samples, such as urine, blood, and breast milk (Barr et al., 2005;Budtz-Jørgensen et al., 2004). Levels of chemicals in these maternal compartments, however, may not accurately reflect the actual dose to the target tissue, including the fetal compartment. For estimating fetal exposure, sampling of amniotic fluid, generally between 15 to 18 weeks of gestation, would be the most direct approach available; however, this is feasible only under very specific health conditions and is not a standard practice. In addition, this type of estimate, while accurate, would likely result in a biased population sampling. Finally, it has not been determined how these levels relate to the more routinely obtained maternal blood matrix.

With birth, other tissue matrices become available. These include meconium (the infant's first stool), umbilical cord blood, umbilical cord tissue, and the infant's nails and hair. Meconium starts accumulating in the fetal bowels from about 12–16 weeks gestation (Bearer, 2003) and has been used as a viable matrix for assessing exposure to drugs of abuse (Lauria et al., 1997;Ostrea et al., 2001), alcohol (Bearer et al., 2003), tobacco (Ostrea et al., 1994), and environmental chemicals (Ramirez et al., 2000;Whyatt and Barr, 2001); again, the relation to more traditional matrices has not been established. Of these matrices, cord blood is the most common matrix monitored. It has been estimated that in the month prior to a child's birth, about 300 quarts of blood are pumped daily from the placenta to the fetal environment. In a recent website report, 287 chemicals were reportedly detected in cord blood (www.ewg.org/reports/bodyburden2). The chemicals included many halogenated organic chemicals, mercury, and polycyclic aromatic hydrocarbons. Although cord blood is exchanged with maternal blood, the concentrations of various chemicals are not necessarily equal. For example, it is known that that the concentration of methylmercury in fetal blood is about 1.7 time that in maternal blood (Amin-Zaki et al., 1974). Sampling of cord blood for chlorpyrifos showed a relationship to indoor air levels as well as outcome measures of length and weight at birth (Whyatt et al., 2004). Burse et al. (2000) utilized umbilical cord tissue to assess *in utero* exposure to selected persistent pesticides and PCBs; one of the difficulties they faced was gravimetrically determining the low lipid content (about 0.2%) of the cords and the variable and unmeasured presence of residual red blood cells in the vessels when the cord tissue was taken. For selenium, Lorenzo et al. (2005) reported positive correlations between cord blood and newborns' hair, between placenta and umbilical cord, and between cord blood and maternal blood. For more details on the matrices available and volumes required for assessing *in utero* and early postpartum exposure, the reader is referred to Barr et al. (2005).

III.D. Assessment of Neurodevelopmental and Related Outcomes

Compared to the measurement of exposure, outcome assessment is comparatively straightforward, since a number of standard developmental test batteries with good psychometric properties are readily available. Since the early work of Hanninen (1966) examining carbon disulfide exposure, the field of human behavioral neurotoxicology has embraced a neurobehavioral approach to the assessment of outcome following exposure, although the specific neurocognitive domains that are targeted vary from study to study. Over a decade ago, Anger et al. (1994) provided a list of functional domains in adults that could be affected by chemicals. These included learning and memory, coding, sustained attention, higher intellectual function, strength, coordination, speed, vision, somatosensory function, and affect. Over the past four decades, this assessment approach to neurodevelopmental outcomes has been extended downward to examine very young children following their exposure *in utero* and excellent outcome measures are now available across the lifespan.

One of the earlier efforts in this area came from the World Health Organization (WHO), which recommended the use of the Neurobehavioral Core Test Battery (NCTB). This battery was comprised of a variety of tasks including digit symbol, digit span, Benton visual retention,

pursuit aiming II, simple reaction time, Santa Ana, and a profile of mood states. This original conceptualization has influenced the selection of neurobehavioral tools for the past 35 years (Anger, 2003). In the context of developing a testing strategy for the National Children's Study, a workshop sponsored by National Institute of Child Health and Development (NICHD) and the Environmental Protection Agency (EPA) examined the questions of what, how, and when to measure neurobehavioral function in children. They considered batteries of tests and stand-alone metrics for use in longitudinal studies examining multiple functional domains. A major emphasis in this report was placed on testing functional competence, cognitive, sensory, motor and social/emotional development in the first four years of life (http:// nationalchildrensstudy.gov/events/workshops/Neurobehavioral_Development.cfm). Paule (2005) also argued for task continuity across species to identify species similarities and differences, as well as to facilitate extrapolation of experimental animal data to humans.

While NCTB has driven much of the application methodology in pediatric neurotoxicology, its applicability to large epidemiological studies has been questioned. The direct assessment of these tasks to exposed individuals and the use of trained personnel create a labor-intensive situation for studying outcome. This is particularly true for large epidemiological studies and studies where cultural and linguistic issues are present. As a result, a number of other approaches have been employed to address these concerns. For example, in 1996 the Agency for Toxic Substances and Disease Registry designed the Pediatric Environmental Neurobehavioral Test Battery (PENTB) for children ages 1 to 16 years. The PENTB was constructed for cross-sectional studies. For children younger than 4 years of age, four informant-based scales are employed: the Parenting Stress Index (PSI), the Personality Inventory for Children (PIC), and the Vineland Adaptive Behavior Scales (VABS). For children ages 4 through 16, the PENTB evaluates cognitive, motor, sensory, and affective domains (Zeitz et al., 2002).

Another approach has embraced computerized assessment strategies and a variety of these have been developed. They require less time and expertise and provide for rapid scoring and data entry strategies. For example, tasks for The National Center for Toxicological Research Operant Test Battery and its variants (Paule, 1990;Paule et al., 1999) were selected based on their association with brain functions (e.g., short-term memory, attention, learning, time perception, motivation, color and position discrimination). For assessing older children, Iregren and Letz (1991) suggested the use of a Minimum Common Core Computerized Battery comprised of symbol digit, finger tapping, and simple reaction time. Otto et al. (1996) used a computerized battery, the Neurobehavioral Evaluation System (NES2), in pediatric neurotoxicology studies. This has been recently updated to the NES-3 (Proctor et al., 2000). A number of other computerized batteries have been developed, including the Cambridge Neuropsychological Test Automated Battery (CANTAB®), CogSport, HeadMinder, ImPACT, Automated Neuropsychological Assessment Metric, Cognitive Evaluation Protocol, and CNS Vital Signs. Other available batteries for use with both humans and non-human primates include the National Center for Toxicological Research (NCTR) Operant Test Battery (Paule et al., 1999) and the Behavioral Assessment and Research System (BARS) (Anger et al., 1996). Computerized batteries have been developed which combine various facets of other batteries (e.g., Rohlman et al., 2001). *However, at present, there is no one neurobehavioral function test battery that has been recommended across age ranges or across cultures for any specific neurotoxic chemicals because none of the current batteries tests all functional domains of the nervous system function*. As an overall guiding principle, choice of cognitive outcome measures should be based on the biological evidence of nervous system structural and functional development. To illustrate, during the infancy period (birth to 2 years of age), research suggests the developmental salience of three specific cognitive dimensions: (1) speed of processing, as assessed on tasks like habituation and visual expectancy performance (Colombo, 1993;Dougherty and Haith, 1997); (2) recognition memory, as assessed on tasks

like novelty preference and conjugate reinforcement (Rovee-Collier and Barr, 2001;Fagen and Ohr, 2001); and (3) behavioral inhibition, as assessed on measures like the A not B task. As the child develops other specific cognitive processes, such as the development of selective attention (Ruff and Rothbart, 1996), processing speed and working memory become increasingly important (Schneider, 2002).

III.D.1. Non-Cognitive Approaches to Developmental Outcomes—The majority of human behavioral neurotoxicology studies have focused primarily on measures of cognitive competence in children. For example, with regard to lead, most of the human developmental studies referenced in a recent review used measures of general cognitive function or cognitive-related measures such as visual-motor behavior, executive function, and attention, though some studies have reported an increased risk for anti-social behavior and impulsivity in children (Hubbs-Tait et al., 2005). The same review also reported virtually no evidence of non-cognitive outcomes as a function of exposure to mercury, manganese, or cadmium (Hubbs-Tait et al., 2005).

In recent years, developmental researchers have emphasized the multi-dimensional nature of children's development, with specific reference to the concept of competence (Masten and Coatsworth, 1998;Masten and Powell, 2003). In addition to the acquisition of culturally- and age-appropriate specific cognitive abilities, the development of competence in children also involves the acquisition of intrapersonal characteristics and interpersonal skills that help the child to meet both major culturally salient developmental goals and to deal effectively with environmental challenges (Yates et al., 2003). This conceptual shift to a multi-dimensional approach to assessing children's development in part reflects the fact that normality of development in one domain does not necessarily mean that there will be normality across all domains (Lester et al., 1995). A major implication of the concept of competence for research in behavioral teratology and toxicology is the need to go beyond a primary focus on cognitive outcomes, such as developmental quotient or IQ scores (Bellinger, 1995;Cory-Slechta, 1990;Rice, 2005;Weiss, 2000) to assess other critical developmental domains that also define competence. Some of these domains include: 1) temperament, including reference to indices of self-regulation and reactivity; 2) development of a secure attachment by the young child; 3) development of motivation by the young child to be actively involved with their environment; and 4) developing the ability to elicit and use other persons as resources and establishing appropriate social interaction patterns with caregivers and peers (Wachs, 1999;Masten and Powell, 2003;Werner, 1990). Children who are strong in these characteristics possess a range of flexible coping strategies to deal with day-to-day stressors, are more likely to demonstrate better performance in school, and are less likely to exhibit behavioral problems.

For example, one major dimension of temperament is self-regulation. This dimension of non-cognitive competence has unique linkages to specific cognitive executive functions, and refers to the child's ability to control his or her own behavior and emotions (Rothbart and Bates, 2006). After the first year of life, indices of self-regulation (e.g., the ability to inhibit ongoing behavior, selectively attending to cues, joint attention) become more stable and increasingly influence an individual's competence (Bates, 2001;Rothbart and Bates, 2006). Dimensions of self-regulation in infancy and childhood can be assessed either by validated parent report measures (e.g., Rothbart et al., 2001) or laboratory-based behavioral assessments (e.g., Kochanska et al., 2000). These measures have been demonstrated to be important in the context of environmental exposures; some studies have reported self-regulation deficits in infants and young children exposed to environmental chemicals (Burnette et al., 1999;Chasnoff et al., 1987;Mayes and Bornstein, 1995;Mendelsohn et al., 1998). Other studies indicate that infants exposed to certain environmental chemicals are harder to test and less likely to complete testing due to deficits in self-regulation skills, such as high distractibility or low consolability when distressed (Fagen and Ohr, 2001;Mayes and Bornstein, 1995;Struthers and Hansen, 1992).

From these data, models have been developed for environmental exposure and disturbances in the development of regulation processes (Mayes, 2002).

A recent monograph from the Institute of Medicine (2004) presents a series of criteria that can be used to evaluate the utility of different procedures for assessing different aspects of infant competence (e.g., age appropriateness of instrument, degree to which scores on the instrument are predictive of later development, evidence that the instrument is sensitive to toxic exposure, whether performance on the instrument can be linked to perturbations in brain function/brain development). The Institute of Medicine (2004) monograph also stresses the importance of utilization of converging measurements of the same construct (aggregation) as a means of maximizing assessment sensitivity. The conclusions drawn in this monograph can serve as a guide to both dimension and instrument selection when assessing both cognitive and non-cognitive competence in early life. Choice of non-cognitive outcome measures should be based on the biological evidence of nervous system structural and functional development.

III.E. Statistical Considerations

While a review of statistical methodologies in neurodevelopmental epidemiology studies was beyond the scope of the Workshop, the Panel emphasized a number of fundamentals such as the importance of laying out a statistical design for the study in advance and ensuring that the statistical methods be transparent to the reviewer and reader. The complexity and expense of most epidemiological studies confines them to confirming suspected associations rather than generating new hypotheses. Consequently, the formulation of *a priori* hypotheses based on prior empirical data or theoretical considerations is essential. After formulating a hypothesis, the statistical design for the primary analysis of the study should be determined in advance. An *a priori* primary analysis plan should provide the investigators the primary answer for whether the hypothesized associations are in fact present or not. This primary analysis plan should be adhered to and the statistical methods and results should be reported in enough detail that they could be replicated (i.e., transparency). This minimizes the possibility of identifying obscure effects or interactions as the result of conducting multiple tests. Secondary analyses can then confirm the primary study or not, answer secondary questions, explicate limitations and explore other possible associations. Secondary analyses can be very useful, but should be identified as such. For example, there may be susceptible populations to specific environmental chemicals. Secondary analyses of such susceptible groups can prevent a Type II error (failure to detect a real effect when one is present because it is confined to a particular group of individuals).

In both planning and evaluating the results of a confirmatory study, it is critical that sample size be taken into account; if the sample size is too small, important effects on small sub-populations can be completely masked. It is less well-recognized that small studies can also have a greater probability of Type I errors. One solution is to take a two-prong strategy: (1) cluster sampling, which (unlike random sampling) ensures representation of small subpopulations; and (2) calculating sample sizes in advance, based on *a priori* established probabilities of Type I and Type II errors. A sample size calculation is a prerequisite for any confirmatory study. In addition, in order to appropriately power studies, one needs a strong knowledge base to identify effects that should be considered. Most notable is the use of mechanism or mode of action for any specific compound underlying a specific endpoint under study.

III.E.1. Confounders, Covariates, Mediators, and Effect Modifiers—The goal of environmental epidemiology is to evaluate the relationship between exposure and a disease or other outcome. This is nearly always accomplished by the use of statistical modeling in which the outcome (dependent variable) is modeled as a function of the exposure (independent

variable). However, there is always a concern that factors other than the exposure could explain an observed relationship. Epidemiologic methods rely on the appropriate use of confounder measurements to avoid or adjust for confounders. While this appears to be rather straightforward, assessment of not only known confounders but of determining how these and other factors change over time and the impact of such changes become critical factors in neurodevelopmental studies. With the expansion of studies to examine effects that occur over a lifetime, an additional effort may be needed to identify influences not previously considered. For these reasons, the Panel reconsidered the current terminology and definitions.

Briefly, a covariate is a variable associated with the outcome of interest. Its inclusion in the model may increase the precision of the parameter estimates. A confounder is related to both exposure and outcome, and affects the estimate of the association between the two. As such, a confounder affects the estimate of the association between the two. Unless the statistical model adjusts for such confounding variables, a biased estimate of the association can result; either the appearance of an effect can occur where none exists or a real association can be attenuated. There are many factors (covariates) that affect developmental outcomes, although they are typically unrelated to exposure; however this may not be the case in every situation and with every environmental exposure; for example some exposures may be associated with SES.

Confounding variables that are commonly considered in studies of neurodevelopmental effects in children include participant age, race, gender, socioeconomic status (SES), and maternal variables such as age, education, and IQ. While each of the variables appears to be distinct, the variables are a mix of multiple potential confounding factors. For example, while age is an important variable in itself, it could also encompass other aspects such as the age at first exposure, the age-period of exposure duration, or the developmental stage of the child when exposure occurred. In addition, in the case of *in utero* exposure, the duration of exposure could be significantly influenced by the pregnancy duration inclusive of premature birth. This would influence the accumulated dose level and whether or not the exposure occurred during a normal period of *in utero* development. For children, additional variables include health status, nutrition and diet, educational experiences, and social experience. In any developmental study, the influence of the parents on the offspring is significant. This presents as a collage of postnatal confounders including genetic background, prenatal environment, parent/child interactions, mental heath, home environment, and health status. For example, while SES may not be a direct risk factor, many of the factors related to SES may increase the risk for the outcome under study. In addition, confounders like SES are not necessarily stagnant and can change over time. For this reason, such confounders require repeated assessment over the course of the study. In the case of multiple confounders, evaluation of a change with or without adjustment for other confounders can make a large difference in the final estimate. Even when adjustments are made, one must continue to be concerned with the remaining (residual) confounding. Given the impact of confounders in the final interpretation of the study results, it is critical that the criteria used to select confounders for adjustment and how they were measured be reported as part of any analysis plan.

More complicated is the potential need to differentiate between confounders and mediators. If an intervening variable is functioning as a mediator, an indirect causal pathway is established between exposure and outcome. In this case, the mediator serves as the mechanism through which toxic exposure translates into biological deficits. For neurodevelopmental studies, an example would include the case where exposure to an environmental chemical produces a change in an infant's temperament characteristics. Such changes could include an increased level of irritability or a reduced level of responsiveness or attention. These, in turn, can have a significant influence on testing performance or have other long-term behavioral consequences (Bendersky et al., 1996;Chasnoff et al., 1987;Fried, 1989;O'Connor et al., 1993;Ruff and

Rothbart, 1996). Under such conditions, exposure would result in negative developmental consequences, albeit through an indirect causal pathway. Thus, treating either temperament or patterns of parent-child relations as a confounder would not be appropriate. Given that confounders and mediators may show the same pattern of relations with predictors and outcomes, it is important to develop decision rules for interpretation. One such rule would involve the degree to which one can assume a causal link between toxic exposure and mediator, as opposed to a non-causal association. Reviews of statistical approaches to analyze for mediator effects are outside the scope of the workshop and can be found in a number of sources (Evans and Lepore, 1997;MacKinnon et al., 2002).

While studies often attempt to isolate the effects of a single chemical agent, exposure to multiple environmental agents occurs in the population (Cory-Slechta, 2005;Mayes, 2002;Wachs, 2000). The level of relation between exposure to environmental chemicals and developmental outcomes can vary systematically as a function of population characteristics (Bellinger, 2000;Ruff, 1999). In this situation, one is dealing with effect modifier(s) or moderating variable(s). An effect modifier modifies the effect of the exposure on disease. It is the variation in the magnitude of any specific measure of an exposure effect as it varies across levels of another variable. In the absence of identifying and testing for moderators, exposure-outcome relationships can be under- or over-estimated due to the fact that the impact of such exposure will be restricted only to a specific subgroup within the larger sample population (Bellinger, 2000;Davidson et al., 2004;Jacobson and Jacobson, 2002;Rauh et al., 2004;Vreugdenhil et al., 2002;Wachs and Plomin, 1991). For developmental neurotoxicology assessments in the human population, such modifiers may become more difficult to assess as the study duration increases and the number and types of modifiers change. How to handle these changing dynamics within a population and within individuals of a population will be a major challenge for future studies.

IV. Interpretation and Dissemination of Study Information

Epidemiologic studies of neurodevelopmental outcomes place extraordinary demands on the professional, intellectual, ethical, and communicative skills of the investigators. As a part of the standard scientific process, data obtained from one study are used to generate hypotheses to be tested in subsequent studies. While this sequence of events, in general, serves to advance our understanding of a process, hypotheses are just that. Quite often, hypotheses generated from studies on the human population can have a significant impact, in and of themselves, on the population under study. Thus, care must be exercised in the final interpretation of such data and the future use of the data. The Panel stressed the care required in order to advance the science while at the same time responsibly interpreting and communicating study data. For these reasons, the Panel reiterated the need for full, open, and transparent reporting of all aspects of a study design and analysis. In addition, the Panel encouraged the reporting of all results with a discussion inclusive of all of the findings of a study whether they identified a relationship or not. It was acknowledged that the reporting process requires at least as much planning, care, and attention to detail as other aspects of study design and implementation. However, it is often the last aspect to be considered in the effort to publish results. Ideally, report planning begins during the study design phase; this is when the investigators decide what research questions will be addressed and consider what conclusions might be drawn from various study outcomes that are possible. Use of data to address questions for which the study was not designed should proceed with caution because of the potential for misinterpretation or misattribution of the relationship between exposure and adverse effects on the nervous system. The results of such *post hoc* exploratory analysis may be hypothesis-generating, but not necessarily definitive.

A study report reaches many audiences, each with its own core concerns and priorities. In addition, the investigators have a responsibility to the participants of any specific study. Investigators need to consider these constituencies in reporting their findings.

- *Study participants* want to know their individual results, any immediate needs for medical or public health follow-up, and any implications for their health and well being. They also want to know about any heightened risks to their household members and community.

- *Members of the study community* want to know about any heightened risks to the community or any of its members, and any need for public health actions or follow-ups.

- *The public at-large* wants to know the public health implications of the results, and any recommendations for public policies or regulations to protect the population.

- *The scientific community* wants to know the detailed methods and results, and the investigators' interpretation. They seek to replicate the findings, if possible, compare the findings to those obtained by other investigators, and consider any generalizable principles or trends suggested by the aggregated outcomes of multiple investigations.

- *The sponsor or funding agency* wants a coherent summary of the information that was gained, and how that information can be used to improve and promote health, eliminate or reduce risks, improve the environment, and protect public health.

- *The investigators themselves* want to understand every detail and nuance of the study findings, to enable exhaustive analysis and maximize the quality and quantity of useful information obtained from the study. They also want to compare their work to other studies, translate their findings into prudent recommendations for action, and generate hypotheses to be examined by further inquiry.

- *The news media* seek take-home messages, translations of the scientific information into lay-person language, and implications for public policy.

Collective experience has been accumulated, by the authors and many others, from dozens of environmental health studies conducted over two decades across the United States (for review, Amler and Tinker, 1992;Friedman et al., 1999;Amler and Falk, 2000;Brauer et al., 2004;Schwartz et al., 2006). This experience demonstrates a consistent benefit of early and regular contact with study communities as an indispensable aid in report planning. As a part of the responsibility of the investigators, prospective study participants, household members, and community members can be familiarized with study objectives and study design. These participants can learn what information can be reasonably expected from the study results. Report planning is facilitated by the enhanced familiarity of study participants and community with issues. This familiarity can also assist in raising concerns that are likely to surface but are unknown to the investigator.

Results that are statistically significant may be significant only in the mathematical probabilistic sense—that the observed outcomes were unlikely to be due to chance alone. For example, a very small observed difference between two study groups in a neuropsychological test result may be inconsequential in any real sense but may achieve statistical significance simply due to the large study population or the large number of tests performed.

Because statistical significance is dependent on sample size, studies with large-sample sizes can yield statistically significant results that have only trivial clinical importance. This implies that results must be interpreted carefully considering at least three distinct types of conclusions and how they will be reported: the medical implications for individual participants, the potential health impact on the study population, and the acceptance or rejection of the study hypotheses.

IV.A. Reporting Test Results to Individual Study Participants

A frequently encountered question is: how meaningful are study findings as they relate to individual study participants, the community, and the general advancement of environmental toxicology. For example, some of the non-cognitive neuropsychological tests as well as tests for specific cognitive functions, although useful in evaluating populations, do not have well established absolute normal values and thus must be interpreted solely in a comparison between exposed and non-exposed groups (Amler and Gibertini, 1996). This implies that results must be interpreted carefully considering at least three distinct types of conclusions and how they will be reported: the medical implications for individual participants, the potential health impact on the study population, and the acceptance or rejection of the study hypotheses.

All study participants with aberrant test results are entitled to some clinical reporting and counseling, or at least referral to an appropriate competent medical resource. Virtually all such individual interpretation issues are easier to manage when they are explained in advance of the study, detailed in the informed consent documentation, and reinforced throughout the conduct of the investigation.

Therefore, a communications plan is essential, both to alleviate concerns about test results that are statistically at variance but not deemed abnormal or injurious, and to ensure proper reporting of individual results that might require medical attention. Investigators must be cautious to avoid labeling study participants, especially children, and be circumspect in reporting test results to parents. Cognitive test results can be especially provocative because many people equate them with intellectual ability. Although there are few empirical studies on the effects of such labeling (Hastings, 1994;Jellison and Duke, 1994;Sparrow et al., 1993;Townsend et al., 1993), the issue of a child's intellect is clearly emotionally charged for most parents and families. For many, this may be the first occasion on which they are confronted with such information, from which they often infer far-reaching implications. Without appropriate preparation, they may be profoundly disheartened and discouraged by this information.

IV.B. Reporting Test Results to Study Communities and the Public At-Large

Communities and the public at-large, like individual study participants, are interested in whether a study finds any real or potential health impact from the toxic exposures or hazardous chemicals of concern, and if such effects are found, what remedies are available and recommended. The investigators, again, must distinguish carefully between study findings that are statistically significant merely in a probabilistic mathematical sense, and those that have real meaning for the health of the community and the public. This is best accomplished at the time the study protocol is developed, when the investigators should consider *a priori* what effect sizes they will consider meaningful in this latter sense. Reporting of the findings should be confined to the data as it relates to the actual research questions for which the study was designed to address. Speculation regarding the impact of these findings without additional supportive data should be discouraged. With regard to any individual concerns, the investigators must be able to put the findings into perspective in that any test results detecting differences between groups do not necessarily indicate individual abnormalities. Any presentation of the findings should be forthright and disciplined in stating the implications for the individual study participants, the exposed community, and human health overall.

In many instances, a small effect size can be important on a population basis even when it is not large enough to be important for individuals. A frequently cited hypothetical example examines the implications of a 3-point decrease in IQ scores in a population. Although a 3-point decrease in an otherwise healthy individual is not likely to be noticeable, a 3-point decline shifting the IQ distribution curve for an entire population might result in thousands more people with impaired cognitive function, and thus thousands of additional people requiring

intervention services. The same 3-point shift might also result in thousands fewer people with exceptional cognitive ability, and perhaps most able to contribute exceptional leadership and vision to a community. Widely published population estimates of the economic and social impacts of lead poisoning are based on this hypothetical example. However, this hypothetical example assumes that both tails of the IQ distribution curve are unaltered in slope and contour by the 3-point shift. Actual effects in any real-life population would depend on a number of variable factors, such as the pre-morbid distribution of IQ scores, and the impairing effect, if any, of the neurotoxicant on that distribution curve, as well as the IQ-score shift itself.

The complex task of explaining fairly technical study results to communities, and helping people put those results in proper perspective, is aided by a number of published guidelines. Education of the health community in the interpretation of data obtained from epidemiological studies will serve substantially in the accurate communication to the general public as well as to the population under study. In many cases these are drawn from considerable experience in discussing the results of environmental health studies in communities located near hazardous waste sites. Effective principles of risk communication include early community involvement, clear identification of quantifiable hazards, and consideration of emotional and psychological components of risk perception (Amler and Tinker, 1992;Friedman et al., 1999;Brauer et al., 2004). Community members are often frustrated by the absence of clear *answers* to highly specific questions (e.g., "Do we have more children in special education because our water is polluted?), but when adequately informed will often respond constructively to recommended *solutions* (e.g., how to minimize exposure and thereby risk, how to detect clinical problems early and minimize their impact) (Amler and Falk, 2000). Training of investigators in effective methods of risk communication will be extremely helpful to this process.

IV.C. Reporting Test Results to the Scientific Community

Throughout report planning and generation, investigators must be disciplined and intellectually honest in asking themselves: how conclusive and how generalizable are the results, were the original study questions (and others) answered, and what are the implications for the individual study participants, the study community, the nation, and human health overall. Discipline is especially critical in restricting reporting to the actual research questions, how they were tested, and what the study found. Broader speculation should be avoided or at least kept to a minimum and clearly labeled as such. There is a parallel requirement for clinicians and policymakers to learn how to read scientific reports and interpret study findings and conclusions. The appropriate role of the scientist as a uniquely informed advocate for social change is a far more complex issue and beyond the scope of this Panel. Nevertheless, all investigators have an ethical duty to report negative as well as positive findings, and should take great care to neither minimize or exaggerate the findings.

V. Additional Discussion Points

The Executive Summary from this Workshop condenses the discussions of the Panel into 17 specific critical points. In addition, a number of discussion points were broached by the Panel regarding optimizing the design and interpretation of epidemiologic studies for assessing neurodevelopmental effects from *in utero* chemical exposure. These included the importance of understanding and integrating results from previous studies into the design of new neurodevelopmental epidemiology studies. The ability to compare results across studies has been severely limited in the past. This is not due to any specific error in design but rather only in differences in design that prevent direct comparisons. The Panel stressed the importance of developing and maintaining the ability to perform interstudy comparisons. While there are "optimal" study designs, they are often not feasible to conduct and there is a need for both small and moderate size studies to focus on a specific question in depth. This can serve to

provide the foundation for large-scale study designs that can both confirm specific hypotheses and generate new ones. The Panel discussed the importance of a study design that accounts for enrichment effects, including concept of resilience.

With respect to the assessment of exposure and development of the nervous system, the concepts of critical windows of development and of effect come into play. While identifying critical windows would serve to significantly focus such studies, this would require previous knowledge concerning effects occurring at a specific stage of fetal development. The Panel emphasized that unless a critical gestational exposure window can be identified, exposure assessment over the entire period of gestation should be considered. In addition to providing an accurate measure of prenatal exposure, longitudinal measurements may allow identification of an especially sensitive period during pregnancy, particularly if sufficient variability is present.

In addition, the Panel stressed that practical considerations should be considered that affect participant burden, such as the length of time required to administer tests and in what setting (home or clinical setting) any testing may occur. All of these factors can influence the cost of a comprehensive evaluation of neurobehavioral function, but the hypothesis being tested should be the overarching determinant of what, when, and how the metrics are employed. Consequently, the real questions apply to the ultimate purpose of the measurement, when measurement should occur, and whether the measurement can be performed reliably. Further, as noted earlier, there is a need for the assessment of multiple neurobehavioral processes including measures of non-cognitive competence. For future studies, the Panel emphasized the inclusion of all information possible into the selection of outcome measures for example assessment of specific cognitive functions related to the development of competence in young children. Theoretical brain-based models can be used in the development of such assessment. Most importantly, the forward momentum for the field can be enhanced by the establishment of a system for neurodevelopmental surveillance. This would allow for tracking the outcomes from *in utero* exposure across early developmental time periods to determine whether central nervous system injuries may be lying silent until challenged. Such efforts will significantly improve the ability to identify those exposures that present a health risk to the fetus.

Other concepts and new directions discussed by the Panel included the examination of specific, *a priori* defined susceptible populations and the use of high risk populations. The expanded use of medical records, when possible, to confirm diagnoses, and the concept of early detection and screening procedures, and the use of dosimetry technology were also discussed. The Panel concluded by formulating the 17 points listed in the Executive Summary to continue to stimulate scientific discussions on contemporary issues with respect to this important area of inquiry.

References

Amin-Zaki L, Elhassani S, Majeed MA, Clarkson TW, Doherty RA, Greenwood M. Intra-uterine methylmercury poisoning in Iraq. Pediatrics 1974;54:587–595. [PubMed: 4480317]

Amler, RW.; Gibertini, M. Pediatric Environmental Neurobehavioral Test Battery. Atlanta: US Department of Health and Human Services, Agency for Toxic Substances and Disease Registry; Sep. 1996

Amler, RW.; Tinker, T. ATSDR—ASTHO National Health Risk Communications Training Program for State Health Agency Personnel. Atlanta: US Department of Health and Human Services, Public Health Service, Agency for Toxic Substances and Disease Registry; Dec. 1992

Amler RW, Falk HF. Opportunities and challenges in community environmental health evaluations. Environ Epidemiol Toxicol 2000;2:51–55.

Anger WK. Neurobehavioural tests and systems to assess neurotoxic exposures in the workplace and community. Occup Environ Med 2003;60:531–538. [PubMed: 12819291]

Anger WK, Letz R, Chrislip DW, Frumkin H, Hudnell K, Russo JM, Chappell W, Hutchinson L. Neurobehavioral test methods for environmental health studies of adults. Neurotoxicol Teratol 1994;18:489–497. [PubMed: 7845332]

Anger WK, Rohlman DS, Sizemore OJ, Kovera CA, Gibertini M, Ger J. Human behavioral assessment in neurotoxicology: producing appropriate test performance with written and shaping instructions. Neurotoxicol Teratol 1996;18:71–79.

Barr DB, Wang RY, Needham LL. Biologic monitoring of exposure to environmental chemicals throughout the life stages: requirements and issues for consideration for the National Children's Study. Environ Health Perspect 2005;113:1083–1091. [PubMed: 16079083]

Bates, J. Adjustment style in childhood as a product of parenting and temperament. In: Wachs, TD.; Kohnstamm, G., editors. Temperament in Context. Mahwah, NJ: Erlbaum Associates; 2001. p. 173-200.

Bearer CF, Jacobson JL, Jacobson SW, Barr D, Croxford J, Molteno CD, Viljoen DL, Marais AS, Chiodo LM, Cwik AS. Validation of a new biomarker of fetal exposure to alcohol. J Pediatr 2003;143:463–469. [PubMed: 14571221]

Bearer CF. Meconium as a biological marker of prenatal exposure. Ambul Pediatr 2003;3:40–3. [PubMed: 12540253]

Bellinger D. Interpreting the literature on lead and child development. Neurotoxicol Teratol 1995;17:201–212. [PubMed: 7542726]

Bellinger D. Effect modification in epidemiologic studies of low-level neurotoxicant exposures and health outcomes. Neurotoxicol Teratol 2000;22:133–140. [PubMed: 10642122]

Bendersky, M.; Alessandri, S.; Lewis, M. Emotions in cocaine-exposed infants. In: Lewis, M.; Sullivan, M., editors. Emotional Development in Atypical Children. Mahwah, NJ: Erlbaum Associates; 1996. p. 89-108.

Brauer M, Bert Hakkinen PJ, Gehan BM, Shirname-More L. Communicating exposure and health effects results to study subjects, the community and the public: strategies and challenges. J Expo Anal Environ Epidemiol 2004;14(7):479–483. [PubMed: 15292907]

Brogan DJ, Denniston MM, Liff JM, Flagg EW, Coates RJ, Brinton LA. Comparison of telephone sampling and area sampling: response rates and within-household coverage. Am J Epidemiol 2001;153:1119–1127. [PubMed: 11390332]

Budtz-Jørgensen E, Grandjean P, Jørgensen PJ, Weihe P, Keiding N. Association between mercury concentrations in blood and hair on methymercury-exposed subjects at different ages. Environ Res 2004;95:385–393. [PubMed: 15220072]

Burnette, M.; Ebbert, R.; Hawver-Edwards, C.; Soma, S.; Canfield, R. Pediatric lead exposure is associated with poor impulse control during free play. Paper presented to the American Psychological Society; Denver, Colorado. 1999.

Burse VW, Najam AR, Williams CC, Korver MP, Smith BF Jr, Sam PM, Young SL, Needham LL. Utilization of umbilical cords to assess *in utero* exposure to persistent pesticides and polychlorinated biphenyls. J Exp Anal Environ Epi 2000;10:776–778.

Caetano R, Ramisetty-Mikler S, McGrath C. Characteristics of non-respondents in a US national longitudinal survey on drinking and intimate partner violence. Addiction 2003;98:791–797. [PubMed: 12780367]

Chasnoff I, Burns A, Burns W. Cocaine use in pregnancy: Perinatal morbidity and mortality. Neurotoxicol Teratol 1987;9:291–293. [PubMed: 3683346]

Colombo, J. Infant cognition: Predicting later intellectual functioning. Thousand Oaks, CA: Sage Publications, Inc; 1993.

Cory-Slechta, D. Bridging experimental animal and human behavioral toxicology studies. In: Russell, R.; Flattau, P.; Pope, A., editors. Behavioral Measures of Neurotoxicity: Report of a Symposium. Washington, DC: National Academy Press; 1990. p. 137-158.

Cory-Slechta D. Studying toxicants as single chemicals: Does this strategy adequately identify neurotoxic risk? NeuroToxicology 2005;26:491–510. [PubMed: 16112317]

Davidson P, Myers G, Shamlaye C, Cox C, Wilding G. Prenatal exposure to methylmercury and child development: influence of social factors. Neurotoxicol Teratol 2004;26:553–559. [PubMed: 15203177]

Dougherty T, Haith M. Infant expectations and reaction time as predictors of childhood speed of processing and IQ. Dev Psychol 1997;33:146–155. [PubMed: 9050399]

Evans, G.; Lepore, S. Moderating and mediating processes in environment-behavior research. In: Moore, G.; Marans, R., editors. Advances in Environment, Behavior and Design. 4. New York: Plenum; 1997. p. 255-285.

Fagen, J.; Ohr, P. Learning and memory in infancy: Habituation, instrumental conditioning, and expectancy formation. In: Singer, L.; Zeskind, PS., editors. Biobehavioral Assessment of the Infant. New York: Guilford Press; 2001. p. 233-273.

Fried, P. Postnatal consequences of maternal marijuana use in humans. In: Hutchings, D., editor. Prenatal Abuse of Licit and Illicit Drugs. 562. Annals of the New York Academy of Sciences New York: Academy of Sciences; 1989. p. 123-132.

Friedman, SM.; Dunwoody, S.; Rogers, CL. Communicating uncertainty: Media coverage of New and Controversial Science. Lawrence Erlbaum Associates; Mahwah, NJ: 1999.

Funkhouser E, Macaluso M, Wang X. Alternative strategies for selecting population controls: comparison of random digit dialing and targeted telephone calls. Ann Epidemiol 2000;10:59–67. [PubMed: 10658689]

Gordis, L. Epidemiology. Philadelphia, PA: WB Saunders Company; 2000. p. 63-81.p. 204-219.

Grandjean P, Weihe P. Neurobehavioral effects of intrauterine mercury exposure: potential sources of bias. Environ Res 1993;61:176–183. [PubMed: 8472672]

Grandjean P, Weihe P, White RF, Debes F, Araki S, Yokoyama K, Murata K, Sorensen N, Dahl R, Jorgensen PJ. Cognitive deficit in 7-year-old children with prenatal exposure to methylmercury. Neurotoxicol Teratol 1997;19:417–428. [PubMed: 9392777]

Greenland S. Multiple-bias modeling for analysis of observational data. J R Statist Soc A 2005;168:267–306.

Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet 2002;359(9302):248–252. [PubMed: 11812579]

Hanninen L. Psychological tests in the diagnosis of carbon disulfide poisoning. Work Environ Health 1966;2:16–20.

Hartman, D. Assessing growth in longitudinal investigations: Selected Measurement and Design Issues. In: Teti, D., editor. Handbook of Research Methods in Developmental Sciences. Malden, MA: Blackwell; 2005. p. 319-339.

Hastings R. On "good" terms: labeling people with mental retardation. Mental Retardation 1994;32:363–365. [PubMed: 7984122]

Hernan M, Hernandez-Diaz S, Robins J. A structural approach to selection bias. Epidemiology 2004;15:615–625. [PubMed: 15308962]

Hubbs-Tait L, Nation J, Krebs N, Bellinger D. Neurotoxicants, micronutrients and social environments: Individual and combined effects on children's development. Psychol Sci Public Interest 2005;6:57–121.

Institute of Medicine. Infant Formula: Evaluating the Safety of New Ingredients. Washington DC: National Academies Press; 2004.

Iregren A, Letz R. Computerized testing in neurobehavioral toxicology. Arbete och Halsa 1991;35:29–34.

Jacobson J, Jacobson S. Breastfeeding and gender as moderators of teratogenic effects on cognitive development. Neurotoxicol Teratol 2002;24:349–358. [PubMed: 12009490]

Jellison J, Duke R. The mental retardation label: music teachers' and prospective teachers' expectations for children's social and music behaviors. J Music Therapy 1994;31:166–85.

Kalsbeek W, Heiss G. Building bridges between populations and samples in epidemiological studies. Annu Rev Publ Health 2000;21:147–169.

Kochanska G, Murray K, Harlan E. Effortful control in early childhood: Continuity and change, antecedents, and implications for social development. Dev Psychol 2000;36:220–2232. [PubMed: 10749079]

Lauria MR, Qureshi F, Jacques SM, Kurtzhals P, Ostrea EM, Gonik B, Sorokin Y. Meconium drug screening of stillborn infants: a feasibility study. Fetal Diagn Ther 1997:248–51. [PubMed: 9354887]

Lemeshow, S.; Stroh, G. Sampling techniques for evaluating health parameters in developing countries. Washington, DC: National Academy Press; 1988. p. 1-5.

Lester, B.; Freier, K.; LaGasse, L. Prenatal cocaine exposure and child outcome: What do we really know?. In: Lewis, M.; Bendersky, M., editors. Mothers, Babies, and Cocaine: The Role of Toxins in Development. Hillsdale, NJ: Erlbaum Associates; 1995. p. 19-39.

Little RJ, Vartivarian S. On weighting the rates in non-response weights. Stat Med 2003;22:1589–1599. [PubMed: 12704617]

Lorenzo AMJ, Bermejo BA, Cocho de Juan JA, Fraga BJM, Bermejo BP. Selenium levels in related biological samples: human placenta, maternal and umbilical cord blood, hair and nails. J Trace Elem Med Biol 2005;19:49–54. [PubMed: 16240672]

MacKinnon D, Lockwood C, Hoffman J, West S, Sheete V. A comparison of methods to test mediation and other intervening variable effects. Psychol Methods 2002;7:83–104. [PubMed: 11928892]

Maclure M, Schneeweiss S. Causation of bias: the episcope. Epidemiol 2001;121:14–22.

Masten A, Coatsworth J. The development of competence in favorable and unfavorable environments. Am Psychol 1998;53:205–220. [PubMed: 9491748]

Masten, A.; Powell, J. A resilience framework for research, policy, and practice. In: Luthar, S., editor. Resilience and Vulnerability: Adaptation in the Context of Childhood Adversities. New York: Cambridge University Press; 2003. p. 1-25.

Mayes L. A behavioral teratogenic model of the impact of prenatal cocaine exposure on arousal regulatory systems. Neurotoxicol Teratol 2002;24:385–95. [PubMed: 12009493]

Mayes, L.; Bornstein, M. Developmental dilemmas for cocaine-abusing parents and their children. In: Lewis, M.; Bendersky, M., editors. Mothers, Babies, and Cocaine: The Role of Toxins in Development. Hillsdale, NJ: Erlbaum Associates; 1995. p. 251-72.

Mendelsohn A, Dreyer B, Fierman A, Rosen C, Legano L, Kruger H, Lim S, Courtlandt C. Low level lead exposure and behavior in early childhood. Pediatrics 1998;101:464–465.

Montes A, Walden CE, Knopp RH, Cheung M, Chapman MB, Albers JJ. Physiologic and supraphysiologic increases in lipoprotein lipids and apoproteins in late pregnancy and postpartum. Possible markers for the diagnosis of "prelipemia". Arteriosclerosis 1984;4:407–417. [PubMed: 6431954]

O'Connor M, Sigman M, Kasari C. Interactional model for the association among maternal alcohol use, mother-infant interaction and infant cognitive development. Infant Behav Dev 1993;16:177–192.

Ostrea EM Jr, Knapp DK, Romero A, Montes M, Ostrea AR. Meconium analysis to assess fetal exposure to nicotine by active and passive maternal smoking. J Pediatr 1994;124:471–6. [PubMed: 8120724]

Ostrea EM Jr, Knapp DK, Tannenbaum L, Ostrea AR, Romero A, Salari V, Ager J. Estimates of illicit drug use during pregnancy by maternal interview, hair analysis, and meconium analysis. J Pediatr 2001;138:344–348. [PubMed: 11241040]

Otto DA, Skalik I, House DE, Hudnell HK. Neurobehavioral evaluation system (NES): comparative performance of second-, fourth-, and eighth-grade Czech children. Neurotoxicol Teratol 1996;18:421–428. [PubMed: 8866533]

Paule MG. Use of the NCTR Operant Test Battery in nonhuman primates. Neurotoxicol Teratol 1990;12:413–418. [PubMed: 2247027]

Paule MG. Chronic drug exposures during development in nonhuman primates: Models of brain dysfunction in humans. Frontiers Bioscience 2005;10:2240–2249.

Paule MG, Chelonic JJ, Buffalo EA, Blake DJ, Casey PH. Operant test battery performance in children: correlation with IQ. Neurotoxicol Teratol 1999;21:223–230. [PubMed: 10386825]

Proctor SP, Letz R, White RG. Validity of a computer-assisted neurobehavioral test battery in toxicant encephalopathy. Neurotoxicol 2000;21:703–714.

Ramirez GB, Cruz MCV, Pagulayan O, Ostrea E, Dalisay C. The Tagum study I: Analysis and clinical correlates of mercury in maternal and cord blood, breast milk, meconium, and infants' hair. Pediatrics 2000;106:774–781. [PubMed: 11015522]

Rauh V, Whyatt R, Garfinkel R, Andrews H, Hoepner L, Reyes A, Diaz D, Camann D, Perera F. Developmental effects of exposure to environmental tobacco smoke and material hardship among inner-city children. Neurotoxicol Teratol 2004;26:373–385. [PubMed: 15113599]

Rice D. Assessing the effects of environmental toxicant exposure in developmental epidemiological studies: Issues for risk assessment. NeuroToxicology 2005;26:483–489. [PubMed: 16112316]

Rice D, Barone S Jr. Critical periods of vulnerability for the developing nervous system: Evidence from humans and animal models. Environ Health Persp 2000;108(suppl 3):511–533.

Richter LM, Norris SA, De Wet T. Transition from Birth to Ten to Birth to Twenty: the South African cohort reaches 13 years of age. Paediatr Perinat Epidemiol 2004;18:290–301. [PubMed: 15255883]

Rohlman DS, Bailey SR, Anger WK, McCauley L. Assessment of neurobehavioral function with computerized tests in a population of Hispanic adolescents working in agriculture. Environ Res 2001;85:14–24. [PubMed: 11161647]

Rothbart, M.; Bates, J. Temperament. In: Eisenberg, N., editor. Handbook of child psychology. 6. New York: Wiley; In press

Rothbart, M.; Chew, K.; Gartstein, M. Assessment of temperament in early development. In: Singer, L.; Zeskind, PS., editors. Biobehavioral Assessment of the Infant. New York: Guilford Press; 2001. p. 190-208.

Rothman, KJ.; Greenland, S., editors. Modern Epidemiology. 2. Philadelphia: Lippincott-Raven Publishers; 1998.

Rovee-Collier, C.; Barr, R. Infant learning and memory. In: Bremner, G.; Fogel, A., editors. Blackwell Handbook of Infant Development. Malden, MA: Blackwell Publishing; 2001.

Ruff, H.; Rothbart, M. Attention in early development: Themes and Variations. New York: Oxford University Press; 1996.

Ruff H. Population-based data and the development of individual children: The case of low to moderate lead levels and intelligence. Dev Behavioral Pediatrics 1999;20:42–49.

Schneider, W. Research on memory development: Past and present. In: Hartup, W.; Silbereisen, R., editors. Growing Points in Developmental Science. New York: Psychology Press; 2002. p. 143-166.

Schwartz DA, Sassaman AP, Collman GW. Environmental health sciences and the community. Environ Health Persp 2006;114(2):A80.

Sparrow W, Shinkfield A, Karnilowicz W. Constraints on the participation of individuals with mental retardation in mainstream recreation. Mental Retardation 1993;31:403–11. [PubMed: 8152386]

Struthers J, Hansen R. Visual recognition memory in drug-exposed infants. J Dev Behav Pediatr 1992;13:108–111. [PubMed: 1577955]

Townsend M, Wilton K, Vakilrad T. Children's attitudes toward peers with intellectual disability. J Intellectual Disability Res 1993;37:405–11.

Vreugdenhil H, Lanting C, Mulder P, Boersma E, Weisglas-Kuperus N. Effects of prenatal PCB and dioxin background exposure on cognitive and motor abilities in Dutch children at school age. J Pediatr 2002;140:48–56. [PubMed: 11815763]

Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. II. Types of controls. Am J Epidemiol 1992;135:1029–1041. [PubMed: 1595689]

Wachs TD. The nature and nurture of child development. Food Nutrition Bull 1999;20:7–22.

Wachs, TD. Necessary but not sufficient: The respective roles of single and multiple influences on human development. Washington, D.C: American Psychological Association Press; 2000.

Wachs, TD.; Plomin, R. Conceptualization and measurement of organism environment interaction. Washington, D.C: American Psychological Association Press; 1991.

Ware NC, Lee VE. Sex differences in choice of college science majors. Am Educ Res J 1988;25:593–614.

Weiss B, Landrigan PJ. The developing brain and the environment: an introduction. Environ Health Persp 2000;108 (Suppl 3):373–4.

Weiss B. Workshop to identify critical windows of exposure for children's Health. Environ Health Persp 2000;108 (Supp 3):375–381.

Werner, E. Protective factors in individual resilience. In: Meisels, S.; Shonkoff, J., editors. Handbook of Early Childhood Intervention. Cambridge: Cambridge University Press; 1990. p. 97-116.

Whyatt RM, Barr DB. Measurement of organophosphate metabolites in postpartum meconium as a potential biomarker of prenatal exposure: a validation study. Environ Health Persp 2001;109:417–420.

Whyatt RM, Rauh V, Barr DB, Camann DE, Andrews HF, Garfinkel R, Hoepner LA, Diaz D, Dietrich J, Reyes A, Tang D, Kinney PL, Perera FP. Potential insecticide exposures, birth weigh and length among an urban minority cohort. Environ Health Persp 2004;112:1125–1132.

Yates, TM.; Egeland, B.; Sroufe, LA. Rethinking resilience: A developmental process perspective. In: Luthar, S., editor. Resilience and Vulnerability: Adaptation in the Context of Childhood Adversities. New York: Cambridge University Press; 2003. p. 243-266.

Zeitz P, Kakolewski K, Imtiaz R, Kaye W. Methods of assessing neurobehavioral development in children exposed to methyl parathion in Mississippi and Ohio. Environ Health Persp 2002;110(Suppl 6):1079–1083.

Table 1

Topics and Associated Questions for the Expert Panel.

Panel Topics and Questions

Study Designs

Q1: What are the components of optimal experimental design and methodologies to assess the Likelihood that *in utero* exposure to an environmental chemical can result in adverse neurodevelopmental effects in newborns that continue into childhood?

Q2: What are the statistical issues that must be addressed to conclude with adequate confidence that an *in utero* exposure to a specific environmental chemical can result in adverse neurodevelopmental effects?

Participant Selection

Q1: What is the most appropriate and valid way to select exposed and control groups for studies attempting to demonstrate an association between *in utero* exposure to a specific environmental chemical and adverse neurodevelopmental effects?

Exposure Assessment

Q1: Which specific measurement tools and biomarkers are best suited and validated for assessing the nature, extent, and patterns of *in utero* exposure to a particular environmental chemical? How might these differ from tools and biomarkers used to assess post-natal exposure?

Q2: To avoid exposure misclassification or misleading estimates when assessing potential exposure, how frequently should exposure be estimated (i.e., what temporal units should be used for serial exposure measurements)? If appropriate temporal units are chemical- or tissue- specific, what data or criteria should be used to determine the optimal units? How should critical timeframes – critical windows of vulnerability in neurodevelopment - be taken into account when designing an exposure assessment approach?

Q3: How should potential aggregate exposure from multiple routes (inhalation, ingestion, dermal) be addressed? How should potential cumulative exposure to multiple chemicals be addressed?

Cognitive and Non-Cognitive Measurement Tools

Q1: What specific measurement tools/tests are best suited and validated for assessing the variety of potential neurodevelopmental and behavioral deficits? What is the known sensitivity, specificity and predictive value of each endpoint being measured? How reproducible is/are the measurement(s)?

Q2: Are there particular sampling strategies or data collection methods that are especially relevant to detecting potential neurodevelopmental effects from *in utero* exposure? What sampling and analysis strategies can be employed to avoid Type II (failure to detect a real effect) errors?

Q3: What is the relationship between the estimated window of exposure and the *nature* of a potential effect, and how might this affect the selection of tests?

Q4: How might data and methods from the field of molecular epidemiology be used to enhance traditional epidemiologic approaches?

Statistical Issues and Confounders

Q1: What guidance can be offered with respect to selecting and measuring potential confounders? What criteria should be applied when selecting control variables for inclusion in a multivariate analysis? How should potential mediating factors be identified and analyzed?

Q2: Are there cultural aspects to neurodevelopmental tests that should be considered before use?

Reporting of Findings

Q1: How is clinical significance versus population significance defined and reported?

Q2: How do researchers address the issue of labeling of children based on study results and how should the study results be reported to parents?