TOXICOLOGICAL SCIENCES **58,** 222–234 (2000)

FORUM

Cognitive Tests: Interpretation for Neurotoxicity? (Workshop Summary)

William Slikker, Jr.,* Barbara D. Beck,†,[1] Deborah A. Cory-Slechta,‡ Merle G. Paule,§ W. Kent Anger,¶ and David Bellinger‖

National Center for Toxicological Research, Division of Neurotoxicology, 3900 NCTR Road, Jefferson, Arkansas 72079; †Gradient Corporation, 238 Main Street, Cambridge, Massachusetts 02142; ‡Department of Environmental Medicine, Box EHSC, University of Rochester Medical School, Rochester, New York 14642; §Behavioral Toxicology Laboratory, Division of Neurotoxicology, HFT-132, National Center for Toxicological Research, 3900 NCTR Road, Jefferson, Arkansas 72079–9502; ¶Center for Research on Occupational and Environmental Toxicology, Oregon Health Sciences University, Portland, Oregon 97201; and ‖Children's Hospital, Neuroepidemiology Unit, 300 Longwood Avenue, Carnegie 208, Boston, Massachusetts 02115

Received June 2, 2000; accepted September 5, 2000

The appropriate use and interpretation of cognitive tests presents important challenges to the toxicologist and to the risk assessor. For example, intelligence cannot be measured directly; rather intelligence is quantified indirectly by scoring responses (i.e., behaviors) to specific situations (problems). This workshop, "Cognitive Tests: Interpretation for Neurotoxicity?" provided an overview on the types of cognitive tests available and described approaches by which the validity of such tests can be assessed. Unlike many tools available to the toxicologist, cognitive tests have a particular advantage. Being noninvasive and species-neutral, the same test can be performed in different mammalian species. This enhances one's ability to assess the validity of test results. Criteria for test validity include comparable responses across species as well as similar disruption by the same neurotoxicant across species. Test batteries, such as the Operant Test Battery, have indicated remarkable similarity between monkeys and children with respect to performance of certain tasks involving, for example, short-term memory. Still, there is a need for caution in interpretation of such tests. In particular, cognitive tests, especially when performed in humans, are subject to confounding by a range of factors, including age, gender, and, in particular, education. Moreover, the ability of such tests to reflect intelligence must be considered. Certain aspects of intelligence, such as the ability to plan or carry out specific tasks, are not well reflected by many of the standard tests of cognition. Nonetheless, although still under development, cognitive tests do hold promise for reliably predicting neurotoxicity in humans.

Key Words: **cognitive tests; behavioral neurotoxicology; metals; solvents; neurobehavioral test battery.**

If the nervous system is the last scientific frontier, then cognitive function is the most distant outpost. Well-educated and competent chemists, pharmacologists, and even toxicologists, have inquired whether one can really measure chemical effects on cognitive function. This healthy skepticism is not all based on ignorance, but is fueled by the inability to determine the molecular weight of a memory trace or the molecular structure of a learning paradigm. How do we overcome this challenge that memory, learning, and other cognitive functions are not directly measurable in a quantitative fashion? In addition to mandating that experimental psychology be incorporated into every neuroscience curriculum, brain researchers must do a better job of describing their science in a logical and quantitative manner. Based on over 50 years of solid scientific investigations, cognitive researchers have established very reproducible, sensitive, and quantitative approaches to assess memory, learning, and attention functions in both animals and humans. Behavioral toxicologists must describe their behavioral paradigms in logical terms and their findings in plain language.

As with much of toxicology, the questions concerning extrapolation of findings from animals to humans are of paramount importance. In this area of species extrapolation, cognitive investigators have a real advantage because of the noninvasive and species-neutral nature of their tests. It is in this area of cross-species extrapolation that modern cognitive function research can make a lasting impact.

Although there has been progress in the application of species-neutral, cognitive function assessment tasks, several impediments have limited greater progress, including (1) the use of language-based (written or spoken) assessment tools for human subjects that are not applicable to animal assessments, (2) the expense and technical/computer skills necessary to conduct operant testing procedures, and (3) the diverse training history and discipline preservation of the researchers or physicians performing the cognitive function assessments. All of

[1] To whom correspondence should be addressed. Fax: (617) 395-5001. E-mail: bbeck@gradientcorp.com.

these impediments can be overcome, however, with enlightenment and a unified commitment.

Cognitive tests should provide for measurement of the total functional output of the nervous system, yet cognitive function cannot be directly observed. The intelligence quotient (IQ), one of the commonly used indices of cognitive abilities in humans, cannot be observed directly but is measured by scoring responses (behaviors) to specific situations (problems). The lack of direct measures makes interpretation of cognitive tests problematic. The extent to which IQ, for example, reflects factors other than intelligence, such as socioeconomic status, has been hotly debated; the difficulty in interpreting cognitive function tests in humans exposed to neurotoxicants is, thus, complicated by a range of factors, such as the adequacy of control of confounders.

Difficulty in test interpretation is also an issue when species other than humans are used for neurotoxicological assessment. To overcome these problems, sophisticated assessment tools have been developed for looking into selected aspects of complex brain function (cognition) and their alteration by toxicant exposure. Advances in the interpretation of animal cognitive function tests has resulted from data generated from the use of carefully designed operant and non-operant problem solving tasks, especially those that can be modeled in animals in exactly the same way as they are in humans. Examples of these tasks include delayed matching-to-sample (short-term memory); repeated acquisition (learning); temporal discrimination (timing ability); condition and position responding (color or position discrimination), and progressive ratio (motivation). With respect to causation, interpretation of cognitive test results in exposed humans is enhanced by consistency with results from animal species. Challenges still remain regarding interpretation of cognitive findings with respect to adverseness of effect. Although improvements are still underway, cognitive function tests, especially those that maintain continuity across species, are quantifiable and can be automated, and these hold promise for reliably predicting neurotoxicity. The aim of this workshop, which was held at the 1999 Society of Toxicology meeting in New Orleans, was to discuss: the range of cognitive tests available; their use in different species, including humans, for predicting neurotoxicity; and the appropriate interpretation of such tests with respect to overall function and general toxicity.

Assessment of Complex Cognitive Function in Rodents and Extrapolation across Species (Deborah A. Cory-Slechta)

In the realm of toxicology, measures of behavioral function are critical for assessment of the neurobehavioral effects of toxic compounds, for the elaboration of mechanisms of action of exogenous agents, and for defining the risks associated with such exposures. Often, these behavioral assessments are derived from experimental animal studies and their results extrapolated to human populations. In other cases, it may be possible to include both human and nonhuman subjects in the assessment process.

Comparison of Methodologies for Animal versus Human Neurobehavioral Testing

Typically, animal studies utilize experimental paradigms designed to evaluate explicit behavioral domains, and are typically based on operant conditioning methods. In contrast, behavioral evaluations carried out in human populations have relied on standardized test instruments (pencil and paper or computer-based), even though the same experimental methods can be used to test people. This dichotomy of approach has risen in part because of the differential training associated with human vs. experimental animal psychology (clinical and experimental psychology, respectively). It has resulted in the evolution of two parallel paths of research, which unfortunately have only infrequent intersections, somewhat different theoretical formulations of the research, and publication of findings in different scientific journals.

This reliance on different instruments for human vs. experimental animal studies complicates the ability to extrapolate across species for several reasons. First, it can be difficult to compare behavioral functions across these different instruments, since they may include measurement of similar but also of dissimilar behavioral functions. Whereas an operant learning paradigm can be devised to explicitly differentiate learning processes from other behavioral domains, an IQ test is a much more global instrument measuring multiple behavioral capabilities while purporting to provide an indication of native intellect. Further, the behavioral functions that standardized tests are said to measure are often not sufficiently defined, operationally, to permit a generalized and universally held understanding of the specific behavioral processes involved. An additional complication is that the different measuring instruments, as typically used in human vs. experimental animal studies, may involve very different limits and levels of sensitivity, and these limits may be poorly defined.

Alternative Approach for Cross-Species Comparisons

An alternative approach, the utilization of the same experimental paradigms in humans and experimental animals, offers several key advantages to furthering the goals defined above for behavioral toxicity and risk assessment. First, it can minimize the obscurities associated with attempts to equate the nature of the behavioral deficits in clinical vs. experimental methods, since the behaviors assessed and the outcome measures would be identical. The utilization of experimental behavioral methods designed to evaluate specific behavioral domains, moreover, would permit, in the case of toxic exposure effects, a more precise delineation of the behavioral deficits and thus provide guidance both to the underlying neurobiological mechanisms and the potential therapeutic strategies.

Such an approach would also permit the comparison of effect levels of a toxicant across species. In so doing, a direct evaluation of the need for specific safety factors in risk assessment can be examined and/or altered as needed. Correspondingly, comparisons of actual exposure-effect data allow determinations of comparable levels of neurotoxicity across species (e.g., comparative ED10 values) and thus provide metrics with respect to differences in species sensitivity (Benignus *et al.*, 1998). A similarity of behavioral performances when the same paradigms are used across species also indicates the phylogenetic continuity of the behavioral domains being assessed, and such continuity further validates experimental animal models and cross-species extrapolations. One impediment, voiced by some, to the strategy of using common behavioral methods in human and experimental animal studies is the lack of a "normative" database for many of the experimental methods used, which may leave residual questions about what constitutes normal impaired performance. This is a potential problem but one that can be systematically remedied. A continuity of behavioral processes further validates experimental animal models and cross-species extrapolations.

A strategy embodying these approaches in the area of neurodegenerative diseases exemplifies their potential utility. First, behavioral deficits can be characterized in human populations with known neurodegenerative disease states using explicit experimentally based methods. Successful reproduction of this pattern of behavioral deficits, using the same behavioral methods in experimental animal models as might be achieved using various tools (e.g., lesions, microinjections, in vivo neurochemistry, gene transfer, or deletion) becomes the next goal, since it defines the underlying neurobiological substrates of the disease. Further, it provides an experimental model for the development of potential therapeutic strategies and assessment of their efficacy. Successful therapeutic outcomes could then be taken back to the affected human population.

Future Directions

Similar benefits could be envisaged to advance neurotoxicology research and risk assessment. Studies of neurotoxicant effects in human populations frequently suffer from the absence of any specific predictions about expected behavioral impairments, and they seldom include control procedures for "false positive" behavioral deficits. Experimental animal studies could be utilized to precisely define the patterns of behavioral deficits expected from a neurotoxicant and provide the bases for such hypotheses, both for domains expected to be sensitive and those that would not be impacted by this neurotoxicant. It could also permit a comparison of exposure effect levels. Additionally, data bases from experimental animal studies may make it possible to differentiate which behavioral deficits arise from which specific neurotoxicant exposures in the case of multiple or mixed human exposure scenarios.

At least two criteria should be met for the utilization of the same experimental behavioral approaches across species to prove beneficial. The first is that there should be continuity of behavioral performance across species. Secondly, these behavioral performances should be sensitive to disruption by pharmacological, toxicological, environmental, or other neurobiological risk factors. Indeed, for various measures of cognitive function, such criteria are clearly satisfied. Some of these experimental methods actually originated in human subjects and have been increasingly applied and adapted for experimental animal studies, attesting to the feasibility of this approach.

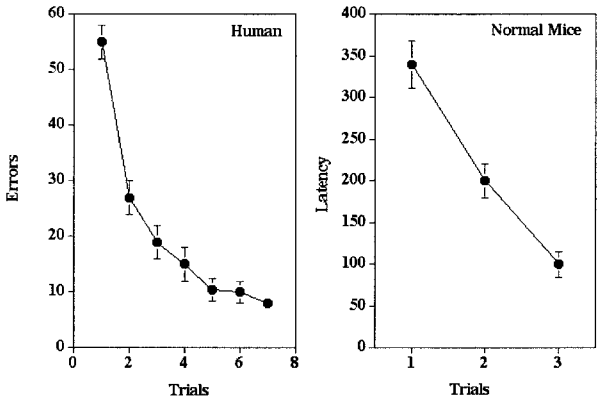

FIG. 1. Decrease in errors across trials, indicative of learning in a repeated-acquisition paradigm in normal human subjects (left column; modified from Kelly *et al.*, 1977) and decreases in latencies to complete the response sequence in normal C57B1 mice (right column; unpublished data from Brooks *et al.*, 2000). Data points represent group mean ± SE values.

One such example is the multiple schedule of repeated learning and performance (Cohn *et al.*, 1993, 1996; Cohn and Paule, 1995; Cory-Slechta, 1994). This behavioral paradigm, first utilized by Boren in humans (Boren, 1963; Boren and Devine, 1968), provides for a separate determination of learning and rote performance of an already learned sequence of responses in the same subject in the same test session. This is achieved by using separate components of the session for the learning vs. performance baselines, with different environmental stimuli signaling to the subject which component is operative. The same baseline has been utilized in several species, including nonhuman primates and rats (Kelly *et al.*, 1997; Moerschbaecher *et al.*, 1985; Thompson, 1977, 1980). More recently, the technique has been adapted for use in normal and genetically engineered mice (Brooks *et al.*, 2000). For all of these species, acquisition of a new response chain, i.e., learning, can be documented within a session, as evidenced by an increase in accuracy, or as a decrease in errors, or in latencies to complete the response chain across the session. Figure 1 depicts such comparable functions in normal humans (Kelly *et al.*, 1997) and genetically-engineered mice (Brooks *et al.*, 2000). In such studies, accuracy remains high in the performance component both within and across sessions, as expected, given that it is an already acquired response chain. Furthermore, as documented in these same studies, the para-

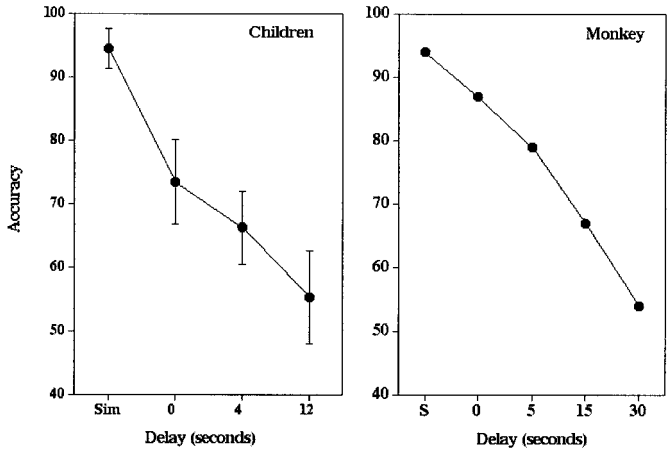

FIG. 2. Decreases in accuracy with increasing delay values in a delayed matching-to-sample paradigm in normal children (left column; 6–12 years-of-age; unpublished data from Brockel and Cory-Slechta) and in normal monkeys (right column; modified from Parker *et al.*, 1997). Data points represent group mean ± SE values.

digm has been found to be sensitive to disruption by pharmacologic agents, lesions, and genetic engineering (Brooks *et al.*, 2000; Cohn and Cory-Slechta, 1993, 1994a,b; Cohn *et al.*, 1993; Kelly *et al.*, 1997).

Similarly, delayed matching-to-sample is a behavioral paradigm that has been used to evaluate memory (remembering) that evidences cross-species continuity as well as sensitivity to behavioral disruption (Paule *et al.*, 1998). The paradigm requires the subject to remember which of two stimuli was presented prior to the beginning of a delay period and to respond correctly to it following the delay. Typically a range of delay values is presented within a session, generating accuracy by delay function for each session. A zero-second delay is used to control for non-mnemonic effects. Typically, accuracy declines as the delay period is lengthened, because it is harder to "remember" the stimulus that was presented before the delay. Such a response pattern has been observed across species, including humans, nonhuman primates (see Fig. 2), and rodents, and has also been shown to be sensitive to various pharmacological compounds as well as to lesion effects (Fray and Robbins, 1996; Heyser *et al.*, 1993; Joel *et al.*, 1997; Penetar and McDonough, 1983).

Another important component of cognitive function is attention. Attention is a global behavioral construct that encompasses numerous behavioral responses. Clinically, attention deficit disorder is diagnosed using three such classes: hyperactivity, impulsivity (self control), and inattention. The ability to sustain attention has received considerable attention and has been examined using sustained attention or vigilance paradigms that require the subject to report target stimulus presentations that occur at unpredictable intervals. Accuracy of reporting is a function of the delay between such presentations, the salience of the stimuli, and other parameters. A continuity in vigilance or sustained attention performance is also seen across species, including human, nonhuman primates, and rodents (Bushnell, 1998). The ability to sustain attention, moreover, can be disrupted across species by various techniques, including psychopharmacological agents, lesions, and exposures to neurotoxic compounds (Brockel and Cory-Slechta, 1999; Bushnell *et al.*, 1994; Callahan *et al.*, 1993; Dick *et al.*, 1992; Kelly *et al.*, 1991; Lane and Phillips-Bute, 1998; Rueckert and Grafman, 1998; Walkowiak *et al.*, 1998).

In summary, the approach of utilizing the same experimental behavioral paradigms across species is clearly feasible. It may offer the potential to significantly accelerate our understanding of the neurotoxicity of chemical agents and thereby define their associated mechanisms of action, devise therapeutic strategies where appropriate, and advance the understanding of the risks they pose to human health.

Assessment of Complex Brain Function in Both Nonhuman Primates and Humans Using Identical Behavioral Tasks (Merle G. Paule)

The use of animals as surrogates for predicting the toxic potential of chemicals in humans is critical. When the toxicity of interest relates to the functional capabilities of the brain, it is important to appropriately model aspects of brain function in animals so that measures relevant to humans can be obtained. This need has led to the identification and application of a variety of relatively complex behavioral tasks for use in both nonhuman primates and humans. Maintenance of task continuity across species allows for direct observations of interspecies similarities and differences in a variety of functional domains and should assist in the extrapolation of data from lab animals to humans.

The Operant Test Battery

The National Center for Toxicological Research (NCTR) Operant Test Battery (OTB) represents one of the first instruments that has undergone relatively extensive testing to determine the relevance and utility of the measures obtainable by its use. The functions modeled by the tasks that comprise the OTB include motivation, color and position discrimination, time estimation, short-term memory, and learning. All tasks are administered using a behavioral or intelligence "panel" on which a variety of levers, press-plates, and visual stimuli are located (see Paule *et al.*, 1988 and Schulze *et al.*, 1988 for details). Both monkeys and humans (children) are assessed using identical apparatus with monkey subjects "working" for banana-flavored food pellets and children "working" for nickels. Additionally, each task is presented for a specific amount of time, allowing for determination of a percent task completed measure that is useful in comparing treatment effects across tasks.

For the motivation task, subjects must operate (hence the term "operant") a response lever in a progressive fashion in order to continue to receive reinforcers: the first reinforcer

might, for example, "cost" 2 lever presses, the second, 4 lever presses, the third, 6, etc. In this way, the work required for subsequent reinforcers is escalated throughout a session and metrics of a subject's motivation to work for their reinforcers are obtained. These include response rate and number of responses made for the last reinforcer earned. Data from monkeys indicate that individual subjects are, from session to session, quite consistent in their apparent level of motivation.

For the color and position discrimination task, 3 press-plates, aligned horizontally, are used. Initially, the center plate is illuminated with one of 4 colors (red, yellow, blue, or green). Subjects respond to this plate (push it) to continue the trial. Immediately after the colored plate is operated, it is extinguished and the 2 side plates are illuminated white. If the color had been red or yellow, a response on the left plate nets a reinforcer. If it had been blue or green, then a response to the right plate would have resulted in reinforcer delivery. Latencies to respond to the colored plates and the choice position plates, as well as accuracy of choice responding are obtained.

For the time estimation task, subjects must depress and hold a response lever for at least 10 s, but no more than 14 s, in order to obtain a reinforcer. Releasing the lever too early or too late has no programmed consequences, but subjects may initiate another trial immediately. Data from this task are often characterized by lever hold durations that form a Gaussian distribution with the mean lever hold times occurring within the 10–14-s window. Mean lever hold times are thought to indicate a subject's timing accuracy, whereas the spread (standard deviation) of the response population is thought to indicate timing precision (Paule *et al.*, 1999b).

For the short-term memory task, a white-on-black geometric symbol (square, circle, triangle, etc.) illuminates the center of 3 press-plates. Subjects are required to press this plate to continue the trial, after which it is immediately extinguished and one of 6 randomly chosen delays begins. When the delay times out, all 3 plates are illuminated, each with a different geometric symbol, one of which matches the initial "sample" symbol. A choice response to the matching symbol results in reinforcer delivery. By determining accuracy of matching for a variety of delays, it is thought that metrics of memory decay or forgetting can be obtained. Response accuracy at very short delays is thought to reflect processes closely associated with the discrimination and encoding of visual information, whereas accuracy at longer delays is thought to be more relevant to the rate of decay of memory or to information retrieval (Paule *et al.*, 1998).

For the learning task, subjects are presented with 4 response levers. Initially, a simple one-lever task is required wherein subjects must determine which one of the 4 levers is the correct one for that level of task difficulty. Once the one-lever sequence has been mastered, the response requirements are incremented so that an additional lever must be pressed before pressing the previously learned lever. In this way, the length of the required response sequence is incremented as subjects demonstrate mastery of the shorter response chains, with the ultimate goal of presenting subjects with 6 lever response sequences.

Each of these tasks has face validity in that the rules associated with the correct performance of each task appear, at face value, to result in the production of behaviors that one would think appropriate for the function being modeled. In addition, each can be said to have content validity, since these procedures have been widely accepted by experts as reasonable instruments for the assessment of the specific functional domains mentioned earlier. Discriminant validity (a type of construct validity) can be demonstrated when the performance of a given task is not highly correlated with the performance of tasks that are supposed to measure other functions or constructs. It has been demonstrated, using monkey data, that performance in the NCTR OTB tasks are not highly correlated with each other and, in some cases, not at all (see Paule, 2000). Thus, discriminant validity has been addressed, at least to a limited degree.

The relevance of OTB measures to more traditional measures of human brain function has been addressed in studies wherein OTB measures obtained from children were correlated with measure of full scale, verbal, and performance IQ in the same subjects (see Paule *et al.*, 1999a). The findings clearly show that the performance of several OTB tasks (e.g., color and position discrimination, time perception, short-term memory, and learning,) is significantly correlated with IQ, even though the correlation coefficients are not large. Thus, OTB behaviors provide data that are relevant to human brain function. Perhaps just as important is the observation that several OTB endpoints do not correlate at all with IQ: this demonstrates that operant behaviors can provide data about brain function that are not attainable using traditional IQ assessments.

Cross-Species Comparisons of OTB Performance

Comparisons of the OTB performance of children with that of adolescent monkeys has also been examined and remarkable between-species similarities have been observed. Equality of task performance between monkeys (4 years old) and children has been demonstrated, albeit, the similarities are dependent upon task, endpoint, and the children's ages. For example, accuracy and rate of forgetting in the short-term memory task is nearly identical for four-year-old children and adolescent monkeys, whereas in this same task, choice response latency (time to make responses to choice stimuli) in monkeys is equivalent to that for thirteen-year-old children. Monkey subjects appear to be just as motivated to work for banana-flavored food pellets as six- to eight-year-old children are to work for nickels. In the color and position discrimination task, monkey accuracy is equivalent to that for eight- to nine-year-old children.

Drug studies in monkeys have provided data for several

TABLE 1
Agents for Which Comparable Behavioral Effects Have Been Observed in Both Monkeys and Humans

Drug	Effect
Delta-9-tetrahydrocannabinol	Overestimate time passage
Marijuana smoke	Short-term memory impairment (acute effect)
	Amotivational syndrome (chronic effect)
Chlorpromazine	Decrease response initiation
Diazepam	Learning and memory impairments
Morphine	Decrease response rates
Atropine	Learning disruption
Pentobarbital	Overestimate time passage

Note. Adapted from Paule, 2000.

compounds that have also been assessed in human adults (see Table 1). While the tasks used to assess drug effects in humans were not identical to those in the OTB, they were designed to assess the same or similar functions. These comparative data are beginning to shed light on the predictive validity of the monkey model (Paule, 2000). For example, delta-9-tetrahydrocannabinol (THC), the main psychoactive ingredient in marijuana smoke, causes both humans and monkeys to overestimate the passage of time (eight seconds "feels" like ten seconds). Marijuana smoke causes deficits in performance of short-term memory tasks in both species. Likewise, for both monkeys and humans, chlorpromazine decreases response initiation, diazepam impairs learning and memory, morphine causes a general decrease in rate of responding, and atropine disrupts learning. In studies on the effects of chronic drug exposure, data from the monkey model indicate that, as reported for adolescent and young adult human subjects, chronic marijuana smoke exposure produces an "amotivational" syndrome (Paule, 2000).

Thus, where appropriate data exist, it seems that the monkey model is quite predictive of drug effects in adult humans. As the use of monkey-appropriate tasks becomes more widespread with human subjects, it will be possible to collect data on a variety of experimental manipulations using exactly the same instruments in both species. Direct comparisons will, thus, be more readily accomplished and the predictive validity of the monkey animal model will be more directly assessable. Remarkable interspecies comparabilities portend the utility of such an approach in neurotoxic risk assessment procedures. In addition, ongoing studies using similar approaches in rodents indicate that additional animal models may prove useful (Ferguson and Paule, 1996; Mayorga *et al.*, 2000a,b; Popke *et al.*, 2000).

Human Neurobehavioral Test Methods for Studying Neurotoxicity in Working Populations (W. Kent Anger)

The dangerous effects of workplace exposure to neurotoxic chemicals were readily apparent in the 1800s and early 1900s when workers died from high-concentration exposures to dangerous chemicals such as lead (Hunter, 1969), and the link between those exposures and their effects was readily established. At the beginning of the 21st century, however, evidence of neurotoxicity is rarely demonstrated by neuropathological evidence of adverse effects, except in developing countries where serious overexposures continue (e.g., Weiss, 1983). Clinical psychologists were the first to apply tests of neurobehavioral performance to identify adverse effects in humans exposed to neurotoxic chemicals (Hänninen *et al.*, 1966), and these have become the methodological staple in research or clinical studies for detecting and characterizing human neurotoxicity in industrialized nations.

Acute exposure studies of human subjects were widely reported in the 1980s, but the danger of intentionally exposing people to neurotoxic chemicals with unknown properties, and the expense of assuring their safety, has led to a decline in this type of research. Consequently, the bulk of the research now being conducted is in the workplace or community where ongoing exposures are believed to be safe, or at least not demonstrated to be hazardous. While it is possible to design acute exposure studies in a workplace, such studies are virtually always conducted against a background of chronic exposures. Cognitive measures have begun to dominate this field of research because they have repeatedly revealed differences in exposed workers as compared to unexposed comparison groups (e.g., Anger *et al.*, 1998).

Core Tests

The cognitive test methods have evolved over the years, but a core of tests has been used in many of the studies. In 1983, early leaders in this field were assembled by the World Health Organization (WHO) to propose a core of tests for future neurotoxicology research (Johnson *et al.*, 1987). The battery was termed the Neurobehavioral Core Test Battery (NCTB), and its 7 core tests (Digit Symbol, Digit Span, Benton Visual Recognition, Simple Reaction Time, Aiming, Santa Ana, and Profile of Mood States) or functionally comparable tests have been used extensively in neurobehavioral research since that time. The Digit Symbol, Digit Span, and Benton tests are primarily cognitive tests, although a motor component is found in the Digit Symbol. The Digit Symbol, perhaps because it draws on such a wide array of neurobehavioral functions, is the most productive test in terms of demonstrated findings in human neurotoxicology research (Anger, 1990). Table 2 lists, for chemical substances with extensively replicated findings, those cognitive functions most frequently associated with performance deficits in exposed workers when compared to unexposed referents. The functional tests most often used to assess those deficits (and to reveal statistically significant differences) are listed under each function.

In the 1980s, computerized neurobehavioral testing systems emerged to increase the efficiency (one-on-one testing is very

TABLE 2
Cognitive Functions Associated with Deficits from Extended Duration Neurotoxicant Exposures in Humans, and Most Frequently Used Tests of Those Functions

Function	Solvent mixtures	Carbon disulfide	Perchloroethylene	Styrene	Toluene	OP*** pesticides	Mercury (inorganic)	Lead (inorganic)
Reasoning/ intelligence	Block Design	Block Design	—	Block Design	Block Design	—	Raven Progressive Matrices	Wechsler Tests
Learning	—	—	—	—	—	—	—	Paired Associates
Complex function (coding)	Digit Symbol/ SD*	Digit Symbol	Digit Symbol	Digit Symbol	Digit Symbol	Digit Symbol/SD*	Digit Symbol	Digit Symbol
Memory	Benton	—	—	—	—	—	Rey Tests, Bender, Sternberg	Rey Tests, Sternberg
Attention/vigilance	Bourdon Wiersma/ CPT**	—	—	CPT**	—	CPT**	—	CPT**
Attention	Digit Span	Digit Span	Digit Span, D2	Digit Span	Digit Span	Digit Span	Digit Span	Digit Span

Note. *SD, symbol digit; **CPT, continuous performance test; ***organophosphate; — indicates deficit not reported.

expensive in personnel time) and reliability (technically trained humans are invariably variable in test administration) of this type of research. These methods were widely adopted, and the Neurobehavioral Evaluation System 2 (NES 2) became the dominant testing system through the 1990s (Letz, 1990), although others emerged to expand the available options. The Cambridge Neuropsychological Test Automated Battery (CANTAB) system provided sophisticated cognitive tests that were developed in animal studies (Fray and Robbins, 1996), as did Stollery's Automated Computerized Test or ACT system (Stollery, 1996). The behavioral assessment and research system (BARS) focused on improved instructions for established tests, drawing in new tests from the animal neurotoxicology literature, and a nine-button response unit that is placed over the computer keyboard for the computer naïve (Anger *et al.*, 1996). The addition of a computerized system for diverse psychological tests (Kovera *et al.*, 1996) added a new dimension to the field that had been lost after the departure of the early clinicians from behavioral neurotoxicology (Anger, 1990).

Validity

Cross-sectional comparisons between exposed and unexposed groups are required to identify nonreversing adverse chemical effects on cognition. Such research designs are subject to possible bias due to selection of unequal comparison groups. Replication in different studies of the same effect associated with the same chemical exposure is therefore necessary to establish a neurotoxic effect. For at least lead, mercury, and a handful of solvents, there is substantial replication of findings in studies from different countries using similar or in many cases the same tests, differing only in the translated instructions from the original format, usually English (Table 2). Anger (1990) identified 185 such studies in a comprehensive review through the end of the 1980s. More studies continue to be reported (e.g., see Anger *et al.*, 1998 and Dick, 1995 for reviews).

The data in Table 2 constitute criterion validity. That is, those findings demonstrate that groups exposed to neurotoxic chemicals consistently have deficits on cognitive tests based on performance, as compared to unexposed comparison groups selected to be similar to the exposed groups in other ways. Another measure of validity is the ability of these tests of neurotoxicity to detect the functional deficits found in neurologic diseases. White *et al.* (1996) administered neurobehavioral tests from the NES2 to 73 Parkinson patients in stages 1–3 of the Hoehn and Yahr scale (Gancher, 1997) and compared their performance to friends and family members of the patients; a similar comparison with 61 multiple sclerosis patients was also reported. These reports demonstrate that the neurobehavioral tests used to detect chemical neurotoxicity also detect differences produced by neurologic diseases. Similarly, BARS tests were administered to 15 Parkinson patients in stages 1–2 of the Hoehn and Yahr scale (Gancher, 1997) and age-, education- and gender-matched controls. Figure 3 reveals that the greater differences in this motor neuron disease lie in the motor tests such as tapping, but cognitive differences, as reflected in performance tests such as Selective Attention, are also seen, demonstrating the sensitivity of cognitive measures (Camicioli *et al.*, in press).

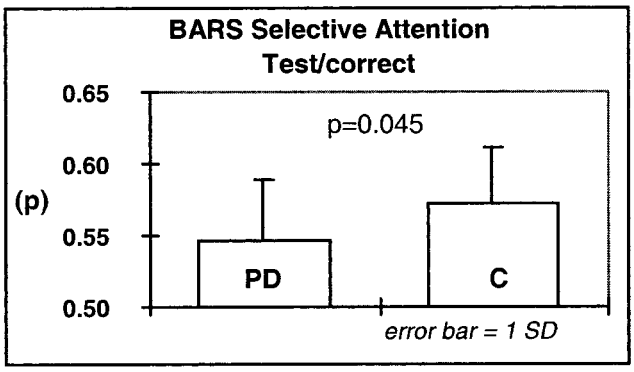

FIG. 3. BARS Tapping and Selective Attention Test performance in Parkinson patients (PD) and age-, education-, and gender-matched controls (C).

Human Subject Factors

Human subject variables, including age, education, gender, and motivation, have a demonstrated effect on cognitive test performance. These factors must strongly influence the selection of appropriate comparison subjects (controls or referents) and cautions for research throughout the world. There is evidence that education has the greatest impact on performance in the cognitive tests used in neurotoxicity research (Anger *et al.*, 1997). Shortly after the development of the WHO-recommended NCTB, a cross-cultural feasibility study was conducted in 10 countries on 4 continents (Anger *et al.*, 1993) using the NCTB. That study revealed comparable performance in such culturally diverse populations as in Poland, France, Italy, USA, Canada, and China in people of comparable age and gender. Unexpectedly, the one Latino population, from Nicaragua, performed at a much lower level on the cognitive tests. However, the Nicaraguan farm workers had a mean of 3 years of education, compared to 10 years and above in the other populations. Subsequent research in Latino populations using the same tests has demonstrated that the number of years of education has a large impact on performance on the tests used in neurotoxicity assessments (Anger *et al.*, 1997). In recent research, this has been extended to repeated cognitive testing. The same series of BARS computerized tests was administered 4 times over several weeks to 9 Caucasian adults educated in the U.S. (mean age 41.0 years; mean education 13.9 years), 9 Hispanic youth educated primarily in the U.S. (mean age 16.0 years; mean education in the U.S. 10.2 years), and 9 Hispanic youth educated primarily in South America (mean age 16.0; mean education in South America 4.1 years). Figure 4 reveals the constant performance in the adults and the improvement of both youth groups over the 4 sessions on the Selective Attention Test, a test of attention drawn from animal research. The education system and its cultural heritage may produce a large difference in performance, as it did in this study, and these effects may not be erased with repeated administration of the cognitive tests. The same trends were seen in all the cognitive tests in this study (Rohlman *et al.*, in press). Clearly, educational background has a large impact on the cognitive tests used in this field, as does age, and to a lesser extent, gender (Anger *et al.*, 1997). Thus, cognitive research in human populations must control these factors in the groups under study, not in the analysis.

New Methods/Future Directions

Any dynamic field has constantly evolving methods, and human behavioral neurotoxicology is no exception. The primary direction is that new systems are largely computerized and more attention is being paid to the manner in which responses are measured. For computerized tests, the keyboard is an inadequate input device from the standpoint of response speed, comfort, and reliability. Two prominent alternative examples are the use of a stylus or pencil-like device by the Neurobehavioral Evaluation System-3 (NES 3) and the nine-button input device implemented for the BARS testing system.

The necessity of conducting most occupational and community research in the field, and away from the clinic or research laboratory, has led to the use of field portable equipment or laptop computers, whenever possible. The increased reliability, consistency, and accuracy of the computerized systems make these testing instruments ideal for neurotoxicity research. Equally important is the efficiency introduced by test computerization, which has reduced the cost of conducting field studies significantly. There is an urgent need to study extended-duration exposures to neurotoxic chemical agents, and also physical agents such as heat overexposure. There are hundreds of neurotoxic chemical substances in the workplace or the environment with some potential for human exposure (Anger, 1990; Anger and Johnson, 1985; Spencer and Schaumburg, 2000); this necessitates the expansion of this research to establish adverse effect potential and exposure thresholds in humans. The increased efficiency of computerized systems will allow more studies to be conducted with limited resources, and the cognitive measures that have emerged as highly sensitive are optimally suited to computerized methods.

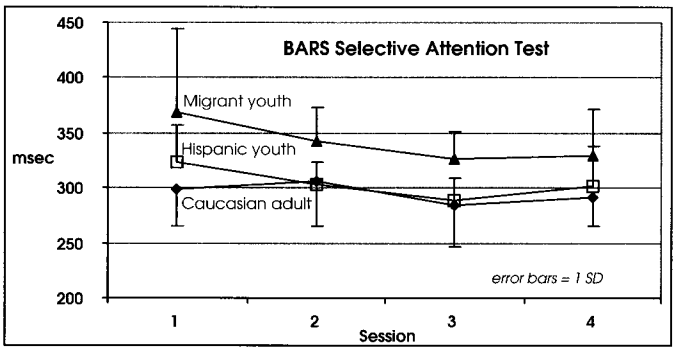

FIG. 4. Performance on the BARS Selective Attention Test in 4 sessions separated by several weeks.

Issues in the Interpretation of Human Neurotoxicity Data (David Bellinger)

Children's neurobehavioral test scores have served as critical endpoints in policy debates about exposure standards for environmental neurotoxicants, such as lead and methyl mercury. Because this is likely to be true in future discussions of other chemical exposures, it is important to consider the limitations of such test scores, particularly pitfalls in their interpretation.

Apical Test Scores Represent Final Common Pathways for the Expression of Diverse Cognitive Patterns

The assessment battery typically used in a neurotoxicant study consists of a global or apical test, supplemented by tests thought to assess particular aspects of cognition (e.g., language, visual-spatial skills, memory, and fine motor function). Historically, however, it is apical test scores (e.g., full-scale IQ) rather than domain-specific test scores that have received the most attention, most likely because they can more readily be incorporated into risk assessment and cost-benefit analyses.

Because apical tests integrate performance on a diverse set of tasks, a variety of cognitive patterns will result in the same score. Gardner (1993) argued that traditional, psychometrically based IQ tests focus on logical-mathematical and linguistic skill intelligence. For instance, three children, one with Down syndrome, one with autism, and one with Williams syndrome, may all have the same full-scale IQ score yet have remarkably little in common in terms of their cognitive strengths and weaknesses. Because of this, a full-scale IQ score is of limited clinical interest or utility. The fact is that logical-mathematical and linguistic skills are considered to be the *sine qua non*. Although the sensitivity of an apical test to a neurotoxicant exposure might be high, its specificity is likely to be low.

This principle extends to tests purported to assess a particular domain of function. For instance, one might assume that a child's score on the design copying test, called the Developmental Test of Visual-Motor Integration (VMI) (Beery, 1989) depends largely on his or her visual-motor integration skills. In fact, a child can score poorly on this test for many other reasons, including deficits in visual perception (seeing part-whole relationships), planning and sequencing (ability to organize behaviors involved in reproducing designs), grapho-motor control (ability to control the pencil tip), behavioral modulation (self-monitoring of progress in implementing a plan), and motivation. A low score does not convey any information about which one (or more) of these aspects of performance was the source of a child's difficulties. The examiner may formulate hypotheses by observing how the child went about the task. Although this information may be included in the clinical report, it is only the test score that is entered into the research database. Nevertheless, lower VMI scores among children with higher lead burdens have led to the inference that lead selectively affects visual-motor integration skills and thus right hemisphere, specifically parietal lobe, function. Although adults with acquired lesions in the right parietal lobe (due to trauma or cerebrovascular accidents) have difficulty on design copying tasks (Lezak, 1995), children chronically exposed to low levels of lead may achieve poor VMI scores for entirely different reasons. Only detailed test batteries, constructed with the explicit goal of evaluating the competing hypotheses, will provide the data needed to address this issue.

Another problem in interpreting neurobehavioral test scores is that two tests purporting to assess the same domain of function may in fact assess quite different aspects of function within that domain. This problem is germane in efforts to evaluate the concordance of study findings. For example, the Faroe Islands study reported that prenatal methyl mercury exposure and memory were inversely related (Grandjean *et al*., 1997), while the Seychelles Islands study reported that they were not (Davidson *et al*., 1998). Are these findings truly discrepant, or were the memory assessments used so different that the data are not comparable? In the Seychelles study, memory was operationalized as performance on the memory scale of the McCarthy Scales of Children's Abilities. This is composed of 4 subtests, each involving memory for different types of material: a pictured array of objects, tone sequences, word strings in specific orders, sentences, and digit strings. In the Faroe study, memory was operationalized as performance on the California Verbal Learning Test-Children, a word list learning task that consists of five learning trials, short and long, free and cued, recall trials, an interference trial, and a recognition trial. Nothing that we know about the neuropsychological toxicity of methyl mercury would lead us to expect that performance on these two quite dissimilar tests would be affected to the same extent.

Even when the same test is administered in two studies, scores may not be directly comparable if different approaches are taken to scoring performance. To use another example from the methyl mercury literature, the Faroe study reported a positive association between prenatal exposure and children's error scores on the Bender-Gestalt Test (Grandjean *et al*., 1997), whereas the Seychelles study did not (Davidson *et al*., 1998). In the Faroe study, the Gottingen scoring system was used, while in the Seychelles study the simpler Koppitz system was used. In a direct comparison, children's error scores, derived using the Gottingen system, were significantly associated with lead burden, whereas scores assigned to the identical test protocols using the Koppitz scoring system were not (Trillingsgaard *et al*., 1985), suggesting that the Gottingen system is more sensitive than the Koppitz system to lead exposure.

Apical Test Scores Convey More about the Product than the Process of Cognition, And Fail to Assess Important Aspects of "Intelligence."

Test items tend to focus on what a child knows rather than on what he or she does when faced with the task of learning

new material. Furthermore, many items are scored pass or fail, ignoring qualitative aspects of performance. Little account is generally taken of errors or error types. Thus, we learn whether exposure to a neurotoxicant is related to a child's present competence with respect to a skill within a particular domain, but little about how he or she achieved that competence.

A more general issue is that apical tests, such as IQ, by no means "cover the ballpark", and exclude consideration of many dimensions of behavior pertinent to an individual's success in life. Gardner (1993) argued that traditional, psychometrically based IQ tests focus on logical-mathematical intelligence and hypothesized the existence of several other semi-autonomous "intelligences," including linguistic, musical, spatial, bodily-kinesthetic, and personal. The fact that logical-mathematical skills are considered to be the *sine qua non* of intelligence probably reveals more about the values of modern industrialized society than about the human brain.

An important dimension of function that is poorly represented in apical tests is "executive functions." This refers to the ability to independently structure moment-to-moment behavior, involving the orderly planning and sequencing of behavior, selecting goals, anticipating outcomes, and monitoring ongoing performance. These skills are expressed variously as the ability to attend to several stimuli simultaneously, to resist distraction, and to follow multi-step directions. To a large extent, these skills are not elicited in the highly structured one-to-one interaction in which an IQ is administered. The diagnosis of attention deficit hyperactivity disorder (ADHD) illustrates this point insofar as it is generally made on the basis of historical information provided by parents and teachers about a child's "free range" behavior in less structured settings. For most children with ADHD, the structure provided by an examiner is sufficient to keep their behavior task-focused, reducing the opportunities to observe vulnerabilities in the ability to sustain attention and to self-organize.

Another important endpoint that is poorly assessed by apical measures is "learning disabilities" (LD). Sometimes LD appears to be equated with a low IQ. In fact, the definition of LD that is most commonly applied in educational settings requires a significant discrepancy between ability (as measured by IQ) and achievement in an academic domain such as reading or mathematics (Reynolds, 1985). A child with LD is one who performs below the level expected based on his or her IQ. Thus, IQ alone is not sufficient for determining whether a child is learning-disabled.

Neurobehavioral Test Scores Are Final Common Pathways for the Expression of Many Influences Other than Neurotoxicants

Under the best of circumstances, when the critical correlates/determinants of performance on neurobehavioral tests have been measured well, approximately 50% of the variance in such scores can be accounted for. Among the reasons why it is important to consider these factors in evaluating neurotoxicant effects are the following:

- Error variance in the endpoint can be reduced, boosting the statistical power of the hypothesis tests involving neurotoxicant exposure. The impact of this can be substantial given that a low-level neurotoxicant exposure is unlikely to account for more than 5% of test score variance.
- Confounding bias can be addressed. This is a methodological artifact that occurs when a correlate/determinant of test score is also associated with neurotoxicant exposure. The critical confounders may be somewhat toxicant-specific. For lead, socioeconomic status is critical (Brody *et al*., 1994). For methyl mercury, the major concern is chemicals, such as organochlorines and polyunsaturated fatty acids, to which an individual may also be exposed *via* nursing or consumption of seafood (Grandjean *et al*., 1995).
- Effect modification (or interaction) can be assessed. This occurs when the magnitude of a neurotoxicant's effect varies depending on the context in which exposure occurs, because some other factor increases or decreases susceptibility (e.g., age, sex, socioeconomic status, behavioral history, nutritional status, or another neurotoxicant exposure). This issue has received much less attention than confounding bias. The analytic approach typically used rests on the dubious assumption that a single point estimate best describes the impact of a neurotoxicant, regardless of the characteristics of the host or the setting in which exposure occurs (Bellinger, 2000).

Supplementary Tests of Specific Domains Are Often Not Well Matched Psychometrically, Impeding Efforts to Identify a "Signature" Behavioral Injury

The tests used to assess different domains need to be comparable in discriminating power (specifically their true-score variance) in order for direct comparisons of scores across domains to be informative (Chapman and Chapman, 1973). Otherwise, increased exposure may appear to be more strongly related to performance in one domain than in another, simply because the tests used to assess the two domains differ in their psychometric characteristics. Observing a neurotoxicant effect on a test of language skills that includes items that vary widely in difficulty, but not on a test of visual-spatial skills that consists exclusively of easy tasks, provides little useful information as to whether language or visual-spatial skills are more vulnerable to the neurotoxicant. In human studies, this problem is rarely discussed when interpreting the pattern of findings on the different tests included in a battery.

In human neuropsychological work, many of the most commonly used tests were developed in the context of clinical evaluation, where the primary goal is to identify clinically significant impairments among patient groups (e.g., individuals with CNS insults such as traumatic brain injury, stroke, tumors, infections, etc.). Many domain-specific tests were not designed to detect small differences in performance within the normal

range. In this respect, apical tests, which were developed to permit inferences about individual differences along the entire spectrum of performance, may be better suited than domain-specific tests to the task of detecting subtle impacts of toxicant exposures on neuropsychological function (Bellinger, 1995).

In conclusion, because a child's scores on conventional neuropsychological tests are strongly predictive of his or her later success in meeting school and workplace challenges, it is likely that such tests will continue to play an important role in risk assessments of environmental neurotoxicants. Therefore, it is critical that the strengths and limitations of such tests be openly acknowledged.

Summary and Conclusions (Barbara D. Beck)

The Workshop on "Cognitive Tests: Interpretation for Neurotoxicity?" identified many of the difficulties and challenges associated with the use of specific tests to assess neurotoxicity in humans. However, the presentations also brought out approaches to enhance the validity of such tests with respect to evaluating toxicant exposure, as well as appropriate interpretation of such tests with respect to performance in humans.

Importantly, the validity of such tests can be enhanced by performance of the same battery across animal species. Behavioral deficits can be characterized in humans with known degenerative disease states and the same tests replicated in experimental animal models, developed, for example, by gene manipulation. Conversely, animal models of neurotoxicant exposure can be developed for identification of patterns of behavioral deficits with subsequent application to humans exposed typically to low levels of neurotoxicants. The utility of such approaches depends upon a continuity of behavioral performance across species and the selection of performances that are affected by various risk factors, including drugs and exogenous chemicals. Several performance measures including specific measures of cognition meet these criteria.

A specific test has been developed at the National Center for Toxicological Research, the Operant Test Battery (OTB). The OTB fulfills several of the above-described criteria for utility. This test battery, which involves the operation of a response level in order to receive reinforcers, has been evaluated in both primates (i.e., four-year-old monkeys) and in humans (children), and has been used to measure learning, memory, and other measures of brain function. The validity of the OTB is reflected in the similarity of response across species on a range of parameters, such as accuracy and rate of forgetting in the short-term memory task. These similarities enhance one's confidence that the response of the monkey to toxicant (e.g., tetrahydrocannibol or chlorpromazine) exposure is a good predictor of response to toxicant exposure in humans.

The use of computerized testing systems has facilitated the expansion of neurobehavioral tests outside the laboratory setting and into workplace and environmental settings. Over time, these systems have become more user-friendly. Examples include the Cambridge Neuropsychological Test Automated Battery (CANTAB) and the Behavioral Assessment and Research System (BARS). As with the OTB, these test systems have been based on tests developed in animal systems. The criteria for validity of such tests include demonstration of consistent patterns of deficits across neurotoxicant (e.g., solvent mixtures, mercury) exposed populations in different countries and an ability to discriminate deficits associated with neurodegenerative diseases, such as Parkinson's. Nonetheless, such tests are subject to confounding from a number of factors such as age, gender and, in particular, education. Thus conclusions regarding causation and magnitude of impact must consider the extent to which these factors have been adequately addressed.

It is important to be careful in the interpretation of cognitive tests, especially in the context of risk assessment for neurotoxicant exposures and subsequent risk management decisions. For example, tests of apical or global function, such as the IQ test, may show sensitivity with respect to neurotoxicant exposure, but may have little specificity. This is because a variety of cognitive deficits may produce the same overall reduction in score, in which case the test provides little or no information on which functions are impaired. Even a test designed to measure a specific performance, e.g., visual-motor integration, may reflect other performances, such as motivation. Of particular importance in interpreting cognitive tests is assessing their ability to accurately reflect "intelligence." Important components of intelligence, such as artistic ability or ability to plan and carry out tasks, are not well reflected in the standard IQ tests. A particular challenge lies in the fact that many of the commonly used neuropsychological tests were developed in the context of clinically significant impairment (e.g., brain trauma) and may be of limited utility in detecting subtle difference in performance in the normal range.

In conclusion, this workshop illustrated some of the challenges associated with the use of cognitive tests for cross-species comparisons. At present there is no "gold standard" for such tests, although the OTB has clearly demonstrated some success in cross-species comparisons involving rodents, monkeys, and humans. At least one human neurotoxicity test (BARS) has adopted the OTB test. Nevertheless, the available test batteries differ in their breadth of applicability, the types of study subjects, and other factors. As such, the information developed among different test batteries should be viewed as complementary.

There are several potential areas for improvements in the use of cognitive tests, especially as applied to risk assessment. For example, better definitions of exposure-response relationships across species would help in quantifying interspecies differences in responsiveness. In addition, more dialogue between risk assessors and neurotoxicologists is needed to facilitate the appropriate interpretation of tests in terms of adverseness of effect (e.g., what is normal vs. impaired performance), potential clinical significance, and

importance of confounding factors. An improved understanding would help in selection of the most relevant endpoints for risk assessment purposes. It is clear that many challenges remain for the risk assessor and the risk manager, who must decide how to consider such tests with respect to success in life, and make decisions regarding permissible exposure limits accordingly.

ACKNOWLEDGMENTS

W.K.A. acknowledges the contributions of Drs. Diane S. Rohlman and Daniel Storzbach to the development of the methods and research reported under the subhead: "Human Neurobehavioral Test Methods for Studying Neurotoxicity in Working Populations." This work was supported by NIEHS 5R21 ES08707–02 and EPA Cooperative Agreement CR 822789–01–0.

REFERENCES

Anger, W. K. (1990). Worksite behavioral research: Results, sensitive methods, test batteries, and the transition from laboratory data to human health. *Neurotoxicology* **11,** 627–717.

Anger, W. K., Cassitto, M. G., Liang, Y.-X., Amador, R., Hooisma, J., Chrislip, D. W., Mergler, D., Keifer, M., Hörtnagl, J., Fournier, L., Dudek, B., and Zsögön, E. (1993). Comparison of performance from three continents on the WHO-recommended Neurobehavioral Core Test Battery. *Environ. Res.* **62,** 125–147.

Anger, W. K., and Johnson, B. L. (1985). Chemicals affecting behavior. In *Neurotoxicology of Industrial and Commercial Chemicals* (J. O'Donoghue, Ed.), pp. 51–148. CRC Press, Boca Raton, FL.

Anger, W. K., Rohlman, D. S., Sizemore, O. J., Kovera, C. A., Gibertini, M., and Ger, J. (1996). Human behavioral assessment in neurotoxicology: Producing appropriate test performance with written and shaping instructions. *Neurotoxicol. Teratol.* **18,** 371–379.

Anger, W. K., Sizemore, O. J., Grossmann, S. J., Glasser, J. A., Letz, R., and Bowler, R. (1997b). Human neurobehavioral research methods: Impact of subject variables. *Environ. Res.* **73,** 18–41.

Anger, W. K., Storzbach, D., Amler, R. W., and Sizemore, O. J. (1998). Human behavioral neurotoxicology: Workplace and community assessments. In *Environmental and Occupational Medicine,* 3rd ed. (W. Rom, Ed.), pp. 709–731. Lippencott-Raven, Philadelphia.

Beery, K. (1989). *Developmental Test of Visual-Motor Integration, Administration, Scoring, and Teaching Manual,* 3rd. ed. Modern Curriculm Press, Ohio.

Bellinger, D. (1995). Interpreting the literature on lead and child development: The neglected role of the "experimental system". *Neurotoxicol. Teratol.* **17,** 201–212.

Bellinger, D. Effect modification in epidemiologic studies of low-level neurotoxicant exposures and health outcomes. *Neurotoxicol. Teratol.* **20,** 133–140.

Benignus, V. A., Boyes, W. K., and Bushnell, P. J. (1998). A dosimetric analysis of behavioral effects of acute toluene exposure in rats and humans. *Toxicol. Sci.* **43,** 186–195.

Boren, J. J. (1963). Repeated acquisition of new behavioral chains. *Am. Psychol.* **17,** 421.

Boren, J. J., and Devine, D. D. (1968). The repeated acquisition of behavioral chains. *J. Exp. Anal. Behav.* **11**, 651–660.

Brockel, B. J., and Cory-Slechta, D. A. (1999). The effects of post-weaning, low-level lead exposure on sustained attention: A study of target densities, stimulus presentation rate, and stimulus predictability. *Neurotoxicology* **20,** 921–933.

Brody, D., Pirkle, J., Kramer, R. A., Flegal, K. M., Matte, T. D., Gunter, E. W., and Paschal, D. C. (1994). Blood lead levels in the U.S. population. Phase 1 of the Third National Health and Nutrition Examination Survey (NHANES III, 1988 to 1991). *JAMA* **272,** 277–283.

Brooks, A. I., Cory-Slechta, D. A., Murg, S. L., and Federoff, H. J. (2000). Repeated acquisition and performance chamber for mice: A paradigm for assessment of spatial learning and memory. *Neurobiol. Learn. Mem.* **74,** 241–258.

Bushnell, P. J. (1998). Behavioral approaches to the assessment of attention in animals. *Psychopharmacology* **138,** 231–259.

Bushnell, P. J., Kelly, K. L., and Crofton, K. M. (1994). Effects of toluene inhalation on detection of auditory signals in rats. *Neurotocicol. Teratol.* **16,** 149–160.

Callahan, M. J., Kinsora, J. J., Harbaugh, R. E., Reeder, T. M., and Davis, R. E. (1993). Continuous ICV infusion of scopolamine impairs sustained attention of rhesus monkeys. *Neurobiol. Aging* **14,** 147–151.

Camicioli, R., Grossman, S. J., Spencer, P. S., Hudnell, K., and Anger, W. K. (in press). Discriminating mild Parkinsonism: Methods for epidemiological research. *Movemt. Disord.*

Chapman, L. J., and Chapman, J. P. (1973). The measurement of differential deficit. *J. Psych. Res.* **14,** 303–311.

Cohn, J., and Cory-Slechta, D. A. (1993). Subsensitivity of lead-exposed rats to the accuracy-impaired and rate-altering effects of MK-801 on a multiple schedule of repeated learning and performance. *Brain Res.* **600,** 208–218.

Cohn, J., and Cory-Slechta, D. A. (1994a). Assessment of the role of dopamine systems in lead-induced learning impairments, using a repeated-acquisition and performance baseline. *Neurotoxicology* **15,** 913–926.

Cohn, J., and Cory-Slechta, D. A. (1994b). Lead exposure potentiates the effects of NMDA on repeated learning. *Neurootxicol. Teratol.* **16,** 455–465.

Cohn, J., Cox, C., and Cory-Slechta, D. A. (1993). The effects of lead exposure on learning in a multiple repeated-acquisition and performance schedule. *Neurotoxicology* **14,** 329–346.

Cohn, J., MacPhail, R. C., and Paule, M. G. (1996). Repeated acquisition and the assessment of centrally acting compounds. *Brain Res. Cogn. Brain Res.* **3,** 183–191.

Cohn, J., and Paule, M. G. (1995). Repeated acquisition of response sequences: The analysis of behavior in transition. *Neurosci. Biobehav. Rev.* **19,** 397–406.

Cory-Slechta, D. A. (1994). The impact of NMDA receptor antagonists on learning and memory functions. *Psychopharmacol. Bull.* **30,** 601–612.

Davidson, P., Myers, G., Cox, C., Axtell, C., Shamlaye, C., Sloane-Reeves, J., Cernichiari, E., Needham, L., Choi, A., Wang, Y., Berlin, M., and Clarkson, T. (1998). Effects of prenatal and postnatal methylmercury exposure from fish consumption on neurodevelopment: Outcomes at 66 months of age in the Seychelles Child Development Study. *JAMA* **280,** 701–707.

Dick, R. B. (1995). Neurobehavioral assessment of occupationally relevant solvents and chemicals in humans. In *Handbook of Neurotoxicology* (L. W. Chang and R. S. Dyer, Eds.), pp. 217–322. Marcel Dekker, New York.

Dick, R. B., Krieg, E. F., Jr, Setzer, J., and Taylor, B. (1992). Neurobehavioral effects from acute exposures to methyl isobutyl ketone and methyl ethyl ketone. *Fundam. Appl. Toxicol.* **19,** 453–473.

Ferguson, S. A., and Paule, M. G. (1996). Effects of chlorpromazine and diazepam on time estimation behavior and motivation in rats. *Pharmacol. Biochem. Behav.* **53,** 115–122.

Fray, P. J., and Robbins, T. W. (1996). CANTAB battery: proposed utility in neurotoxicology. *Neurotoxicol. Teratol.* **18,** 499–504.

Gancher, S. T. (1997). Scales for the assessment of movement disorders. In *Handbook of Neurologic Rating Scales* (R. M. Herndon, Ed.), pp. 81–106. Demos Vermande, New York.

Gardner, H. (1993). *Frames of Mind: The Theory of Multiple Intelligences*. Basic Books, New York.

Grandjean, P., Weihe, P., Needham, L., Burse, V., Patterson, D., Sampson, E.,

Jorgensen, P., and Vahter, M. (1995). Relation of a seafood diet to mercury, selenium, arsenic, and polychlorinated biphenyls and other organochlorine concentrations in human milk. *Environ. Res.* **71,** 29–38.

Grandjean, P., Weihe, P., White, R. F., Debes, F., Araki, S., Yokoyama, K., Murata, K., Sorensen, N., Dahl, R., and Jorgensen, P. J. (1997). Cognitive deficit in 7-year-old children with prenatal exposure to methylmercury. *Neurotoxicol. Teratol.* **19,** 417–428.

Hänninen, H. (1966). Psychological tests in the diagnosis of carbon disulfide poisoning. *Work Environ. Health* **2,** 16–20.

Heyser, C. J., Hampson, R. E., and Deadwyler, S. A. (1993). Effects of delta-9-tetrahydrocannabinol on delayed match to sample performance in rats: Alterations in short-term memory associated with changes in task-specific firing of hippocampal cells. *J. Pharmacol. Exp. Ther.* **264,** 294–307.

Hunter, D. (1969). *The Diseases of Occupations*, 2nd ed. Little Brown, Boston.

Joel, D., Weiner, I., and Feldon, J. (1997). Electrolysis lesions of the medial prefrontal cortex in rats disrupt performance on an analog of the Wisconsin Card Sorting Test, but do not disrupt latent inhibition: implications for animal models of schizophrenia. *Behav. Brain Res.* **85**, 187–201.

Johnson, B. L., Baker, E. L., EI Batawi, M., Gilioli, R., Hänninen, H., Seppäläinen, A. M., Xintaras, C. (Eds.) (1987). *Prevention of Neurotoxic Illness in Working Populations*. John Wiley, New York.

Kelly, T. H., Foltin, R. W., and Fischman, M. W. (1991). The effects of repeated amphetamine exposure on multiple measures of human behavior. *Pharmacol. Biochem. Behav.* **38,** 417–426.

Kelly, T. H., Foltin, R. W., Serpick, E., Fischman, M. W. (1997). Behavioral effects of alprzolam in humans. *Behav. Pharmacol.* **8,** 47–54.

Kovera, C. A., Anger, W. K., Campbell, K. A., Binder, L. M., Storzbach, D., Davis, K. L., and Rohlman, D. S. (1996). Computer-administration of questionnaires: A health-screening system (HSS) developed for veterans. *Neurotoxicol. Teratol.* **18,** 511–518.

Lane, J. D., and Phillips-Butre, B. G. (1998). Caffeine deprivation affects vigilance performance and mood. *Physiol.Behav.* **65,** 171–175.

Letz, R. (1990). The neurobehavioral evaluation system: An international effort. In *Advances in Neurobehavioral Toxicology: Applications in Environmental and Occupational Health*. (B. L. Johnson, W. K. Anger, A. Durao, and C. Xintaras, Eds.), pp. 489–495. Lewis Publishing, Chelsea, MI.

Lezak, M. (1995). *Neuropsychological Assessment*, 3rd ed. Oxford University Press, New York.

Mayorga, A. J., Fogle, C. M., and Paule, M. G. (2000a). Adaptation of a primate operant test battery to the rat: Effects of chlorpromazine. *Neuortox. Teratol.* **22,** 31–39.

Mayorga, A. J., Popke, E. J., Fogle, C. M., and Paule, M. G. (2000b). Similar effects of amphetamine and methylphenidate on the performance of complex operant tasks in rats. *Behav. Brain Res.* **109,** 59–68.

Moerschbaecher, M., Thompson, D. M., and Winsauer, P. J. (1985). Effects of opioids and phencyclidine in combination with naltrexone on the acquisition and performance of response sequences in monkeys. *Pharmacol. Biochem. Behav.* **22,** 1061–1069.

Parker, A., Eacott, M. J., and Gaffan, D. (1997). The recognition memory deficit caused by mediodorsal thalamic lesion in non-human primates: A comparison with rhinal cortex lesion. *Eur. J. Neurosci.* **9,** 2423–2431.

Paule, M. G. (1988). Acute effects of delta-9-tetrahydrocannabinol (THC) in rhesus monkeys as measured by performance in a battery of complex operant tests. *J. Pharmacol. Exp. Ther.* **245,** 178–186.

Paule, M. G. (2000). Validation of a behavioral test battery for monkeys. In *Methods of Behavioral Analysis in Neuroscience* (J. J. Buccafusco, Ed.), pp. 281–294. CRC Press LLC, Boca Raton, FL.

Paule, M. G., Bushnell, P. J., Maurissen, J. P., Wenger, G. R., Buccafusco, J. J., Chelonis, J. J., and Elliott, R. (1998). Symposium overview: The use of delayed matching-to-sample procedures in studies of short-term memory in animals and humans. *Neurotox. Teratol.* **20,** 493–502.

Paule, M. G., Chelonis, J. J., Buffalo, E. A., Blake, D. J., and Casey, P. H. (1999a). Operant test battery performance in children: correlation with IQ. *Neurotoxicol. Teratol.* **21,** 223–230.

Paule, M. G., Cranmer, J. M., Wilkins, J. D., Stern, H. P., and Hoffman, E. L. (1988). Quantitation of complex brain function in children: Preliminary evaluation using a nonhuman primate behavioral test battery. *Neurotoxicology* **9,** 367–378.

Paule, M. G., Meck, W. H., McMillan, D. E., McClure, G. Y., Bateson, M., Popke, E. J., Chelonis, J. J., and Hinton, S. C. (1999b). Symposium overview: The use of timing behaviors in animals and humans to detect drug and/or toxicant effects. *Neurotoxicol. Teratol.* **21,** 491–502.

Penetar, D. M., and McDonough, J. H. (1983). Effects of cholinergic drugs on delayed match-to-sample performance of rhesus monkeys. *Pharmacol. Biochem. Behav.* **19,** 963–967.

Popke, E. J., Mayorga, A. J., Fogle, C. M., and Paule, M. G. (2000). Effects of acute nicotine on several operant behaviors in rats. *Pharmacol. Biochem. Behav.* **65,** 247–254.

Reynolds, C. (1985). Critical measurement issues in learning disabilities. *J. Special Educ.* **18,** 451–476.

Rohlman, D. S., Gimenes, L. S., Ebbert, C. A., Anger, W. K., Bailey, S. R., and McCauley, L. (in press). Smiling faces and other rewards: Using the behavioral assessment and research system (BARS) with unique populations. *Neurotoxicology*.

Rueckert, L., and Grafman, J. (1998). Sustained attention deficits in patients with lesions of posterior cortex. *Neuropsychologia.* **36**, 653–660.

Schulze, G. E., McMillan, D. E., Bailey, J. R., Scallet, A. C., Ali, S. F., Slikker, W., Jr., and Beery, K. (1989). *The VMI: Administration, Scoring, and Teaching Manual*, 3rd Ed. Modern Curriculum Press, Cleveland.

Schulze, G. E., McMillan, D. E., Bailey, J. R., Scallet, A. C., Ali, S. F., Slikker, W., Jr., and Paule, M. G. (1988). Acute effects of delta-9-tetrahydrocannabinol (THC) in rhesus monkeys as measured by performance in a battery of complex operant tests. *J. Pharmacol. Exp. Ther*. **245,** 178–186.

Spencer, P. S., and Schaumburg, H. H. (Eds.). (2000). *Experimental and Clinical Neurotoxicology*, 2nd. ed. Oxford University Press, New York.

Stollery, B. T. (1996). The Automated Cognitive Test (ACT) system. *Neurotoxicol. Teratol*. **18,** 493–497.

Thompson, D. M. (1977). Development of tolerance to the disruptive effects of cocaine on repeated acquisition and performance of response sequences. *J. Pharacol. Exp. Ther.* **20,** 294–302.

Thompson, D. M. (1980). Selective antagonism of the rate-decreasing effect of d-amphetamine by chlorpromazine in a repeated-acquisition task. *J. Exp. Anal. Behav*. **34,** 87–92.

Trillingsgaard, A., Hansen, O., and Beese, I. (1985). The Bender-Gestalt Test as a neurobehavioral measure of preclinical visual-motor integration deficits in children with low-level lead exposure. In: *Neurobehavioral Methods in Occupational and Environmental Health Reports from the Second International Symposium* (P. Grandjean, Ed.) World Health Organization, Copenhagen.

Walkowiak, J., Altmann, L., Kramer, U., Sveinsson, K., Turfeld, M., Weishoff-Houben, M., and Winneke, G. (1998). Cognitive and sensorimotor functions in 6-year-old children in relation to lead and mercury levels: adjustment for intelligence and contrast sensitivity in computerized testing. *Neurotoxicol. Teratol.* **20,** 511–521.

Weiss, B. (1983). Behavioral toxicology and environmental health science: Opportunity and challenge for psychology. *Amer. Psychol.* **38,** 1174–1187.

White, R. F., Diamond, R., Krengel, M., Lindem, K., and Feldman, R. G. (1996). Validation of the NES2 in patients with neurologic disorders. *Neurotoxicol. Teratol.* **18,** 441–448.