National Institute of Environmental Health Sciences

NIEHS Report on Evaluating Features and Application of Neurodevelopmental Tests in Epidemiological Studies

NIEHS 01

June 2022

NIEHS Report on Evaluating Features and Application of Neurodevelopmental Tests in Epidemiological Studies

NIEHS Report 01

June 2022

National Institute of Environmental Health Sciences
Public Health Service
U.S. Department of Health and Human Services
ISSN: 2768-5632

Research Triangle Park, North Carolina, USA

Foreword

The National Institute of Environmental Health Sciences (NIEHS) is one of 27 institutes and centers of the National Institutes of Health, part of the U.S. Department of Health and Human Services. The NIEHS mission is to discover how the environment affects people in order to promote healthier lives. NIEHS works to accomplish its mission by conducting and funding research on human health effects of environmental exposures, developing the next generation of environmental health scientists, and providing critical research, knowledge, and information to citizens and policymakers, to help in their efforts to prevent hazardous exposures and reduce the risk of preventable disease and disorders connected to the environment. NIEHS is a foundational leader in environmental health sciences and committed to ensuring that its research is directed toward a healthier environment and healthier lives for all people.

The NIEHS Report series began in 2022. The environmental health sciences research described in this series is conducted primarily by the Division of the National Toxicology Program (DNTP) at NIEHS. NIEHS/DNTP scientists conduct innovative toxicology research that aligns with real-world public health needs and translates scientific evidence into knowledge that can inform individual and public health decision-making.

NIEHS reports are available free of charge on the NIEHS/DNTP website and cataloged in PubMed, a free resource developed and maintained by the National Library of Medicine (part of the National Institutes of Health).

Table of Contents

Tables

About This Report

Authors

Roberta F. White[1], Joseph M. Braun[2], Leonid Kopylev[3], Deborah Segal[3], Christopher A. Sibrizzi[4], Alexander J. Lindahl[4], Pamela A. Hartman[4], John R. Bucher[5]

[1]Boston University, Boston, Massachusetts, USA

[2]Brown University, Providence, Rhode Island, USA

[3]U.S. Environmental Protection Agency, Washington, District of Columbia, USA

[4]ICF, Fairfax, Virginia, USA

[5]Division of the National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA

Boston University, Boston, Massachusetts, USA
Contributed to conception and design, and contributed to drafting of report
Roberta F. White, Ph.D., A.B.P.P./C.N., Lead Author

Brown University, Providence, Rhode Island, USA
Contributed to conception and design, and contributed to drafting of report
Joseph M. Braun, R.N., M.S.P.H., Ph.D., Lead Author

Division of the National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA
Contributed to conception or design and contributed to drafting of report
John R. Bucher, Ph.D., Project Lead

U.S. Environmental Protection Agency, Washington, District of Columbia, USA
Contributed to conception or design and contributed to drafting report
Leonid Kopylev, Ph.D., Project Co-lead
Deborah Segal, M.E.H.S., Project Co-lead

ICF, Fairfax, Virginia, USA
Contributed to conception or design and contributed to drafting report
Christopher A. Sibrizzi, M.P.H., Lead Work Assignment Manager

Extracted data and contributed to drafting report
Alexander J. Lindahl, M.P.H.

Contributed to drafting report
Pamela A. Hartman, M.E.M.

Contributors

Division of the National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA
Provided oversight for external peer review
Sheena L. Scruggs, Ph.D.
Mary S. Wolfe, Ph.D.

Critically reviewed draft report
Mamta V. Behl, Ph.D.
Kyla W. Taylor, Ph.D.

Kelly Government Services, Research Triangle Park, North Carolina, USA
Supported external peer review
Elizabeth A. Maull, Ph.D. (retired from NIEHS, Research Triangle Park, North Carolina, USA)

U.S. Environmental Protection Agency, Washington, District of Columbia, USA
Critically reviewed draft report
Krista Christensen, Ph.D.
Elizabeth G. Radke, Ph.D.

ICF, Fairfax, Virginia, USA
Provided contract oversight
David F. Burch, M.E.M.
Jessica A. Wignall, M.S.P.H.

Extracted data
Yousuf Ahmad, M.P.H.
Kathleen A. Clark, B.A.
Lindsey M. Green, M.P.H.
Camryn R. Lieb, B.A.
Alessandria J. Schumacher, B.A.

Prepared and edited report
Sarah K. Colley, M.S.P.H.
Jeremy S. Frye, M.S.L.S
Kaitlin A. Geary, B.S.
Tara Hamilton, M.S.
Courtney R. Lemeris, B.A.
Rachel C. McGill, B.S.

Supported external peer review
Canden N. Byrd, B.S.
Blake C. Riley, B.S.
Megan C. Rooney, B.A.

Peer Review

The Division of the National Toxicology Program (DNTP) at the National Institute of Environmental Health Sciences (NIEHS) conducted an external peer review of the draft *NIEHS Report on Evaluating Features and Application of Neurodevelopmental Tests in Epidemiological Studies* by letter in May 2021 by the experts listed below. Reviewer selection and document review followed established DNTP practices. The reviewers were charged to:

(1) Peer review the draft NIEHS Report on Evaluating Features and Application of Neurodevelopmental Tests in Epidemiological Studies.

(2) Comment on whether the draft document and draft database are clearly stated and objectively presented.

DNTP carefully considered reviewer comments in finalizing this report.

Peer Reviewers

Kim N. Dietrich, Ph.D.
Professor Emeritus, Department of Environmental Health, Division of Epidemiology and Biostatistics
University of Cincinnati College of Medicine
Cincinnati, Ohio, USA

Nancy Fiedler, Ph.D.
Deputy Director and Professor, Environmental and Occupational Health Sciences Institute, Exposure Science and Epidemiology
Rutgers University
New Brunswick, New Jersey, USA

Publication Details

Publisher: National Institute of Environmental Health Sciences

Publishing Location: Research Triangle Park, NC

ISSN: 2768-5632

DOI: https://doi.org/10.22427/NIEHS-01

Report Series: NIEHS Report Series

Report Series Number: 01

Official citation: White RF, Braun JM, Kopylev L, Segal D, Sibrizzi CA, Lindahl AJ, Hartman PA, Bucher JR. 2022. NIEHS report on evaluating features and application of neurodevelopmental tests in epidemiological studies. Research Triangle Park, NC: National Institute of Environmental Health Sciences. NIEHS Report 01.

Acknowledgments

This work was supported by the Intramural Research Program (ES103316, ES103318, and ES103319) at the National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health and performed for NIEHS under contracts GS00Q14OADU417 (Order No. HHSN273201600015U) and HHSN271201800012I.

Conflict of Interest

Individuals identified as authors in the About This Report section have certified that they have no known real or apparent conflict of interest related to neurodevelopmental tests in epidemiological studies.

Abstract

Psychometric tests are routinely used to assess facets of neurodevelopment in epidemiological studies and are a tool for estimating the potential effects of toxicants on the nervous system. When assessing the validity and reliability of results from a series of epidemiological studies, the specific psychometric tests utilized and factors affecting their administration in human populations present unique challenges. This report describes the historical application of psychometric tests to the study of the neurodevelopmental toxicity of methylmercury and defines neurodevelopmental domains that are assessed with these instruments. Principles are proposed for evaluating the validity and reliability of psychometric tests and for identifying potential sources of bias in the selection and administration of these tests in human populations.

Preface

Exposures to an increasing number of substances in our environment are being recognized as affecting human neurological development. The purpose of this report is to provide information to assist in determining if a given psychometric test is adequate for assessing a specific neurobehavioral domain or trait in human studies of neurotoxicants. Psychometric tests used in epidemiology studies that were reviewed as part of the U.S. Environmental Protection Agency's Integrated Risk Information System assessment of methylmercury were selected for review. The reviewed tests included those that were widely used in this literature, as well as those used frequently in studies of other neurotoxicants.

The psychometric tests were assigned to broad neurodevelopmental domains, and principles were developed and used to evaluate 81 psychometric tests for reliability, validity, normative data, and standardized methods for administering the tests. This report provides examples of factors that could present problems and introduce bias in the test results. The examples explain elements that would lead to different levels of bias and cover many aspects of study performance. We hope this information is useful to regulatory agencies charged with interpreting and evaluating the quality of epidemiology studies using these psychometric tests, and to the research community in designing epidemiology studies to assess factors affecting neurobehavior.

Introduction

This report presents basic principles for reviewing neurodevelopmental tests used in research that assess associations of exposure to known or putative neurotoxic chemicals during neurodevelopment. The emphasis is on the validity and reliability of psychometric tests used to assess neurodevelopment and methodological aspects of administering and interpreting them in epidemiological studies. Psychometric tests are used extensively because they provide a valid and noninvasive means to quantitatively assess brain function or dysfunction. Referenced throughout the document are psychometric tests designed to assess specific neurodevelopmental traits, and the scores obtained from these tests, which are used to quantify and compare the quality of performance on tests that assess these neurodevelopmental traits in and across individuals or populations.

The impetus for these principles and this document grew out of the need for assistance in assessing the quality of and potential biases in these tests when applying systematic review methodology to pediatric environmental epidemiology literature. Specifically for this document, this was done to aid in the development of the in-process U.S. Environmental Protection Agency (EPA) Integrated Risk Information System (IRIS) toxicological review of methylmercury (MeHg). To develop these principles, an evaluation was conducted on psychometric tests used to estimate the effects of MeHg on neurodevelopmental outcomes (see Process of Selecting Tests section). MeHg was a model toxicant to develop this document and associated principles because the developmental neurotoxicity of MeHg has been extensively studied for over 50 years in cohorts around the world, and the literature evaluating neurodevelopmental effects of MeHg captures many representative psychometric tests used in epidemiological research on many other neurotoxicants.

The work described here is important because it recognizes the challenges involved in conducting and reviewing population-based studies of neurodevelopmental toxicity due to a wide array of neurodevelopmental outcomes, the differing psychometric properties of available tests, and the variations in methods used to administer, score, and interpret test results in individual studies. To address this challenge, it is essential to understand the role that psychometric tests play in assessing neurodevelopment, the validity and reliability of these tests, and the least biased methods to administer, score, and interpret psychometric tests.

While some sub-disciplines of epidemiology, such as molecular epidemiology, have similar principles or “best practices” to maintain high validity and reliability by conducting external and internal quality assurance and quality control procedures (e.g., duplicate measurements, standard calibration materials, and interlaboratory proficiency programs) (Gallo et al. 2011), this document is not intended to serve as a set of best practices for the administration of psychometric tests when studying developmental neurotoxicants. Rather, its goal is to provide the necessary background information for understanding psychometric principles and how psychometric tests can be validly and reliability developed, evaluated, and employed when studying neurotoxicity in epidemiological studies. Ultimately, the principles proposed here are designed to assist in systematic reviews of epidemiological studies of potentially neurotoxic chemical exposures and aid in the design of future research studies.

This document is organized as follows. The first section (Psychometric Tests) provides background on psychometric tests and neurodevelopment. The second section (Methodology for Identifying Psychometric Tests and Extracting Test Information) describes the methods for identifying and selecting psychometric tests and extracting information on psychometric features of the selected tests. Then, the principles discussed in this document are presented in two main parts. In Part 1 (Principles for Evaluating Psychometric Tests), the neurodevelopmental domains and the instruments that have been used to assess neurodevelopment in MeHg research are described. Psychometric features, strengths, and weaknesses of these instruments are assessed with the goal of providing information that can be used to determine if an instrument is more or less suitable for addressing a research question of interest. In Part 2 ("Potential Sources of Bias Related to Selection and Administration of Neurodevelopment Assessments in Epidemiological Studies"), principles related to the application of psychometric tests are described along with potential sources of bias related to using the tests to assess neurodevelopment in epidemiological research.

Psychometric Tests

Background

The psychometric outcomes that were evaluated in this document derive from three traditions that have developed over the last 100+ years in the fields of psychology and psychometric assessment. These include the design and evolving development of quantified psychometric tests, the growth of the field of neuropsychological testing, and the exponential increase in neurodevelopmental assessment using psychometric tools in research studies (Lezak 1976).

Throughout this document, these tests are purposely referred to as "neurodevelopmental" and are used in the life-span sense of development/neurodevelopment. While the term "developmental" is often understood to refer to health and functioning in the prenatal and childhood periods, changes in brain structure, function, and organization begin at conception and continue across the life span. Thus, this document uses a life-span developmental psychology approach to characterize brain health and function from conception through death. Indeed, this approach aligns with the approach adopted by many epidemiological studies that were established to study the effect of early life neurotoxicant exposures on neurodevelopment from infancy through childhood and adolescence and into adulthood. Thus, tests are included and discussed in this document that assess adult function, as the administration of childhood tests would be inappropriate at older ages. This approach is consistent with the principle that very early exposures can affect brain structure, organization, and function across the life span and possibly in different ways at different points in the life span.

Psychometric test development began in the early 20th century when psychologists introduced the theory of *g* or general intelligence, a term for an individual's overall level of cognitive skills that affects their intellectual functioning. Early tests from Piaget, Binet, and Wechsler were developed to measure *g*. Although primitive compared with the sophisticated tests used contemporaneously, they included many features still used in contemporaneous tests, including several subtests; standardized instructions and test materials; scoring rules, and basic norms allowing for comparisons of performance among examinees of the same age, including child and adult populations. These tests (especially Stanford-Binet and Wechsler) gave rise to the concept of IQ as an entity, generally quantified as a standard score of 100 as average, with a standard deviation (SD) of 15 or 16, depending on the test. It should be noted that such scores are not meant to serve as the sole measures of an individual's aptitude or potential for future success given that they are influenced by a variety of factors (Spreen and Strauss 1991). Systematic differences in test performance may exist if individuals come from a population that is distinct from the test's target population or population used to design the test in terms of race/ethnicity, culture, socioeconomic status, etc. Thus, the scores are a kind of benchmark but are not a concrete entity.

Tests of academic achievement (reading, writing, mathematical skills) also began to appear in the educational psychology literature as school psychologists and personnel required means of assessing whether students were progressing at the expected rate in academic knowledge for age and grade level. These tests also grew in sophistication and became standardized over time.

Other tests that attempted to assess child development/neurodevelopment also began to appear that would allow pediatricians and other clinicians to determine if a child was on a normal

neurodevelopmental trajectory or was behind or ahead for his/her age. Tests were also developed to determine if children had typical patterns of cognitive, behavioral, emotional, social, psychiatric, or personality traits (Spreen and Strauss 1991).

Around the mid-20th century, the field of clinical neuropsychology emerged and grew dramatically for several decades. In this field, psychometric tests are used to assess brain function to determine if it is normal or abnormal. If abnormalities are noted, the neuropsychologist may opine on the specific brain structures or systems that are dysfunctional or their likely neurological cause. It had been known since the 1800s that some parts of the brain were responsible for specific kinds of behavior. Discoveries of these relationships stemmed from brain autopsy findings from symptomatic patients. For example, patients with lesions in certain parts of the frontal lobes developed the inability to speak (expressive aphasia), while patients with lesions in posterior portions of the temporal lobes could speak but could not comprehend the speech of others (receptive aphasia). As more associations between specific brain regions and behavioral and cognitive functions became apparent, the notion of structure-function relationships within the brain became more prominent. An impetus for this field was the absence of neuroimaging techniques that are now available by which clinicians could visualize brain tumors, strokes, and other kinds of lesions related to neurological illnesses (e.g., multiple sclerosis plaques, effects of traumatic brain injury). When patients presented to physicians with new acute neurological symptoms, it was important to diagnose how the brain was malfunctioning and which parts of the brain might be affected in order to determine whether a neurosurgical intervention might help or cure the patient (White 1992).

Although neurologists and other clinicians could assess behaviors to try to localize structural damage in the brain, the techniques were generally specific to individual clinicians and not systematically quantified or comparable among clinicians. Individual neuropsychologists and the field of neuropsychology applied standardized psychometric methods to the evaluation of specific aspects of brain function in attempts to determine whether a neurological condition might be present and associated with abnormal test scores, where the lesions/abnormalities in the brain might be located, and possible underlying diagnoses (e.g., dementias, seizure disorders). In fact, neurosurgeons sometimes relied on the lesion location diagnoses of neuropsychologists to treat patients. Some neuropsychologists approached this work by developing new tests designed to assess specific aspects of brain function. For example, the Halstead-Reitan group developed a battery of 10 tests that assessed motor function on both sides of the body, conceptual reasoning skills, working memory, and other skills. Other neuropsychological groups developed their own tests of brain function but also drew from existing psychometric tests that had already been developed, standardized, and normed to determine if these instruments could aid in assessing the functional integrity of the brain and locating lesions in specific brain areas or systems. Thus, the existing IQ tests and their subtests became included in this clinical and research endeavor. As time has passed, an extensive literature has developed describing what these various kinds of tests demonstrate about specific types of brain damage and neurological disorders, validating the tests applied as measures of CNS function and, in fact, as specific indicators of certain kinds of disordered (or functional) brain-behavior relationships. This knowledge continues to evolve given the capacity to “watch” the brain and its systems function as participants undergo tests and carry out behaviors through methods like functional brain imaging (White and Reuben In Press).

Initially, the field of neuropsychology was focused heavily on adults. This probably arose from the greater number of adult patients with neurological disorders. It also could be related,

however, to the fact that localized lesions with predictable behavioral anomalies are more common in adults. When a lesion or disorder appears in the developed adult brain, structure-function relationships are well established, and it is often clear where the problem might be. Neurodevelopment is more complicated. During child development, the brain is more plastic with regard to how specific structures mediate behaviors, is capable of compensating when a structure or brain area is diseased (or even absent), can be influenced by acute or chronic exposures during susceptible periods, and develops and expresses behavior in a dynamic fashion given the circumstances occurring at any stage of development. Because of all these considerations, structure-function relationships in early brain development are more diffuse and less "focal" than in adults, and insults to the developing brain—both toxicants and other neurological conditions—may have different effects than would insults on the mature brain. Given these circumstances, neurodevelopmental assessments have used a combination of existing tests for children (e.g., IQ, developmental, academic); adaptations of adult tests for children; and specialized tests that have been developed, standardized, normed, and validated in clinical populations (e.g., Developmental Neuropsychological Assessment, Behavior Assessment System for Children, Child Behavior Checklist, tests that assess clinical conditions in children) (White 2004).

Almost all psychometric tests provide raw scores that can be converted into standard or scaled scores (mean = 10, SD = 3), T-scores (mean = 50, SD = 10), standard scores (mean = 100, SD = 15), and corresponding percentiles using normative or reference data. This practice allows participants' scores to be compared with one another after removing the effects of age, sex, or other characteristics on each participant's raw score. For instance, even within narrow age intervals, older children have higher average raw scores on tests of mental and psychomotor development than younger children. By using standard scores, the traits of participants who are different ages (or other characteristics) can be compared. Moreover, an individual's performance over time can be examined in relation to reference or normative data. Such comparisons between groups or within individuals over time are often made in terms of number of points or in units of SDs away from the average score.

Note that normative data are not required for comparison of scores between individuals when appropriate measures are taken to validly analyze raw scores; however, normative data are necessary and routinely used when making decisions about an individual's health care, clinical diagnosis, vocational services, education, legal status, etc. In addition, when appropriate normative data are available, they can be used as the outcome in statistical models.

There is a longstanding tension regarding the interpretation of results from population-based studies that examine neurodevelopmental toxicity. Generally, this friction arises from the lack of consensus regarding the "clinical significance" of effects observed in epidemiological studies. In clinical situations, it is appropriate to refer to "abnormalities" or "deficits" in a functional area when an examinee performs poorly on a test (usually 1–2 SDs away from the average score). In epidemiological studies, effect sizes often fall short of being clinically significant in that they occur within the "normal range" of test performance and are typically less than half of an SD away from the average score. However, subtle shifts in a continuous trait can have profound impacts on the tails of a distribution in a population (Needleman 1990; Rose 1985; 2001; Weiss 2000). For instance, a 5-point decrease in a population's IQ would nearly double the number of people classified as intellectually disabled (Braun 2016). Finally, it is important to refer to subclinical or preclinical findings of lower scores in these situations as "decrements" in

performance or "diminished" performance, rather than "deficits" or "impairments" (White and Reuben In Press).

Test Domains

In the fields of neurodevelopmental psychometric tests and neuropsychology, psychologists often talk about "domains" of behavior. These domains are generally highly functional in nature—that is, they focus on a particular kind of activity such as attention. They are helpful in some ways simply because they allow investigators to group tests into meaningful categories. However, some other issues about the domains are important to consider, which are discussed below.

While these descriptions of neurodevelopment traits are presented as if they are isolated attributes, it is acknowledged that they are dimensional traits that interact with other domains in determining behavior. Indeed, this type of framework (e.g., Research Domain Criteria) has been adopted by the National Institute of Mental Health to complement the Diagnostic and Statistical Manual approach to characterizing mental health and neurodevelopmental disorders (Morris and Cuthbert 2012).

First, domains roughly map onto the functioning of specific brain regions, but there is generally not a one-to-one correspondence between a domain and a brain region. For instance, while the capacity to pay attention and to monitor or inhibit behaviors is strongly related to functioning of the frontal lobe, other brain regions (e.g., striatum) may also be involved in these behaviors. Similarly, learning and memory functions are mediated by the limbic system, especially the hippocampus and related structures, although other brain structures play a role. Second, many psychometric instruments that are applied in the fields of neurodevelopment, neuropsychological assessment, and developmental neurotoxicity are omnibus tests for which performance is determined by many different kinds of behavioral processes at the same time and do not fit neatly into a functional domain.

Related, some individual tests require highly complex integration of many kinds of functions, and even tests or subtests considered to be domain-focused often rely on several brain regions for successful performance. These kinds of tests may better fit under a category such as "multi-determined tests," but this is difficult to do because all tests rely on more than one type of ability (i.e., no test is a pure measure of the trait). Interestingly, the tests with many functional demands are often sensitive to an insult such as a neurotoxicant exposure, although they generally do not reveal very much about what portions of the brain the insult has affected. For example, coding tests require respondents to look at a code that pairs symbols with digits and to write in the appropriate digit in a blank space below the symbol according to the code. This task requires an examinee to recognize visual symbols, write quickly and accurately, scan visual arrays quickly, form associations to remember pairs of symbols, inhibit interference from outside stimuli, follow rows and columns, and so on. These multiple determinants of response quality affect the ultimate score—a deficit in any one of them can reduce the score. This is less true of tests that require only paying attention, or writing quickly, or remembering data.

Finally, there are subdomains associated with each domain. For example, "attention" can include the ability to recognize stimuli, capacity to ignore irrelevant stimuli, speed of responding to stimuli, ability to repeat back numbers in order, and so on. Verbal tests can include the ability to

provide abstract definitions of vocabulary words, comprehension of spoken language, and/or comprehension of written language. Toxicant-related decrements in performance within each of these subdomains can have different implications for the effects of the exposure on specific structures or systems within the brain during neurodevelopment.

The domains defined here refer to those that were used to evaluate the psychometric outcomes included in the neurodevelopmental research under discussion. Appendix A contains a list of the domains and the tests categorized within each domain (White 2004).

Omnibus Tests

General intelligence/IQ: Tests measuring general intelligence purport to assess the examinee's overall level of cognitive or intellectual abilities, usually by applying a variety of types of intellectual challenges.

Academic achievement: Tests in this domain evaluate the child's ability to carry out academic skills such as reading, spelling, vocabulary, arithmetic, and more complex abilities.

Developmental: These tests assess how successfully an infant or child is acquiring age-appropriate verbal, motor, and social skills.

Neuropsychological assessment batteries: These tests assess a variety of the domain-specific functions described below and may or may not include an overall test score (usually they do not).

Clinical Assessment Instruments

Clinical conditions: These tests assess a specific diagnostic outcome or set of outcomes (e.g., autism spectrum disorder, attention-deficit disorder, anxiety disorders, neuropsychiatric conditions) to produce a criterion-based clinical diagnosis.

Mental status: Mental status examinations are screening tools used to determine whether an individual's cognitive function is within expected limits for age.

Domain-specific Tests

Attention: This domain includes evaluation of capacity to monitor incoming stimuli, inhibit responses to irrelevant stimuli, hold small bits of information for immediate use ("pre-memory"), and listen to instructions and communications. The prefrontal cortex prominently mediates performance on these tasks.

Executive function: Executive function refers to the capacity to manipulate complex stimuli, reason abstractly, develop effective strategies for task completion, and problem solve. It is broadly defined by three subdomains: working memory, cognitive flexibility, and inhibition. Working memory refers to the capacity to hold and simultaneously manipulate information and data from stimuli for task completion. Cognitive flexibility is the ability to adjust behavior in the face of changing demands and goals. Inhibition includes both the ability to ignore irrelevant stimuli or suppress a triggered behavior to sustain efforts to complete a goal. Relevant brain structures include the prefrontal cortex and subcortical white matter connections.

Motor function: Fine motor control, speed, accuracy, and coordination are the key components of motor function. This domain is usually assessed using the hand (manual motor skills). Relevant

brain structures that support completion of these tasks include the motor cortex of the frontal lobes, the cerebellum, and the extrapyramidal system (e.g., basal ganglia).

Learning and memory: Learning and memory comprise several processes that include coding visual or verbal information into short-term memory stores, retaining information longer term, recalling newly learned information spontaneously, and recognizing newly learned information that may or may not be recalled spontaneously. Retrograde memory refers to the ability to recall information learned in the more distant past. Brain structures related to learning and memory include the limbic system, specifically the hippocampus, and frontal cortex.

Social-emotional: This domain encompasses expressions of affect or mood (temporary or chronic), including anxiety, depression, and irritability; ability to control strong emotions or reactions to events or people; communication patterns; traits such as tendency to externalize (often associated with attention-deficit/hyperactivity disorder [ADHD] diagnosis) or internalize (often associated with diagnosis of depression); reciprocal social behaviors, repetitive and restricted behaviors (often associated with autism spectrum disorder [ASD]); and personality traits. Relevant brain structures include the prefrontal cortex and limbic system.

Verbal/language: Verbal skills include recognition of word meanings, ability to define vocabulary words abstractly, and comprehension of verbalizations. This domain is sometimes subdivided into verbal and language, with language skills referring more directly to skills associated with aphasia, such as confrontation naming, repetition, simple comprehension, and simple writing and reading. Most aspects of language/verbal function are mediated by the dominant cerebral hemisphere (usually left hemisphere, though aspects of complex interpretation of verbal information and appreciation of verbally expressed humor often involve the nondominant hemisphere).

Visuospatial function: This domain includes the ability to evaluate pictures and drawings with missing details, appreciate and replicate visual designs and drawings, recognize gestalts, detect embedded figures, understand maps and navigate directions, perform facial and abstract design recognition, and complete puzzle and block design assembly. (Tests within this domain sometimes have a significant motor component.) Relevant brain structures include the parietal and occipital lobes, the cerebellum, the extrapyramidal system, and subcortical white matter connections.

Processing speed: Time to complete tasks is assessed by this domain, which in recent years has been added to IQ and other omnibus tests. No parent tests assessing processing speed were identified in the epidemiological studies from the in-progress EPA IRIS toxicological review of MeHg, resulting in no evaluation table for this domain in Appendix B. Subtests belonging to omnibus tests that assess processing speed have been identified and listed for this domain. Relevant brain structures include the frontal and prefrontal cortex, cerebellum and basal ganglia, and the white matter connections within the brain.

Methodology for Identifying Psychometric Tests and Extracting Test Information

This section describes the methods for identifying and selecting psychometric tests and extracting information on psychometric features of the selected tests. Test selection and information extraction were conducted to support an evaluation of the psychometric features of each test to determine the adequacy of tests in assessing neurodevelopmental or central nervous system (CNS) function in studies of neurotoxicants. Test selection and information extraction were conducted first and are described here. The test evaluation process and test evaluation results (found in Appendix B) are referenced in this section but are described in detail in the Principles for Evaluating Psychometric Tests section of the document.

Process of Selecting Tests

Specific psychometric tests for evaluation were selected and information on relevant test features was extracted using a set of studies identified from the systematic review and dose-response analysis for the in-progress EPA IRIS toxicological review of MeHg (see protocol for more details https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=345309; see Introduction for description of the focus on the MeHg literature in this document).

Peer-reviewed and published epidemiological studies compiled by EPA by July 2019 as part of the in-progress EPA IRIS toxicological review of MeHg were reviewed and a list of psychometric tests was developed (n = 134 tests) based on these studies. Next, a multistage process was implemented to identify information to extract for describing the psychometric features of the identified tests. If accessible, physical manuals or electronic copies of test manuals for the identified tests were reviewed as a first step, as these manuals provided the best available primary sources of information. As a secondary source of information, several editions of two academic textbooks on neuropsychological testing, *A Compendium of Neuropsychological Tests* (Spreen and Strauss 1991; 1998; Strauss et al. 2006) and *Neuropsychological Assessment* (Lezak 1995; Lezak et al. 2004; Lezak et al. 2012), were manually searched to identify relevant test features in those sources. If access to a test manual was available and/or usable information was identified in these academic textbooks, further information was generally not sought. Tests with information from manuals or academic textbooks were automatically included in the extraction and evaluation process.

In certain cases, when a physical or electronic copy of a manual or information from one of these academic textbooks was not available, peer-reviewed literature (original research and literature reviews) was identified by searching online journal databases for the test name, any related abbreviations, and relevant keywords (e.g., “psychometric,” “valid*,” “reliability”). EBSCO host was used to search multiple online databases, including Medline, CINAHL Complete, PsycARTICLES, PsycINFO, PsycBOOKS, and PsycEXTRA. Titles and abstracts were screened, and full-text articles were obtained if the abstract discussed psychometric properties, factor analyses, or comparisons with other neuropsychological tests. Studies on groups of people with specific conditions that may affect test performance (e.g., deaf patients) were not included. Ultimately, tests were included in the extraction and evaluation process if enough information was available to assess a majority of the evaluation principles (see Principles for Evaluating Psychometric Tests section) following the steps outlined above. After retrieving sources of

information for each test, relevant information was extracted and summarized in an Excel spreadsheet (titled "DNT_Test_Information_Extraction_Database.xlsx"; referred to in this document as the "extraction table"; Appendix C).

Some tests were excluded from the extraction and evaluation process due to idiosyncratic features or insufficient information. Tests were identified as idiosyncratic if they appeared in a single MeHg study, were used only in a specific population, and/or were only used in studies later determined to not be conducive to dose-response analysis for the in-progress EPA IRIS toxicological review of MeHg. Tests were categorized as having insufficient information if no manual was available and secondary sources, including peer-reviewed literature, did not provide information to assess a majority of the evaluation principles.

Of the 134 tests initially identified, 81 were included in the extraction and evaluation process (see Appendix A for the lists of included and excluded tests). Thirty-three tests with idiosyncratic features and 20 tests with insufficient information were excluded from the extraction and evaluation process. The number of included tests by information source(s) included:

- Eight tests with information from manuals only
- Twenty tests with information from manuals and academic textbooks
- Two tests with information from manuals and peer-reviewed literature
- Eighteen tests with information from academic textbooks only
- Three tests with information from academic textbooks and peer-reviewed literature
- Thirty tests with information from peer-reviewed literature only

The included tests (n = 81 tests) were organized into broad domains based on a framework previously developed by the co-author Dr. Roberta White (White 2004; White et al. 2009; White 2011). These broad domains included omnibus tests (intelligence quotient [IQ], academic achievement, developmental, and neuropsychological assessment batteries); clinical assessment instruments (clinical conditions and mental status assessments); and domain-specific functional tasks (attention, executive function, learning and memory, motor function, social-emotional, verbal/language, and visuospatial tests). Note that some tests may assess multiple domains.

Subtests and subscales within tests that were used in the epidemiological studies from the in-progress EPA IRIS toxicological review of MeHg were also identified; however, information extraction and test evaluation were conducted at the parent-test level and not for specific subtests, given resource constraints that inhibited the assessment of specific features of subtests. While each test evaluated in this document has been categorized by domain, there are cases for which subtests/subscales within omnibus tests assess domain-specific functions. In these cases, for each domain, the identified subtests or subscales are listed in a footnote below their respective domain-specific evaluation table in Appendix B. Please see the Principles for Evaluating Psychometric Tests section for further detail on the evaluation process, the evaluation results (Appendix B), and a description of the how the evaluations may be applicable for these cases.

Test Information Extraction

The data in the extraction table (Appendix C) for each test include the following:

- Test name and any alternative names
- Test domain
- Test publication date
- Time required for administration
- Appropriate age range for test administration
- Original publication language(s)
- Availability of the test in other languages
- Availability of culturally adapted version(s) of the test
- Source population or culture from which the test was developed
- Test reliability (internal consistency and test-retest)
- Test validation (content, construct, criterion)
- Sensitivity and specificity of the test
- Description of quantitative outcomes provided by the test
- Test standardization or normative data
- Applicability for use as a screening tool for clinical diagnoses
- Training requirements and qualifications for test administrators, and the availability of specific instructions or a test manual
- Appropriate test environment

The sample populations used to develop the test and its norms were described in the extraction table (Appendix C) to the extent possible based on the source material. Direct quotes from the sources of information were extracted when appropriate. If no information was found for an extracted topic, it was recorded that no information was present in the available source materials. The extraction table was used to inform the test evaluation process. During the peer-reviewed literature search process, supplemental information was found for several tests for which information from manuals or academic textbooks had already been extracted. In these cases, this supplemental information was added to the corresponding test extractions.

Potential Limitations of Test Selection and Data Extraction Approach

One limitation to the selection of tests is that the scope was narrowed to studies identified from in-progress EPA IRIS toxicological review of MeHg. Studies of other well-characterized neurotoxicants (e.g., lead, organophosphate pesticides) might yield additional tests that are relevant to assessing developmental neurotoxicity. While other well-studied neurotoxicants could have been included (e.g., lead), they were beyond the charge's scope for developing this document. The approach used to select studies and extract data on properties of psychometric tests was limited by a lack of information for 20 of the 134 tests initially identified from in-progress EPA IRIS toxicological review of MeHg. In addition, conducting complete extraction and evaluations at the subtest level for the original 81 included studies was beyond the scope of this effort. Thus, it is possible that subtests may be rated differently from their parent test. Moreover, because the selection of tests was based on the tests used in epidemiological studies identified for in-progress EPA IRIS toxicological review of MeHg, test selection may not reflect the latest versions of tests used for studying other neurodevelopmental toxicants. Despite these limitations, the tests selected and evaluated in this document are considered to be a representative selection of tests for effects in the associated domains that would be applicable for studying an array of neurodevelopmental toxicants, and the principles for evaluating the tests could be applied to future research.

One issue that limited access to test-specific information was that information was often not available for older tests or for those tests that have not been used extensively in research. Thus, in general, lesser-used tests and older tests had less information or a poorer quality of information related to the features included in the extraction table (Appendix C). In addition, the sources of information varied among included tests (e.g., test manuals available for some tests and peer-reviewed literature only for others), which led to variation in the availability and quality of information. As a result, the features of individual psychometric tests were not systematically evaluated using the same types of information sources. The more commonly used tests and those with multiple and contemporaneous editions (e.g., the omnibus IQ tests such as the Wechsler, Stanford-Binet, and Kaufman scales) were originally developed and evaluated using robust processes conducted by the test developers (often commercial entities). Thus, they may have more complete information because of their well-funded, rigorous development and the detailed technical manuals that accompany the tests. This does not mean that other tests lack reliability or validity, only that information on these tests may be more difficult to obtain. In some cases, the manuals of tests with multiple versions did not contain information on specific topics included in the extraction table when a specific version was being reviewed. In these cases, data from manuals for other editions of the test were used or data were reported that relied on the expert knowledge and judgement of the evaluator, Dr. Roberta White. When data were not available from manuals or other literature (and therefore not reflected in the extraction table), evaluator expert knowledge and judgement were utilized, and such ratings were noted in the evaluation tables.

Data Availability

Data relevant for evaluating neurodevelopmental tests in epidemiological studies are included in the extraction table available in the NTP Chemical Effects in Biological Systems (CEBS) database: https://doi.org/10.22427/NIEHS-DATA-NIEHS-01 (NTP 2022).

Part 1: Principles for Evaluating Psychometric Tests

The purpose of this section is to provide principles that can be used by scientists and regulators to determine if a psychometric test is adequate for assessing neurodevelopment or CNS function (including specific neurobehavioral domains or traits) or to aid in the selection of psychometric tests for research studies. While this document emphasizes developmental neurotoxicity studies of chemical exposures, these principles could be extended to other exposures (e.g., psychosocial stress, nutrition, etc.). The major principles for evaluating psychometric tests, which are described below, are those commonly used by psychometrists for this purpose. They include specific methods for evaluating aspects of the four overarching psychometric criterion areas: reliability, validity, standardized administration methods, and normative data associated with specific tests. It is critical to note that these criteria apply only to the features of the psychometric tests themselves and not to the application of the test in a research setting. (Part 2 of this document proposes criteria for test application.)

The assessment of test-specific aspects of reliability, validity, standardized test administration, and normative data for the tests featured in Appendix B were based on information derived from a combination of sources described in the Introduction to this document and summarized in the extraction table (Appendix C). Evaluation of specific aspects or subcriteria for each of the four approaches to understanding the psychometric integrity and properties of the tests was completed independently by both of the evaluators (Dr. Roberta White and Dr. Joseph Braun) using the data summarized in the extraction table.

The ratings for each subcriterion were the following: adequate, deficient, not applicable, or not present (i.e., not enough information available to the evaluator). For the normative data subcriteria, separate ratings were determined for adults and children (adulthood was defined as beginning at age 18 years). It should be noted that the evaluative ratings used do not necessarily dictate whether or not a test is appropriate for an individual study. For example, if a test or outcome is being used to evaluate a specific brain function in a unique population, but the test lacks adequate population norms, it can be used if its raw scores are appropriately analyzed. In addition, some criteria were difficult to rate because, in some cases, multiple sources of information or studies in the peer-reviewed literature on a test were consulted that varied considerably in quality or level of detail or contained different results across studies. Once independent ratings were determined by both evaluators, a consensus meeting was held to finalize them through discussion between the evaluators. Discussion was needed to arrive at a consensus rating for approximately 20% of the ratings. A final review of ratings was completed by the evaluators to ensure consistent application of evaluation criteria. Additional test descriptions and rating justification notes were added to the evaluation tables when needed. Explanatory notes for rating justifications were added for each instance of a deficient or not present rating and for some adequate ratings that were not clearly derived from the material provided in the extraction table (Appendix C). When data were not available in manuals or other literature (and therefore not reflected in the extraction table), the evaluators based their ratings on their own knowledge and noted this in the evaluation tables.

While all four psychometric criterion areas (reliability, validity, standardized administration methods, and normative data) are important in evaluating psychometric tests, it should be noted that reliability, validity, and standardized administration methods are considered most important

in selecting psychometric tests for research studies and in determining the adequacy of psychometric tests to assess neurodevelopment or CNS function in epidemiological research. It is recommended that scientists and regulators consider the strength of the normative data only for tests that are considered adequate regarding reliability, validity, and standardized administration methods.

Given the above considerations, the evaluation ratings for normative data are presented separately from the ratings for reliability, validity, and standardized administration methods because the adequacy of normative data is most relevant once a valid and reliable test has been developed. Moreover, adequacy of normative data is only applicable in epidemiological studies that use normative scores as outcomes rather than raw scores. In addition, many domain-specific tests are used to test hypotheses regarding specific skills and abilities to assess specific brain systems and are often not developed with the same resources as larger omnibus tests, resulting in limitations to or a lack of normative data. These tests can still be valid and reliable, but the scores they produce might need to be adjusted for age, sex, or other factors predictive of raw scores.

Appendix B contains the evaluation ratings and notes for the tests by domain. For each domain, the ratings for reliability, validity, and standardized administration methods are provided in one table, and the ratings for normative data are provided in a second table. Subtests and subscales within omnibus tests assessing domain-specific functions that have been identified in epidemiological studies from the in-progress EPA IRIS toxicological review of MeHg are listed below their respective domain-specific evaluation tables. The evaluation tables provide the publication date of each test, as age of the test at the time a study was completed can be an important factor in considering whether a test was appropriately applied. For example, older tests may contain items or questions that no longer persist in general knowledge (e.g., naming outdated technology or household items, such as a record player). However, because test age is not static (i.e., it depends upon lag time between date of test and date of study, as well as reasons investigators chose the test), evaluation criteria were not developed for this variable. Factors related to the age of a test at the time it was employed in a study are considered in some detail in Part 2 of this document.

This document does not provide an overall designation of adequate or inadequate (or any other ranking) for specific tests. The complexities of choosing, applying, and interpreting neurobehavioral methods in research settings prevents simplistic summary evaluations. Some users of this document may consider making this designation when they are trying to determine if a study using a given test will be included in a meta-analysis, if the results related to a test are to be used for policy decisions, or if the test will be administered as part of a research study. Thus, the relative importance of the four criteria (and subcriteria) in making these types of decisions will differ with the goal of the end user. For example, researchers selecting a test for administration in a research study might weigh the availability of specific, applicable normative data more heavily than the other domains because they are conducting a study in a culturally unique population. As another example, scientists selecting tests for inclusion in a meta-analysis of a specific neurobehavioral domain might place more weight on the validity of a test if they want to ensure that only results from tests accurately measuring the specific domain are included.

Some caveats to applying the criteria in this document should be noted. First, designations of adequate or deficient are applied to the criteria without additional gradations. This is because there are standards available to designate a test as adequate or deficient for some aspects of the

psychometric criteria, but additional gradations are not available or widely used (White and Proctor 1992; White et al. 1994). While alternative methods (e.g., risk-of-bias analysis) could provide finer gradations of each criterion, systematic approaches that could do this for psychometric tests are not available. Thus, a binary designation allows users to determine if a given test meets the criteria as described below in a reasonable enough fashion that it would be acceptable for use in population-based research. This approach is consistent with clinical and research practice as some psychometric tests are used in clinical or research settings when the test has known inadequacies. For instance, this situation can arise when there are no better alternatives for assessing a given domain or when a test assesses a highly specific cognitive process.

Many psychometric tests include both an omnibus assessment of a neurobehavioral function or successful neurodevelopment as well as subtests assessing specific domains related to that function. Therefore, some specific summary or subscale scores within an omnibus test may be adequate or deficient while others are not. In general, focusing on the summary scores (e.g., IQ measures, domain summaries) from tests is recommended for most purposes. In cases for which a test provides multiple domain or trait scores but no summary measure(s), using the overall pattern of adequacy/deficiency across domain or trait scores is recommended to determine the adequacy of a given criterion for the test as a whole. If the goal of the user is to apply a limited number of an instrument's subtests (one or more) to assess specific domain functioning, only information relevant to the subscale(s) of interest should be considered by the user. This information is generally available in test manuals and can also be found in the peer-reviewed or gray literature.

The major principles as noted above for evaluating psychometric tests (reliability, validity, standardized administration methods, and normative data) are described below.

Reliability

For a psychometric test to be reliable, its results should be consistent across time (test-retest reliability), across items (internal reliability), and across raters (inter-rater reliability). Part 2 discusses inter-rater reliability of the document because it is not an intrinsic feature of a test. Thus, internal reliability demands that the individual items on a given test should measure the same domain(s) or trait(s) (i.e., internal consistency). Reproducibility, or test-retest reliability, requires that consistent scores would be obtained from the same individual upon repeated testing.

To assess the internal reliability of a test, items within the test should be correlated with each other to ensure internal consistency. To assess the test-retest reliability of a test, it should be administered in a standardized manner to the same person twice, and the score(s) from the repeated measurements should be consistent.

When assessing internal consistency, a high correlation among items on domain-specific subscales indicates that the test items measure the same trait (e.g., as indicated by having a high split-half reliability). The most popular criterion used to assess internal consistency was developed by Sattler (2001). He recommended that tests with reliability coefficients <0.6 (e.g., correlations mentioned above) be deemed unreliable. Moreover, for research purposes, Sattler (2001) suggested that tests with reliability coefficients ≥ 0.6 and <0.7 be considered marginally reliable and those with coefficients ≥ 0.7 be considered relatively reliable.

For test-retest reliability, high correlations between repeated administrations of a test to the same person within an appropriate time interval ensures that the test can consistently measure trait(s) assessed by the instrument in an individual. Test-retest reliability is generally assessed by intraclass correlation coefficient (ICC; ideally >0.4), Pearson correlation coefficient (ideally >0.3), or Cohen's kappa coefficient (>0.4).

Determining adequacy: When evaluating a given psychometric test, it must have internal consistency reliability coefficients of ≥0.6 (e.g., Cronbach's alpha, ICC) to be considered "adequate." Test-retest reliability should meet one of the following criteria as indicated above: ICC >0.4, Pearson correlation coefficient >0.3, or Cohen's kappa coefficient >0.4. Some deviations for subtests are acceptable if summary scales or the majority of subtests are at least marginally reliable.

Validity

Validity is typically assessed across three broad domains: content, construct, and criterion validity. Each is distinct but ultimately all are related to a test's ability to measure what it is designed to measure. It is critical to note that validity is not a static, "all or none" metric and is re-evaluated as a test is used in varied clinical practice and research settings over time.

Content validity is the extent to which the test items, tasks, and questions assess the trait that the test is designed to measure. This can be thought of as a sampling issue, wherein the test content should be representative of the population of all possible test content that could measure that trait. Content validity is assessed by evaluating test themes, theoretical models, scientific evidence supporting a test, domain definition, domain operationalization, item selection, and item review. Review of content validity is often qualitative in nature and relies on expert evaluation and judgment; however, quantitative techniques like factor analytic approaches are often used to refine test content and confirm content validity.

Construct validity is the degree to which the test estimates the trait of interest using the items selected for the test. It usually pertains to complex traits (e.g., intelligence). Note that a construct is theoretical and requires accumulation of evidence from several sources beyond correlation of tests purported to measure constructs such as intelligence. Construct validity is evaluated with formal construct definitions, correlations with other tests that measure the same (convergent validity) and different (divergent validity) construct(s), and factor analysis. Construct validity is quantitatively assessed using results (typically correlation coefficients) from well-designed studies that administer the test of interest to normative and clinical samples of individuals. There are no strict thresholds to establish construct validity, but minimum correlation coefficients of 0.3 have been proposed (Lezak 1995; Lezak et al. 2004; Lezak et al. 2012). Correlations with related tests reflect convergent validity, while relationships to tests that measure other traits should be low, establishing discriminant validity.

Finally, criterion validity assesses the ability of a psychometric test to predict an individual's performance or outcome now (concurrent validity) or in the future (predictive validity). It requires identification of an appropriate criterion for comparison (e.g., clinical disease related to the trait), assessment of the test and criterion, calculation of classification accuracy, or correlation with other tests/criteria.

Determining adequacy: Content, construct, and criterion validity should be separately evaluated for each test.

(1) Content validity: Qualitatively determine that the test is theoretically grounded, had item content appropriately identified from a large item pool that was expertly judged and curated, and has defined and theoretically justified domains. Factor analysis can be used to confirm that included items are specific to the domain(s) of interest.

(2) Construct validity: Must show validity through positive correlations with other measures of same construct or similar test (i.e., convergent validity). Ideally, the test should not be correlated with unrelated constructs (i.e., divergent validity). Factor analysis can be used to support any summary or subtest scales.

(3) Criterion validity: Criterion should be well defined; must be reasonably accurate in association with or for predicting criterion (e.g., kappa > 0.6).

Standardized Administration Methods

Psychometric tests must be administered in a rigorous and standardized fashion. This precision is critical in population-based studies when groups of participants with different levels of exposure are being compared with one another, as non-standardized administration could introduce random or systematic bias. When comparing results from one study with another, it is also critical to ensure that data were collected in the same fashion (i.e., the studies carried out the same test in the same way).

Well-designed psychometric tests include explicit guidelines regarding test material presentation/organization, instructions to participants, instructions to test administrators on scoring participant responses and calculating test scores, and explicit phrasing for oral instructions and/or verbal questions. Some psychometric tests use stimulus material (e.g., pictures, blocks), and the same standardized materials must be used across test sessions to ensure consistent responses from subjects. In addition, the materials used in the test should be identical to those described in the test administration manual or provided by the publisher of the test.

Finally, psychometric tests often require administration by trained personnel or supervision by a clinical psychologist, neuropsychologist, or other appropriate professional. The interpretation of test results or feedback to parents/guardians and affected communities must be conducted by persons with professional credentials appropriate to the outcomes and the setting in which the study is conducted. The required qualifications of the test administrator should be indicated in the test manual or a document of standardized test procedures. An exception to this can be self-administered questionnaire instruments that are completed by individuals about themselves (self-reports) or others (teacher or parent ratings of children).

Determining Adequacy: The following rules should be used to evaluate the adequacy of a test's administration instructions. Part 2 notes specific administration factors relevant to studies of developmental neurotoxicity.

(1) The test must have a manual or published paper that provides explicit and clear instructions on how test materials should be administered, how responses are scored, and how normative scores are calculated.

(2) Tests that use stimulus materials should include standardized materials for administration.

(3) Test manuals should explicitly state the qualifications necessary to administer a test, with the exception of questionnaire instruments.

Normative Data

Almost all psychometric test manuals provide normative data that allow conversion of participants' raw scores into scaled scores (mean = 10, SD = 3), T-scores (mean = 50, SD = 10), or standard scores (mean = 100, SD = 15) with corresponding percentiles. These converted scores and percentiles are calculated based on data from a reference (or normative) population. Typically, test developers administer the test to a sample of hundreds or thousands of participants drawn from the target population of interest; normative scores, as well as corresponding percentiles, are derived from these individuals. This scoring is often conducted by calculating means and SDs of raw scores for specific ages of children or adults.

In culturally distinct populations for which test items and raw scores are determined to be valid and reliable, normative data are not necessary. Thus, the adequacy of normative data does not need to be assessed. Raw scores or study-specific normative scores can be used in statistical models when certain assumptions are met, and appropriate statistical techniques are used. The nature of some raw scores may preclude using them as the outcome in regression models. For instance, the Bayley Scales have infants or children complete a different set and number of items based on their age. Thus, the raw scores may not be equivalent across individuals. While a variety of methods could be used to create new scores (e.g., summed scores, PCA-derived scores), they have various strengths and limitations (McNeish and Wolf 2020). A full evaluation of these methods is beyond the scope of this document.

It is important to note that the sample size and representativeness of normative data for some domain-specific tests of neurobehavioral function are smaller and less generalizable, respectively, than those for commonly used omnibus tests such as intelligence tests or tests of overall neurodevelopment that have been developed by large psychological service companies.

The importance of adequate normative data for tests depends heavily on why the researcher is utilizing the test and how the outcomes are scored. Some tests, especially those that assess highly specific neuropsychological functions, are used because they allow for evaluation of specific relationships between structural brain function and a predictor such as exposure to a toxicant. Other tests are applied to an experimental situation because no standardized tests are available for the population being evaluated. In these situations, and in other circumstances when the normative data available for a test are not appropriate according to the standards listed below, the instrument may be legitimately applied, but outcome data must be appropriately analyzed. For example, raw scores might be adjusted for relevant confounders such as age, gender, educational attainment, or parental education. Several features of a test's normative data should be evaluated.

(1) First, differences in native language, even different dialects, can affect an individual's performance on a psychometric test. Thus, normative data should ideally be derived from participants with the same language and, if possible, the same dialect.

(2) Different cultures, races, and ethnicities may be exposed to different educational materials or have different socioeconomic backgrounds. These factors can affect test

performance. Thus, normative data should be representative of these subgroups. In some cases, researchers develop normative data for which a test is adapted to specific cultures, languages, or subgroups. When examining the normative data associated with a test, it is important to consider the sample sizes for subgroups (e.g., racial/ethnic minorities), as this affects the precision of normative data for these subgroups.

(3) Almost all psychometric tests, particularly those administered to infants and children, are age-standardized to account for age-related neurodevelopment. Thus, age-specific normative data should be available for specific age groups. Moreover, the age ranges within the age bins used for standardization should be examined to ensure that they are granular enough and include data from enough children to accurately capture age-related differences in neurodevelopment.

(4) The process for generating normative data should be systematic in terms of participant recruitment, representativeness, test administration, and score/percentile derivation.

Determining adequacy: The following criteria should be met for a psychometric test to have adequate normative data.

(1) Normative data should be based on sample sizes of at least 1,000 for omnibus tests and should adequately represent the population for which the test was intended. Smaller sample sizes may be appropriate for domain-specific tests. (In the evaluation tables, a sample size of 250 was considered adequate for domain-specific tests.)

(2) Normative data should be appropriate for the culture and language of the participants to which the test is being administered.

(3) Normative data should be derived in a systematic fashion and not from convenience samples.

(4) Age-specific normative data should be derived. Adequacy of age-specific information available for tests is judged by several factors. The age-specific norms should be appropriate for the population to which the test is administered and of an adequate size to derive stable means and percentiles—measured in weeks or months for infants, months for younger children, and years in older children (late adolescence) and adults. Age bands in adulthood should not be too wide (e.g., greater than 5 years) in later adulthood, when declines can occur for many tests. In addition, the number of participants included in the determination of the age-specific norm should be adequate; generally, this ranges from 30 to 100 depending on the kind of test. Large population omnibus measures such as IQ tests should average about 100, whereas 30 may be adequate for domain-specific tests. Finally, the age bins or age ranges used should be appropriate for the trait being evaluated. For example, tests of cognitive abilities generally require much narrower age bands than tests of social/emotional traits. When evaluating the age-specific normative data, the evaluators considered all three criteria in determining adequacy.

Part 2: Potential Sources of Bias Related to Selection and Administration of Neurodevelopmental Assessments in Epidemiological Studies

The results of neurodevelopmental assessments can be biased by factors related to the attributes of psychometric test(s), features of the study participants, and/or aspects of test administration. Outcome misclassification resulting from measurement error of an outcome by a psychometric test or clinical diagnosis can introduce systematic or random bias depending on whether it is related to the exposure of interest. Non-differential (i.e., random error) outcome misclassification occurs when the bias is unrelated to the exposure of interest and is expected to attenuate the association between the exposure and outcome toward the null. Differential (i.e., systematic error) outcome misclassification occurs when the bias in the outcome is related to the exposure of interest and is expected to change the direction and magnitude of the observed association relative to the true association.

To illustrate these two types of biases, imagine a study estimating the effect of MeHg exposure on children's IQ that uses two geographically separate populations similar in all regards, except that one population consumes high levels of MeHg-contaminated fish and the other does not. The investigators in this study intend to use study location (i.e., MeHg-contaminated fish consumption) as a proxy of MeHg exposure. A single researcher assesses the IQ of children at the two study sites. Furthermore, imagine that consumption of contaminated fish causes reductions in child IQ. Non-differential misclassification could arise if the researcher inadvertently scores some children's IQ two points higher than they should have at both study locations. In the absence of other biases, this random error creates "noise" in the data and would be expected to attenuate the association between MeHg (i.e., study location) and IQ toward the null. Differential misclassification could arise if the researcher scored children's IQ two points lower at the study location that consumed MeHg contaminated fish but did not do so at the other study location. In the absence of other biases, this systematic error would be expected to cause an overestimation of the association between MeHg and IQ.

An overview is presented below that describes how various factors could influence psychometric test performance or scores. Details are provided about the features that a study would have to include to minimize the influence of these factors. Central to evaluating the potential bias from outcome misclassification, criteria are presented for deciding whether a study adequately addresses a given factor that could create bias (Table 1). The specific evaluation levels that were used are *very low, low, moderate*, and *high* likelihood of bias (Table 2). *Uncertain* can be used in cases in which the study authors do not provide sufficient information about a factor. Additional details about each factor are provided in Table 1.

Factors were identified that influence neurodevelopmental test performance iteratively in consultation with collaborators at DNTP and EPA. The process included starting with a list of factors known to influence test performance given past experience (e.g., participant age, culture, and test conditions). Additional factors specific to studies of MeHg were added based on discussions with EPA collaborators (e.g., score derivation) or from prior assessments of the literature (e.g., blinding).

In some cases, fewer than four levels were applied to each criterion. For some factors, only *very low* and *low* were selected because it was determined that the factors were unlikely to affect the interpretation of the results. In other cases, only *very low* and *high* were selected for likelihood of bias when it was determined that the factor did not have finer gradations. Finally, instances were noted for which the absence of information about a given factor warrants a rating of moderate risk of bias because the failure to address this factor increases the likelihood of biasing a study's results or reduces the validity of the psychometric test(s) administered to the participants. In other cases, when a given factor is not discussed, there is not necessarily a high likelihood of bias that would systematically distort psychometric test performance. However, the failure to address this factor could increase random error.

Given that there are a variety of ways to administer psychometric tests—examiner-administered, questionnaire, or computerized—some of the factors discussed below do not apply to certain types of tests. Examiner-administered tests are those for which an examiner is directly administering the test to a participant and is also responsible for scoring responses. Questionnaires include participant, teacher, or parent ratings of the participant's thoughts, behaviors, or feelings. Computerized tests are those that are completed via a laptop, desktop computer, or tablet after standardized instructions are provided by an examiner or the software.

Application of the evaluative criteria for each factor should be applied to *each* psychometric test that was used for data analysis in a study (or manuscript). Most, if not all, cohort studies examining developmental neurotoxicity have administered a variety of examiner-administered, questionnaire-based, and computerized tests to participants. Thus, it is possible for there to be low risk of bias for one test but a higher risk of bias for another test within the same study.

As noted at the beginning of this document, the evaluation ratings provided in Part I are not meant to provide a definitive determination of whether a psychometric test should be used in a particular setting. Rather, they provide a summary of what is known about each test's psychometric properties to guide the reader in making such determinations. At the broadest level, an assessment of an outcome's usefulness and validity will need to consider whether a given factor will affect the internal and external validity of a result. A study's results could be internally valid even in the presence of biases related to psychometric test administration. For instance, the normative scores used as the outcomes from a psychometric test in a study may not be derived from a reference population that is highly similar to the one used in the study, but the direction and magnitude of the reported association might be similar regardless of how the scores were treated. Approaching the evaluation in this way will help guide decisions about whether a given result is informative for a meta-analysis or risk assessment.

Note that evaluators will need to apply some discretion regarding the context of the specific psychometric test, study location, participants, and other factors when applying the principles described below. For example, with regard to study location, some countries or institutions may only require master's level training in psychology or neuropsychology to supervise study staff administering psychometric tests, but in other countries, doctoral-level training may be required. As another example, specifically with regard to psychometric test translation, translation and back-translation are important for all tests, with pilot testing of the translated versions conducted before the study begins.

Four factors are less applicable to questionnaires—examiner training, examiner blinding, number of examiners, and test environment. It is possible that these factors could introduce random error into the measurement of a trait. For example, if an examiner "guides" the respondent to indicate abnormal functioning by saying that this is what the study is looking for in a particular exposure situation, the outcome could be biased. This can be avoided by providing scripts to examiners to use when they hand questionnaires to respondents. In addition, participant responses on a questionnaire might vary if they are in a quiet room at the study clinic versus a noisy room in their residence. However, systematic bias related to aspects of test administration seems less likely using self-administered questionnaires.

Table 1: Summary of Factors Influencing Neurodevelopmental Test Performance and the Likelihood of Bias

Factor	Likelihood of Bias				
	Very Low	**Low**	**Moderate**	**High**	**Uncertain**
Test Selection	The psychometric test was selected based on existing scientific knowledge about the effects of a given toxicant on nervous system development. This could include knowledge obtained from experimental animal, acute toxicity, case-series, or epidemiological (e.g., occupational or population-based) studies.	The psychometric test was selected because it has been demonstrated to be sensitive to other neurotoxicants or because the study authors are attempting to replicate a prior finding of a study with low likelihood of bias in this domain. The test was selected for convenience or because it was previously used to examine other hypotheses.	N/A	The test is no longer considered a valid measure of the traits it was designed to assess when it was originally developed.	The study authors do not state why the psychometric test was selected.
Age of Participant	The psychometric test is appropriate for the age range of participants being examined in the study.	The psychometric test is age-appropriate for the participants being studied, but the outcome(s) being examined might be more reliably assessed at other ages.	The age-appropriateness of the test or age of the participants is not stated or cannot be determined.	The psychometric test is not appropriate for the age range of participants being examined in the study.	N/A

Factor	Likelihood of Bias				
	Very Low	**Low**	**Moderate**	**High**	**Uncertain**
Test Culture	The psychometric test was developed for the culture of the participants being examined in the study. In cases when the test is being adapted to other cultures, there is evidence from validation or pilot studies in the target population demonstrating the validity and reliability of the adapted test.	In the case of multisite studies, a culturally appropriate test was administered to the participants being examined in the study, exposure was not related to study site, and the investigators took efforts to minimize potential differences in psychometric test performance related to study site.	The psychometric test is not culturally appropriate for the study participants being examined or there is no evidence presented that it is valid or reliable in the population being studied. The cultural appropriateness of the test is not stated.	In the case of multisite studies, there are cultural differences across study sites that could affect psychometric test scores among participants being examined in the study, and study site is related to exposure.	N/A
Test Language	The psychometric test was developed for the primary language of the participants being examined in the study. In cases when a test was translated to another language, translation and back-translation is evident. There is evidence from validation or pilot studies in the target population demonstrating the validity and reliability of the translated test.	The test was translated and back-translated; there is no evidence demonstrating validity or reliability of the translated test in the target population. The test was translated, but not back-translated; there is evidence of validity and reliability of the translated version.	The psychometric test was adapted for another language, but only translated and not back-translated for comparison to the original test, and there is no evidence of validity or reliability of the test in the new language. The study authors do not state if/how the test was translated and whether validity studies were performed.	The psychometric test was not administered in the same language spoken by the study participants.	N/A

Factor	Likelihood of Bias				
	Very Low	**Low**	**Moderate**	**High**	**Uncertain**
Test Score Derivation	Normalized test scores are appropriate for the source population of the study participants (i.e., developed for this source population). Alternatively, raw scores are used in analyses wherein norms do not exist and raw scores are adjusted for factors known to influence test performance (e.g., age and sex).	Normalized scores from a comparable target population are applied to the study participants (e.g., United States versus Canada).	Normalized scores from a noncomparable target population are applied to the study participants (e.g., United States versus China). Raw scores are used and there are no adjustments for factors predictive of the measured trait.	N/A	The study authors do not state which scores were used or do not provide details about score derivation.
Standard Test Administration	The psychometric test was developed for the population being studied and there is evidence that it was administered in a standardized manner.	In cases when tests were adapted for the study population, the adapted version is administered in a standardized fashion. There is evidence that the adapted version is reliable across study sites and examiners. Moreover, there is evidence that the adapted test measures the same construct (e.g., cognition, motor skills) in the target population as in the population for which the test was developed (i.e., validity).	In cases when a test was adapted for the study population, the adapted version is administered in a standardized fashion, but there is no evidence of the reliability or validity of the adapted test.	There is no evidence that the adapted test has been administered in a standardized fashion within the study. Reliability and validity of the adapted version has not been demonstrated.	N/A

Factor	Likelihood of Bias				
	Very Low	**Low**	**Moderate**	**High**	**Uncertain**
Examiner Training	A Ph.D.-level investigator with training in psychometric tests trains study staff on how to administer tests and assesses examiner validity and reliability at baseline about every 12 months. Examiners have adequate training and experience administering tests and have special training or experience for unique subpopulations (e.g., children).	A Ph.D.-level investigator trains staff initially, but there is no additional training or testing to ensure continued validity and reliability. Examiners have adequate training and experience administering tests but may not have special training or experience for unique subpopulations (e.g., children).	The study authors do not report features of examiner training and experience.	No staff training by a Ph.D.-level investigator is conducted or the examiners are inexperienced with administering psychometric tests.	N/A
Examiner Blinding	The examiners (or reporters) are blind to each study participant's exposure at the time of examination(s).	In the case of multisite studies for which participants were selected on the basis of exposure, the investigators make efforts to minimize the potential for an examiner's knowledge of participant exposure to influence examinations. This includes minimizing examiner effects (i.e., high interrater reliability), providing adequate and repeated refresher training to examiners, and controlling for study site in statistical analyses if warranted.	For single site studies, the examiners (or reporters) are not blind to each study participant's exposure at the time of examination(s). The study authors do not report on the blinding of examiners or reporters.	In the case of multisite studies for which participants were selected on the basis of exposure, the investigators did not make efforts to minimize potential for examiner's knowledge of participant exposure to influence examinations.	N/A

Factor	Likelihood of Bias				
	Very Low	**Low**	**Moderate**	**High**	**Uncertain**
Number of Examiners	There is one examiner who administers all psychometric tests and has high reliability. If there are multiple examiners, the interrater reliability is high and examiners are assessed for validity and reliability about every 12 months. Longitudinal assessments are done by the same examiner(s) when feasible.	N/A	There are multiple examiners with uncertain interrater reliability or the interrater reliability is low with any number of examiners. The study authors do not state the number of examiners.	N/A	N/A
Test Environment	The described test conditions are optimized to obtain the best estimate of an individual's psychometric test score. Rooms are quiet, well lit, and free of distractions. Accommodations are made for pediatric participants.	Test conditions vary, but attempts are made to standardize and optimize the conditions (e.g., tests administered in schools or participant homes).	Test conditions are not standardized or optimized. The study authors do not state the test conditions.	In the case of multisite studies for which participants are selected from sites on the basis of exposure, the testing environment is related to study site and the subject's performance on the test.	N/A

Factor	Likelihood of Bias				
	Very Low	**Low**	**Moderate**	**High**	**Uncertain**
Longitudinal Measures	The same or a comparable psychometric test is administered by the same examiner (when feasible) at least 12 months apart. Because specific tests may be revised during a longitudinal study, the version of the instrument used must be considered. Best practices can include continuing to use the version of the test initially applied, especially if the adjusted raw score outcomes are utilized or the study aim is to determine if the predictor variables cause test performance to change over time.	The same or a comparable psychometric test is administered by different examiners and used over time, or less than 12 months elapsed between examinations. A comparable test can include a newer version of the test applied earlier to the population, especially if normative outcomes are utilized or if the test undergoes very little change during revision. Sometimes, however, a revised version of a test differs too much from a previously applied version and may not be truly comparable.	Noncomparable psychometric tests are administered on different occasions but are treated as interchangeable. Statistical methods are not appropriate for analyzing repeated measures data.	N/A	The study authors do not state whether the measures are comparable over time. They do not report length of time between measures, number of examiners, or statistical methods.

Factor	Likelihood of Bias				
	Very Low	**Low**	**Moderate**	**High**	**Uncertain**
Clinical Diagnosis	Neurodevelopmental disorder diagnoses are made according to established criteria, confirmed in a subset of individuals using other diagnostic methods or validated using other methods, and do not vary across space and time.	N/A	Diagnoses are not confirmed using another method or diagnostic methods varied over space and time and these differences are not taken into account. It is not clear how diagnoses were made, whether they were confirmed using other methods, or if diagnostic criteria varied over space and time.	Diagnoses are not made according to established criteria.	N/A

N/A = not applicable.

Table 2: Likelihood of Bias for Individual Factors Related to Psychometric Test Administration in Epidemiological Neurodevelopmental Toxicity Studies

Likelihood of Bias	How to Interpret
Very Low	Appropriate study conduct related to the factor and minor deficiencies are not expected to influence the results.
Low	The study has some limitations, but limitations are not likely to be severe or have a notable impact on the results.
Moderate	Identified biases or deficiencies are interpreted as likely to have a notable impact on the results or prevent reliable interpretation of study findings.
High	A judgment that the study conduct relating to the factor introduced a serious flaw that is interpreted to be the primary driver of any observed effect or makes the study uninterpretable. Study is not used without exceptional justification.
Uncertain	The study authors did not present information about the factor and an evaluation of it cannot be conducted.

Test Attributes

Several attributes need to be considered when evaluating the use or features of a psychometric test. In Part 1 of this report, details are provided about the psychometric properties and features of specific tests. This section describes more broadly how specific test characteristics influence their use and interpretation in population-based studies.

Selection of Psychometric Tests

Selection of appropriate psychometric tests for research assessing toxicant effects on the developing nervous system requires an understanding of the toxicant and research question. The selection of tests will depend on the toxicant of interest. Different toxicants affect different nervous system pathways, and thus affect specific domains or subdomains of neurobehavioral function. A null result may not indicate a lack of neurotoxicity of the chemical being investigated, but rather that the toxicant does not affect the biological pathways related to the domain, subdomain, or behavioral outcome measured by the test(s) chosen for the study (i.e., lack of sensitivity).

Ideally, to safeguard against this situation, psychometric tests should be selected based on a review of existing scientific knowledge (e.g., experimental studies in animals, acute toxicity studies or case-series, occupational studies) about the nervous system pathways and structures that are affected by the exposure of interest. Knowledge about mechanisms of neurotoxic action on the nervous system and its structures can be used to select outcomes that are most likely to detect neurotoxic effects. It is important to note that, for some toxicants, there may be no or limited prior literature available to select outcomes that would be most sensitive. In this case, researchers might consider the effects of toxicants with similar structure and modes of action in their choice of outcomes. Alternatively, they might assess a broad range of domains using outcomes with previously documented sensitivity to various neurotoxicants, with the intent of identifying which facets of neurodevelopment are potentially affected by the toxicant of interest.

Less ideally, psychometric tests are selected out of convenience because they are familiar to the researchers, used by other researchers, readily available, or easy to use. In addition, some cohort

studies that were designed to evaluate questions about neurodevelopment, but not specifically neurotoxicity, are adapted for use in the field of neurotoxicology. These cohorts often have large sample sizes, relevant covariates, and available biospecimens or environmental samples that can be used to assess neurotoxicant exposure during periods of heightened susceptibility (e.g., pregnancy, childhood). In these cases, the neurodevelopmental outcomes may have been selected to address other research questions or as general assessments of neurodevelopment (e.g., IQ tests, personality tests) without a specific a priori reason to use them for studying a given toxicant. In some cases, this situation may underlie null findings. In general, these sources of data can be helpful in the absence of resources to create new cohorts or conduct new studies.

The "age" of a test is an additional consideration for test selection because psychometric tests are often replaced by newer versions of a test or are phased out of use. Evaluators should consider the length of time between test development and test administration. "Aged" tests may no longer be relevant to contemporaneous populations (e.g., content or language) or their validity and the theoretical constructs on which they are based may no longer be consistent with the latest advances in psychometric testing and neurodevelopment. Assuming that a "dated" test remains valid, it may continue to be used in cases when the test users have already conducted repeated assessments using the same instrument. However, the normed outcomes (e.g., scaled scores) may need to be updated or raw scores, rather than normed scores, may need to be used as outcomes with adjustment for age, sex, or other factors that explain variance in the raw scores. Given the difficulty in determining if an individual test is "dated" at the time of its administration, this factor was not considered as an evaluative criterion in this document.

Appropriateness of Tests for Assessment

The previous section describes how psychometric properties of individual tests might influence the ability to validly and reliably assess a given domain or subdomain of neurodevelopment. How these features interact with attributes of the population being studied is considered below.

Participant Age

The age of the participants at the time of test administration is important to consider because virtually all psychometric tests are developed for specific age ranges, and the results of some tests may vary as a function of age. For instance, some tests of cognitive abilities have been developed specifically for preschool children (e.g., Wechsler Preschool and Primary Scale of Intelligence [WPPSI]), whereas others have been developed for school-age children (e.g., Wechsler Intelligence Scale for Children [WISC]) (Wechsler 2002; Wechsler 2003). For a given test, scores might be age-standardized to allow comparisons of psychometric traits across children of different ages.

In some cases, the appropriate age for administering a given test is stated in years. Unless stated otherwise, it should be assumed that this is inclusive from the first date of the first year to the last date of the last year (e.g., a test designed for 2- to 5-year-olds is appropriate for children who are greater than or equal to 2 years of age and less than 6 years of age).

Failure to adhere to age ranges recommended by the test developers can lead to invalid results. An egregious example would be to administer tasks that require reading to pre-literate children. Thus, it is essential that the selected psychometric test is designed and validated for the age group of interest and that the test is only administered to age-appropriate subjects.

The values and variability of some psychometric test scores vary as a function of age. For instance, parent-reported child anxiety can increase as children age (Braun et al. 2017). In addition, an environmental toxicant may not be associated with some domains of child cognitive development if the presence or variance of the trait associated with the domain is absent or low, respectively. For example, some executive functions develop later in childhood than others. Moreover, the reliability of a specific domain could be lower at younger ages (e.g., infancy) when performance on a given test may be sensitive to recent feeding or sleeping. For instance, the Mental Development Index of the Bayley Scales of Infant Development-II (BSID-II), a measure of infant and toddler cognitive abilities, has fair reproducibility in the first 3 years of life, whereas the full-scale IQ of the Wechsler instruments, another measure of cognitive abilities, has excellent reproducibility at later ages (Braun et al. 2017). At earlier ages, the variability in test scores arises because of the highly dynamic nature of development and potential for other factors (e.g., lack of sleep, hunger) to have a greater effect on test performance. This variability does not negate the validity of neurodevelopmental assessments at earlier ages but may increase the variance in measures taken at earlier ages (i.e., random error).

Preferably, the age of test administration is selected based on a priori knowledge of the specific domain being assessed in order to reduce within-person variance or age-related changes in the trait associated with the domain. Alternatively, longitudinal measures of the domain can be collected to reduce within-person variance as well as to examine the effect of an environmental toxicant exposure on trajectories of a domain.

A final point to consider with regard to age is the length of the psychometric test battery. Scores obtained during a lengthy test battery might not reflect a participant's true abilities. For instance, infants and toddlers usually will not tolerate more than 75 minutes of testing, whereas school-aged children can tolerate 2–3 hours. In addition to length of the test battery and age-appropriateness of the test, investigators ideally should provide information about whether participants received adequate breaks between tests or tests were completed during multiple visits.

Culture

Psychometric test results could be biased to specific cultural groups in ways that affect performance. Even tests like the Raven's Progressive Matrices, which are thought to be "culturally fair," favor participants who have formal schooling that teaches them to organize data into rows and columns (Nisbett et al. 2012). Other examples include using test items (e.g., analogies) that might be unfamiliar to certain cultures.

Ideally, psychometric tests are selected so that cultural biases are minimized. Initially, pilot testing of psychometric tests in the study population of interest can provide data to ensure that there are not specific items or composite/subscale scores that are lower than expected, that the participants understand test expectations, and that items are correctly ordered for difficulty. In the absence of these problems, comparison of normed scores in the study participants to other reference populations can aid in identifying cultural bias. Lower-than-average normalized scores in the study participants may be indicative of cultural bias. Finally, in multisite studies, it is important to consider whether there are cultural differences across study sites that might influence test performance. This can result in systematic bias if study site is related to toxicant exposure(s).

Language

Some psychometric tests have been translated into different languages to facilitate measurement of various domains in populations around the world; however, test translation may introduce bias if individual test items are not appropriately translated. In some cases, test developers provide validated test translations that are readily available (e.g., Pearson products).

Ideally, the psychometric test was developed for the setting-specific language or the test publishers have endorsed specific versions of the test in other languages. Less ideally, the test is translated by the investigators into the setting-specific language. If this is done, the test should first be translated to the new language by one bilingual translator and then back-translated to the original language by another bilingual translator and piloted before applied to the study population.

Score Derivation

Raw scores from psychometric tests are usually converted to scaled scores, T-scores, or standard scores using data derived from a reference population—often a representative sample of the U.S. population. Scaling of scores is done to account for age- and sometimes sex-specific differences in the means and variances of raw scores, thus allowing comparison of scores across different subgroups. It is important to note that subgroup-specific means and variances may be specific to the reference population and not generalizable to other populations. Thus, derived scores can vary in different populations through two mechanisms:

(1) If the mean of raw scores from the study participants is higher/lower than the reference population's raw scores, then the mean of the scaled, T-, or standard scores will be higher/lower.

(2) If the variance of raw scores from the study participants is higher/lower than the reference population's raw scores, then the variance of the scaled, T-, or standard scores will be higher/lower.

The presence of a shift in the mean and/or variance of raw scores can be assessed if the study authors report the mean and variance of the scaled, T-, or standard scores in their entire group of study participants and ideally across subgroups that are used to derive scaled, T-, or standard scores. However, if the raw score mean or variance is correlated with the exposure, then the exposure-outcome association could be biased in an unpredictable way. This would be difficult to verify unless the study authors report the raw score means and variances by exposure and covariates.

Ideally, the study participants are drawn from the same or a comparable source population as the reference sample. In other cases, population-specific references may be available (e.g., Canadian reference for Wechsler IQ tests). If a comparable reference is not available, the raw scores from the instrument can be used as the outcome. Ideally, a model using the raw score as the outcome would adjust for the same variables that are used to derive scaled, T-, or standard scores (e.g., age and sex). There is the potential for moderate risk of bias if adjustments are not made for these variables when the variable(s) are related to raw score values. Most critical would be if these factors are related to exposure. Less critical is when these variables are not related to the exposure, but still explain variation in the raw scores. In the latter case, adjusting for them can result in more precise estimates of the exposure-outcome relationship.

As noted in the Normative Data section, the nature of some raw scores may preclude using them as the outcome in regression models. For example, the Bayley Scales have infants or children complete a different set and number of items based on their age. Thus, the raw scores may not be equivalent across individuals. While a variety of methods could be used to create new scores (e.g., summed scores, PCA-derived scores), they have various strengths and limitations (McNeish and Wolf 2020), and a full evaluation of these methods is beyond the scope of this document.

Factors Related to Test Administration

Standardized Test Administration

Psychometric tests need to be administered in a standardized fashion to ensure that variations in administration procedures do not unduly influence test scores. Standardized administration of psychometric tests enhances the confidence that test results and findings are comparable across individuals, study sites, and populations.

The standardized administration of many psychometric tests is detailed in their respective manuals. The goal of providing detailed test administration instructions is to ensure that psychometric tests are administered in the same manner by any test user (e.g., psychometrist, psychologist). These instructions define standard test procedures for components of the test, the test materials used in administering the test, and scoring of participant responses. These manuals often provide word-for-word instructions and questions ("scripts") to be used in administering the test. The test materials generally include test stimuli, answer sheets, and test administration booklets that should be used for every administration. Finally, the test manuals provide methods and formulas for scoring each item in the test.

Deviations from standardized test administration across participants, study sites, examiners, or populations can produce results that are not comparable with each other. Some deviations from standard procedure as defined by the test manual are described in this document and are common in epidemiological research when psychometric tools are applied to populations aside from the one(s) on which the test was normed. These include changes in the test stimuli or order of stimulus presentation, usually due to cultural factors, and the language in which the test is administered. In these cases, there should be evidence that the administration of the adapted tests is standardized across participants, study sites, and examiners. Moreover, there should be evidence that any adaptations do not threaten the validity and reliability of the test.

Test Examiner

An individual's psychometric test performance can be influenced by factors related to the test examiner. Therefore, a number of specific factors should be in place to ensure that any potential examiner effects on test performance are minimized.

The following factors are potential sources of bias for cases when the examiner is actively administering a test to the participant and is responsible for directly engaging the participant, presenting stimuli, rating the participant's response, and scoring the responses (e.g., IQ tests). Standardized instructions should be used when instructing participants on how to complete questionnaires (e.g., behavioral rating forms) or computerized tests (e.g., continuous performance tasks).

Training

Persons directly administering psychometric tests should have adequate training in and experience with administering psychometric tests to subjects similar to the study participants. Given the intensive nature of administering more complex psychometric tests (e.g., BSID, NEPSY, Brazelton, NNNS), some researchers and clinical professionals believe that they should only be administered by licensed practitioners. Examiners should receive training from a Ph.D.-level investigator with expertise in neuropsychology, clinical psychology, educational psychology, or other closely related discipline. This "expert-level" trainer should have the requisite qualifications and experience to ensure that they can validly and reliably administer the specific tests. Examiners should have specialized training or experience when working with unique subpopulations (e.g., children). More complex or in-depth tests (e.g., NEPSY) require that a highly experienced trainer provides examiners with didactic instruction on the test, hands-on practice to administer the test, and assessment of validity/reliability in subjects similar to the study's source population.

In addition to having adequate training and experience, examiners should routinely be assessed for drift over time. Test drift can arise if examiners change the way they administer or score specific items over time, resulting in additional variability in test scores. Ideally, a Ph.D.-level investigator who is trained in psychometric test administration assesses each examiner's validity and reliability about every 12 months. Reliability can be assessed using test-retest correlations with the same examiner administering the test to the same individuals at the same age at the time of testing. A less ideal approach is to have a Ph.D.-level investigator provide refresher training on the specific tests. Finally, multisite studies should ensure that all examiners receive the same initial training to ensure validity and reliability, are routinely checked for validity and reliability, and obtain additional training as needed.

Blinding

There is evidence from randomized controlled trials that non-blinded trials report exaggerated effect sizes relative to blinded trials; however, the magnitude of these differences varies according to specific treatments and outcomes (Bello et al. 2014; Cuijpers et al. 2015; Hrobjartsson et al. 2014; Saltaji et al. 2018). It is conceivable that this phenomenon is present in observational studies of neurotoxicants in which examiners or reporters (e.g., caregivers) are not blind to the exposure status of participants. This may arise in studies where staff have access to exposure level data for participants or when exposure biomarker results are reported back to participants or participants' parents/caregivers (Brody et al. 2014). Indeed, some investigators have reported chemical biomarker results back to participants, especially in community-based participatory research. Ideally, all examiners (or reporters) in observational studies of neurotoxicants should be blind to each participant's exposure status. This might not be possible in multisite studies for which sites were selected on the basis of exposure. The magnitude of the effect that blinding would have in observational studies of some neurotoxicants is unclear. It is important to note that certain psychometric tests (e.g., computerized tests that involve little interaction with an examiner, self-reported behaviors) may be less susceptible to this source of bias.

Number of Examiners

Studies using a larger number of examiners could produce variability in psychometric test scores due to differences in test administration and scoring across examiners compared with studies

using a smaller number of examiners. Inter-examiner reliability is typically assessed by having the study's examiners administer the test to the same individuals at the same age; these data are used to calculate inter-examiner correlations. Ideally, studies should strive for high inter-examiner reliability as measured by Pearson correlations, intraclass correlations coefficients, etc. (Strauss et al. 2006). Moreover, in longitudinal studies, ideally the same examiner(s) should assess participants at subsequent visits if the same psychometric test is administered again. Examiner effects can be assessed by quantifying the reliability of two examiners assessing the same child or by comparing the mean instrument scores across examiners. It is important to note that statistically adjusting for examiners will not correct for differences in interrater validity or reliability. For instance, if an examiner invalidly administered a psychometric test to participants, then statistical adjustment for the examiner will not make these measurements "more" valid as they were never valid.

Test Administration, Environment, and Conditions

Test performance can be affected by the environment and conditions of test administration. As Strauss notes, optimal conditions are those that facilitate the participants having their best performance possible, whereas standardized conditions ensure that the conditions of the test are as similar as possible across repeated test administrations (Strauss et al. 2006). Therefore, in epidemiological studies, test environments and conditions should be standardized and optimized to ensure that all participants are given the same opportunity for their best performance with as little variation as possible in the method of administration across subjects.

Environmental factors like noise and temperature can influence test performance. Sources of environmental noise (e.g., traffic, conversations) can adversely affect performance on a wide range of psychometric tests (Bhang et al. 2018; Klatte et al. 2013; Shield and Dockrell 2008). Environmental conditions such as excessive heat can also adversely affect psychometric test performance (Klatte et al. 2013; Shield and Dockrell 2008). Conditional factors, like the location of the test (e.g., study clinic versus the participant's home) or presence of a parent or other caregiver in the room might cause test performance to vary.

Ideally, the test environment is a quiet, well-lit, and private room that is free of distractions. For pediatric participants, age-appropriate furnishings should be provided so that children are able to sit properly and engage with test materials. It can be reasonable to administer psychometric tests in the home or school environment if the environment is modified for the specific test and provisions are made to standardize conditions as much as possible. For assessments of children less than 2 years of age, it is desirable and often necessary for the parent to be present. At older ages, children should be assessed without the caregiver to reduce the potential for that caregiver to influence a child's performance. Investigators should note instances when test conditions were not optimal and consider excluding these results from statistical analyses.

Participant-level characteristics also need to be considered when administering psychometric evaluations. Time-varying factors like sleep, hunger, and current infections can affect performance on psychometric tests. For instance, children with or recovering from otitis media may have temporary decreases in hearing ability, which in turn can adversely affect test performance. Other factors like loss of mobility, amputation, and visual or hearing impairments can hinder assessment of specific domains or subdomains.

Ideally, participants are healthy enough to have tests administered to them or are able to return for assessments if it is determined that they are too sick to perform optimally. In cases of permanent disabilities that directly affect the ability to complete the test, these participants should not be assessed or should be excluded from statistical analyses. For pediatric assessments, efforts should be made to assess infants/children when they are well rested and have been fed. For older children, breaks should be offered during longer neurodevelopmental assessment batteries.

Finally, it is important to minimize errors that could arise in the scoring and entry of psychometric test data. Recent versions of some tests now have scannable forms, computerized data entry, and automated scoring. While not infallible, these advancements can improve quality control, as they reduce the potential for human errors in data entry and test scoring.

Longitudinal Studies

Some studies examine the impact of a neurotoxicant on repeated measures of a psychometric test. This can be done to increase statistical power and precision or to determine if the potential effect of a neurotoxicant persists, emerges, or wanes over time (Braun et al. 2017). There are several factors related to the timing of test administration, type of psychometric test, participant follow-up, and statistical analyses that should be considered when evaluating studies with repeated psychometric measures (White 2004).

The amount of time between assessments should be considered because individual test scores can improve due to practice effects on a given instrument. While some instruments like the WPPSI and WISC recommend at least 1 year between administrations (Wechsler 2002; Wechsler 2003), practice effects have also been found to linger for many years (Calamia et al. 2012). In addition, the number of examiners and examiner drift should be considered because different examiners may administer or score tests differently or change their administration/scoring practices over time. This can be curtailed by videotaping or observing sample examiner test administration every few months during a study and/or each time that a new data collection cycle begins.

In some cases, different psychometric instruments can be used to assess a given domain across ages. For instance, the WPPSI is designed to assess cognitive abilities in children 2.5–7.25 years of age, whereas the WISC is designed to assess these same cognitive abilities in children ages 6–16 years. Both Wechsler instruments are intentionally designed to overlap in ages and assess comparable domains, but this is not the case for all psychometric tests. For less comparable instruments, it is important to consider whether different tests assess the same neurodevelopmental domain and are interchangeable.

Ideally, an examiner administers the same or comparable psychometric tests that assess the same neurodevelopmental domain and the same examiner repeatedly assesses a given participant. If testing is carried out frequently, parallel or equivalent tests that assess a domain can be used instead of repeating a subtest. Moreover, examiner drift over time should be periodically assessed. While there are no clear rules about the length of time required to eliminate practice effects, tests should be administered at least 12 months apart (Calamia et al. 2012; Scharfen et al. 2018).

Finally, appropriate statistical techniques must be used to examine repeated measures data (e.g., linear mixed effect models or linear regression with generalized estimating equations). Failure to account for repeated measures within the same individuals will result in inappropriately small standard errors because traditional methods treat repeated observations as independent.

Clinical Diagnoses

Some studies examine associations between neurotoxicant exposure and risk of diagnosis with a neurodevelopmental disorder (e.g., ADHD, autism spectrum disorders [ASD], or learning disabilities [LD]) (Anderko et al. 2010; Engel et al. 2018; Shelton et al. 2014). Often these studies rely on registries maintained by disease monitoring networks or governments to identify cases with specific neurodevelopmental disorders (Anderson and Burnett 2017; CDC 2012; Engel et al. 2018). In other cases, investigators conduct detailed neurodevelopmental assessments and medical chart reviews to confirm or deny a clinical diagnosis (Hertz-Picciotto et al. 2006).

Case identification and verification are critical issues to consider when evaluating developmental neurotoxicity studies using clinical diagnoses. Ideally, the probability of an individual being diagnosed (or not diagnosed) with a neurodevelopmental disorder should be the same regarding exposure, as well as socioeconomic status, geography, and calendar time. It is important to note that the probability of being diagnosed is not the same as the probability of having the disorder as there could be over- or under-diagnoses in subpopulations.

With regard to case identification, it is critical that standardized definitions of disease be applied to all participants (e.g., Diagnostic and Statistical Manual criteria). Case verification may be accomplished using administrative records (e.g., International Classification of Disease [ICD] codes) or expert review of medical charts, including psychometric tests and other diagnostic information. Individuals reviewing individual participant information should have a master's or doctoral degree in clinical psychology or related discipline and be blinded to the exposure status of the participants. Methods of case verification should be the same across clinics, providers, or reviewers who make the diagnosis. In addition, the criteria for establishing diagnosis should be stable over time; otherwise, changes must be considered.

Reassuringly, for ASD and ADHD in the United States and Europe, there is evidence that registry-based measures of diagnosis have high levels of agreement with other diagnostic measures of these disorders (Lauritsen et al. 2010; Linnet et al. 2009; Rimvall et al. 2014; Suren et al. 2012). In addition, there is high agreement of parental report of ASD with psychometric instruments used to diagnose the disorder (Roberts et al. 2013). Moreover, for ASD, there is evidence that registries are effective at ascertaining a high proportion of cases (Nicholas et al. 2012). In addition, registry-based diagnoses of ASD have a high degree of agreement with external expert reviewers (e.g., clinical psychologist) and a high degree of temporal stability (Bakian et al. 2015; Wiggins et al. 2012).

Ideally, a study will use the same criteria to establish a diagnosis across clinics or providers and over time while verifying there are no geographical and temporal deviations in how diagnoses are made. Diagnoses should be based on data from at least two sources in a subset of participants (e.g., registry-based diagnosis, caregiver reports of relevant symptoms, external expert review).

References

Anderko L, Braun J, Auinger P. 2010. Contribution of tobacco smoke exposure to learning disabilities. J Obstet Gynecol Neonatal Nurs. 39(1):111-117. http://dx.doi.org/10.1111/j.1552-6909.2009.01093.x

Anderson PJ, Burnett A. 2017. Assessing developmental delay in early childhood — Concerns with the Bayley-III scales. Clin Neuropsychol. 31(2):371-381. http://dx.doi.org/10.1080/13854046.2016.1216518

Bakian AV, Bilder DA, Carbone PS, Hunt TD, Petersen B, Rice CE. 2015. Brief report: Independent validation of autism spectrum disorder case status in the Utah Autism and Developmental Disabilities Monitoring (ADDM) Network Site. J Autism Dev Disord. 45(3):873-880. http://dx.doi.org/10.1007/s10803-014-2187-6

Bello S, Krogsboll LT, Gruber J, Zhao ZJ, Fischer D, Hrobjartsson A. 2014. Lack of blinding of outcome assessors in animal model experiments implies risk of observer bias. J Clin Epidemiol. 67(9):973-983. http://dx.doi.org/10.1016/j.jclinepi.2014.04.008

Bhang SY, Yoon J, Sung J, Yoo C, Sim C, Lee C, Lee J, Lee J. 2018. Comparing attention and cognitive function in school children across noise conditions: A Quasi-experimental study. Psychiatry Investig. 15(6):620-627. http://dx.doi.org/10.30773/pi.2018.01.15

Braun JM. 2016. Early-life exposure to EDCs: Role in childhood obesity and neurodevelopment. Nat Rev Endocrinol. 13(3):161-173. http://dx.doi.org/10.1038/nrendo.2016.186

Braun JM, Yolton K, Stacy SL, Erar B, Papandonatos GD, Bellinger DC, Lanphear BP, Chen A. 2017. Prenatal environmental chemical exposures and longitudinal patterns of child neurobehavior. Neurotoxicology. 62:192-199. http://dx.doi.org/10.1016/j.neuro.2017.07.027

Brody JG, Dunagan SC, Morello-Frosch R, Brown P, Patton S, Rudel RA. 2014. Reporting individual results for biomonitoring and environmental exposures: Lessons learned from environmental communication case studies. Environ Health. 13(1):40. http://dx.doi.org/10.1186/1476-069X-13-40

Calamia M, Markon K, Tranel D. 2012. Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. Clin Neuropsychol. 26(4):543-570. http://dx.doi.org/10.1080/13854046.2012.680913

Centers for Disease Control and Prevention (CDC). 2012. Prevalence of autism spectrum disorders--Autism and Developmental Disabilities Monitoring Network, 14 sites, United States, 2008. MMWR Surveill Summ. 61(3):1-19.

Cuijpers P, Karyotaki E, Andersson G, Li J, Mergl R, Hegerl U. 2015. The effects of blinding on the outcomes of psychotherapy and pharmacotherapy for adult depression: A meta-analysis. Eur Psychiatry. 30(6):685-693. http://dx.doi.org/10.1016/j.eurpsy.2015.06.005

Engel SM, Villanger GD, Nethery RC, Thomsen C, Sakhi AK, Drover SSM, Hoppin JA, Zeiner P, Knudsen GP, Reichborn-Kjennerud T et al. 2018. Prenatal phthalates, maternal thyroid

function, and risk of attention-deficit hyperactivity disorder in the Norwegian mother and child cohort. Environ Health Perspect. 126(5):057004. http://dx.doi.org/10.1289/EHP2358

Gallo V, Egger M, McCormack V, Farmer PB, Ioannidis JP, Kirsch-Volders M, Matullo G, Phillips DH, Schoket B, Stromberg U et al. 2011. STrengthening the Reporting of OBservational studies in Epidemiology--Molecular Epidemiology STROBE-ME: an extension of the STROBE statement. J Clin Epidemiol. 64(12):1350-1363. https://doi.org/10.1016/j.jclinepi.2011.07.010

Hertz-Picciotto I, Croen Lisa A, Hansen R, Jones Carrie R, van de Water J, Pessah Isaac N. 2006. The CHARGE study: An epidemiologic investigation of genetic and environmental factors contributing to autism. Environ Health Perspect. 114(7):1119-1125. http://dx.doi.org/10.1289/ehp.8483

Hrobjartsson A, Emanuelsson F, Skou Thomsen AS, Hilden J, Brorson S. 2014. Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. Int J Epidemiol. 43(4):1272-1283. http://dx.doi.org/10.1093/ije/dyu115

Klatte M, Bergstrom K, Lachmann T. 2013. Does noise affect learning? A short review on noise effects on cognitive performance in children. Front Psychol. 4:578. http://dx.doi.org/10.3389/fpsyg.2013.00578

Lauritsen MB, Jorgensen M, Madsen KM, Lemcke S, Toft S, Grove J, Schendel DE, Thorsen P. 2010. Validity of childhood autism in the Danish Psychiatric Central Register: Findings from a cohort sample born 1990-1999. J Autism Dev Disord. 40(2):139-148. http://dx.doi.org/10.1007/s10803-009-0818-0

Lezak MD. 1976. Neuropsychological assessment. New York, NY: Oxford University Press.

Lezak MD. 1995. Neuropsychological assessment, 3rd ed. New York, NY: Oxford University Press.

Lezak MD, Howieson DB, Loring DW, Hannay HJ, Fischer JS. 2004. Neuropsychological assessment, 4th ed. New York, NY: Oxford University Press.

Lezak MD, Howieson DB, Bigler ED, Tranel D. 2012. Neuropsychological assessment, 5th ed. New York, NY: Oxford University Press.

Linnet KM, Wisborg K, Secher NJ, Thomsen PH, Obel C, Dalsgaard S, Henriksen TB. 2009. Coffee consumption during pregnancy and the risk of hyperkinetic disorder and ADHD: A prospective cohort study. Acta Paediatr. 98(1):173-179. http://dx.doi.org/10.1111/j.1651-2227.2008.00980.x

McNeish D, Wolf MG. 2020. Thinking twice about sum scores. Behav Res Methods. 52(6):2287-2305. https://doi.org/10.3758/s13428-020-01398-0

Morris SE, Cuthbert BN. 2012. Research Domain Criteria: Cognitive systems, neural circuits, and dimensions of behavior. Dialogues Clin Neurosci. 14(1):29-37. https://doi.org/10.31887/DCNS.2012.14.1/smorris

National Toxicology Program (NTP). 2022. NIEHS 1: Chemical Effects in Biological Systems (CEBs) data repository. Research Triangle Park, NC: U.S. Department of Health and Human Services, National Institute of Environmental Health Sciences, National Toxicology Program. https://doi.org/10.22427/NIEHS-DATA-NIEHS-01.

Needleman HL. 1990. What can the study of lead teach us about other toxicants? Environ Health Perspect. 86:183-189. https://doi.org/10.1289/ehp.9086183

Nicholas JS, Carpenter LA, King LB, Jenner W, Wahlquist A, Logan S, Charles JM. 2012. Completeness of case ascertainment for surveillance of autism spectrum disorders using the Autism developmental disabilities monitoring network methodology. Disabil Health J. 5(3):185-189. http://dx.doi.org/10.1016/j.dhjo.2012.03.004

Nisbett RE, Aronson J, Blair C, Dickens W, Flynn J, Halpern DF, Turkheimer E. 2012. Intelligence: New findings and theoretical developments. Am Psychol. 67(2):130-159. http://dx.doi.org/10.1037/a0026699

Rimvall MK, Elberling H, Rask CU, Helenius D, Skovgaard AM, Jeppesen P. 2014. Predicting ADHD in school age when using the Strengths and Difficulties Questionnaire in preschool age: A longitudinal general population study, CCC2000. Eur Child Adolesc Psychiatry. 23(11):1051-1060. http://dx.doi.org/10.1007/s00787-014-0546-7

Roberts AL, Lyall K, Hart JE, Laden F, Just AC, Bobb JF, Koenen KC, Ascherio A, Weisskopf MG. 2013. Perinatal air pollutant exposures and autism spectrum disorder in the children of nurses' health study II participants. Environ Health Perspect. 121(8):978-984. http://dx.doi.org/10.1289/ehp.1206187

Rose G. 1985. Sick individuals and sick populations. Int J Epidemiol. 14 1:32-38.

Rose G. 2001. Sick individuals and sick populations. Int J Epidemiol. 30(3):427-432; discussion 433-424. https://doi.org/10.1093/ije/30.3.427

Saltaji H, Armijo-Olivo S, Cummings GG, Amin M, da Costa BR, Flores-Mir C. 2018. Influence of blinding on treatment effect size estimate in randomized controlled trials of oral health interventions. BMC Med Res Methodol. 18(1):42. http://dx.doi.org/10.1186/s12874-018-0491-0

Sattler JM. 2001. Assessment of children: Cognitive applications, 4th ed. San Diego, CA: Jerome M. Sattler, Inc.

Scharfen J, Jansen K, Holling H. 2018. Retest effects in working memory capacity tests: A meta-analysis. Psychon Bull Rev. 25(6):2175-2199. https://doi.org/10.3758/s13423-018-1461-6

Shelton JF, Geraghty EM, Tancredi DJ, Delwiche LD, Schmidt RJ, Ritz B, Hansen RL, Hertz-Picciotto I. 2014. Neurodevelopmental disorders and prenatal residential proximity to agricultural pesticides: The CHARGE Study. Environ Health Perspect. 122(10):1103-1109. http://dx.doi.org/10.1289/ehp.1307044

Shield BM, Dockrell JE. 2008. The effects of environmental and classroom noise on the academic attainments of primary school children. J Acoust Soc Am. 123(1):133-144. http://dx.doi.org/10.1121/1.2812596

Spreen O, Strauss E. 1991. A compendium of neuropsychological tests: Administration, norms, and commentary, 1st ed. New York, NY: Oxford University Press.

Spreen O, Strauss E. 1998. A compendium of neuropsychological tests: Administration, norms, and commentary, 2nd ed. New York, NY: Oxford University Press.

Strauss E, Sherman EMS, Spreen O. 2006. A compendium of neuropsychological tests: Administration, norms, and commentary, 3rd ed. New York, NY: Oxford University Press.

Suren P, Bakken IJ, Aase H, Chin R, Gunnes N, Lie KK, Magnus P, Reichborn-Kjennerud T, Schjolberg S, Oyen AS et al. 2012. Autism spectrum disorder, ADHD, epilepsy, and cerebral palsy in Norwegian children. Pediatrics. 130(1):e152-158. http://dx.doi.org/10.1542/peds.2011-3217

Wechsler D. 2002. Wechsler preschool and primary scale of intelligence. San Antonio, TX: The Psychological Corporation.

Wechsler D. 2003. Wechsler Intelligence Scale for Children - 4th ed: Administration and scoring manual. San Antonio, TX: PsychCorp by Harcourt Assessment, Inc.

Weiss B. 2000. Vulnerability of children and the developing brain to neurotoxic hazards. Environ Health Perspect. 108 Suppl 3(Suppl 3):375-381. https://doi.org/10.1289/ehp.00108s3375

White RF. 1992. Clinical syndromes in adult neuropsychology: The practitioner's handbook. Amsterdam: Elsevier.

White RF, Proctor S. 1992. Research and clinical criteria for the development of neurobehavioral test batteries. J Occup Med. 34:140-148. http://dx.doi.org/10.1097/00043764-199202000-00013

White RF, Cohen RF, Gerr F, Green R, Lezak M, Lybarger J, Mack J, Silbergeld E, Valciukas J. 1994. Criteria for progressive modification of neurobehavioral batteries. Neurotoxicol Teratol. 16:511-524. http://dx.doi.org/10.1016/0892-0362(94)90130-9

White RF. 2004. Neuropsychological assessments in children from a longitudinal perspective for the National Children's Study. White Paper for NIH, Fall, 2004. http://nationalchildrensstudy.gov/research/analytic_reports/upload/Neuropsychological-Assessments-in-Children-from-a-Longitudinal-Perspective-for-the-National-Children-s-S.

White RF, Campbell R, Echeverria D, Knox SS, Janulewicz P. 2009. Assessment of neuropsychological trajectories in longitudinal population-based studies of children. J Epidemiol Community Health. 63(Suppl 1):i15-16. http://dx.doi.org/10.1136/jech.2007.071530

White RF. 2011. Ch. 24 Cognitive disorders in adults. The Oxford Handbook of clinical psychology. Oxford: Oxford University Press.

White RF, Reuben AS. In Press. Environmental toxicities including lead. In: Brown GG, King TZ, Haaland KY, Crosson B, editors. APA Handbook of Neuropsychology: Vol 1 Neurobehavioral Disorders and Conditions: Accepted Science and Open Questions. American Psychological Association.

Wiggins LD, Baio J, Schieve L, Lee LC, Nicholas J, Rice CE. 2012. Retention of autism spectrum diagnoses by community professionals: Findings from the Autism and Developmental Disabilities Monitoring Network, 2000 and 2006. J Dev Behav Pediatr. 33(5):387-395. http://dx.doi.org/10.1097/DBP.0b013e3182560b2f

Appendix A. Psychometric Tests Considered for Evaluation

Table of Contents

Tables

A.1. Psychometric Tests Included for Evaluation

Table A-1 lists the neurodevelopmental domains in alphabetical order and their associated psychometric tests. The evaluation ratings and notes for each test are detailed in Appendix B, where they are sorted by test type and domain.

Table A-1. Psychometric Tests Included for Evaluation by Domain

Domain	Test Name
Academic Achievement	KeyMath Diagnostic Arithmetic Test[4]
	KeyMath Diagnostic Arithmetic Test-Revised (KeyMath R)[4]
	Wechsler Individual Achievement Test (WIAT)[6]
	Woodcock-Johnson III Test of Achievement (WJ III ACH)[1]
Attention	Conners' Continuous Performance Test-II (CPT-II)[1]
	Test of Everyday Attention for Children (TEA-CH)[4]
	Test of Variables of Attention (TOVA)[4]
Clinical Conditions	Autism Spectrum Quotient (AQ)[6]
	Barkley Adult ADHD Rating Scale-IV[6]
	Childhood Asperger's Syndrome Test (CAST)[6]
	Childhood Autism Rating Scale (CARS)[6]
	Conners' Parent Rating Scale-Revised (CPRS-R)[1]
	Conners' Teacher Rating Scale-Revised (CTRS-R)[1]
	Spence Children's Anxiety Scale (SCAS)[6]
Developmental	Bayley Scales of Infant Development (BSID-1)[4]
	Bayley Scales of Infant Development, Second edition (BSID-II)[2]
	Bayley Scales of Infant Development-III (BSID-3)[6]
	Brazelton Newborn Assessment Scale (NBAS)[6]
	Denver Development Screening Test (DDST)[6]
	Denver Development Screening Test-II (DDST II)[6]
	Fagan Test of Infant Intelligence[6]
	Gesell Developmental Schedules[6]
	Griffith Mental Development Scales (GMDS)[6]
	Kyoto Scale of Psychological Development (KSPD; K-test)[6]
	Neonatal Intensive Care Unit Network Neurobehavioral Scale (NNNS)[6]
	Prechtl General Movement Assessment (GMA)[6]
Executive Function	Stroop Color-Word Test[4]
	Trail-making Test[4]
	Verbal Fluency Test[5]
	Wisconsin Card Sorting Test (WCST)[2]

Domain	Test Name
General Intelligence/IQ	Kaufman Assessment Battery for Children (K-ABC)[4]
	Kaufman Brief Intelligence Test (K-BIT)[4]
	McCarthy Scales of Children's Abilities (MSCA)[2]
	Raven's Coloured Progressive Matrices (Raven's CPM)[2]
	Raven's Standard Progressive Matrices (Raven's SPM)[2]
	Stanford-Binet Intelligence Scale: Fourth Edition (S-B 4)[2]
	Stanford-Binet Intelligence Scale: Fifth Edition (S-B 5)[2]
	Wechsler Intelligence Scale for Children (WISC)[6]
	Wechsler Intelligence Scale for Children-Revised (WISC-R)[2]
	Wechsler Intelligence Scale for Children-III (WISC-III)[1]
	Wechsler Intelligence Scale for Children-IV (WISC-IV)[2]
	Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R)[1]
	Wechsler Adult Intelligence Scale-Revised (WAIS-R)[2]
	Wechsler Adult Intelligence Scale-III (WAIS-III)[2]
Learning and Memory	California Verbal Learning Test (CVLT)[2]
	California Verbal Learning Test – Children (CVLT-C)[4]
	Wechsler Memory Scale-Revised (WMS-R)[2]
	Wechsler Memory Scale-III (WMS-III)[2]
Mental Status	Mini-Mental State Examination (MMSE)[5]
Motor Function	Bruininks-Oseretsky Test of Motor Proficiency (BOTMP)[6]
	Grooved Pegboard[4]
	Movement Assessment Battery for Children (MABC)[6]
Neuropsychological Assessment Batteries	Developmental Neuropsychological Assessment (NEPSY)[2]
	Neurobehavioral Evaluation System (NES)[6]
Social-Emotional	Beck Depression Inventory (BDI)[4]
	Beck Depression Inventory-Second Edition (BDI-II)[4]
	Behavior Assessment System for Children, 2nd ed. (BASC-2)[1]
	Child Behavior Checklist (CBCL)[4]
	Children's Communication Checklist (CCC)[6]
	Difficulties in Emotion Regulation Scale (DERS)[6]
	Disruptive Behavior Disorders Scale (DBD)[6]
	Early Childhood Behavior Questionnaire[6]
	EAS Temperament Survey for Children[6]
	Profile of Mood States (POMS)[2]
	Social and Communication Disorders Checklist (SCDC)[6]

Domain	Test Name
	Social Communication Questionnaire[6]
	State-Trait Anxiety Inventory (STAI)[3]
	Strengths and Difficulties Questionnaire (SDQ)[3]
	Vineland Adaptive Behavior Scales[5]
Verbal/Language	Boston Naming Test (BNT)[4]
	Boston Naming Test-2 (BNT-2)[4]
	MacArthur-Bates Communicative Development Inventories (CDI)[6]
	Peabody Picture Vocabulary Test-Revised (PPVT-R)[2]
	Peabody Picture Vocabulary Test-III (PPVT-III)[2]
	Preschool Language Scales-3 (PLS-3)[1]
	Speech and Language Assessment Scale (SLAS)[6]
	Test of Language Development (TOLD)[6]
Visuospatial Function	Bender Visual-Motor Gestalt Test[2]
	Bender Visual-Motor Gestalt Test II[2]
	Developmental Test of Visual-Motor Integration (VMI)[4]
	Finger Identification Test[4]

[1]Test information was obtained from manuals only.
[2]Test information was obtained from manuals and academic textbooks.
[3]Test information was obtained from manuals and peer-reviewed literature.
[4]Test information was obtained from academic textbooks only.
[5]Test information was obtained from academic textbooks and peer-reviewed literature.
[6]Test information as obtained from peer-reviewed literature only.

A.2. Psychometric Tests Excluded from Extraction and Evaluation

Table A-2 lists the psychometric tests that were excluded from the extraction and evaluation process and the reason for exclusion.

Table A-2. Psychometric Tests Excluded from Evaluation[1]

Reason for Exclusion from Evaluation	Test Name
Idiosyncratic[2]	Aberrant Behavior Checklist
	Ages & Stages Communication Scale (Parent Assessment)
	Ameil-Tyson and Gosselin Exam
	Audiometric evaluation using Modified Hughson-Westlake Procedure
	Autism Diagnostic Interview-Revised
	Autism Diagnostic Observation Schedules (Revised & Generic)
	Autism Treatment Evaluation Checklist for Parents
	Burt Recognition Test
	Burt Recognition Test-Revised
	Cambridge Neuropsychological Test Automated Battery (CANTAB)
	Clinical Global Impression-Severity Scale
	Comprehensive Developmental Inventory for Infants & Toddlers (CDIIT)
	Cook-Medley Hostility Index-Youth Version
	Dale & Bishop Grammar Rating
	Delayed Spatial Alternation Test
	Everts Behavior Rating Scale
	Go/No-Go Response Inhibition Paradigm
	Healthy Behavior Questionnaire
	Huttenlocher Motor Tasks
	Infant Behavior Questionnaire
	Infant Behavior Questionnaire-Revised
	Test of Attentional Performance for Children (KITAP)
	Localisation of Tactile Stimuli Test
	Mullen Scales of Early Learning
	Neurobehavioral Core Test Battery
	Neurological Examination for Subtle Signs
	Parent's Evaluation of Developmental Status
	Sheriden Gardiner Letter Matching Test
	Social Responsiveness Scale
	Static Motor Steadiness Test

Reason for Exclusion from Evaluation	Test Name
	Stycar Miniature Toy Test
	Test of Haptic Matching
	Toddler Temperament Scale
Insufficient information[3]	A-not-B Test
	Abnormal and Repetitive Behavior Scale
	Children's Category Test
	Clay Diagnostic Survey
	Conner's Continuous Performance Test (CCPT)
	Differential Reinforcement of Low Rate Schedules
	Frontal Assessment Battery
	Halstead-Reitan Battery
	Hong Kong List Learning Test
	Neurobehavioral Evaluation System (NES2)
	Neurobehavioral Evaluation System (NES3)
	Santa Ana Form Board
	Social Maturity Scale
	Twenty Statements about Language-Related Difficulties List
	Visual Expectation Paradigm
	Visual Recognition Memory (VRM) Paradigm
	Wide Range Assessment of Memory and Learning
	Wide Range of Assessment of Visual-Motor Abilities
	Woodcock-Johnson Test of Achievement (W-J-II)
	WPS Electronic Tapping Test

[1]Test names appear primarily as they are reported in studies included in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg. Additional variations of test names may be utilized in the neuropsychology literature.
[2]Tests were categorized as idiosyncratic if they appeared in a single publication considered in the in-progress EPA Toxicological review of Methylmercury, were used only in a specific population, and/or were used in a study that was not conducive to dose-response analysis (n = 33).
[3]Tests were categorized as having insufficient information if no manual was available and secondary sources, including peer-reviewed literature, did not provide enough information to assess a majority of the evaluation principles (n = 20).

Appendix B. Test Evaluation Tables

Table of Contents

Tables

B.1. Omnibus Tests

B.1.1. General Intelligence/IQ

Table B-1. Adequacy or Deficiency of Factors Affecting the Reliability, Validity, and Administration Standards of Psychometric Tests Used to Assess General Intelligence/IQ in Developmental Neurotoxicity Studies[1,2]

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
Kaufman Assessment Battery for Children (K-ABC); 1983	A	A	A	D[3]	A	A	NP[3]	This IQ test is normed for children 2.5–12.5 years of age. It is used less often than Wechsler scales in both research and clinical settings, contributing to less knowledge about criterion validity. It has 16 subtests, 6 of which are nonverbal. **Construct validity:** The correlations between K-ABC and other measures of IQ are low, except for the Achievement outcome. **Examiner qualifications:** Data were not present in the sources consulted for the extraction table.[4] **Outcomes, domain-specific subscales, and subtests:** The test comprises four domains—*Achievement*, *Mental Processing*, *Sequential Processing*, and *Simultaneous Processing*—and consists of 16 subtests. Achievement is based on six subtests (Expressive Vocabulary, Faces and Places, Arithmetic, Riddles, Reading/Decoding, Reading /Understanding); Sequential Processing is a composite of three subtests (Hand Movements, Number Recall, Triangles); and Simultaneous Processing is a composite of seven subtests (Magic Window, Face Recognition, Gestalt Closure, Word Order, Matrix Analogies, Spatial Memory, Photo Series). The Sequential

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
								Processing and Simultaneous Processing scales are combined to represent the Mental Processing Composite score. Composite standardized scores have a mean of 100 and standard deviation (SD) of 15. Each subtest standardized score has a mean of 10 and SD of 3. The subscales for this test that have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg are listed in the footnotes of evaluation tables for other domains, where applicable. Mental Processing appears in the *Executive Function* domain evaluation table. Subtests might be categorized in other domains but did not appear as outcomes in the literature queried for the in-progress EPA IRIS toxicological review of MeHg.
Kaufman Brief Intelligence Test (K-BIT); 1990	A	A	A[3]	A	A	A	NP[3]	The K-BIT is a quick screening test for IQ (15–30 minutes) and is normed for ages 4–90 years. This test has been used for IQ screening in research. It has fewer subtests than its parallel Wechsler test (i.e., Wechsler Abbreviated Scale of Intelligence—WASI [1999] and WASI-II [2011]). **Content validity:** Sources consulted and summarized in the extraction table provided little information on theoretical or content bases for the test; however, item choice was presumed to have been driven by the tests from which the K-BIT was derived (Kaufman Adult Intelligence Test and the Kaufman Assessment Battery for Children). **Examiner qualifications:** Data were not present in the sources consulted for the extraction table.

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
								Outcomes, domain-specific subscales, and subtests: The K-BIT consists of two subtests—Vocabulary Performance and Matrices. Scores from these tests are summed to provide a Composite IQ score (standardized mean = 100, SD = 15). The Vocabulary Performance test provides a Verbal Intelligence score, and the Matrices test provides a Nonverbal Intelligence score, each with a standardized mean of 100 and SD of 15. The subtests for the K-BIT have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are listed in the footnotes of evaluation tables for other domains, where applicable. These include the Vocabulary Performance subtest (*Verbal* domain) and Matrices subtest, also called Nonverbal Intelligence (*Visuospatial* domain).
McCarthy Scales of Children's Abilities (MSCA); 1972[3]	A	A	A	A	NP[3]	A	A	This MSCA was developed for children 2.5–8.5 years old. The test and its normative sample are generally considered outdated. **Year of publication:** The year is sometimes listed as 1970. **Criterion validity:** Although this test predicts scores on other tests, its capacity to predict developmental outcomes was not described in the sources consulted and summarized in the extraction table. **Outcomes, domain-specific subscales, and subtests:** This test uses 18 short subtests to produce a General Cognitive Index, which would be considered the IQ outcome (standardized mean = 100, SD = 16). The 18 subtests can be categorized into domains

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
								producing several subscales, each of which has a standardized score with a mean of 50 and SD of 10. The subscales for this test have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are listed in the footnotes of evaluation tables for other domains, where applicable. These include the Verbal Index (*Verbal* domain), Perceptual Index (*Visuospatial* domain), Quantitative Index (*Academic Achievement* domain), Memory Index (*Learning and Memory* domain), and Motor Index (*Motor Function* domain). An Executive Function Index is also listed among these subscales in the in-progress EPA IRIS toxicological review of MeHg but was not reported in the test manual.
Raven's Coloured Progressive Matrices (Raven's CPM); 1965	A	A	NP[3]	NP[3]	A	A	A	Unlike other tests listed in this evaluation table, Raven's CPM is not a classic omnibus IQ test as defined in this document under domains. It is an older test developed in the United Kingdom. Although it directly measures executive function, it has often been used as a brief method to evaluate general intelligence. The test produces an "IQ"-like outcome (although the norms do not apply to most populations) and, given its availability for decades, has been used by neurotoxicologists and other researchers. It was developed for children 5–11 years old but has also been used with adult populations. **Content validity:** Data were not present in the sources consulted for the extraction table.

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
								Construct validity: Data were not present in the sources consulted for the extraction table. **Outcomes, subscale, and subtest scores:** IQ outcomes are classified using five gradations from superior (>95th percentile) to defective (<5th percentile).
Raven's Standard Progressive Matrices (Raven's SPM); 1977	A	A	A	A	A	A	A	Unlike other tests listed in this evaluation table, Raven's SPM is not a classic omnibus IQ test as defined in this document under domains. It is an older test developed in the United Kingdom. Although it directly measures executive function, is has often been used as a brief method to evaluate general intelligence. The test produces an IQ-like outcome (although the norms do not apply to most populations) and, given its availability for decades, has been used by neurotoxicologists and other researchers. It is often used in mercury studies to evaluate parent intelligence as a control measure, not outcomes in children. It was developed for children 6–13.5 years old but has been used with adult populations. **Outcomes, subscale, and subtests scores:** Percentile scores (5th, 10th, 25th, 50th, 75th, 90th, and 95th) are determined based on total raw score for all items.
Stanford-Binet Intelligence Scale: Fourth Edition (S-B 4); 1986	A	A	A	A	A	A	A	Like the Wechsler scales, the Stanford-Binet scales have been used extensively both to assess general intelligence and to evaluate children with learning problems in school. The S-B 4 can be useful for individuals with very high or very low IQ as it is less subject to floor and ceiling effects compared with the Wechsler scales. The test can also be useful in longitudinal studies because many

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
								subtests can be given across the full age range, which is 2–90 years. **Outcomes, subscale, and subtest scores:** The S-B 4 provides a Full-Scale IQ score (standardized mean = 100, SD = 16). In addition, there are Verbal IQ and Nonverbal IQ scores and five factor indices (Fluid Reasoning, Knowledge, Quantitative Reasoning, Visual-Spatial Processing, and Working Memory), all with standardized means of 100 and SDs of 16. Subtest scores are standardized to a mean of 50 and SD of 8. The subtests for the S-B 4 have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are listed in the footnotes of evaluation tables for other domains, where applicable. These include Total Copying (*Visuospatial* domain) and Copying Test Recall and Bead Memory (*Learning and Memory* domain).
Stanford-Binet Intelligence Scale: Fifth Edition (S-B 5); 2003	A	A	A	A	A[3]	A	A	The S-B 5 is a newer version of the S-B 4 and has similar advantages. It should be noted that some of the S-B 4 subtests are not included in the S-B 5, so the S-B 5 cannot be used as an updated version for these subtests. **Criterion validity:** This test has predictive validity for levels of intelligence and other outcomes similar to the S-B 4. **Outcomes, subscale, and subtest scores:** The S-B 5 provides Full-Scale, Verbal, and Performance IQs, all with a standardized mean of 100 and SD of 15. In addition, five factor indices (Fluid Reasoning, Knowledge, Quantitative Processing, Visual-Spatial

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
								Processing, and Working Memory) are standardized to scores with a mean of 100 and SD of 15. Subtest scaled scores have a mean of 10 and SD of 3. The subtests for the S-B 5 have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are listed in the footnotes of evaluation tables for other domains, where applicable. These include the Knowledge Index (*Verbal* domain), Visuospatial Processing Index (*Visuospatial* domain), Quantitative Processing Index (*Academic Achievement* domain), and the Working Memory Index (*Executive Function* domain).
Wechsler Intelligence Scale for Children (WISC); 1949	A	A	A[3]	A	A[3]	A	NP[3]	Information on the WISC from the peer-reviewed literature was limited. Some of the conclusions below are based on experience using the test.[4] **Content validity:** The source consulted for the extraction table suggests there is no available information on content validity; however, personal experience[4] indicates adequate content validity. This version of the WISC is the basis for all future versions of the WISC. **Criterion validity:** The source consulted for the extraction table suggests criterion validity is inadequate; however, personal experience[4] indicates adequate criterion validity in pediatric populations. **Examiner qualifications:** Data were not present in the sources consulted for the extraction table.

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
Wechsler Intelligence Scale for Children-Revised (WISC-R); 1974	A	A	A	A	A[3]	A	A	This widely used standard test of general abilities is a revision of the original Wechsler Intelligence Scale for Children (WISC). The WISC-R includes 12 subtests, and 10 of 12 (5 verbal, 5 nonverbal) are used to determine IQ. Because of its widespread use, The WISC-R has been applied to the domain-specific functioning associated with neuropsychological assessment. It is designed for children aged 6–16 years. For all editions of the WISC, there is some overlap with the WAIS (adult version of Wechsler test) at age 16. In general, for persons of average or above-average intelligence, the WAIS tests are more appropriate for 16-year-olds. This is not true for children of less-than-average intelligence. **Criterion validity:** The WISC-R correlates with many other established measures of intelligence and predicts learning disabilities and other outcomes in children. **Outcomes, subscale, and subtest scores:** Outcome measures include Full-Scale, Verbal, and Performance IQ scores, all with a standardized mean of 100 and SD of 15. The 12 subtest scaled scores have a mean of 10 and SD of 3. The subtests for the WISC-R have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are listed in the footnotes of evaluation tables for other domains, where applicable. These include the Verbal IQ outcome and the Similarities subtest (*Verbal* domain) and the Performance IQ outcome and Block Designs subtest (*Visuospatial* domain).

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
Wechsler Intelligence Scale for Children-III (WISC-III); 1991	A	A	A	A	A	A	A	This revision of the WISC-R is similar to its predecessor but adds a new subtest, Symbol Search. The WISC-III further advances the idea of domain-specific assessment by including four index scores—Verbal Comprehension, Perceptual Organization, Freedom from Distractibility, and Processing Speed. **Outcomes, subscale, and subtest scores:** Outcome measures include Full-Scale, Verbal, and Performance IQ scores, all with a standardized mean of 100 and SD of 15. The four index scores also have a standardized mean of 100 and SD of 15, and the 13 subtest scaled scores have a mean of 10 and SD of 3. The index scores and subtests for the WISC-III have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are listed in the footnotes of evaluation tables for other domains, where applicable. These include the Verbal IQ outcome, Information subtest, and Vocabulary subtest (*Verbal* domain); the Performance IQ outcome and Block Designs subtest (*Visuospatial* domain); and the Coding subtest (*Motor Function* domain).
Wechsler Intelligence Scale for Children-IV (WISC-IV); 2003	A	A	A[3]	A	A[3]	A	A	The WISC-IV removed three subtests and added five new subtests relative to the WISC-III for a total of 15 subtests. It includes four index scores, but one has changed from Freedom from Distractibility to Processing Speed. **Content validity:** This test has content validity similar to that of prior versions of the WISC. It is also based on psychometric

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
								theory, which has affected the index outcome revisions and subtest additions. **Criterion validity:** This test has criterion validity similar to that of prior versions of the WISC. **Outcomes, subscale, and subtest scores:** Outcome measures include Full-Scale, Verbal, and Performance IQ scores (standardized mean = 100, SD = 15). The index scores have a mean of 100 and SD of 15, and the 15 subtest scaled scores have a mean of 10 and SD of 3. The index scores for the WISC-IV have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are listed in the footnotes of evaluation tables for other domains, where applicable. These include the Verbal Comprehension Index (*Verbal* domain), Processing Speed Index (*Miscellaneous* domain), and the Perceptual Reasoning Index (*Visuospatial* domain).
Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R); 1989	A	A	A	A	A	A	A	The WPPSI scales are designed for children aged 3 years to 7 years, 3 months. This test has some age overlap with the WISC scales at ages 6 years to 7 years, 3 months. In this age range, the WPPSI-R is considered too easy for some children or populations, making the WISC more appropriate. The test has 11 subtests that all contribute to the Full-Scale IQ, with 5 contributing to the Performance IQ and 6 contributing to the Verbal IQ. **Outcomes, subscale, and subtest scores:** Outcome measures include Full-Scale, Verbal, and Performance IQ scores (standardized mean = 100, SD = 15). Age-

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
								scaled scores (standardized mean = 10, SD = 3) are available for the subtests at 3-month age intervals. The index scores for the WPPSI-R have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are listed in the footnotes of evaluation tables for other domains, where applicable. These include Verbal IQ (*Verbal* domain) and Performance IQ (*Visuospatial* domain).
Wechsler Adult Intelligence Scale-Revised (WAIS-R); 1981	A	A	A	A	A[3]	A	A	This test is a revision of the 1955 Wechsler Adult Intelligence Scale (WAIS), which was based on the Wechsler-Bellevue Intelligence Scale (W-B) and contains the same 11 subtests in revised form. It is standardized for persons aged 16 years to 74 years, 11 months. It has ceiling and floor effects, meaning that IQ scores above 135 or below 70 may not be as precise as high or low scores on other intelligence tests. Because it has been widely applied, a great deal is known about the relationship between performance on WAIS-R subtests and outcomes such as localized brain damage and neuropsychiatric disorders. **Criterion validity:** This test's criterion validity derives at least in part from that of WAIS and W-B, which correlate highly with clinician ratings of individual intelligence, "empirical studies of several groups of known intellectual level" (manual, p. 49), and WAIS correlations with academic success (WAIS manual, p. 50). **Outcomes, subscale, and subtest scores:** Administration of the WAIS-R allows derivation of Full-Scale, Verbal, and

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
								Performance IQs (standardized mean = 100, SD = 15). In addition, each subtest can be scored according to age-appropriate norms with a standardized mean of 10 and SD of 3. The subtests for the WAIS-R have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are listed in the footnotes of evaluation tables for other domains, where applicable. These include the Block Design subtest (*Visuospatial* domain).
Wechsler Adult Intelligence Scale-III (WAIS-III); 1997	A	A	A	A	A	A	A	This revision of the WAIS-R is designed for ages 16–89 years. Because its norms start at age 16 years, it can be an appropriate test for older children of average or above-average intelligence and can be used in place of the WISC scales. There are 14 subtests included in the WAIS-III—11 revisions of WAIS-R subtests and 3 new subtests. This version of the Wechsler adult scales has a somewhat lower floor (45) and higher ceiling (155) than the WAIS-R, but still does not measure extremes of intelligence. **Outcomes, subscale, and subtest scores:** Administration of the WAIS-III allows derivation of Full-Scale, Verbal, and Performance IQs (standardized mean = 100, SD = 15) as in prior adult scales. In addition, there are four subscale outcomes (Verbal Comprehension Index, Perceptual Organization Index, Working Memory Index, and Processing Speed Index), with standardized means of 100 and SDs of 15. Subtests have age norms that produce age-scaled scores with standardized means of 10 and SDs of 3.

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
								The index scores and subtests for the WAIS-III have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are listed in the footnotes of evaluation tables for other domains, where applicable. These include the Vocabulary subtest, Similarities subtest, and Comprehension subtest (*Verbal* domain); the Block Design subtest, Matrix Reasoning subtest, Picture Completion subtest, Object Assembly subtest, and Picture Arrangement subtest (*Visuospatial* domain); Digit Symbol/Coding subtest (*Motor Function* domain); and the Processing Speed Index (*Miscellaneous* domain).

[1]A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2]Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3]See Notes column.
[4]Adequacy and deficiency are determined by the professional experience and knowledge of the co-author Dr. Roberta F. White.
*Subtests or scales within tests that belong to a different domain may be applicable to general intelligence/IQ. This includes the Bayley Scales of Infant Development-II: Mental Development Index (*Developmental* domain).

Table B-2. Adequacy or Deficiency of Factors Affecting the Normative Data of Psychometric Tests Used to Assess General Intelligence/IQ in Developmental Neurotoxicity Studies[1,2]

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
Kaufman Assessment Battery for Children (K-ABC); 1983	A	NA	A	NA	A	NA	None.
Kaufman Brief Intelligence Test (K-BIT); 1990	NP[3]	NP[3]	A	A	NP[3]	NP[3]	**Subject age (child & adult):** The standardization sample was divided into four wide age bands (4–6, 7–19, 20–44, 45–92), but the sources cited in the extraction table state that it is not clear if these were used for the age-standardized scores or if they were based on more precise data. Since the sample size was 2,022 and ages of the sample were "proportional," it is not clear how many people were in each age band, but there would be many age bands needed for the children and adults. Ultimately, it is unclear from the sources consulted for the extraction table if age bands were further divided for normative scores. **Sample size (child & adult):** 2,022 participants in total comprised the sample. It is not clear if age bands were sufficiently well populated for children. Assuming 10-year age bands, there may have been a large enough sample of adults.
McCarthy Scales of Children's Abilities (MSCA); 1972[3]	A	NA	A	NA	A	NA	**Year of publication:** The year is sometimes listed as 1970.
Raven's Coloured Progressive Matrices (Raven's CPM); 1965	A	D[3]	D[3]	D[3]	D[3]	D[3]	**Subject age (adult):** Adult norms exist for persons 60–85 years and do not represent younger adults. **Population representation (child & adult):** Child and adult norms are based on populations from a single area in the United Kingdom.

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
							Sample size (child & adult): Child norms appear to be based on 627 children. It is not clear what the age ranges are for the adult standardization sample. Adult norms are only based on 271 individuals from the United Kingdom.
Raven's Standard Progressive Matrices (Raven's SPM); 1977	A	A	D[3]	D[3]	A	A	**Population representation (child & adult):** Child samples are from limited areas of the United Kingdom. Adult normative samples include military personnel and civilians (undefined).
Stanford-Binet Intelligence Scale: Fourth Edition (S-B 4); 1986	A	A	A	A	A	A	None.
Stanford-Binet Intelligence Scale: Fifth Edition (S-B 5); 2003	A	A	A	A	A	A	None.
Wechsler Intelligence Scale for Children (WISC); 1949	A	NA	A	NA	A	NA	None.
Wechsler Intelligence Scale for Children-Revised (WISC-R); 1974	A	NA	A	NA	A	NA	None.
Wechsler Intelligence Scale for Children-III	A	NA	A	NA	A	NA	None.

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
(WISC-III); 1991							
Wechsler Intelligence Scale for Children-IV (WISC-IV); 2003	A	NA	A	NA	A	NA	None.
Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R); 1989	A	NA	A	NA	A	NA	None.
Wechsler Adult Intelligence Scale-Revised (WAIS-R); 1981	A	A	A	A	A	A	None.
Wechsler Adult Intelligence Scale-III (WAIS-III); 1997	A	A	A	A	A	A	None.

[1]A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2]Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3]See Notes column.
[4]The link to the extraction table is provided in Appendix C.

B.1.2. Academic Achievement

Table B-3. Adequacy or Deficiency of Factors Affecting the Reliability, Validity, and Administration Standards of Psychometric Tests Used to Assess Academic Achievement in Developmental Neurotoxicity Studies[1,2]

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
KeyMath Diagnostic Arithmetic Test; 1976[3]	A	NP[3]	A	A	A[3]	A[3]	NP[3]	**Year of publication:** The KeyMath, American edition was first published by American Guidance Service in 1976. The Canadian edition was published by Psycan in 1979, and a supplemental norms table was published by American Guidance Service in 1983.This test is appropriate for children "preschool" age through grade 6. **Test-retest reliability:** Data were not present in the sources consulted for the extraction table.[4] **Criterion validity:** No specific information on criterion validity is reported in the sources consulted and summarized in the extraction table; however, the information on predictive validity indicates that criterion validity is adequate. **Instructions/manual:** Data in the extraction table come from sources other than the manual, and ratings reflect statements from those sources. There are manuals available that could be consulted. **Examiner qualifications:** Data were not present in the sources consulted for the extraction table.

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
KeyMath Diagnostic Arithmetic Test-Revised (KeyMath R); 1991	A	NP[3]	A	A	A[3]	A[3]	NP[3]	This test is appropriate for children ages 5 years, 6 months through 15 years, 5 months (or grades K through 8). **Test-retest reliability:** Data were not present in the sources consulted for the extraction table. **Criterion validity:** No specific information on criterion validity is reported in the sources consulted and summarized in the extraction table; however, the information on predictive validity indicates that criterion validity is adequate. **Instructions/manual:** Data in the extraction table come from sources other than the manual, and ratings reflect statements from those sources. There are other manuals that could be consulted. **Examiner qualifications:** Data were not present in the sources consulted for the extraction table.
Wechsler Individual Achievement Test (WIAT); 1992	A	A	A	A	D[3]	A	A	The WIAT is designed for use on children grades K through 12. **Criterion validity:** Correlations between WIAT test performance and school grades are poor.

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
Woodcock-Johnson III Test of Achievement (WJ III ACH); 2001	A	A	A	A	A[3]	A	A	Administration of individual subtests within the WJ III varies depending upon age. Norms are available by month from age 24 months to 19 years and by year from ages 20 to 90+ years. **Criterion validity:** No specific information on criterion validity is reported in the sources consulted and summarized in the extraction table; however, the information on discriminant validity indicates that criterion validity is adequate. **Outcomes, subscales, and subtest scores:** Outcomes are scored with a standardized mean of 100 and standard deviation (SD) of 15. The index scores and subtests for the WJ III that have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg are Antonyms, Concept Formation, Decision Speed, Memory for Words, Numbers Reversed, Spatial Relations, Synonyms, Verbal Analogies, Visual Matching, Applied Problems, Calculation, Letter-Word, Math Fluency, and Passage Comprehension (*Academic Achievement* domain).

[1]A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2]Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3]See Notes column.
[4]The link to the extraction table is provided in Appendix C.
*Subtests or scales within tests that belong to a different domain may be applicable to academic achievement. These include the Kaufman Assessment Battery for Children (KAB-C): Achievement Index (*IQ* domain); McCarthy Scales of Children's Abilities (MSCA): Quantitative Index (*IQ* domain); and Stanford-Binet, 5th Edition: Quantitative Processing Index (*IQ* domain).

Table B-4. Adequacy or Deficiency of Factors Affecting the Normative Data of Psychometric Tests Used to Assess Academic Achievement in Developmental Neurotoxicity Studies[1,2]

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
KeyMath Diagnostic Arithmetic Test; 1976[3]	NP[3]	NA	A	NA	A	NA	**Year of publication:** The KeyMath, American edition, was first published by American Guidance Service in 1976. The Canadian edition was published by Psycan in 1979, and a supplemental norms table was published by American Guidance Service in 1983. **Subject age (child):** Data were not present in the sources consulted for the extraction table.[4]
KeyMath Diagnostic Arithmetic Test-Revised (KeyMath R); 1991	A[3]	NA	A	NA	A	NA	**Subject age (child):** Grade-level normative bands are used instead of age-based normative bands. Although there may be some age misclassification (e.g., a 16-year-old middle school student), this is thought to occur infrequently.
Wechsler Individual Achievement Test (WIAT); 1992	A	NA	A	NA	A	NA	None.
Woodcock-Johnson III Test of Achievement (WJ III ACH); 2001	A	NA	A	NA	A	NA	None.

[1] A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2] Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3] See Notes column.
[4] The link to the extraction table is provided in Appendix C.

B.1.3. Developmental

Table B-5. Adequacy or Deficiency of Factors Affecting the Reliability, Validity, and Administration Standards of Psychometric Tests Used to Assess the Developmental Domain in Developmental Neurotoxicity Studies[1,2]

Test; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/Manual	Examiner Qualifications	
Bayley Scales of Infant Development (BSID-1);1969[3]	A	A	A[3]	A	A[3]	A[3]	A[3]	This test assesses aspects of cognition labeled "mental" in the test manual. It assesses motor and social behaviors in very young children—ages 2 months to 2 years of age. **Content validity:** Evaluation is based on correlations with BSID-2. **Criterion validity:** Criterion validity depends on age. When the test is given at very young ages, correlations with both later-in-life BSID scores and later IQs can be very low. **Instructions/manual:** Data in the extraction table[4] come from text (Spreen and Strauss), and ratings reflect statements from that text. There is a manual with explicit instructions.[5] **Examiner qualifications:** These are not well defined. This test is difficult to give without training from an experienced administrator, and the instructions allow for a great deal of leeway, which can ultimately affect scores. **Outcomes, subscale, and subtest scores:** The index scores and subtests for the BSID-1 have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are the Mental Development Index (MDI) (*IQ* domain) and Psychomotor Index (*Visuospatial* domain), which represent the overall outcome scores for the test with standardized means of 100 and standard deviations (SDs) of 15.

Test; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/Manual	Examiner Qualifications	
Bayley Scales of Infant Development, Second edition (BSID-II); 1993[3]	A	A	A	A	A	A	A[3]	The age range for the test is 1 month to 42 months. **Year of publication:** The publication year of 1993 is from the Psychological Corporation by author Nancy Bayley. The original BSID was copyrighted in 1969. This test is a revision of the BSID-I. **Examiner qualifications:** Examiners require extensive training. **Outcomes, subscale, and subtest scores:** The Mental and Motor Index scores have a standardized mean of 100 and SD of 15. The index scores for this test have been identified in the epidemiologic literature for in-progress EPA IRIS toxicological review of MeHg and are listed in the footnotes of evaluation tables for other domains, where applicable. These include the Mental Development Index (*IQ* domain) and the Motor Index (*Motor Function* domain).
Bayley Scales of Infant Development-III (BSID-3); 2006	A	A	A	A	A	A	A[3]	The BSID-III represents the second revision of the test and is appropriate for children aged 1 month to 42 months. **Examiner qualifications:** Examiners require extensive training. **Outcomes, subscale, and subtest scores:** The test includes three domains—*Cognitive*, *Language*, and *Motor*—which have subscale scores. Domains and subtests can be scored with age-adjusted normative outcomes of several types.
Brazelton Newborn Assessment Scale (NBAS); 1973	D[3]	D[3]	A	D[3]	A[3]	A	A[3]	This test was developed for infants 1 day to 1 month, although some items can be administered through 10 weeks old. It is used clinically and may be more valuable in clinical situations in which testing is done by an experienced clinician.[5] **Internal consistency:** Data suggest inter-item consistency ranges from 0.15 to 0.52. Low correlations may in part reflect heterogeneity of test items. (In addition, the review of quantitative

Test; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/Manual	Examiner Qualifications	
								outcomes suggests that data analysis is challenging due to the correlations of the behavioral items on the inventory and nature of the data.) **Test-retest reliability:** This test measures moment-to-moment state behaviors in children and appears to have face validity, but this makes it unreliable in some ways.[5] **Construct validity:** This test does not correlate well with other tests. **Criterion validity:** The test correctly predicted mild and moderate disability among preterm infants. **Examiner qualifications:** Research assistants with experience in the care and assessment of young infants can reliably administer these scales with extensive training from a certified examiner. **Outcomes, subscale, and subtest scores:** This test produces 6 main cluster scores ranging from 1 to 9, 7 supplementary scores ranging from 1 to 9, and 18 reflex/motor scores ranging from 0 to 3. The index scores for this test have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are Autonomic Stability, Habituation, Motor, Orientation, Range of State, Reflexes, and Regulation of State (*Developmental* domain).
Denver Development Screening Test (DDST); 1967	NP[3]	A	A[3]	A[3]	D[3]	A[3]	A[3]	DDST is designed for children from birth to 6 years. **Internal consistency:** Data were not present in the sources consulted for the extraction table. **Content validity:** This test has inherent content validity as a measure of developmental milestones. **Construct and criterion validity:** Data in the extraction table show good correlations with other developmental tests in 236 pediatric patients. A study on validity in 2,569 children is described in the extraction table, but the outcomes are not clearly described. Low sensitivity and under-referral rate

Test; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/Manual	Examiner Qualifications	
								indicate that the instrument is not consistently predictive of developmental deficits, which supports a criterion validity rating of deficient. **Instructions/manual:** Data in the extraction table come from validation papers, and ratings reflect statements from these papers. There is a test manual. **Examiner qualifications:** Examiner qualifications are discussed but are not specific. **Outcomes, subscale, and subtest scores:** The four scored developmental areas include gross motor, fine motor, language, and social and personal skills. They are scored as normal, questionable, or abnormal. The scores for this test have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are listed in the footnotes of evaluation tables for other domains, where applicable. They include Communication (language area, *Verbal* domain); Fine Motor Skills and Gross Motor Skills (*Motor Function* domain); and Social Skills (social/emotional area, *Social/Emotional* domain).
Denver Development Screening Test-II (DDST II); 1992	A	A	A[3]	A[3]	A[3]	A[3]	A[3]	This test is designed for neonates through 6 years of age. **Content validity:** The content validity rating is based on DDST. **Construct and criterion validity:** Validity and correlations with other developmental tests and with diagnostic outcomes are adequate in summaries provided in the extraction table for children at older ages, but not for children 2 years of age or younger. These studies are based on highly specific populations. **Instructions/manual**: Data in the extraction table come from validation papers, and ratings reflect statements from these papers. There is a test manual.

Test; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/Manual	Examiner Qualifications	
								Examiner qualifications: Examiner qualifications are discussed but are not specific.
Fagan Test of Infant Intelligence; 1985	D[3]	D[3]	A[3]	A[3]	D[3]	A[3]	A[3]	This test is designed for infants ages 27–52 weeks and is corrected for prematurity. **Reliability:** According to sources consulted for the extraction table, internal consistency and test-retest reliability were low. This is probably due to the variety of test items and ages of infants and may not be a surmountable issue for this type of test. **Validity (in general):** Studies reported in extraction table show adequate values but are based on small sample sizes. **Criterion validity:** Per the data in the extraction table, the test predicts normal development adequately but is deficient for abnormal development. **Instructions/manual:** There are explicit instructions. **Examiner qualifications:** Examiners are trained using a training manual and videotape. **Outcomes, subtest, and subscale scores:** The scores for this test have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are listed in the footnotes of evaluation tables for other domains, where applicable. This includes the Visual Attention outcome (*Attention* domain) and Visual Recognition Memory outcome (*Learning and Memory* domain), which have associated "novelty scores" determined by the computer when the baby responds to stimuli.
Gesell Developmental Schedules; 1925[3]	NP[3]	NP[3]	A[3]	A	NP[3]	A	NP[3]	This test was developed for children from neonate to 56 weeks. **Year of publication:** The Gesell Developmental Schedules appear to have been republished or revised in 1940 and 1949.

Test; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/Manual	Examiner Qualifications	
								Internal consistency: Data were not present in the sources consulted for the extraction table. **Test-retest reliability:** This is not discussed in the papers consulted for the extraction table. The data summarized in the extraction table indicate that test performance at 3-month intervals starting at 3 months was not well correlated with that at later ages (which is more a measure of validity), but the sources consulted do not provide data on test-retest reliability at the same age. **Content validity:** At the time this test was developed, it had content validity; however, this test was based on theory from an earlier time, and what is known about neurodevelopment has evolved since then. **Criterion validity:** Data were not present in the sources consulted for the extraction table. **Examiner qualifications:** Data were not present in the sources consulted for the extraction table. It is known that examiners require extensive training.
Griffith Mental Development Scales (GMDS); 1970	NP[3]	A	A	A	NP[3]	NP[3]	A[3]	The age range for this test is birth to 1 year, 11 months. Since it was published and normed around 1970, the timeframe in which it was applied to research must be considered. **Internal consistency:** Data were not present in the sources consulted for the extraction table. **Criterion validity**: Data were not present in the sources consulted for the extraction table. **Instructions/manual:** Data were not present in the sources consulted for the extraction table. Instructions may be available in papers that were not consulted for the extraction table.[5] **Examiner qualifications:** Information on examiner qualifications was not directly available from sources consulted for the extraction table; however, "Griffiths certified examiners" are mentioned,

Test; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/Manual	Examiner Qualifications	
								indicating that expert examiners trained other examiners. **Outcomes, domain-specific subscales, and subtests:** Scores for this test as a whole (General Quotient) and for the five subscales are based on standardized mean of 100 and SD of 15. The subscales for this test have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are listed in the footnotes of evaluation tables for other domains, where applicable. These include Hand-eye Coordination Index and Locomotion Index (*Motor Function* domain), Hearing and Speech Index (*Verbal* domain), Hand-eye Performance Index (*Visuospatial* domain), and Personal/Social Skills subscale (*Social-Emotional* domain).
Kyoto Scale of Psychological Development (KSPD; K-test); 2002	A	NP[3]	A[3]	A[3]	A[3]	A[3]	A	It appears that this test was developed for Japanese children ages 0–5 years. **Test-retest reliability:** Data were not present in the sources consulted for the extraction table. **Validity (in general):** Studies in Japan show that the test correlates with another Japanese developmental test and predicts developmental abnormalities of various types. **Criterion validity:** The sources consulted for the extraction table describe a study of children with known developmental disorders. As a group, they performed below average on the developmental quotient (DQ) outcome. **Instructions/manual:** The sources consulted for the extraction table cite administration manuals in Japan.

Test; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/Manual	Examiner Qualifications	
Neonatal Intensive Care Unit Network Neurobehavioral Scale (NNNS); 2004	A	NP[3]	A	A	A	A	A[3]	Gestational ages 30–48 weeks is the target age range for this test. **Test-retest reliability:** Data were not present in the sources consulted for the extraction table. **Examiner qualifications:** Research assistants with experience in the care and assessment of young infants can reliably administer these scales with extensive training from a certified examiner.
Prechtl General Movement Assessment (GMA); 1977[3]	NP[3]	A[3]	A[3]	A	A[3]	A[3]	NP[3]	Age of administration is 0–4 months. **Year of publication:** The year is sometimes listed as 2004, specifically to general movements (GMs) or the General Movements Assessment (GMsA). **Internal consistency:** Data were not present in the sources consulted for the extraction table. **Test-retest reliability:** The extraction table mentions an intra-rater study that is small but indicates adequacy. **Content validity:** Content validity is based on theory. **Criterion validity:** Predictive validity has been indicated as adequate in a few small studies. This test seems to predict cerebral palsy well. **Instructions/manual:** Data in the extraction table come from validation papers, and ratings reflect statements from these papers. A formal assessment system appears to exist and could be consulted. **Examiner qualifications:** The sources consulted for the extraction table report that videos are used for scoring, but do not indicate who is qualified to rate them.

[1] A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2] Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3] See Notes column.
[4] The link to the extraction table is provided in Appendix C.
[5] Adequacy and deficiency are determined by the professional experience and knowledge of the co-author Dr. Roberta F. White.

Table B-6. Adequacy or Deficiency of Factors Affecting the Reliability, Validity, and Administration Standards of Psychometric Tests Used to Assess the Developmental Domain in Developmental Neurotoxicity Studies[1,2]

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
Bayley Scales of Infant Development (BSID-1); 1969[3]	NP[3]	NA	A[3]	NA	A	NA	**Year of publication:** The year is sometimes listed as 1969. The manual may be available at Psychological Corporation. **Subject age:** The sources consulted for the extraction table do not provide information on the age bands in the normative sample. **Population representation:** The standardization sample is noted in the extraction table as being reasonably representative of the U.S. population, but is otherwise not described.
Bayley Scales of Infant Development, Second edition (BSID-II); 1993[3]	A	NA	A	NA	A	NA	**Year of publication:** The year of 1993 is from the Psychological Corporation by author Nancy Bayley. The original BSID was copyrighted in 1969.
Bayley Scales of Infant Development-III (BSID-3); 2006	A	NA	A	NA	A	NA	None.
Brazelton Newborn Assessment Scale (NBAS); 1973	NP[3]	NA	NP[3]	NA	NP[3]	NA	There was inadequate information on test norms in the sources consulted for the extraction table.
Denver Development Screening Test (DDST); 1967	NP[3]	NA	D[3]	NA	A	NA	**Subject age (child):** Age bands are not described in validation studies per the extraction table but may be in the manual. **Population representation (child):** The standardization sample is limited to Denver children. The representativeness was not described in sources consulted for the extraction table.

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
Denver Development Screening Test-II (DDST II); 1992	NP[3]	NA	D[3]	NA	A	NA	**Subject age (child):** Age bands were not described in the sources consulted for the extraction table but may be available in the manual. **Population representation (child):** Several specific population normative studies are described in the extraction table. The initial sample was 2,096 children "from all over Colorado."
Fagan Test of Infant Intelligence; 1985	NP[3]	NA	D[3]	NA	D[3]	NA	**Subject age (child):** The sources consulted for the extraction table did not provide information on age bands used in the normative sample. **Population representation (child):** The normative sample was not designed to represent a population. **Sample size (child):** The normative sample size was relatively small (<250 children).
Gesell Developmental Schedules; 1925[3]	A[3]	NA	D[3]	NA	A[3]	NA	**Year of publication:** The Gesell Developmental Schedules appears to have been republished or revised in 1940 and 1949. **Subject age (child):** Age bands of 4-week intervals from newborn to 56 weeks were used. **Population representation (child):** The standardization sample consisted of 107 middle class infants from North America. **Sample size (child):** Although the sample size only consisted of 107 infants, they were repeatedly tested, which yielded repeated measures of the same children.

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
Griffith Mental Development Scales (GMDS); 1970	NP[3]	NA	NP[3]	NA	A	NA	**Subject age (child):** Infants and children ages 6, 12, and 24 months were included in the standardization sample. There is no information in the sources consulted and summarized in the extraction table on the age bands used to calculate the outcomes. **Population representation (child):** Information on the representativeness of the sample from the United Kingdom was not available from sources consulted and summarized in the extraction table.
Kyoto Scale of Psychological Development (KSPD; K-test); 2002	NP[3]	NA	NP[3]	NA	A	NA	**Subject age (child):** Normative studies are not detailed in the sources consulted for the extraction table. **Population representation (child):** Normative study populations are not detailed in the sources consulted for the extraction table.
Neonatal Intensive Care Unit Network Neurobehavioral Scale (NNNS); 2004	NP[3]	NA	D[3]	NA	A	NA	**Subject age (child):** No age band data are present in the sources consulted for the extraction table. **Population representation (child):** A normative study of children in Boston is described in the extraction table, which would not be considered representative.
Prechtl General Movement Assessment (GMA); 1977	NP[3]	NA	D[3]	NA	A[3]	NA	**Year of publication:** The year is sometimes listed as 2004, specifically to general movements (GMs) or General Movements Assessment (GMsA). **Subject age, population representation, sample size (child):** The data in the extraction table do not suggest that true norms or normative samples exist. A sample of 233 infants for whom 783 video recordings were available is described; all children were at high risk for neurodevelopmental disorders.

[1]A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2]Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3]See Notes column.

B.1.4. Neuropsychological Assessment Batteries

Table B-7. Adequacy or Deficiency of Factors Affecting the Reliability, Validity, and Administration Standards of Psychometric Tests Used to Assess Neuropsychological Assessment Batteries in Developmental Neurotoxicity Studies[1,2]

Test; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
Developmental Neuropsychological Assessment (NEPSY); 1998	NP[3]	A	A	A	A	A	A[3]	This test is designed for children 3–12 years of age. **Internal consistency:** Data were not present in the sources consulted for the extraction table. To be meaningful, this would have to be calculated separately for each subtest or domain, as the subtests and subscales measure different constructs associated with brain function and brain damage. **Examiner Qualifications:** This test is a difficult instrument to master when all or many subtests are administered (some studies select only one or two subtests, as there is not an overall omnibus score result). The use of multiple NEPSY subtests requires extensive training and supervision by a developmental neuropsychologist. **Outcomes, subscale, and subtest scores:** Composite domain scores (standardized mean = 100, standard deviation [SD] = 15) are calculated from the standard scores associated with subtests (mean = 10, SD = 3). The applicable subtests depend upon the age of the child. The domain composite scores for this test have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are listed in the footnotes of evaluation tables for other domains, where applicable. These include the Attention subscale and the Executive Function subscale.

Test; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
Neurobehavioral Evaluation System (NES); 1985	A	A	A	A[3]	A[3]	A[3]	A	The NES is designed for adults; some subtests have been adapted for children. This is a computer-administered test designed for research; outcomes are the scores on tasks which, in research, are adjusted for relevant variables. **Construct validity:** The NES is correlated with analogous noncomputerized tests. **Criterion validity:** The Adequate rating is based on personal knowledge and research.[4] Specific subtests have been found to be associated with toxicant exposure and certain neurological disorders. **Instructions/manual:** The NES is a computer-administered test with an examiner present.

[1]A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2]Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3]See Notes column.
[4]Adequacy and deficiency are determined by the professional experience and knowledge of the co-author Dr. Roberta F. White.

Table B-8. Adequacy or Deficiency of Factors Affecting the Normative Data of Psychometric Tests Used to Assess the Neuropsychological Assessment Batteries Domain in Developmental Neurotoxicity Studies[1,2]

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
Developmental Neuropsychological Assessment (NEPSY); 1998	A	NA	A	NA	A	NA	None.
Neurobehavioral Evaluation System (NES); 1985	NP[3]	NP[3]	NP[3]	NP[3]	NP[3]	NP[3]	**Norms (child & adult):** The NES does not utilize normative data; outcomes are raw scores, adjusted for factors including age, gender, and education.

[1] A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2] Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3] See Notes column.

B.2. Clinical Assessment Instruments

B.2.1. Clinical Conditions

Table B-9. Adequacy or Deficiency of Factors Affecting the Reliability, Validity, and Administration Standards of Psychometric Tests Used to Assess Clinical Conditions in Developmental Neurotoxicity studies[1,2]

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
Autism Spectrum Quotient (AQ); 2001	A	A	A	A	A	A	A[3]	This test has a child version that starts at age 4, an "adolescent" version for ages 9.8–15.4 years, and an adult version for ages 16 and above. **Examiner qualifications:** This is a self-administered questionnaire.
Barkley Adult ADHD Rating Scale-IV; 2011	A	A	A[3]	A	A[3]	A	A[3]	The age range for the test is 15–89 years. **Content validity:** The rating is based on Diagnostic and Statistical Manual (DSM) criteria for developmental disorders. **Criterion validity:** Convergent and divergent validities are demonstrated. **Examiner qualifications:** This is a self-reported measure.
Childhood Asperger's Syndrome Test (CAST); 2002	NP[3]	A	A[3]	A[3]	A	A	A[3]	The age range for this test is 4–11 years. **Internal consistency:** Data were not present in the sources consulted for the extraction table. **Content validity**: This test uses items that reflect Asperger's symptoms. **Construct validity:** The results from the criterion validity studies suggest construct validity. **Examiner qualifications:** This is a parent-completed questionnaire.
Childhood Autism Rating	A	NP[3]	A	A[3]	A	A	A[3]	This test is designed for children up to 10 years of age.

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
Scale (CARS); 1980								**Test-retest reliability:** Data were not present in the sources consulted for the extraction table. **Construct validity:** The results from the criterion validity studies suggest construct validity. **Examiner qualifications:** Sources consulted for the extraction table indicate that examiners have been trained by the test authors and that clinical psychologists have carried out the test in at least one study.
Conners' Parent Rating Scale-Revised (CPRS-R); 1997[3]	A	A	A	A	A	A[3]	A	The parent and teacher scales are essentially the same test and manual. Parents rate children 3–17 years of age. **Year of publication:** This test was also published in 2000 and 2001. **Instructions/manual:** It is unclear how explicit the manual is regarding instructions to raters. **Outcomes, subscales, and subtest scores:** The index scores and subtests for the CPRS-R have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are ADHD Index (T-scores based on normative sample) (*Clinical Conditions* domain), and Behavioral Index, Externalizing Problems, Hyperactivity Index, and Oppositional Index (*Social-Emotional* domain).
Conners' Teacher Rating Scale-Revised (CTRS-R); 1997[3]	A	A	A	A	A	A[3]	A	The parent and teacher scales are essentially the same test and manual. Teachers rate children 3–17 years of age. **Year of publication:** This test was also published in 2000 and 2001. **Instructions/manual:** It is unclear how explicit the manual is regarding instructions to raters.

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
Spence Children's Anxiety Scale (SCAS); 1994	A	A	A	A	NP[3]	A	A[3]	The age range for this test is 8–12 years. **Criterion validity:** Data were not present in the sources consulted for the extraction table. **Examiner qualifications:** This is a self-report questionnaire completed by children.

[1]A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2]Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3]See Notes column.
*Subtests or scales within tests that belong to a different domain may be applicable to clinical conditions. This includes the Conners' Parent Rating Scale-R: ADHD Index (*Attention* domain).

Table B-10. Adequacy or Deficiency of Factors Affecting the Normative Data Standards of Psychometric Tests Used to Assess Clinical Conditions in Developmental Neurotoxicity Studies[1,2]

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
Autism Spectrum Quotient (AQ); 2001	NP[3]	D[3]	NP[3]	A[3]	NP[3]	D[3]	**Subject age, population representation, and sample size (child):** Data are not present in the sources consulted for the extraction table. **Subject age (adult):** Adult samples are obtained from a narrow age range of young adults with no age bands. **Population representation (adult):** The normative population includes autism cases plus randomly selected controls from East Anglia. **Sample size (adult):** The adult norms are based on a sample of 58 adults with autism spectrum disorders and 174 controls.
Barkley Adult ADHD Rating Scale-IV; 2011	NA	A[3]	NA	A[3]	NA	A	**Subject age (adult):** Age bands are not discussed in the extraction table. The test is normed for adults 18–89 years. **Population representation (adult):** Normative data are drawn from a U.S. census-matched sample. The test is representative for the United States.
Childhood Asperger's Syndrome Test (CAST); 2002	D[3]	NA	D[3]	NA	D[3]	NA	**Subject age, population representation, and sample size (child):** The sources consulted for the extraction table suggest normative studies have not been conducted. The instrument was developed in the United Kingdom.
Childhood Autism Rating Scale (CARS); 1980	D[3]	NA	D[3]	NA	A	NA	**Subject age (child):** Over half of the normative age sample was less than 6 years of age, and only 11% were aged 10 years or older. Thus, the number of children in some of the age bins appears to be insufficient.

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
							Population representation (child): Normative data use a convenience sample with very specific characteristics.
Conners' Parent Rating Scale-Revised (CPRS-R); 1997	A	NA	A	NA	A	NA	None.
Conners' Teacher Rating Scale-Revised (CTRS-R); 1997	A	NA	A	NA	A	NA	None.
Spence Children's Anxiety Scale (SCAS); 1994	NP[3]	NA	D[3]	NA	A[3]	NA	**Subject age (child):** Age bands are not reported in the sources consulted for the extraction table. **Population representation (child):** Normative data use a convenience sample in Brisbane. **Sample size (child):** Confirmatory factor analytic studies have been done on approximately 1,400 children.

[1]A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2]Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3]See Notes column.

B.2.2. Mental Status

Table B-11. Adequacy of Factors Affecting the Reliability, Validity, and Administration Standards of Tests Used to Assess Mental Status in Developmental Neurotoxicity Studies[1,2]

Test; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
Mini-Mental State Examination (MMSE); 1975	NP[3]	A	A[3]	A[3]	A	A	NP[3]	The MMSE is typically administered to adults and not children. **Internal consistency:** Internal consistency may be limited due to heterogeneity of items measuring different constructs. **Content validity:** The adequate rating is based on personal knowledge; the test assesses common mental status domains.[4] **Construct validity:** The adequate rating is based on correlation with other similar tests. **Examiner qualifications:** This test can be given by a wide variety of practitioners.[4]

[1] A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2] Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3] See Notes column.
[4] Adequacy and deficiency are determined by the professional experience and knowledge of the co-author Dr. Roberta F. White.

Table B-12. Adequacy or Deficiency of Factors Affecting the Normative Data of Psychometric Tests Used to Assess Mental Status in Developmental Neurotoxicity Studies[1,2]

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
Mini-Mental State Examination (MMSE); 1975	NA	NP[3]	NA	NP[3]	NA	NP[3]	**Norms (adult):** No information on normative data was provided in the sources consulted for the extraction table.

[1] A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2] Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.

B.3. Domain-specific Tests

B.3.1. Attention

Table B-13. Adequacy or Deficiency of Factors Affecting the Reliability, Validity, and Administration Standards of Psychometric Tests Used to Assess Attention in Developmental Neurotoxicity Studies[1,2]

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/Manual	Examiner Qualifications	
Conners' Continuous Performance Test-II (CPT-II); 2000[3]	A	A	A	A	A	A[3]	NP[3]	There are multiple versions of tests assessing the Continuous Performance Test (CPT) paradigm (including auditory versions). CPT is an older type of test, and several versions have been used in the neurotoxicology literature. The Conners' version is appropriate for ages >6 years. **Year of publication:** This test was also published in 2004 and printed in Canada in 2006. **Instructions/manual:** Manual and instructions exist but are not detailed with regard to instructions. **Examiner qualifications:** Data were not present in the sources consulted for the extraction table. **Outcomes, subscales, and subtests (all assessing aspects of attention):** Most outcomes are scored using T-scores (standardized mean = 50, SD = 10). The index scores and subtests for the CPT-II have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are COPT-Hit Reaction Time, CPT-d Prime T, and CPT Hit Reaction Time (HRT) and are all normed to T-scores (standardized mean = 50, SD = 10).

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/Manual	Examiner Qualifications	
Test of Everyday Attention for Children (TEA-CH); 1999	NP[3]	A	A	A	A	A	NP[3]	This test is designed for children ages 6–16 years. **Internal consistency:** Data were not present in the sources consulted for the extraction table. **Examiner qualifications:** Data were not present in the sources consulted for the extraction table. **Outcomes, subscale, and subtest scores:** The index scores and subtests for this test have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are Map Mission, Sky Search, and Walk/Don't Walk, which are described below. All outcomes have standardized means of 10 and SDs of 3 and assess specific aspects of attention. Map Mission**:** Examinees search a visually cluttered display for target stimuli. This subtest is designed to measure selective and focused attention. Sky Search: For this subtest, the examinee circles targets on a plastic sheet while being distracted. It is designed to measure selective and focused attention. Walk/Don't Walk: For this subtest, children listen to verbal instructions and must decide whether to move forward or not. The test includes both "go" and "no-go" paradigms and is designed to measure sustained attention and response inhibition.
Test of Variables of Attention (TOVA); 1992	A	A	A	A	A	A	NP[3]	This test is designed for persons aged 4–80 years. **Examiner qualifications:** Data were not present in the sources consulted for the extraction table. It is primarily a computer-administered test.

[1]A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2]Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3]See Notes column.
*Subtests or scales within tests that belong to a different domain may be applicable to attention. These include the Developmental Neuropsychological Assessment (NEPSY): Attention domain subscale (*Neuropsychological Assessment Batteries* domain); Fagan Test of Infant Intelligence (FTII): Visual Attention outcome (*Developmental* domain); and Wechsler Memory Scale-III (WMS-III): Spatial Span Forward (*Learning and Memory* domain).

Table B-14. Adequacy or Deficiency of Factors Affecting the Normative Data of Psychometric Tests Used to Assess Attention in Developmental Neurotoxicity Studies[1,2]

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
Conners' Continuous Performance Test-II (CPT-II); 2000	A	A	A	A	A	A	None.
Test of Everyday Attention for Children (TEA-CH); 1999	D[3]	NA	NP[3]	NA	A	NA	**Subject age (child):** Some age bands had only a few participants (N = 13–30), and 1-year age bands for younger children are too wide. **Population representation (child):** Data were not present in the sources consulted for the extraction table.
Test of Variables of Attention (TOVA); 1992	A	D[3]	D[3]	D[3]	A	A	**Subject age (adult)**: The norms are based on pooled samples of 1,596 individuals aged 4–80 years, of whom 1,349 were children and 250 were adults aged 20–80 years. This is a small sample for adults in a wide age range (6 age bins, the oldest of which consisted of elderly adults). **Population representation (child & adult):** All participants in the normative sample were from Minnesota and 99% were white.

[1] A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2] Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3] See Notes column.

B.3.2. Executive Function

Table B-15. Adequacy or Deficiency of Factors Affecting the Reliability, Validity, and Administration Standards of Psychometric Tests Used to Assess Executive Function in Developmental Neurotoxicity Studies[1,2]

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
Stroop Color-Word Test; 1978/2003[3]	NP[3]	A	A[3]	A	A	A[3]	NP[3]	**Year of publication:** The Golden versions of the test for adults and children were published in 1978 and 2003, respectively. This test has several versions. The Victoria version is designed for ages 18–94 years, the Golden version for ages 5–90 years, and the Trenerry version for ages 18–50 years. **Internal consistency:** Data were not present in the sources consulted for the extraction table. **Content validity:** This is an old paradigm with inherent content validity. **Instructions/manual:** Several versions are mentioned in the sources consulted for the extraction table. **Examiner qualifications:** Data were not present in the sources consulted for the extraction table.
Trail-making Test; 1944	NP[3]	A	A	A	A	A	NP[3]	This is an older test developed initially for adults and later applied to children. There are multiple versions, with the Reitan version used more often in the past. There is now a child version with a manual. For these ratings, it is assumed studies are using the Reitan version. **Internal consistency:** The cited sources mention correlations between Parts A and B, but this is not a real reliability measure, as the two parts have different instructions and response requirements. **Examiner qualifications:** Data were not present in the sources consulted for the extraction table. **Outcomes, domain-specific scores and subtests:** There are a variety of scoring methods for this test, including pass/fail, time to

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
								completion for each of the two test forms (Pattern A and Pattern B) and/or errors on each, and differences in completion time for the two forms. The scores for this test have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are listed in the footnotes of evaluation tables for other domains, where applicable. They include "Pattern A" and "Pattern B," but the actual outcomes are unclear. Both are useful for assessing the working memory aspect of executive function, although Pattern A is sometimes considered to be more of an attention test.
Verbal Fluency Test; 2001	A	A	A	A	A	A	A	This is an older test initially developed to identify fluency deficits in aphasia. It has since been adapted as a more general test of executive function or working memory and is known more generally as the Controlled Oral Word Association Test (although Verbal Fluency Test will be used in the older literature). It is sometimes considered a test of verbal function. Norms exist for ages 8–89 years.
Wisconsin Card Sorting Test (WCST); 1993	A	A	A	A	A	A	A	This test has multiple versions (see extraction table).

[1]A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2]Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3]See Notes column.
*Subtests or scales within tests that belong to a different domain may be applicable to executive function. These include the Developmental Neuropsychological Assessment (NEPSY): Executive function domain subscale (*Neuropsychological Assessment Batteries* domain); Kaufman Assessment Battery for Children (K-ABC): Mental Processing, Sequential Processing, Simultaneous Processing (*IQ* domain); McCarthy Scales of Children's Abilities (MSCA) [an executive function scale is presented in the mercury literature; however, this scale is not discussed in the MSCA manual] (*IQ* domain); Stanford-Binet, 5th ed.: Fluid Reasoning Index, Working Memory Index (*IQ* domain); Wechsler Intelligence Scale for Children-R. (WISC-R): Digit span subtest (*IQ* domain); Wechsler Intelligence Scale for Children-III (WISC-III): Digit span subtest (*IQ* domain); Wechsler Intelligence Scale for Children—IV (WISC-IV): Working Memory Index (*IQ* domain); Wechsler Adult Intelligence Scale-III (WAIS-III): Arithmetic, digit span, and letter-number sequencing subtests (*IQ* domain); Wechsler Memory Scale-R (WMS-R): Spatial span (*Learning and Memory* domain); and Wechsler Memory Scale-III (WMS-III): Spatial span backward (*Learning and Memory* domain).

Table B-16. Adequacy or Deficiency of Factors Affecting the Normative Data of Psychometric Tests Used to Assess Executive Function in Developmental Neurotoxicity Studies[1,2]

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
Stroop Color-Word Test; 1978/2003[3]	NP[3]	A[3]	D[3]	D[3]	NP[3]	A[3]	**Year of publication:** The Golden versions of the test for adults and children were published in 1978 and 2003, respectively. **Subject age (child):** The sources consulted for the extraction table indicate no norms for children. **Subject age (adult):** Age bands for normative data depend on the version of the test. Some are adequate. **Population representation (child & adult):** The population (Canadian) is highly specific and may depend on the version used. **Sample size (child):** Data were not present in the sources consulted for the extraction table. **Sample size (adult):** Sample sizes for normative data depend on the version of the test. Some are adequate.
Trail-making Test; 1944	NP[3]	A	NP[3]	A	NP[3]	A	**Norms (child):** Data were not present in the sources consulted for the extraction table.
Verbal Fluency Test; 2001	A	A	A	A	A	A	This test has multiple versions (see extraction table).
Wisconsin Card Sorting Test (WCST); 1993	A	A	A	A	A	A	This test has multiple versions (see extraction table).

[1] A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2] Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3] See Notes column.

B.3.3. Motor Function

Table B-17. Adequacy of Factors Affecting the Reliability, Validity, and Administration Standards of Tests Used to Assess Motor Function in Developmental Neurotoxicity Studies[1,2]

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
Bruininks-Oseretsky Test of Motor Proficiency (BOTMP); 1978	A	A	A	A	NP[3]	A	A	This test is normed for ages 3–18 years. **Criterion validity:** Data were not present in the sources consulted for the extraction table. Highly specific outcomes are available for this test, with motor functions divided into subdomains.
Grooved Pegboard; 1963	NA[3]	A	A[3]	A	A[3]	A	NP[3]	There are several sets of norms for this test. Those in the manual include ages 5–60 years, but other norms extend to at least age 85 (ages 20–85). This test is often scored by time instead of normative outcome. **Internal consistency:** This criterion is not applicable as the test involves moving pegs (all stimuli and responses are the same). **Content validity:** This is not discussed in the extraction table but, by definition, pegboard tests assess motor coordination (and, in this case, manual motor speed). **Criterion validity:** This is not presented in the extraction table, but the test outcomes predict handedness, unilateral lesions affecting motor areas, and other endpoints. **Examiner qualifications:** Data were not present in the sources consulted for the extraction table. **Outcomes, domain-specific scores, and subtests:** The scores for this test have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and include mean time for each hand,

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
								a raw score based on time to completion of the task.
Movement Assessment Battery for Children (MABC); 1992	NP[3]	A	A	A	NP[3]	A	A	This test is designed for children ages 4–12 years. **Internal consistency:** Data were not present in the sources consulted for the extraction table. **Criterion validity:** Data were not present in the sources consulted for the extraction table.

[1]A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2]Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3]See Notes column.
*Subtests or scales within tests that belong to a different domain may be applicable to motor function. These include the Bayley Scales of Infant Development-II (BSID-II): Motor Index (*Developmental* domain); Denver Developmental Screening Test (DDST): Fine motor skills area and gross motor skills area outcomes (*Developmental* domain); Griffith Mental Development Scales (GMDS): Hand-eye Coordination Index and Locomotion Index (*Developmental* domain); McCarthy Scales of Children's Abilities (MSCA): Motor Index (*IQ* domain); Wechsler Intelligence Scale for Children-III (WISC-III): Coding subtest (*IQ* domain); and Wechsler Adult Intelligence Scale-III (WAIS-III): Coding/Digit symbol subtest (*IQ* domain).

Table B-18. Adequacy or Deficiency of Factors Affecting the Normative Data of Psychometric Tests Used to Assess Motor Function in Developmental Neurotoxicity Studies[1,2]

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
Bruininks-Oseretsky Test of Motor Proficiency (BOTMP); 1978	A	NA	D[3]	NA	A	NA	**Population representation (child):** The sample population for norms is from a small town with primarily white children.
Grooved Pegboard; 1963	NP[3]	A	NP[3]	NP[3]	NP[3]	A	There are several sets of norms for this test. **Subject age (child):** According to the extraction table, norms for younger children should be used cautiously, although they exist beginning at age 5. Age bands were not present in the sources consulted for the extraction table. **Population representation (child & adult):** Data were not present in the sources consulted for the extraction table. **Sample size (child):** Data were not present in the sources consulted for the extraction table.
Movement Assessment Battery for Children (MABC); 1992	A[3]	NA	NP[3]	NA	NP[3]	NA	**Subject age (child):** The age bands are rather wide, especially the age band from 4 to 6 years of age. **Population representation (child):** Data were not present in the sources consulted for the extraction table. **Sample size (child):** Data were not present in the sources consulted for the extraction table.

[1] A: adequate, D: deficient, NP: not present in test manuals or other materials reviewed, NA: not applicable.
[2] Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3] See Notes column.

B.3.4. Learning and Memory

Table B-19. Adequacy or Deficiency of Factors Affecting the Reliability, Validity, and Administration Standards of Psychometric Tests Used to Assess Learning and Memory in Developmental Neurotoxicity Studies[1,2]

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
California Verbal Learning Test (CVLT); 1983, 1987	A	A	A	A	A	A	A	The CVLT is a list learning test that includes multiple measures of learning, short- and long-term recall, recognition, free recall, and interference effects. Normative outcomes are available for persons aged 17–80 years.
California Verbal Learning Test – Children (CVLT-C); 1994	A	A	A	A	A	A	A	The children's version of the CVLT is structured similarly to the adult version, assessing multiple aspects of verbal learning and memory. It is designed for children ages 5 years through 16 years, 11 months.
Wechsler Memory Scale-Revised (WMS-R); 1987	A	A	A	A	A[3]	A	A	This test was developed for individuals ages 16–74 years. **Criterion validity:** The rating of adequate is based on personal knowledge of how scores on the test are associated with clinical diagnoses and specific sites of brain damage.[4] **Outcomes, domain-specific scores, and subtests:** The scores for this test have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are listed in the footnotes of evaluation tables for other domains, where applicable. This includes Spatial Span (Working Memory), a test that has forward and backward span conditions. The outcomes for subtests such as Spatial Span are age-standardized scores (mean = 10, standard deviation [SD] = 3). Scales are calculated by summing the appropriate subtest scores and determining the appropriate age-standardized scores (mean = 100, SD = 15).

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
Wechsler Memory Scale-III (WMS-III); 1997	A	A	A	A	A[3]	A	A	This test was normed for individuals 16–89 years of age. **Outcomes, domain-specific scores, and subtests:** The scores for this test have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and are listed in the footnotes of evaluation tables for other domains, where applicable. This includes Spatial Span Forward (Attention) and Spatial Span Backward (Executive Function/Working Memory). The outcomes for subtests such as Spatial Span are age-standardized scores (mean = 10, SD = 3). Scales are calculated by summing the appropriate subtest scores and age-standardized scores (mean = 100, SD = 15).

[1]A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2]Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3]See Notes column.
[4]Adequacy and deficiency are determined by the professional experience and knowledge of the co-author Dr. Roberta F. White.
*Subtests or scales within tests that belong to a different domain may be applicable to learning and memory. These include the Fagan Test of Infant Development: Visual recognition memory score (*Developmental* domain); McCarthy Scales of Children's Abilities (MSCA): Memory Index (*IQ* domain); Stanford-Binet-4th edition.: Bead Memory test (*IQ* domain) [Copying Test Recall has also appeared as an outcome in the mercury literature, but this would have been a raw score adjusted for relevant variables, as the standard Copying Test does not have a recall condition and does not produce a scaled score]; and Stanford-Binet, 5th edition.: Memory Index (*IQ* domain).

Table B-20. Adequacy or Deficiency of Factors Affecting the Normative Data of Psychometric Tests Used to Assess Learning and Memory in Developmental Neurotoxicity Studies[1,2]

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
California Verbal Learning Test (CVLT); 1983, 1987	NA	A	NA	D[3]	NA	A	**Population representation (adult):** Normative sample subjects were drawn from only three cities in the United States.
California Verbal Learning Test – Children (CVLT-C); 1994	NP[3]	NA	A	NA	A	NA	**Subject age (child):** The sources consulted did not provide information on age bands used for the normative sample.
Wechsler Memory Scale-Revised (WMS-R); 1987	A	A	A	A	A	A	None.
Wechsler Memory Scale-III (WMS-III); 1997	A	A	A	A	A	A	None.

[1] A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2] Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3] See Notes column.

B.3.5. Social-Emotional

Table B-21. Adequacy or Deficiency of Factors Affecting the Reliability, Validity, and Administration Standards of Psychometric Tests Used to Assess the Social-Emotional Domain in Developmental Neurotoxicity Studies[1,2]

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
Beck Depression Inventory (BDI); 1961	A	A	A	A	A[3]	A[3]	A[3]	The extraction table indicates that the test is given to adolescents and adults but not to children. **Criterion validity:** It is well established that the test correlates well with other depression measures; however, these data are not present in the extraction table. **Instructions/manual:** Data in the extraction table come from the Spreen and Strauss text, and ratings reflect statements from that text. It was assumed that there is a test manual that could be consulted. **Examiner qualifications:** This is a self-administered questionnaire.
Beck Depression Inventory-Second Edition (BDI-II); 1996	A	A	A[3]	A	A	A[3]	A[3]	The age range for this test is 13–86 years. **Content validity:** BDI-II includes items drawn from the Diagnostic and Statistical Manual of Mental Disorders (DSM) and correlates with the Structured Clinical Interview for DSM (SCID). **Instructions/manual:** Data in the extraction table come from the Spreen and Strauss text, and ratings reflect statements from that text. It was assumed that there is a test manual that could be consulted. **Examiner qualifications:** This is a self-administered questionnaire.
Behavior Assessment System for Children, 2nd ed. (BASC-2); 2004	A	A	A	A	A	A	A	The BASC-2 is appropriate for ages 2–25 years; however, the upper age limit not clear.

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
Child Behavior Checklist (CBCL); 1998	NP[3]	A	A	A	A[3]	A	A[3]	The age range for this test is not provided in the extraction table. The Child Behavior Checklist (CBCL) is appropriate for ages 2–18 years. **Internal consistency:** Data were not present in the sources consulted for the extraction table. **Criterion validity:** Criterion validity exists in other literature that was not consulted in the extraction table. **Examiner qualifications:** Most forms of this test are questionnaires answered by a parent, teacher, or child. There are interview and observation versions, but no data are present in the extraction table on examiner qualifications for these versions.
Children's Communication Checklist (CCC); 1998	A	A	A	A	A	A	A[3]	The age range for this test appears to be 5–17 years. **Examiner qualifications:** This is a questionnaire filled out by a parent, teacher, or clinician about the child's behavior.
Difficulties in Emotion Regulation Scale (DERS); 2004	A	A	A	A	A	A[3]	A[3]	The age range for this test appears to vary by study (see extraction table) and has included 18–55, 13–17, and 11–15 years of age. **Instructions/manual:** Data in the extraction table come from reliability and validity studies, and ratings reflect statements from that text. It is assumed that there is a test manual that could be consulted. **Examiner qualifications:** This is a self-reported measure.
Disruptive Behavior Disorders Scale (DBD); 1997	A	A	A	A	NP[3]	A	A[3]	The age range for this test is unclear but may include 6–12-year-old children, preschool and school-age children, or preschool through high school children (see extraction table).

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
								Criterion validity: Data were not present in the sources consulted for the extraction table. **Examiner qualifications:** This is a self-reported measure filled out by parents or teachers about the child's behavior.
Early Childhood Behavior Questionnaire; 2002	A	A	A	A	A	A[3]	A[3]	The age range for this test is 18–36 months. **Instructions/manual:** Data in the extraction table come from reliability and validity studies, and ratings reflect statements from that text. It was assumed that there is a test manual that could be consulted, as there is a standard set of questions. **Examiner qualifications:** This is a parent-completed questionnaire.
EAS Temperament Survey for Children; 1984	A	A	A	A	A	A	A[3]	This test appears to be used for children aged 5 months to 9 years (see extraction table). **Examiner qualifications:** This is a parent-completed questionnaire.
Profile of Mood States (POMS); 1971[3]	A	A	A	A	A	A	A[3]	Norms for this test exist for ages 18 and older. **Year of publication:** Revisions are reported in 1992 and 2003, but a fully revised version is "under way." **Examiner qualifications**: This is a standard self-reported questionnaire.
Social and Communication Disorders Checklist (SCDC); 1997	A	A	A	A	A	A	A[3]	The extraction table describes uses of the test at age 5 and at 2.5–17 years. **Examiner qualifications:** This is a parent-completed or teacher-completed questionnaire.
Social Communication Questionnaire; 2003	A	A	A	A	A	A	A[3]	Normative data for this test are based on persons 4–40 years of age. **Examiner qualifications:** This is a primary caregiver-completed questionnaire.

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/ Manual	Examiner Qualifications	
State-Trait Anxiety Inventory (STAI); 1966	D[3]	A[3]	A	A	A[3]	A	A	The STAI was designed for high school students, college students, and adults. There is also a child version of the test (STAI-C [not extracted]). **Internal consistency:** Measures of internal consistency were reported to be below the threshold of acceptable (i.e., <0.6). **Test-retest reliability:** Test-retest reliability appears to be Adequate for the trait anxiety measure; however, test-retest reliability is less robust for the state anxiety variable (this would be expected as state anxiety varies across circumstances and days). **Criterion validity:** The Adequate rating is based on information provided for sensitivity.
Strengths and Difficulties Questionnaire (SDQ); 1994	A	A	A[3]	A	A	A	A[3]	The SDQ can be administered using teacher- and parent-reported answers for ages 3–16 years. For ages 11–16 years, the SDQ may be administered using self-, teacher-, and parent-reported answers. **Content validity:** The Adequate rating is based on factor analyses presented in the extraction table. **Examiner qualifications:** This is a self-administered questionnaire.
Vineland Adaptive Behavior Scales; 1998	A	A	A	A	A	A	A	This test was designed for children aged 3 years to 18 years, 11 months.

[1]A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2]Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3]See Notes column.
*Subtests or scales within tests that belong to a different domain may be applicable to social-emotional. These include the Conners' Parent Rating Scale-Revised: Behavioral Index, Externalizing Problems, Hyperactivity Index, Oppositional Index (*Clinical Conditions* domain); Denver Developmental Screening Test (DDST): Social/emotional area (called Social skills in the literature) (*Developmental* domain); and Griffith Mental Development Scale (GMDS): Personal/Social Skills Index (*Developmental* domain).

Table B-22. Adequacy or Deficiency of Factors Affecting the Normative Data Standards of Psychometric Tests Used to Assess the Social-Emotional Domain in Developmental Neurotoxicity Studies[1,2]

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
Beck Depression Inventory (BDI); 1961	NA	NP[3]	NA	NP[3]	NA	NP[3]	**Norms (adults):** There is no information on normative data in the sources consulted and summarized in the extraction table.
Beck Depression Inventory-Second Edition (BDI-II); 1996	NP[3]	NP[3]	D[3]	D[3]	D[3]	D[3]	Additional normative data may be available from other sources not consulted for the extraction table. **Subject age (child & adult):** Data on age bands are not present in the extraction table. **Population representation & sample size (child & adult):** The normative data are based on a convenience sample of 127 patients from one university who are not representative of the U.S. population.
Behavior Assessment System for Children, 2nd ed. (BASC-2); 2004	A	D[3]	A	D[3]	A	NP[3]	**Subject age (adult):** The normative sample does not report information for ages 21–25 years old—the upper age range for the test as described by the test publishers. **Population representation (adult):** The older age range of adults considered in the normative sample was not described. **Sample size (adult):** Data were not present in the sources consulted for the extraction table for ages 18–21 years.
Child Behavior Checklist (CBCL); 1998	A[3]	NA	A	NA	A	NA	**Subject age (child):** Age bands are wide for norms.
Children's Communication Checklist (CCC); 1998	NP[3]	NA	NP[3]	NA	NP[3]	NA	Data on these criteria are not provided in the sources consulted for the extraction table. The test has mainly been used in the United Kingdom and Holland (see extraction table).

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
Difficulties in Emotion Regulation Scale (DERS); 2004	NA	A	NA	D[3]	NA	A[3]	**Population representation & sample size (adult):** Reliability and validity studies were carried out on convenience samples that were not representative of the population but included adolescents. The original normative sample was 357 college students ages 18–55 years.
Disruptive Behavior Disorders Scale (DBD); 1997	NP[3]	NA	NP[3]	NA	NP[3]	NA	**Norms (child):** Data were not present in the sources consulted for the extraction table.
Early Childhood Behavior Questionnaire; 2002	A[3]	NA	D[3]	NA	D[3]	NA	**Subject age (child):** Age bands were of appropriate width but did not have enough children per age band. **Population representation (child):** The population was not representative of the United States (90% white fathers, 95% white mothers). **Sample size (child):** Ten children were tested at multiple timepoints.
EAS Temperament Survey for Children; 1984	A	NA	D[3]	NA	D[3]	NA	**Population representation (child):** Normative data were derived from small local samples. **Sample size (child):** Sample sizes for norms were 182 children (91 mothers) in the United States and 222 children in Holland.
Profile of Mood States (POMS); 1971	NA	A	NA	A[3]	NA	A	**Population representation (adult):** Several normative studies have been performed on adults. Population representation is considered adequate but should consider which normative sample is being used.
Social and Communication Disorders Checklist (SCDC); 1997	NP[3]	NA	NP[3]	NA	NP[3]	NA	**Norms (child):** Data were not present in the sources consulted for the extraction table.

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
Social Communication Questionnaire; 2003	NP[3]	NA	NP[3]	NA	NP[3]	NA	**Norms (child):** Data were not available in the sources consulted for the extraction table.
State-Trait Anxiety Inventory (STAI); 1966	NA	D[3]	NA	D[3]	NA	A	**Subject age (adult):** Age band from 50 to 69 years of age is too wide, as age-related changes in anxiety may occur. **Population representation (adult):** The normative sample was not representative of the United States, where the normative study was conducted.
Strengths and Difficulties Questionnaire (SDQ); 1994	D[3]	NA	NP[3]	NA	A	NA	**Subject age (child):** Ages 4–7 years is a wide age band for this outcome. **Population representation (child):** Data were not present in the sources consulted for the extraction table; normative data may be based on data sent to test authors.
Vineland Adaptive Behavior Scales; 1998	A	NA	A	NA	A	NA	None.

[1]A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2]Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3]See Notes column.

B.3.6. Verbal/Language

Table B-23. Adequacy of Factors Affecting the Reliability, Validity, and Administration Standards of Tests Used to Assess Verbal/Language Abilities in Developmental Neurotoxicity Studies[1,2]

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/Manual	Examiner Qualifications	
Boston Naming Test (BNT); 1978[3]	A	NP[3]	A	A	A[3]	A[3]	NP[3]	**Year of publication:** This test was also published in 1978. Age range for the test is not stated in the extraction table but is believed to be 6–85 years. **Test-retest reliability:** Data were not present in the sources consulted for the extraction table. **Criterion validity:** Poor test performance is associated with a number of neurological disorders (e.g., aphasic syndromes, Alzheimer's disease, frontal dementia) and with diagnosed verbal learning problems in children.[4] **Examiner qualifications:** Data were not present in the sources consulted for the extraction table, although it is usually administered by speech pathologists and neuropsychologists.[4]
Boston Naming Test-2 (BNT-2); 2001	A	A[3]	A	A	A	A	NP[3]	The age range for the test is 5–13 years and 18 years and older. **Test-retest reliability:** This is rated as adequate given data from adults. **Examiner qualifications:** Data were not present in the sources consulted for the extraction table, but this test is generally administered by psychologists and speech pathologists.[4]
MacArthur-Bates Communicative Development Inventories (CDI); 1993	A	A	A	A	A	A[3]	NP[3]	This test is normed for children 8–30 months of age. **Instructions/manual:** Data in the extraction table come from validation papers, and ratings reflect statements from these papers. There is a test manual. **Examiner qualifications:** The test has direct testing versions and a parent-report version. Data were not

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/Manual	Examiner Qualifications	
								present in the sources consulted for the extraction table.
Peabody Picture Vocabulary Test-Revised (PPVT-R); 1981	A	A	A	A	A	A	A	The age range for this test is 2.5–40 years of age.
Peabody Picture Vocabulary Test-III (PPVT-III); 1997	A	A	A	A	A	A	A	The age range for the PPVT-III is 2.5–90+ years of age.
Preschool Language Scales-3 (PLS-3); 1969[3]	NP[3]	A	A	A	A	A	NP[3]	This test is designed for children ages 2 weeks to 6 years. **Year of publication:** This test was also published in 1979 and 1992. **Internal consistency:** Data were not present in the sources consulted for the extraction table. **Examiner qualifications:** Data were not present in the sources consulted for the extraction table. The scores for this test that have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg are all verbal and include: Auditory Comprehension, Expressive Communication (formerly named Verbal Ability), and Total Language. All are scored with a standardized mean of 100 and standard deviation (SD) of 15.

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/Manual	Examiner Qualifications	
Speech and Language Assessment Scale (SLAS); 1989	NP[3]	NP[3]	A	A	A	A	A[3]	This is a parent-report test of speech and language that assesses children 3–5 years of age. **Internal consistency:** Data were not present in the sources consulted for the extraction table. **Test-retest reliability:** Data were not present in the sources consulted for the extraction table. **Examiner qualifications:** This is a parent-completed questionnaire on speech.
Test of Language Development (TOLD); 1977[3]	A	NP[3]	A	A	A[3]	A	A	This test applies to ages 4–9 years. **Year of publication**: This test was also published in 1978. **Test-retest reliability:** Data were not present in the sources consulted for the extraction table. The scores for this test that have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg are all measures of verbal function and include: Grammar Completion, Grammar Understanding, Oral Vocabulary, Picture Vocabulary, and Sentence Imitation. All subtests are scored with a standardized mean of 10 and SD of 3.

[1]A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2]Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3]See Notes column.
[4]Adequacy and deficiency are determined by the professional experience and knowledge of the co-author Dr. Roberta F. White.
*Subtests or scales within tests that belong to a different domain may be applicable to verbal/language. These include the Denver Developmental Screening Test (DDST): Communication area (*Developmental* domain); Griffith Mental Development Scale (GMDS): Hearing and Speech Index (*Developmental* domain); Kaufman Brief Intelligence Test (K-BIT): The Verbal Intelligence score is equivalent to Vocabulary Performance (*IQ* domain); McCarthy Scales of Children's Abilities (MSCA): Verbal Index (*IQ* domain); Stanford-Binet, 5th edition: Knowledge Index (*IQ* domain); Wechsler Intelligence Scale for Children-R (WISC-R): Verbal IQ, Similarities subtest (*IQ* domain); Wechsler Intelligence Scale for Children-III (WISC-III): Verbal IQ, Information subtest, Vocabulary subtest (*IQ* domain); Wechsler intelligence Scale for Children-IV (WISC-IV): Verbal Comprehension Index (*IQ* domain); Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R): Verbal IQ (*IQ* domain); Wechsler Adult Intelligence Scale—III (WAIS-III): Vocabulary subtest, Similarities subtest, Comprehension subtest (*IQ* domain); Griffith Mental Development Scales (GMDS): Hearing and Speech (*Developmental* domain); and Denver Developmental Screening Test (DDST): Communication (language) area (*Developmental* domain).

Table B-24. Adequacy of Factors Affecting the Normative Data of Psychometric Tests Used to Assess Verbal/Language Abilities in Developmental Neurotoxicity Studies[1,2]

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
Boston Naming Test (BNT); 1978	D[3]	A[3]	NP[3]	NP[3]	D[3]	D[3]	**Subject age (child):** Child norms are based on a very small sample size. **Subject age (adult):** Age bins for adults are in 5-year increments. **Population representation (child & adult):** Data were not present in the sources consulted for the extraction table. **Sample size (child & adult):** Sample sizes for both children and adults are described as small in the extraction table.
Boston Naming Test-2 (BNT-2); 2001	D[3]	A	NP[3]	A[3]	NP[3]	A[3]	**Subject age (child):** In the sources consulted for the extraction table, norms were presented for ages 6–12.5 years only and appeared to be outdated. **Population representation (child):** Data were not present in the sources consulted for the extraction table. **Population representation (adult):** The characterization of population representation in adults is not detailed, but it appears to be adequate. **Sample size (child):** Data were not present in the sources consulted for the extraction table. **Sample size (adult):** Several adult normative samples exist, with one consisting of 663 participants and another of 1,000 participants (see extraction table).

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
MacArthur-Bates Communicative Development Inventories (CDI); 1993	A	NA	NP[3]	NA	NP[3]	NA	**Population representation & sample size (child):** Data were not present in the sources consulted for the extraction table. However, it was noted that the test has been used in several different countries.
Peabody Picture Vocabulary Test-Revised (PPVT-R); 1981	A	D[3]	A	D[3]	A	A	**Subject age (adult):** Normative sample limited to ages 18 or 19 to 40 years. **Population representation (adult):** Adult participants do not appear to be representative.
Peabody Picture Vocabulary Test-III (PPVT-III); 1997	A	A	A	A	A	A	None.
Preschool Language Scales-3 (PLS-3); 1969	A	NA	A	NA	A	NA	None.
Speech and Language Assessment Scale (SLAS); 1989	NP[3]	NA	NP[3]	NA	NP[3]	NA	**Norms (child):** Data were not present in the sources consulted for the extraction table.
Test of Language Development (TOLD); 1977	NP[3]	NA	NP[3]	NA	A	NA	**Subject age (child):** The sources consulted for the extraction table did not provide information on age bands used in the normative data. **Population representation (child):** Data were not present in the sources consulted for the extraction table.

[1]A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2]Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3]See Notes column.

B.3.7. Visuospatial Function

Table B-25. Adequacy or Deficiency of Factors Affecting the Reliability, Validity, and Administration Standards of Psychometric Tests Used to Assess Visuospatial Function in Developmental Neurotoxicity Studies[1,2]

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/Manual	Examiner Qualifications	
Bender Visual-Motor Gestalt Test; 1938	A	A	A	A	A[3]	A	NP[3]	This test has been administered to children 2.5 years through adulthood. **Criterion validity:** The extraction table does not reflect existing data on the test's association with known brain damage and certain developmental disorders. **Examiner qualifications:** Data were not present in the sources consulted for the extraction table. **Outcomes, subscale, and subtest scores:** The scores for this test that have been identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg are copy and recall. It is assumed that raw scores were used as true normative scores do not exist for this version of the Bender.
Bender Visual-Motor Gestalt-Test II; 2003	A	A	A	A[3]	A	A	NP[3]	This test has versions for 4 years, 11 months through 85+ years. **Construct validity:** This test has inherent construct validity as a test of visuo-constructive ability and its relationship to the initial Bender Gestalt test. Correlations with IQ and other tests listed in the extraction table are not high; this is not an indictment of the test, as it measures a specific ability and not general intelligence. **Examiner qualifications:** Data were not present in the sources consulted for the extraction table.

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/Manual	Examiner Qualifications	
Developmental Test of Visual-Motor Integration (VMI); 1967[3]	A	A	A[3]	A	A[3]	A	A	This test is used with children age 2.5 and adults. **Year of publication:** This test was restandardized in 1982 and 1989. **Content validity:** This test has inherent construct validity given its stimuli and task demands. **Criterion validity:** The reported predictive validity was used to determine an Adequate rating for criterion validity. This test is a predictor of verbal/performance discrepancies on IQ tests in children with nonverbal learning disabilities.
Finger Identification Test; 1959[3]	A	A	A[3]	A[3]	NP[3]	A[3]	NP[3]	This test can be administered to ages 6 years through adulthood. **Year of publication:** The test was also published in Contributions to Neuropsychological Assessment, 1983. The test was developed to assess a highly specific aspect of brain dysfunction associated with the left posterior portion of the brain. Finger identification is called finger gnosis (or finger agnosia when referring to deficits in finger naming). **Content validity:** Finger agnosia is a specific and localized neurological function. Therefore, by definition, the test has content validity. **Construct validity:** Data reported in the extraction table indicate that performance is associated with reading and predictive of subsequent reading achievement, and reading is localized in the same brain area as finger naming. The poor correlations with finger dexterity do not argue against the test's validity, as this function is controlled by different brain areas.

Test*; Year of Publication	Reliability		Validity			Administration		Notes
	Internal Consistency	Test-Retest	Content	Construct	Criterion	Instructions/Manual	Examiner Qualifications	**Criterion validity:** Data were not present in the sources consulted for the extraction table. **Instructions/manual:** Data in the extraction table come from text Strauss and Lezak, and ratings reflect statements from that text. Lezak indicates that there is a test manual (Lezak et al. 2004; Strauss et al. 2006). **Examiner qualifications:** Data were not present in the sources consulted for the extraction table.

[1] A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2] Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3] See Notes column.
* Subtests or scales within tests that belong to a different domain may be applicable to visuospatial function. These include K-BIT: Nonverbal intelligence outcome represents score on the Matrices test, a test of visuospatial reasoning (*IQ* domain); Bayley Scale of Infant Development-II (BSID-II): Psychomotor Index (*Developmental* domain); Griffith Mental Development Scale (GMDS): Hand-eye Performance Index (*Developmental* domain); McCarthy Scales of Children's Abilities (MSCA): Perceptual Index (*IQ* domain); Stanford-Binet Intelligence Test, 4th edition: Copying test (*IQ* domain); Stanford-Binet Intelligence Test, 5th edition: Visual-spatial Processing Index (*IQ* domain); Wechsler Intelligence Scale for Children-R: Performance IQ, Block Design subtest (*IQ* domain) ["Visuospatial performance" was identified in the epidemiological literature for the in-progress EPA IRIS toxicological review of MeHg and is likely Performance IQ and not a separate measure, at least it is not a WISC-R measure with that name]; Wechsler Intelligence Scale for Children-III (WISC-III): Performance IQ, Block Design test (*IQ* domain); Wechsler Intelligence Scale for Children-IV (WISC-IV): Perceptual Reasoning Index (*IQ* domain); Wechsler Preschool and Primary Scale of Intelligence-R (WPPSI-R): Performance IQ (*IQ* domain); Wechsler Adult Intelligence Scale-Revised (WAIS-R): Block Design subtest (*IQ* domain); Wechsler Adult Intelligence Scale-III (WAIS-III): Block Design subtest, Picture Completion subtest, Object Assembly subtest, Picture Arrangement subtest, Matrix Reasoning subtest (sometimes classified as executive function) (*IQ* domain); and Fagan Test of Infant Intelligence (FTII): Visual Recognition (*Developmental* domain).

Table B-26. Adequacy or Deficiency of Factors Affecting the Normative Data of Psychometric Tests Used to Assess Visuospatial Function in Developmental Neurotoxicity Studies[1,2]

Test; Year of Publication	Normative Data						Notes
	Subject Age		Population Representation		Sample Size		
	Children	Adults	Children	Adults	Children	Adults	
Bender Visual-Motor Gestalt Test; 1938	NP[3]	D[3]	D[3]	D[3]	D[3]	D[3]	This is an older test, and raw scores were used as outcomes when this test was used historically (it was later revised into Bender Gestalt II). There are no norms in the usual sense of the term. However, the manual presents typical drawings produced by children for each year of age from 3 to 11 years and 11 years through adulthood (see extraction table). **Subject age (child):** Age-specific norms were not present in the sources consulted for the extraction table. **Subject age (adult):** Outcomes for age 11 years through adulthood are not parsed into bins. **Population representation & sample size (child & adult):** True normative data were not collected with now current sampling strategies.
Bender Visual-Motor Gestalt-Test II; 2003	A	A	A	A	A	A	None.
Developmental Test of Visual-Motor Integration (VMI); 1967[3]	A[3]	D[3]	A	D[3]	A	D[3]	**Year of publication:** Test was restandardized in 1982 and 1989. **Subject age (child):** For children, the newer norms are Adequate for age and population representation through age 14 years. **Subject age, population representation, and sample size (adult):** The test author suggests that norms for age 14 can be used to assess test results in older adolescents and adults. Adults are not well represented in norms.
Finger Identification Test; 1959[3]	A	A	NP[3]	NP[3]	D[3]	D[3]	**Year of publication:** The test was also published in Contributions to Neuropsychological Assessment, 1983. **Population representation (child & adult):** Data were not present in the sources consulted for the extraction table. **Sample size (child & adult):** The sample size utilized for this test appears to be small.

[1] A : adequate, D : deficient, NP : not present in test manuals or other materials reviewed, NA : not applicable.
[2] Adequacy and deficiency are defined based on the criteria described in Part 1 of this document.
[3] See Notes column.

B.3.8. Processing Speed

No tests assessing processing speed were identified in the epidemiological studies from the in-progress EPA IRIS toxicological review of MeHg.

Subtests or scales within tests that belong to a different domain may be applicable to processing speed. These include Wechsler Intelligence Scale for Children III (WISC III): Processing Speed Index (measures speed in processing visual material) (*IQ* domain); Wechsler intelligence Scale for Children-IV (WISC-IV): Processing Speed Index (*IQ* domain); Wechsler Adult Intelligence Scale-III (WAIS-III): Processing Speed Index. (*IQ* domain); and Conners' Continuous Performance Test-II (CPT-II): Hit Reaction Time (*Attention* domain).

Appendix C. References for DNT Test Information Extraction Database

For the DNT Test Information Extraction Database, see Appendix D.

C.1. Test Manuals and Textbooks

Bayley N. 1993. Bayley Scales of Infant Development, 2nd ed: Manual. San Antonio, TX: The Psychological Corporation, Harcourt Brace & Company.

Bender L. 1938. A visual motor gestalt test and its clinical use. Research Monographs No 3. New York, NY: The American Orthopsychiatric Association.

Brannigan GG, Decker SL. 2003. Bender Visual-Motor Gestalt Test, 2nd ed: Examiner's manual. Itasca, IL: Riverside Publishing.

Conners CK. 2001. Conners' Rating Scales - revised: Technical manual. North Tonawanda, NY: Multi-Health Systems, Inc.

Conners CK. 2004. Conners' Continuous Performance Test (CPT II): Version 5 for Windows: Technical guidance and software manual. North Tonawanda, NY: Multi-Health Systems, Inc.

Delis DC, Kramer JH, Kaplan E, Ober BA. 1987. California Verbal Learning Test, research edition: Manual, Version 1. San Antonio, TX: The Psychological Corporation, Harcourt Brace & Company.

Dunn LM, Dunn LM. 1981. Peabody Picture Vocabulary Test - revised: Manual for forms L and M. Circle Pines, MN: American Guidance Service.

Heaton RK, Chelune GJ, Talley JL, Kay GG, Curtiss G. 1993. Wisconsin Card Sorting Test manual: Revised and explained. Psychological Assessment Resources, Inc.

Korkman M, Kirk U, Kemp S. 1998. A developmental neuropsychological assessment: Manual. The Psychological Corporation.

Lezak MD. 1976. Neuropsychological assessment. New York, NY: Oxford University Press, Inc.

Lezak MD. 1995. Neuropsychological assessment, 3rd ed. New York, NY: Oxford University Press.

Lezak MD, Howieson DB, Bigler ED, Tranel D. 2012. Neuropsychological assessment, 5th ed. New York, NY: Oxford University Press.

Lezak MD, Howieson DB, Loring DW, Hannay HJ, Fischer JS. 2004. Neuropsychological assessment, 4th ed. New York, NY: Oxford University Press.

Mather N, Woodcock RW. 2001. Woodcock-Johnson III: Examiner's manual. Itasca, IL: Riverside Publishing.

McCarthy D. 1972. Manual for the McCarthy Scales of Children's Abilities. The Psychological Corporation, Harcourt Brace Jovanovich, Inc.

McGrew KS, Woodcock RW. 2001. Woodcock-Johnson III: Technical manual. Itasca, IL: Riverside Publishing.

McNair DM, Lorr M, Droppleman LF, Heuchert JP. 2005. Profile of mood states technical update: Manual. North Tonawanda, NY: Multi-Health Systems, Inc.

Raven JC. 1956. Guide to using the Coloured Progressive Matrices: Sets A, Ab, B (revised order, 1956). London, England: H.K. Lewis & Co. Ltd.

Raven JC. 1977. Manual for Raven's Progressive Matrices and Vocabular Scales - section 3: Standard progressive matrices. London, England: H.K. Lewis & Co. Ltd.

Roid GH. 2003. Stanford-Binet Intelligence Scales, 5th ed: Examiner's manual. Itasca, IL: Riverside Publisher.

Roid GH. 2003. Stanford-Binet Intelligence Scales, 5th ed: Technical manual. Itasca, IL: Riverside Publisher.

Sattler JM. 2001. Assessment of children: Cognitive applications, 4th ed. San Diego, CA: Jerome M. Sattler, Inc.

Spreen O, Strauss E. 1991. A compendium of neuropsychological tests: Administration, norms, and commentary, 1st ed. New York, NY: Oxford University Press.

Spreen O, Strauss E. 1998. A compendium of neuropsychological tests: Administration, norms, and commentary, 2nd ed. New York, NY: Oxford University Press.

Strauss E, Sherman EMS, Spreen O. 2006. A compendium of neuropsychological tests: Administration, norms, and commentary, 3rd ed. New York, NY: Oxford University Press.

Thorndike RL, Hagen EP, Sattler JM. 1986. Stanford-Binet Intelligence Scales, 4th ed: Guide for administration and scoring. Chicago, IL: The Riverside Publishing Company.

Thorndike RL, Hagen EP, Sattler JM. 1986. Stanford-Binet Intelligence Scales, 4th ed: Technical manual. Chicago, IL: The Riverside Publishing Company.

Wechsler D. 1981. WAIS-R manual: Wechsler Adult Intelligence Scale - revised. San Antonio, TX: The Psychological Corporation.

Wechsler D. 1987. Manual for the Wechsler Intelligence Scale for Children - revised. New York, NY: The Psychological Corporation.

Wechsler D. 1987. Wechsler Memory Scale - 3rd ed: Administration and scoring manual. San Antonio, TX: The Psychological Corporation, Harcourt Brace & Company.

Wechsler D. 1987. Wechsler Memory Scale - revised: Manual. San Antonio, TX: The Psychological Corporation, Harcourt Brace Jovanovich, Inc.

Wechsler D. 1989. Wechsler Preschool and Primary Scale of Intelligence - revised: Manual. New York, NY: The Psychological Corporation, Harcourt Brace Jovanovich, Inc.

Wechsler D. 1991. Wechsler Intelligence Scale for Children - 3rd ed: Manual. San Antonio, TX: The Psychological Corporation, Harcourt Brace Jovanovich, Inc.

Wechsler D. 1997. Wechsler Adult Intelligence Scale - 3rd ed (WAIS-III), Wechsler Memory Scale - 3rd ed (WMS-III): Technical manual. San Antonio, TX: The Psychological Corporation, Harcourt Brace & Company.

Wechsler D. 1997. Wechsler Adult Intelligence Scale - 3rd ed (WAIS-III): Administration and scoring manual. San Antonio, TX: The Psychological Corporation, Harcourt Brace & Company.

Wechsler D. 2003. Wechsler Intelligence Scale for Children - 4th ed: Administration and scoring manual. San Antonio, TX: PsychCorp by Harcourt Assessment, Inc.

Williams KT, Wang JJ. 1997. Technical references to the Peabody Picture Vocabulary Test - 3rd ed (PPVT-III). Circle Pines, MN: American Guidance Service.

Woodcock RW, Johnson MB. 1989. Woodcock-Johnson Psycho-Educational Battery - revised: Tests of cognitive ability: Standard and supplemental batteries. Allen, TX: DLM Teaching Resources.

Woodcock RW, Johnson MB. 1990. Woodcock-Johnson Psycho-Educational Battery - revised: Tests of achievement: Standard and supplemental batteries. The Riverside Publishing Company.

Zimmerman IL, Steiner VG, Pond RE. 1992. Preschool Language Scale-3: Examiner's manual. San Antonio, TX: The Psychological Corporation, Harcourt Brace Jovanovich, Inc.

C.2. Peer-reviewed Literature

Ahsan S, Murphy G, Kealy S, Sharif F. 2008. Current developmental surveillance: Is it time for change? Ir Med J. 101(4):110-112.

Akaragian S, Dewa C. 1992. Standardization of the Denver Developmental Screening Test for Armenian children. J Pediatr Nurs. 7(2):106-109.

Albers CA, Grieve AJ. 2007. Test review: Bayley, N. (2006). "Bayley Scales of Infant and Toddler Development--3rd ed". San Antonio, TX--Harcourt Assessment. J Psychoeduc Assess. 25(2):180-190.

Aleksandrowicz MK, Aleksandrowicz DR. 1976. Precursors of ego in neonates: Factor analysis of Brazelton scale data. J Am Acad Child Psychiatry. 15(2):257-268. http://dx.doi.org/10.1016/S0002-7138(09)61486-2

Allen CW, Silove N, Williams K, Hutchins P. 2007. Validity of the social communication questionnaire in assessing risk of autism in preschool children with developmental problems. J Autism Dev Disord. 37(7):1272-1278. http://dx.doi.org/10.1007/s10803-006-0279-7

Allison C, Williams J, Scott F, Stott C, Bolton P, Baron-Cohen S, Brayne C. 2007. The Childhood Asperger Syndrome Test (CAST): Test-retest reliability in a high scoring sample. Autism. 11(2):173-185. http://dx.doi.org/10.1177/1362361307075710

Amod Z, Cockcroft K, Soellaart B. 2007. Use of the 1996 Griffiths Mental Development Scales for infants: A pilot study with a Black, South African sample. J Child Adolesc Ment Health. 19(2):123-130. http://dx.doi.org/10.2989/17280580709486647

Anderko L, Braun J, Auinger P. 2010. Contribution of tobacco smoke exposure to learning disabilities. J Obstet Gynecol Neonatal Nurs. 39(1):111-117. http://dx.doi.org/10.1111/j.1552-6909.2009.01093.x

Anderson PJ, Burnett A. 2017. Assessing developmental delay in early childhood — Concerns with the Bayley-III scales. Clin Neuropsychol. 31(2):371-381. http://dx.doi.org/10.1080/13854046.2016.1216518

Andersson HW. 1996. The Fagan Test of Infant Intelligence: Predictive validity in a random sample. Psychol Rep. 78(3):1015. http://dx.doi.org/10.2466/pr0.1996.78.3.1015

Aoki S, Hashimoto K, Ikeda N, Takekoh M, Fujiwara T, Morisaki N, Mezawa H, Tachibana Y, Ohya Y. 2016. Comparison of the Kyoto Scale of Psychological Development 2001 with the parent-rated Kinder Infant Development Scale (KIDS). Brain Dev. 38(5):481-490. http://dx.doi.org/10.1016/j.braindev.2015.11.001

Arcia E, Ornstein PA, Otto DA. 1991. Neurobehavioral Evaluation System (NES) and school performance. J Sch Psychol. 29(4):337-352. http://dx.doi.org/10.1016/0022-4405(91)90021-I

Arcia E, Otto DA. 1992. Reliability of selected tests from the Neurobehavioral Evaluation System. Neurotoxicol Teratol. 14(2):103-110. http://dx.doi.org/10.1016/0892-0362(92)90058-I

Armstrong K, Iarocci G. 2013. Brief report: The autism spectrum quotient has convergent validity with the social responsiveness scale in a high-functioning sample. J Autism Dev Disord. 43(9):2228-2232. http://dx.doi.org/10.1007/s10803-013-1769-z

Asghar A, Malik TA. 2016. Factor analytic structure and cross-informant agreement for Childhood Disruptive Behaviour Scale. PJPR. 31(1):77-92.

Aylward GP, Verhulst SJ, Colliver JA. 1985. Development of a brief infant neurobehavioral optimality scale: Longitudinal sensitivity and specificity. Dev Neuropsychol. 1(3):265-276. http://dx.doi.org/10.1080/87565648509540313

Baker EL. 1985. A computer-based neurobehavioral evaluation system for occupational and environmental epidemiology: Methodology and validation studies. Neurobehav Toxicol Teratol. 7(4):369-377.

Ball RS. 1977. The Gesell developmental schedules: Arnold Gesell (1880–1961). J Abnorm Child Psychol. 5(3):233-239. http://dx.doi.org/10.1007/BF00913694

Barnard-Brak L, Brewer A, Chesnut S, Richman D, Schaeffer AM. 2016. The sensitivity and specificity of the Social Communication Questionnaire for autism spectrum with respect to age. Autism Res. 9(8):838-845. http://dx.doi.org/10.1002/aur.1584

Baron-Cohen S, Hoekstra RA, Knickmeyer R, Wheelwright S. 2006. The Autism-Spectrum Quotient (AQ): Adolescent version. J Autism Dev Disord. 36(3):343-350. http://dx.doi.org/10.1007/s10803-006-0073-6

Baron-Cohen S, Wheelwright S, Skinner R, Martin J, Clubley E. 2001. The Autism-Spectrum Quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. J Autism Dev Disord. 31(1):5-17. http://dx.doi.org/10.1023/A:1005653411471

Becker PT, Lederman RP, Lederman E. 1989. Neonatal measures of attention and early cognitive status. Res Nurs Health. 12(6):381-388. http://dx.doi.org/10.1002/nur.4770120608

Benasich AA, Bejar, II. 1992. The Fagan Test of Infant Intelligence: A critical review. J Appl Dev Psychol. 13(2):153-171. http://dx.doi.org/10.1016/0193-3973(92)90027-F

Bishop DVM. 1998. Development of the Children's Communication Checklist (CCC): A method for assessing qualitative aspects of communicative impairment in children. J Child Psychol Psychiatr. 39(6):879-891. http://dx.doi.org/10.1017/S0021963098002832

Bishop DVM, Baird G. 2001. Parent and teacher report of pragmatic aspects of communication: Use of the Children's Communication Checklist in a clinical setting. Dev Med Child Neurol. 43(12):809-818. http://dx.doi.org/10.1017/S0012162201001475

Bishop S, Seltzer M. 2012. Self-reported autism symptoms in adults with autism spectrum disorders. J Autism Dev Disord. 42(11):2354-2363. http://dx.doi.org/10.1007/s10803-012-1483-2

Bode MM, D'Eugenio DB, Mettelman BB, Gross SJ. 2014. Predictive validity of the Bayley, 3rd ed at 2 years for intelligence quotient at 4 years in preterm infants. J Dev Behav Pediatr. 35(9):570-575.

Boer F, Westenberg PM. 1994. The factor structure of the Buss and Plomin EAS Temperature Survey (parental ratings) in a Dutch sample of elementary school children. J Pers Assess. 62(3):537. http://dx.doi.org/10.1207/s15327752jpa6203_13

Bölte S, Tomalski P, Marschik PB, Berggren S, Norberg J, Falck-Ytter T, Pokorska O, Jones EJH, Charman T, Roeyers H et al. 2018. Challenges and inequalities of opportunities in European psychiatry research: The example of psychodiagnostic tool availability in research on early autism identification. Eur J Psychol Assess. 34(4):270-277. http://dx.doi.org/10.1027/1015-5759/a000340

Borowitz KC, Glascoe FP. 1986. Sensitivity of the Denver Developmental Screening Test in speech and language screening. Pediatrics. 78(6):1075-1078.

Bos AF, Martijn A, Okken A, Prechtl HFR. 1998. Quality of general movements in preterm infants with transient periventricular echodensities. Acta Paediatr. 87(3):328-335. http://dx.doi.org/10.1111/j.1651-2227.1998.tb01447.x

Botting N. 2004. Children's Communication Checklist (CCC) scores in 11-year-old children with communication impairments. Int J Lang Commun Disord. 39(2):215-227. http://dx.doi.org/10.1080/13682820310001617001

Bould H, Joinson C, Sterne J, Araya R. 2013. The Emotionality Activity Sociability Temperament Survey: Factor analysis and temporal stability in a longitudinal cohort. Pers Individ Dif. 54(5):628-633. http://dx.doi.org/10.1016/j.paid.2012.11.010

Bourdon KH, Goodman R, Rae DS, Simpson G, Koretz DS. 2005. The Strengths and Difficulties Questionnaire: U.S. normative data and psychometric properties. J Am Acad Child Adolesc Psychiatry. 44(6):557. http://dx.doi.org/10.1097/01.chi.0000159157.57075.c8

Bricker D, Squires J. 1989. The effectiveness of parental screening of at-risk infants: The infant monitoring questionnaires. Topics Early Child Spec Educ. 9(3):67-85. http://dx.doi.org/10.1177/027112148900900306

Brown HR, Harvey EA. 2019. Psychometric properties of ADHD symptoms in toddlers. J Clin Child Adolesc Psychol. 48(3):423-439. http://dx.doi.org/10.1080/15374416.2018.1485105

Brown T. 2019. Structural validity of the Bruininks-Oseretsky test of motor proficiency—2nd ed brief form (BOT-2-BF). Res Dev Disabil. 85:92-103. http://dx.doi.org/10.1016/j.ridd.2018.11.010

Brown T, Lalor A. 2009. The Movement Assessment Battery for Children--2nd ed (MABC-2): A review and critique. Phys Occup Ther Pediatr. 29(1):86-103.

Brugha TS, McManus S, Smith J, Scott FJ, Meltzer H, Purdon S, Berney T, Tantam D, Robinson J, Radley J et al. 2012. Validating two survey methods for identifying cases of autism spectrum disorder among adults in the community. Psychol Med. 42(3):647-656. http://dx.doi.org/10.1017/S0033291711001292

Burakevych N, McKinlay CJD, Alsweiler JM, Wouldes TA, Harding JE. 2017. Bayley-III motor scale and neurological examination at 2 years do not predict motor skills at 45 years. Dev Med Child Neurol. 59(2):216-223. http://dx.doi.org/10.1111/dmcn.13232

Buss AH, Plomin R. 1984. Temperament: Early developing personality traits. Hillsdale, NJ: Lawrence Erlbaum Associates.

Campbell JM. 2005. Diagnostic assessment of Asperger's disorder: A review of five third-party rating scales. J Autism Dev Disord. 35(1):25-35. http://dx.doi.org/10.1007/s10803-004-1028-4

Chandler S, Charman T, Baird G, Simonoff E, Loucas T, Meldrum D, Scott M, Pickles A. 2007. Validation of the social communication questionnaire in a population cohort of children with autism spectrum disorders. J Am Acad Child Adolesc Psychiatry. 46(10):1324-1332. http://dx.doi.org/10.1097/chi.0b013e31812f7d8d

Chesnut SR, Wei T, Barnard-Brak L, Richman DM. 2017. A meta-analysis of the social communication questionnaire: Screening for autism spectrum disorder. Autism. 21(8):920-928. http://dx.doi.org/10.1177/1362361316660065

Chorpita BF, Yim L, Moffitt C, Umemoto LA, Francis SE. 2000. Assessment of symptoms of DSM-IV anxiety and depression in children: A revised child anxiety and depression scale. Behav Res Ther. 38(8):835-855. http://dx.doi.org/10.1016/S0005-7967(99)00130-8

Chow SMK, Chan L, Chan CPS, Lau CHY. 2002. Reliability of the experimental version of the Movement ABC. Br J Ther Rehab. 9(10):404-407. http://dx.doi.org/10.12968/bjtr.2002.9.10.13677

Chow SMK, Henderson SE. 2003. Interrater and test-retest reliability of the Movement Assessment Battery for Chinese preschool children. Am J Occup Ther. 57(5):574-577. http://dx.doi.org/10.5014/ajot.57.5.574

Chwen-Jen C, Li C, Li-Yin C. 2003. Developmental status among 3 to 5-Year-Old preschool children in three kindergartens in the Peitou District of Taipei City. J Nurs Res. 11(2):73.

Cirelli I, Graz MB, Tolsa JF. 2015. Comparison of Griffiths-II and Bayley-II tests for the developmental assessment of high-risk infants. Infant Behav Dev. 41:17-25. http://dx.doi.org/10.1016/j.infbeh.2015.06.004

Cohen LG. 1993. Wechsler Individual Achievement Test. Diagnostique. 18(3):255-268. http://dx.doi.org/10.1177/153450849301800306

Costa R, Figueiredo B, Tendais I, Conde A, Pacheco A, Teixeira C. 2010. Brazelton Neonatal Behavioral Assessment Scale: A psychometric study in a Portuguese sample. Infant Behav Dev. 33(4):510-517. http://dx.doi.org/10.1016/j.infbeh.2010.07.003

Croce RV, Horvat M, McCarthy E. 2001. Reliability and concurrent validity of the movement assessment battery for children. Percept Mot Skills. 93(1):275. http://dx.doi.org/10.2466/pms.2001.93.1.275

Darsaklis V, Snider LM, Majnemer A, Mazer B. 2011. Predictive validity of Prechtl's Method on the Qualitative Assessment of General Movements: A systematic review of the evidence. Dev Med Child Neurol. 53(10):896-906. http://dx.doi.org/10.1111/j.1469-8749.2011.04017.x

de la Osa N, Granero R, Penelo E, Ezpeleta L. 2014. Usefulness of the Social and Communication Disorders Checklist (SCDC) for the assessment of social cognition in preschoolers. Eur J Psychol Assess. 30(4):296-303. http://dx.doi.org/10.1027/1015-5759/a000193

De Pauw SSW, Mervielde I, Van Leeuwen KG. 2009. How are traits related to problem behavior in preschoolers? Similarities and contrasts between temperament and personality. J Abnorm Child Psychol. 37(3):309-325. http://dx.doi.org/10.1007/s10802-008-9290-0

De-Andrés-Beltrán B, Rodríguez-Fernández AL, Güeita-Rodríguez J, Lambeck J, Rodríguez-Fernández Á. 2015. Evaluation of the psychometric properties of the Spanish version of the Denver Developmental Screening Test II. Eur J Pediatr. 174(3):325-329. http://dx.doi.org/10.1007/s00431-014-2410-7

Douglas A, Letts L, Liu L. 2008. Review of cognitive assessments for older adults. Phys Occup Ther Geriatr. 26(4):13-43. http://dx.doi.org/10.1080/02703180801963758

Eaves RC, Milner B. 1993. The criterion-related validity of the Childhood Autism Rating Scale and the Autism Behavior Checklist. J Abnorm Child Psychol. 21:481-491. http://dx.doi.org/10.1007/BF00916315

Einspieler C, Bos AF, Libertus ME, Marschik PB. 2016. The general movement assessment helps us to identify preterm infants at risk for cognitive dysfunction. Front Psychol. 7. http://dx.doi.org/10.3389/fpsyg.2016.00406

Einspieler C, Prechtl HFR. 2005. Prechtl's assessment of general movements: A diagnostic tool for the functional assessment of the young nervous system. Ment Retard Dev Disabil Res Rev. 11(1):61-67. http://dx.doi.org/10.1002/mrdd.20051

Einspieler C, Yang H, Bartl-Pokorny KD, Chi X, Zang FF, Marschik PB, Guzzetta A, Ferrari F, Bos AF, Cioni G. 2015. Are sporadic fidgety movements as clinically relevant as is their absence? Early Hum Dev. 91(4):247-252. http://dx.doi.org/10.1016/j.earlhumdev.2015.02.003

Essau CA, Anastassiou-Hadjicharalambous X, Muñoz LC. 2011. Psychometric properties of the Spence Children's Anxiety Scale (SCAS) in Cypriot children and adolescents. Child Psychiatry Hum Dev. 42(5):557-568. http://dx.doi.org/10.1007/s10578-011-0232-7

Essau CA, Muris P, Ederer EM. 2002. Reliability and validity of the Spence Children's Anxiety Scale and the Screen for Child Anxiety Related Emotional Disorders in German children. J Behav Ther Exp Psychiatry. 33(1):1-18. http://dx.doi.org/10.1016/S0005-7916(02)00005-8

Fagan JF, Detterman DK. 1992. The Fagan Test of Infant Intelligence: A technical summary. J Appl Dev Psychol. 13(2):173-193. http://dx.doi.org/10.1016/0193-3973(92)90028-G

Fagan JF, Shepard P. 1986. The Fagan test of infant intelligence. Cleveland, OH: Infantest Corporation. 87.

Field TM, Hallock NF, Dempsey JR, Shuman HH. 1978. Mothers' assessments of term and pre-term infants with respiratory distress syndrome: Reliability and predictive validity. Child Psychiatry Hum Dev. 9(2):75-85. http://dx.doi.org/10.1007/BF01448351

Foreman MD. 1987. Reliability and validity of mental status questionnaires in elderly hospitalized patients... the SPMSQ, MMSE, and CCSE. Nurs Res. 36(4):216-220. http://dx.doi.org/10.1097/00006199-198707000-00004

Frankenburg WK, Camp BW, van Natta PA. 1971. Validity of the Denver Developmental Screening Test. Child Dev. 42(2):475-485. http://dx.doi.org/10.2307/1127481

Frankenburg WK, Dodds J, Archer P, Shapiro H, Bresnick B. 1992. The Denver II: A major revision and restandardization of the Denver Developmental Screening Test. Pediatrics. 89(1):91.

Frankenburg WK, Dodds JB. 1967. The Denver Developmental Screening Test. J Pediatr. 71(2):181-191. http://dx.doi.org/10.1016/S0022-3476(67)80070-2

Frankenburg WK, Fandal AW, Thornton SM. 1987. Revision of Denver Prescreening Developmental Questionnaire. J Pediatr. 110(4):653-657. http://dx.doi.org/10.1016/S0022-3476(87)80573-5

Friend TJ, Channell RW. 1987. A comparison of two measures of receptive vocabulary. Lang Speech Hear Serv Sch. 18(3):231-237. http://dx.doi.org/10.1044/0161-1461.1803.231

Gage GE, Naumann TF. 1965. Correlation of the Peabody Picture Vocabulary Test and the Wechsler Intelligence Scale for Children. J Educ Res. 58(10):466-468. http://dx.doi.org/10.1080/00220671.1965.10883277

Galeote M, Checa E, Sánchez-Palacios C, Sebastián E, Soto P. 2016. Adaptation of the MacArthur-Bates Communicative Development Inventories for Spanish children with Down syndrome: Validity and reliability data for vocabulary. Am J Speech Lang Pathol. 25(3):1-10. http://dx.doi.org/10.1044/2015_AJSLP-15-0007

Gard L, Rosblad B. 2009. The qualitative motor observations in Movement ABC: Aspects of reliability and validity. Adv Physiother. 11(2):51-57. http://dx.doi.org/10.1080/14038190902792346

Gau S-F, Lee CF, Lai MC, Chiu YN, Huang YF, Kao JD, Wu YY. 2011. Psychometric properties of the Chinese version of the Social Communication Questionnaire. Res Autism Spectr Disord. 5(2):809-818. http://dx.doi.org/10.1016/j.rasd.2010.09.010

Geurts H, Hartman C, Verte S, Oosterlaan J, Roeyers H, Sergeant J. 2009. Pragmatics fragmented: The factor structure of the Dutch Children's Communication Checklist (CCC). Int J Lang Commun Disord. 44(5):549-574. http://dx.doi.org/10.1080/13682820802243344

Gill K, Osiovich A, Synnes A, Agnew JA, Grunau RE, Miller SP, Zwicker JG. 2019. Concurrent validity of the Bayley-III and the Peabody Developmental Motor Scales-2 at 18 months. Phys Occup Ther Pediatr. 39(5):514-524. http://dx.doi.org/10.1080/01942638.2018.1546255

Glascoe FP, Byrne KE. 1993. The accuracy of three developmental screening tests. J Early Interv. 17(4):368-379. http://dx.doi.org/10.1177/105381519301700403

Glutting JJ, Youngstrom EA, Ward T. 1997. Incremental efficacy of WISC-III factor scores in predicting achievement: What do they tell us? Psychol Assess. 9:295-301. http://dx.doi.org/10.1037/1040-3590.9.3.295

Goodman R. 2001. Psychometric properties of the Strengths and Difficulties Questionnaire. J Am Acad Child Adolesc Psychiatry. 40(11):1337. http://dx.doi.org/10.1097/00004583-200111000-00015

Gratz KL, Roemer L. 2004. Multidimensional assessment of emotion regulation and dysregulation: Development, factor structure, and initial validation of the Difficulties in Emotion Regulation Scale. J Psychopathol Behav Assess. 26(1):41-54. http://dx.doi.org/10.1023/B:JOBA.0000007455.08539.94

Greer S, Bauchner H, Zuckerman B. 1989. The Denver Developmental Screening Test: How good is its predictive validity? Dev Med Child Neurol. 31(6):774-781. http://dx.doi.org/10.1111/j.1469-8749.1989.tb04073.x

Griffiths A, Morgan P, Anderson PJ, Doyle LW, Lee KJ, Spittle AJ. 2017. Predictive value of the Movement Assessment Battery for Children - 2nd ed at 4 years, for motor impairment at 8 years in children born preterm. Dev Med Child Neurol. 59(5):490-496.

Hadders-Algra M, Philippi H. 2018. Predictive validity of the General Movements Assessment: Type of population versus type of assessment. Dev Med Child Neurol. 60(11):1186-1186. http://dx.doi.org/10.1111/dmcn.14000

Hadley PA, Rice ML. 1993. Parental judgments of preschoolers' speech and language development: A resource for assessment and IEP planning. Semin Speech Lang. 14(4):278-288. http://dx.doi.org/10.1055/s-2008-1064177

Hattie J, Edwards H. 1987. A review of the Bruininks-Oseretsky Test of Motor Proficiency. Br J Educ Psychol. 57(1):104-113. http://dx.doi.org/10.1111/j.2044-8279.1987.tb03065.x

He J, Qiu P, Park KY, Xu Q, Potegal M. 2013. Young Chinese children's anger and distress: Emotion category and intensity identified by the time course of behaviors. Int J Behav Dev. 37(4):349-356. http://dx.doi.org/10.1177/0165025413477006

Heilmann J, Weismer SE, Evans J, Hollar C. 2005. Utility of the MacArthur-Bates Communicative Development Inventory in identifying language abilities of late-talking and typically developing toddlers. Am J Speech Lang Pathol. 14(1):40-51. http://dx.doi.org/10.1044/1058-0360(2005/006)

Hoekstra RA, Bartels M, Cath DC, Boomsma DI. 2008. Factor structure, reliability and criterion validity of the Autism-Spectrum Quotient (AQ): A study in Dutch population and patient groups. J Autism Dev Disord. 38(8):1555-1566. http://dx.doi.org/10.1007/s10803-008-0538-x

Hoffman LM, Loeb DF, Brandel J, Gillam RB. 2011. Concurrent and construct validity of oral language measures with school-age children with specific language impairment. J Speech Lang Hear Res. 54(6):1597-1608. http://dx.doi.org/10.1044/1092-4388(2011/10-0213)

Howieson DB, Lezak MD. 2007. Neuropsychological assessment. In: Bourgeois JA, Hales RE, Yudofsky SC, editors. The American Psychiatric Publishing Board Prep and Review Guide for Psychiatry. Arlington, VA: American Psychiatric Publishing, Inc. p. 55-61.

Ishikawa SI, Sato H, Sasagawa S. 2009. Anxiety disorder symptoms in Japanese children and adolescents. J Anxiety Disord. 23(1):104-111. http://dx.doi.org/10.1016/j.janxdis.2008.04.003

Junge C, Cutler A. 2014. Early word recognition and later language skills. Brain Sci. 4(4):532-559. http://dx.doi.org/10.3390/brainsci4040532

Kambas A, Aggeloussis N. 2006. Construct validity of the Bruininks-Oseretsky Test of Motor Proficiency-Short Form for a sample of Greek preschool and primary school children. Percept Mot Skills. 102(1):65-72.

Ketelaars MP, Cuperus JM, van Daal J, Jansonius K, Verhoeven L. 2009. Screening for pragmatic language impairment: The potential of the children's communication checklist. Res Dev Disabil. 30(5):952-960. http://dx.doi.org/10.1016/j.ridd.2009.01.006

Kooijman V, Junge C, Johnson EK, Hagoort P, Cutler A. 2013. Predictive brain signals of linguistic development. Front Psychol. 4. http://dx.doi.org/10.3389/fpsyg.2013.00025

Koyama T, Osada H, Tsujii H, Kurita H. 2009. Utility of the Kyoto Scale of Psychological Development in cognitive assessment of children with pervasive developmental disorders. Psychiatry Clin Neurosci. 63(2):241-243. http://dx.doi.org/10.1111/j.1440-1819.2009.01931.x

Krengel M, White RF, Diamond R, Leeza R. 1996. A comparison of NES2 and traditional neuropsychological tests in a neurologic patient sample. Neurotoxicol Teratol. 18(4):435-439. http://dx.doi.org/10.1016/0892-0362(96)00022-0

Krohn EJ, Traxler AJ. 1979. Relationship of the McCarthy Scales of Children's Abilities to other measures of preschool cognitive, motor, and perceptual development. Percept Mot Skills. 49(3):783-790. http://dx.doi.org/10.2466/pms.1979.49.3.783

Kwong AKL, Fitzgerald TL, Doyle LW, Cheong JLY, Spittle AJ. 2018. Predictive validity of spontaneous early infant movement for later cerebral palsy: A systematic review. Dev Med Child Neurol. 60(5):480-489. http://dx.doi.org/10.1111/dmcn.13697

Lam HMY. 2011. Assessment of preschoolers' gross motor proficiency: Revisiting Bruininks-Oseretsky Test of Motor Proficiency. Early Child Dev Care. 181(2):189-201. http://dx.doi.org/10.1080/03004430.2011.536640

Lester BM, Tronick EZ. 2004. History and description of the Neonatal Intensive Care Unit Network Neurobehavioral Scale. Pediatrics. 113:634-640.

Lim HC, Chan T, Yoong T. 1994. Standardisation and adaptation of the Denver Developmental Screening Test (DDST) and Denver II for use in Singapore children. Singapore Med J. 35(2):156-160.

Littell WM. 1960. The Wechsler Intelligence Scale for Children: Review of a decade of research. Psychol Bull. 57(2):132-156. http://dx.doi.org/10.1037/h0044513

Liu J, Bann C, Lester B, Tronick E, Das A, Lagasse L, Bauer C, Shankaran S, Bada H. 2010. Neonatal neurobehavior predicts medical and behavioral outcome. Pediatrics. 125(1):e90-e98. http://dx.doi.org/10.1542/peds.2009-0204

Loong NKS. 2016. Psychometric characteristics of a Chinese translation of the Barkley Adult ADHD Rating Scale-IV (BAARS-IV) in Hong Kong. [Hong Kong]: Alliant International University.

Lopez B, Lincoln A, Ozonoff S, Lai Z. 2005. Examining the relationship between executive functions and restricted, repetitive symptoms of autistic disorder. J Autism Dev Disord. 35(4):445-460. http://dx.doi.org/10.1007/s10803-005-5035-x

Luiz DM, Foxcroft CD, Povey JL. 2006. The Griffiths Scales of Mental Development: A factorial validity study. South African Journal of Psychology. 36(1):192-214. http://dx.doi.org/10.1177/008124630603600111

Luiz DM, Foxcroft CD, Stewart R. 2001. The construct validity of the Griffiths Scales of Mental Development. Child Care Health Dev. 27(1):73-83. http://dx.doi.org/10.1046/j.1365-2214.2001.00158.x

Luiz DM, Foxcroft CD, Tukulu AN. 2004. The Denver II Scales and the Griffiths Scales of Mental Development: A correlational study. J Child Adolesc Ment Health. 16(2):77-81. http://dx.doi.org/10.2989/17280580409486573

Lynch R. 2018. The psychometric properties of the Barkley Adult ADHD Rating Scale: IV (BAARS-IV) in a college sample. [Tallahassee, FL]: Florida State University.

Magiati I, Lerh JW, Hollocks MJ, Uljarevic M, Rodgers J, McConachie H, Ozsivadjian A, South M, Van Hecke A, Hardan A et al. 2017. The measurement properties of the Spence Children's Anxiety Scale-Parent Version in a large international pooled sample of young people with autism spectrum disorder. Autism Res. 10(10):1629-1652. http://dx.doi.org/10.1002/aur.1809

Majnemer A, Mazer B. 1998. Neurologic evaluation of the newborn infant: Definition and psychometric properties. Dev Med Child Neurol. 40(10):708-715. http://dx.doi.org/10.1111/j.1469-8749.1998.tb12332.x

Mathiesen KS, Tambs K. 1999. The EAS Temperament Questionnaire--factor structure, age trends, reliability, and stability in a Norwegian sample. J Child Psychol Psychiatry. 40(3):431-439. http://dx.doi.org/10.1111/1469-7610.00460

Meisels SJ. 1989. Can developmental screening tests identify children who are developmentally at risk? Pediatrics. 83(4):578.

Meyer A, Eilertsen DE, Sundet JM, Tshifularo J, Sagvolden T. 2004. Cross-cultural similarities in ADHD-like behaviour amongst South African Primary school children. S Afr J Psychol. 34(1):122-138. http://dx.doi.org/10.1177/008124630403400108

Milne SL, McDonald JL, Comino EJ. 2015. Alternate scoring of the Bayley- III improves prediction of performance on Griffiths Mental Development Scales before school entry in preschoolers with developmental concerns. Child Care Health Dev. 41(2):203-212. http://dx.doi.org/10.1111/cch.12177

Mitsopoulou E, Kafetsios K, Karademas E, Papastefanakis E, Simos P. 2013. The Greek Version of the Difficulties in Emotion Regulation Scale: Testing the factor structure, reliability and validity in an adult community sample. J Psychopathol Behav Assess. 35(1):123-131. http://dx.doi.org/10.1007/s10862-012-9321-6

Molina BS, Pelham WE, Blumenthal J, Galiszewski E. 1998. Agreement among teachers' behavior ratings of adolescents with a childhood history of attention deficit hyperactivity disorder. J Clin Child Psychol. 27(3):330-339. http://dx.doi.org/10.1207/s15374424jccp2703_9

Munsell KL. 2007. A screening battery for identifying at-risk infants: Prediction of outcome on Bayley Scales of Infant/Todder Development-III. [Fresno, CA]: Alliant International.

Muris P, Meesters C, van den Berg F. 2003. The Strengths and Difficulties Questionnaire (SDQ): Further evidence for its reliability and validity in a community sample of Dutch children and adolescents. Eur Child Adolesc Psychiatry. 12(1):1. http://dx.doi.org/10.1007/s00787-003-0298-2

Mutlu A, Livanelioğlu A, Korkmaz A. 2010. Assessment of "general movements" in high-risk infants by Prechtl analysis during early intervention period in the first year of life. Turk J Pediatr. 52(6):630-637.

Newcomer P, Hammill DD. 1978. Using the Test of Language Development with language-impaired children. J Learn Disabil. 11(8):521-524. http://dx.doi.org/10.1177/002221947801100811

Nishiyama T, Suzuki M, Adachi K, Sumi S, Okada K, Kishino H, Sakai S, Kamio Y, Kojima M, Suzuki S et al. 2014. Comprehensive comparison of self-administered questionnaires for measuring quantitative autistic traits in adults. J Autism Dev Disord. 44(5):993-1007. http://dx.doi.org/10.1007/s10803-013-2020-7

Noble Y, Boyd R. 2012. Neonatal assessments for the preterm infant up to 4 months corrected age: A systematic review. Dev Med Child Neurol. 54(2):129-139. http://dx.doi.org/10.1111/j.1469-8749.2010.03903.x

Nordahl-Hansen A, Kaale A, Ulvund SE. 2013. Inter-rater reliability of parent and preschool teacher ratings of language in children with autism. Res Autism Spectr Disord. 7(11):1391-1396. http://dx.doi.org/10.1016/j.rasd.2013.08.006

Nugent JH. 1976. A comment on the efficiency of the revised Denver Developmental Screening Test. Am J Ment Defic. 80(5):570-572.

Owens JS, Storer J, Holdaway AS, Serrano VJ, Watabe Y, Himawan LK, Krelko RE, Vause KJ, Girio-Herrera E, Andrews N. 2015. Screening for social, emotional, and behavioral problems at kindergarten entry: Utility and incremental validity of parent report. School Psych Rev. 44(1):21-40. http://dx.doi.org/10.17105/SPR44-1.21-40

Papay JP, Spielberger CD. 1986. Assessment of anxiety and achievement in kindergarten and first- and second-grade children. J Abnorm Child Psychol. 14(2):279-286. http://dx.doi.org/10.1007/BF00915446

Parmenter BA, Zivadinov R, Kerenyi L, Gavett R, Weinstock-Guttman B, Dwyer MG, Garg N, Munschauer F, Benedict RHB. 2007. Validity of the Wisconsin Card Sorting and Delis–Kaplan Executive Function System (DKEFS) Sorting Tests in multiple sclerosis. J Clin Exp Neuropsychol. 29(2):215-223. http://dx.doi.org/10.1080/13803390600672163

Patterson RL, Osullivan MJ, Spielberger CD. 1980. Measurement of state and trait anxiety in elderly mental health clients. J Behav Assess. 2(2):89-97. http://dx.doi.org/10.1007/BF01338925

Pease D, Rosauer JK, Wolins L. 1961. Reliability of three infant developmental scales administered during the first year of life. J Genet Psychol. 98:295-298. http://dx.doi.org/10.1080/00221325.1961.10534380

Pelletier J, Collett B, Gimpel G, Crowley S. 2006. Assessment of disruptive behaviors in preschoolers: Psychometric properties of the Disruptive Behavior Disorders Rating Scale and School Situations Questionnaire. J Psychoeduc Assess. 24(1):3-18. http://dx.doi.org/10.1177/0734282905285235

Pilowsky T, Yirmiya N, Shulman C, Dover R. 1998. The Autism Diagnostic Interview-Revised and the Childhood Autism Rating Scale: Differences between diagnostic systems and comparison between genders. J Autism Dev Disord. 28(2):143-151. http://dx.doi.org/10.1023/A:1026092632466

Putnam SP, Gartstein MA, Rothbart MK. 2006. Measurement of fine-grained aspects of toddler temperament: The Early Childhood Behavior Questionnaire. Infant Behav Dev. 29(3):386-401. http://dx.doi.org/10.1016/j.infbeh.2006.01.004

Robinson EB, Munir K, Munafo MR, Hughes M, McCormick MC, Koenen KC. 2011. Stability of autistic traits in the general population: Further evidence for a continuum of impairment. J Am Acad Child Adolesc Psychiatry. 50(4):376-384. http://dx.doi.org/10.1016/j.jaac.2011.01.005

Ronald A, Simonoff E, Kuntsi J, Asherson P, Plomin R. 2008. Evidence for overlapping genetic influences on autistic and ADHD behaviours in a community twin sample. J Child Psychol Psychiatr. 49(5):535-542. http://dx.doi.org/10.1111/j.1469-7610.2007.01857.x

Ronald A, Viding E, Happé F, Plomin R. 2006. Individual differences in theory of mind ability in middle childhood and links with verbal ability and autistic traits: A twin study. Soc Neurosci. 1(3):412-425. http://dx.doi.org/10.1080/17470910601068088

Rowe DC, Plomin R. 1977. Temperament in early childhood. J Pers Assess. 41(2):150-156. http://dx.doi.org/10.1207/s15327752jpa4102_5

Rubio-Codina M, Araujo MC, Attanasio O, Muñoz P, Grantham-McGregor S. 2016. Concurrent validity and feasibility of short tests currently used to measure early childhood development in large scale studies. PLoS One. 11(8). http://dx.doi.org/10.1371/journal.pone.0160962

Saemundsen E, Magnusson P, Smari J, Sigurdardottir S. 2003. Autism diagnostic interview-revised and the childhood Autism Rating Scale: Convergence and discrepancy in diagnosing autism. J Autism Dev Disord. 33(3):319. http://dx.doi.org/10.1023/A:1024410702242

Salisbury AL, Fallone MD, Lester B. 2005. Neurobehavioral assessment from fetus to infant: The NICU Network Neurobehavioral Scale and the Fetal Neurobehavior Coding Scale. Ment Retard Dev Disabil Res Rev. 11(1):14-20. http://dx.doi.org/10.1002/mrdd.20058

Sappok T, Brooks W, Heinrich M, McCarthy J, Underwood L. 2017. Cross-cultural validity of the social communication questionnaire for adults with intellectual developmental disorder. J Autism Dev Disord. 47(2):393-404. http://dx.doi.org/10.1007/s10803-016-2967-2

Sappok T, Diefenbacher A, Gaul I, Bölte S. 2015. Validity of the Social Communication Questionnaire in adults with intellectual disabilities and suspected Autism spectrum disorder. Am J Intellect Dev Disabil. 120(3):203-214. http://dx.doi.org/10.1352/1944-7558-120.3.203

Schopler E, Reichler RJ, DeVellis RF, Daly K. 1980. Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS). J Autism Dev Disord. 10(1):91-103. http://dx.doi.org/10.1007/BF02408436

Scott FJ, Baron-Cohen S, Bolton P, Brayne C. 2002. The CAST (Childhood Asperger Syndrome Test): Preliminary development of a UK screen for mainstream primary-school-age children. Autism. 6(1):9-31. http://dx.doi.org/10.1177/1362361302006001003

Scott FJ, Baron-Cohen S, Bolton P, Brayne C. 2002. 'The CAST (Childhood Asperger Syndrome Test): Preliminary development of a UK screen for mainstream primary-school-age children - Erratum. Autism. 6(4). http://dx.doi.org/10.1177/1362361302006001003

Sevin JA, Matson JL, Coe DA, Fee VE, Sevin BM. 1991. A comparison and evaluation of three commonly used autism scales. J Autism Dev Disord. 21(4):417-432. http://dx.doi.org/10.1007/BF02206868

Skuse DH, Mandy WPL, Scourfield J. 2005. Measuring autistic traits: Heritability, reliability and validity of the Social and Communication Disorders Checklist. Br J Psychiatry. 187:568-572. http://dx.doi.org/10.1192/bjp.187.6.568

Snider LM, Majnemer A, Mazer B, Campbell S, Bos AF. 2008. A comparison of the general movements assessment with traditional approaches to newborn and infant assessment: Concurrent validity. Early Hum Dev. 84(5):297-303. http://dx.doi.org/10.1016/j.earlhumdev.2007.07.004

Spence R, Owens M, Goodyer I. 2013. The longitudinal psychometric properties of the EAS Temperament Survey in adolescence. J Pers Assess. 95(6):633-639. http://dx.doi.org/10.1080/00223891.2013.819513

Spence SH. 1998. A measure of anxiety symptoms among children. Behav Res Ther. 36(5):545-566. http://dx.doi.org/10.1016/S0005-7967(98)00034-5

Spielberger CD. 1985. Assessment of state and trait anxiety: Conceptual and methodological issues. S Psychol. 2(4):6-16.

Spironello C, Hay J, Missiuna C, Faught BE, Cairney J. 2010. Concurrent and construct validation of the short form of the Bruininks-Oseretsky Test of Motor Proficiency and the Movement-ABC when administered under field conditions: Implications for screening. Child Care Health Dev. 36(4):499-507. http://dx.doi.org/10.1111/j.1365-2214.2009.01066.x

Spittle AJ, Doyle LW, Boyd RN. 2008. A systematic review of the clinimetric properties of neuromotor assessments for preterm infants during the first year of life. Dev Med Child Neurol. 50(4):254-266. http://dx.doi.org/10.1111/j.1469-8749.2008.02025.x

Stępień-Nycz M, Rostek I, Białecka-Pikul M, Białek A. 2018. The Polish adaptation of the Early Childhood Behavior Questionnaire (ECBQ): Psychometric properties, age and gender differences and convergence between the questionnaire and the observational data. Eur J Dev Psychol. 15(2):192-213. http://dx.doi.org/10.1080/17405629.2017.1292906

Sullivan MC, Miller RJ, Fontaine LA, Lester B. 2012. Refining neurobehavioral assessment of the high-risk infant using the NICU Network Neurobehavioral Scale. J Obstet Gynecol Neonatal Nurs. 41(1):17-23.

Takeda T, Burns GL, Jiang Y, Becker SP, McBurnett K. 2019. Psychometric properties of a sluggish cognitive tempo scale in Japanese adults with and without ADHD. Atten Defic Hyperact Disord. 11(4):353-362. http://dx.doi.org/10.1007/s12402-019-00300-z

Tasbihsazan R, Nettelbeck T, Kirby N. 2003. Predictive validity of the Fagan Test of Infant Intelligence. Br J Dev Psychol. 21(4):585-597. http://dx.doi.org/10.1348/026151003322535237

Teal MB, Wiebe MJ. 1986. A validity analysis of selected instruments used to assess autism. J Autism Dev Disord. 16:485-494. http://dx.doi.org/10.1007/BF01531713

Thal D, DesJardin JL, Eisenberg LS. 2007. Validity of the MacArthur–Bates Communicative Development Inventories for measuring language abilities in children with cochlear implants. J Med Speech Lang Pathol. 16(1):54-64. http://dx.doi.org/10.1044/1058-0360(2007/007)

Treloar JM. 1994. Wechsler Individual Achievement Test (WIAT). Interv Sch Clin. 29(4):242. http://dx.doi.org/10.1177/105345129402900409

Tronick E, Lester BM. 2013. Grandchild of the NBAS: The NICU Network Neurobehavioral Scale (NNNS): A review of the research using the NNNS. J Child Adolesc Psychiatr Nurs. 26(3):193-203.

Valentin T, Uhl K, Einspieler C. 2005. The effectiveness of training in Prechtl's method on the qualitative assessment of general movements. Early Hum Dev. 81(7):623-627. http://dx.doi.org/10.1016/j.earlhumdev.2005.04.003

Van Eck K, Finney SJ, Evans SW. 2010. Parent report of ADHD symptoms of early adolescents: A confirmatory factor analysis of the Disruptive Behavior Disorders Scale. Educ Psychol Meas. 70(6):1042-1059. http://dx.doi.org/10.1177/0013164410378093

Vasilev CA, Crowell SE, Beauchaine TP, Mead HK, Gatzke-Kopp LM. 2009. Correspondence between physiological and self-report measures of emotion dysregulation: A longitudinal investigation of youth with and without psychopathology. J Child Psychol Psychiatr. 50(11):1357-1364. http://dx.doi.org/10.1111/j.1469-7610.2009.02172.x

Vaughn BE, Taraldson B, Crichton L, Egeland B. 1980. Relationships between neonatal behavioral organization and infant behavior during the first year of life. Infant Behav Dev. 3(1):47-66. http://dx.doi.org/10.1016/S0163-6383(80)80006-3

Venetsanou F, Kambas A, Aggeloussis N, Fatouros I, Taxildaris K. 2009. Motor assessment of preschool aged children: A preliminary investigation of the validity of the Bruininks–Oseretsky Test of Motor Proficiency—Short form. Hum Mov Sci. 28(4):543-550. http://dx.doi.org/10.1016/j.humov.2009.03.002

Veselka L, Schermer JA, Just C, Hur YM, Rushton JP, Jeong HU, Vernon PA. 2012. Emotion and behavior: A general factor of personality from the EAS Temperament Survey and the Strengths and Difficulties Questionnaire. Twin Res Hum Genet. 15(5):668-671. http://dx.doi.org/10.1017/thg.2012.21

Wang HQ, Qu CY, Zhao SP. 2007. Standardization of the Griffith Mental Development Scales for children aged 0-7 years in the cities of Shanxi Province. Chin Mental Health J. 21(10):700-703.

Warren SL, Umylny P, Aron E, Simmens SJ. 2006. Toddler anxiety disorders: A pilot study. J Am Acad Child Adolesc Psychiatry. 45(7):859-866. http://dx.doi.org/10.1097/01.0000220852.94392.eb

Weinberg A, Klonsky ED. 2009. Measurement of emotion dysregulation in adolescents. Psychol Assess. 21(4):616-621. http://dx.doi.org/10.1037/a0016669

White RF, Diamond R, Krengel M, Lindem K. 1996. Validation of the NES2 in patients with neurologic disorders. Neurotoxicol Teratol. 18(4):441-448. http://dx.doi.org/10.1016/0892-0362(96)00021-9

White RF, James KE, Vasterling JJ, Letz R, Marans K, Delaney R, Krengel M, Rose F, Kraemer HC. 2003. Neuropsychological screening for cognitive impairment using computer-assisted tasks. Assessment. 10(1):86-101. http://dx.doi.org/10.1177/1073191102250185

Wiart L, Darrah J. 2001. Review of four tests of gross motor development. Dev Med Child Neurol. 43(4):279-285. http://dx.doi.org/10.1017/S0012162201000536

Wijedasa D. 2012. Developmental screening in context: adaptation and standardization of the Denver Developmental Screening Test-II (DDST-II) for Sri Lankan children. Child Care Health Dev. 38(6):889-899. http://dx.doi.org/10.1111/j.1365-2214.2011.01332.x

Williams J, Allison C, Scott F, Stott C, Bolton P, Baron-Cohen S, Brayne C. 2006. The Childhood Asperger Syndrome Test (CAST): Test-retest reliability. Autism. 10(4):415-427. http://dx.doi.org/10.1177/1362361306066612

Williams J, Scott F, Stott C, Allison C, Bolton P, Baron-Cohen S, Brayne C. 2005. The CAST (Childhood Asperger Syndrome Test): Test accuracy. Autism. 9(1):45-68. http://dx.doi.org/10.1177/1362361305049029

Wong BY, Roadhouse A. 1978. The Test of Language Development (TOLD): A validation study. Learn Disabil Q. 1(3):48-61. http://dx.doi.org/10.2307/1510937

Wong HS, Santhakumaran S, Cowan FM, Modi N. 2016. Developmental assessments in preterm children: A meta-analysis. Pediatrics. 138(2):1-12. http://dx.doi.org/10.1542/peds.2016-0251

Yao S, Zhang C, Zhu X, Jing X, McWhinnie CM, Abela JRZ. 2009. Measuring adolescent psychopathology: Psychometric properties of the self-report Strengths and Difficulties Questionnaire in a sample of Chinese adolescents. J Adolesc Health. 45(1):55-62.

Young S, González RA, Mutch L, Mallet-Lambert I, O'Rourke L, Hickey N, Asherson P, Gudjonsson GH. 2016. Diagnostic accuracy of a brief screening tool for attention deficit hyperactivity disorder in UK prison inmates. Psychol Med. 46(7):1449-1458. http://dx.doi.org/10.1017/S0033291716000039

Yu YT, Hsieh WS, Hsu CH, Chen LC, Lee WT, Chiu NC, Wu YC, Jeng SF. 2013. A psychometric study of the Bayley Scales of Infant and Toddler Development – 3rd Edition for term and preterm Taiwanese infants. Res Dev Disabil. 34(11):3875-3883.

Zarokanellou V, Kolaitis G, Vlassopoulos M, Papanikolaou K. 2017. Brief report: A pilot study of the validity and reliability of the Greek version of the Social Communication Questionnaire. Res Autism Spectr Disord. 38:1-5. http://dx.doi.org/10.1016/j.rasd.2017.03.001

Appendix D. Supplemental Files

The following supplemental files are available at https://doi.org/10.22427/NIEHS-DATA-NIEHS-01.

D.1.1. DNT Test Information Extraction Database

DNT Test Information Extraction Database
DNT_Test_Information_Extraction_Database.xlsx

National Institute of
Environmental Health Sciences
Division of the National Toxicology Program
Office of Policy, Review, and Outreach
P.O. Box 12233
Durham, NC 27709

www.niehs.nih.gov/reports

ISSN 2768-5632