This chapter is available only in PDF format (and not in HTML format) because the tables, figures and photos included in this chapter cannot be displayed satisfactorily in the HTML format. Also, the page numbers in the PDF format of this chapter do not correspond to the page numbers in the original report because of formatting changes during conversion of the original report (Word file) into PDF format.

Design and Methods

External Advisory Committee

In January 2000, agreement was reached with ATSDR and EPRI to form an External Advisory Committee (EAC). In addition to representatives of all the federal partners and from EPRI, membership also included five independent scientists, Drs. Peter Fried, David Bellinger, Joan Cranmer, David Otto and Joseph Jacobson. The EAC met twice during the course of the project, once on a telephone conference in mid-2000 prior to the start of data collection and again in Rochester on June 20, 2001, after about one year of data collection. Both meetings addressed methodological issues and resulted in recommendations that led to the final design. The principal suggestions from the June 2001 meeting resulted in the research team modifying recruitment practices to accelerate enrollment. The test battery was also slightly modified to streamline data collection. The EAC also reviewed the preliminary data analysis plan but did not contribute to the formulation of the final plan which emerged following the collection of all data and inspection of bivariates statistics.

Subjects

The subjects consisted of 293 9-16 year old children ($\overline{X}$=12.03) solicited from graduates of the Golisano Children's Hospital at Strong Neonatal Continuing Care Program (NCCP), a program designed for high-risk newborn follow-up during the first 10 years of life. A subset of NCCP patients are evaluated in the Neonatal Continuing Care Clinic (NCCC). The guidelines for admission into the NCCC are shown in Table 2. All NCCC patients experienced NICU hospitalizations as newborns. NCCC enrollment criteria are shown in Table 2. Between 1987 and 1993 (the relevant years for the target age group), 1,016 cases were enrolled in the NCCC. A very high percentage of these children resided in the Rochester metropolitan area and were available for recruitment.

Figure 1 depicts the process of enrollment and subsequent testing. The 1,016 subjects were identified as eligible based on their age.

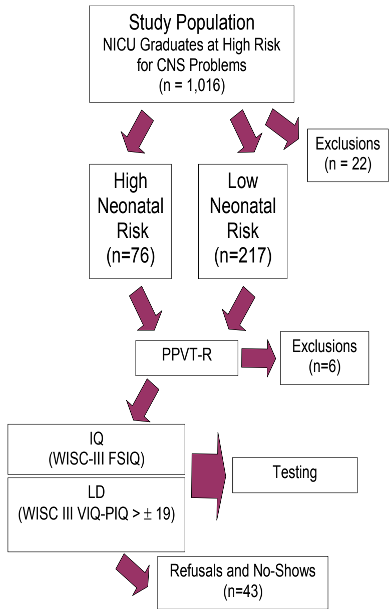

Fig. 1. Enrollment and Testing Flowchart

Table 2
Golisano Children's Hospital at Strong Neonatal Continuing Care Clinic (NCCC) Admission Guidelines

I. NCCC Follow-Up Eligibility Criteria (At least one must be present)

- < 32 weeks gestation (Gestational age as determined by the neonatologist by Ballard exam).
- < 1,250 grams
- SGA/IUGR (small for gestational age intrauterine growth retardation defined as <10^{th} % or < -2 SD on the Usher chart)
- NICU stay >24 days
- Clinical Research Trial (All infants enrolled in clinical research trials)
- Physician Referral (NICU attending or primary care physician)
- Specific Disorder
    - Significant Neurological Problems
        - Seizures
        - Hypoxic ischemic encephalopathy
        - Microcephaly
        - Periventricular Leukomalacia (PVL)
        - Intraventricular Hemorrhage (Grades III and IV)
        - Porencephaly
        - Cerebral Vascular Accident (CVA)
        - Significant Persistent Pulmonary Hypertension (PPHN)
    - Infectious Diseases
        - Meningitis
        - Congenital TORCH Infections (toxoplasmosis rubella, cytomegalovirus, herpes, other)
        - Culture Proven Sepsis *at Birth*
        - Genetic / Metabolic Disorders
    - Surgical Complications
        - Diaphragmatic Hernia
        - Gastroschisis
        - Omphalocele
        - Extracorporeal Membrane Oxygenation (ECM0)
        - Double Volume Exchange Transfusion

II. Timing of First Visit

- Average first visit between 6 and 9 months corrected age
- Specific conditions seen earlier in life:
    - 24-26 weeks gestation are seen at 6 months corrected age
    - Neurological problems (seizures or other) are seen 1 to 2 months post discharge
    - Any baby whose health care provider (CHN, PNP, physician) has expressed developmental or neurological concerns will be seen at any age up to 2 year upon notification of NCCC.

Children with a diagnosis of Down syndrome or with a history of an intra-cranial hemorrhage ≥ Grade IV were excluded (n=22). Each remaining child's name, address, and primary care pediatrician's name were then drawn from the NCCC computerized database. The subjects were solicited through the NCCC secretary, who was not part of the research team, and their names were not turned over to us until they had accepted our offer to participate in the study. Thus, we do not know who refused. The clinic secretary then contacted parents for an initial screening appointment. Children were also sorted by the pediatric practice providing their primary health care. Each pediatrician was then contacted by mail with a list of his or her eligible patients. Any child for whom a pediatrician decided that testing would not be advised was not contacted for a screening appointment (n=3). If parents gave oral consent the child was given an appointment for formal consenting and initial screening. A small number refused oral consent or did not appear for their screening appointment (n=40). At the screening appointment, consent and assent were obtained and each consenting child was given the Peabody Picture Vocabulary Test Revised (PPVT-R). The child's head circumference was measured and an informational questionnaire was completed by the parent. The questionnaire included notation of any serious medical problems (hospitalizations, mental illness, accidents, serious illnesses). Information was also requested about special school services, allergies, medications, need for corrective lens or hearing aids, or food preferences. These latter data were used only for coordinating the testing visit, not for analysis purposes. Any child receiving a score of <55 on the PPVT-R or having a head circumference greater than 3 SDs above or below the average for his or her age was excluded from the study (n=6). Only those children who remained eligible and whose caregivers gave formal consent were then given a second appointment for study testing. A small number of additional eligible and consenting subjects did not appear for their testing appointment (n=3), leaving a final sample of 293. All subjects were tested at the University of Rochester Medical Center. Subjects were paid $40 each for their participation. In addition, each child received a tee-shirt and a $25 gift certificate for music CDs. Parents were provided with a detailed written summary of their child's results.

Risk Categories

Each subject was classified according to three different developmental risk factors. Testers were blinded to the subjects risk status.

Neonatal Risk. The literature indicates that samples of low birth weight children (< 1,500 grams) offer useful populations for determining the sensitivity of the motor and sensory-motor tests. Of a group of 83 very low birth weight children who had normal neurological and intellectual development at 5 years of age, 71% had below average scores for fine motor skills (Goyen et al, 1998). Poor visual contrast sensitivity has also been observed in such children (Powls et al, 1997). Somatosensory function (tactile localization) in low birth weight children seems to be impaired (Maio-Feldman, 1994). NICU graduates are at a higher risk for learning disabilities and may display more neurodevelopmental problems than normal achievers (Blumsack et al, 1997). Children with learning disabilities have deficiencies in both fine and gross motor skills (Bruininks and Bruininks, 1977; Kendrick and Hanten, 1980; O'Brien et al, 1988), in vigilance performance (Swanson, 1983), and in visual memory (Hung et al, 1987).

The medical records for each child enrolled in the NCCC who fell within the eligible age range were searched to determine the presence or absence of high-risk status, defined as (1) evidence of CNS damage based upon CNS imaging; (2) head circumference ≥ 2 SDs above the age mean, or (3) presence of a CNS infection during NICU hospitalization. One of the senior investigators (a pediatric neurologist and a board certified neonatologist) accomplished the classification. A total of 76 children met the criteria for high neonatal risk. The mean IQ for the high neonatal risk group was 98.9 and 97.5 for the low risk group.

IQ. The IQ is an omnibus measure of general cognitive ability and is widely understood to reflect cognitive deficits when it is below average. Lower IQ scores represent various degrees

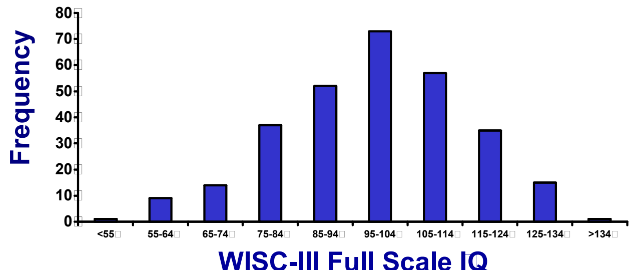

Fig. 2. WISC-III Full Scale IQ Distribution of Cohort (n=293)

of mental retardation, a serious compromise of intelligence and adaptive function. Exposures to various neurotoxicants have been linked to significant adverse effects on IQ. The lower the IQ falls below 70, the more severe the mental retardation. But smaller adverse effects on IQ have been documented for low dose exposures. Therefore, we were interested in recruiting subjects whose IQs were no lower than the range of mild mental retardation, usually considered being > 55 on a standard measure of IQ.

At the conclusion of the screening visit, children who scored at least 56 on the PPVT-R (Dunn & Dunn, 1981) were given an abbreviated version of the *Wechsler Children's Intelligence Test-Third Edition (WISC-III)* to determine the Full Scale IQ, a common psychometric measure of general intelligence. We decided to use an abbreviated version to save testing time. There are several ways to abbreviate the WISC-III; we selected the method that allows for the most subtest scores to be obtained, giving a robust estimate of IQ, but allowing for a valid estimate of variation between verbal and performance abilities (Silverstein, 1987; 1989). The resulting distribution is shown in Figure 2. WISC-III subtest scores are shown in Table 3. The Silverstein modification that we used drops every other item from each subtest administered, and does not include Object Assembly on the Performance Scale. A total of 61 children had WISC-III Full Scale IQ scores of >55 but <85. We subsequently sorted the subjects into high (IQ>84) and low (IQ ≤84) IQ groups. The high IQ group had a mean IQ=104.9 and the low IQ group mean was 74.6.

Table 3
WISC-III Subscale and Subtest Scores

WISC-III Subtest	Mean	SD	Range
Verbal Scale IQ	*101.4*	*16.5*	*55-139*
Information	10.0	3.8	1-19
Similarities	10.7	3.4	1-19
Arithmetic	9.9	3.7	1-19
Vocabulary	10.9	4.4	1-19
Digit Span	9.4	3.2	1-19
Performance Scale IQ	*96.0*	*17.4*	*46-140*
Picture Completion	9.1	3.6	1-18
Picture Arrangement	10.0	4.4	1-19
Block Design	9.1	3.6	1-19
Coding	8.8	3.5	1-18

Learning Disability. All children were classified according to whether or not they had evidence of a learning disability. There are several conventional ways to operationally define learning disability (LD), and considerable disagreement about which is the most appropriate. Most clinical definitions demand establishing a difference between cognitive ability and school achievement, a procedure which we could not easily follow. Sattler (2002) points out that a child with LD can be characterized by a significant difference between verbal ability, as measured by the WISC-III Verbal IQ (VIQ) and performance ability measured by the WISC-III Performance IQ (PIQ). His examples all exceed 10 points. We decided to adopt this latter approach, and with the concurrence of the EAC, established the conservative criterion of a minimum difference of 20 points. Thus, for this study, LD was defined as a greater than 19-point difference between scores from the two major subscales of the WISC-III, the first measuring verbal ability (the Verbal IQ) and the other, performance ability (the Performance IQ). A total of 49 subjects were categorized as learning disabled. The mean IQ for both LD groups was 98.6.

Table 4 shows the Pearson intercorrelations among the three risk factors. The standardization sample mean for the WISC-III Full Scale IQ is 100 (SD=15), and a score of 100 is considered normal.

Table 4
Risk Factor Intercorrelations

Risk Factor	Percent of Cohort or Cohort Mean(SD)	Pearson R (*p*)	
		LD Group N=49	IQ (N=293)
High Neonatal Risk Group (N=76)	26%	-0.00009(0.99)	-0.036 (0.54)
LD Group (N=49)	17%	---	0.027 (0.64)
IQ (N=293)	98.6(16.9)	---	---

There was a high degree of group separation and independence of risk factors, indicating that risk factors co-linearity was limited, and that risk factors might have independent effects on endpoints. While the original design of the study called for an analysis using only neonatal risk as the sole risk indicator, the findings in Table 4 persuaded us to adopt three categories of risk analyze each test and task's capacity to predict each risk category separately.

Tests and Tasks Comprising the Battery

The battery was comprised of tests and tasks to assess subtle variation in performance in electrophysiological, neuropsychological, and sensory motor functions reflective of the developmental domains discussed earlier. **The domain measured by each test or task is noted following the abbreviations given in Table 1.**

All neuropsychological and electrophysiological tests of auditory functions required data on ear-specific hearing sensitivity and middle ear functioning. Hence all children who entered the final cohort were tested using a standard audiometer. Behavioral audiologic pure-tone thresholds testing was completed on each ear at octave intervals from 500-8000 Hz (500, 1000, 2000, 4000 and 8000 Hz). In addition, all subjects received tympanometric testing to assess the pathologic status of the middle ear system. It is known that often children under the age of 5 years and some children older than age 5 have frequent bouts of middle ear effusion. Although the middle ear effusion usually is short lasting, especially in older children, the effusion can cause temporary shifts in pure-tone thresholds and can also affect other tests relying on good hearing such as otoacoustic emissions testing, auditory evoked potentials and tests for central auditory processing abilities.

Auditory Processing

Pitch Pattern Sequence Test (AV). The PPST (Katz, 1978) is test assessed the listener's perception of the sequence of high versus low pitch tones presented in bursts of three tones per test item. The test was presented via audiotape using a high fidelity tape recorder. The PPST included a training session to familiarize the child with the task before test items were given. The test was administered in a quiet room, using a free field, as suggested by Auditec, the developer of PPS. The task required the child to listen to three-tone bursts, identify them verbally or manually as to sequence of pitches, e.g., high low high. Children responded verbally on Trial 1 and manually (hand gestures to indicate *high/low tones in* the sequence) on Trial 2.

Auditory Continuous Performance Test (C). The ACPT (Keith, 1998) is a non-verbal monaural test of auditory vigilance. The test assessed the ability to sustain auditory attention over time by listening and responding to the recurrence of the stimulus *dog* in a continuous stream of foil words. The child was asked to raise his/her thumb when the taped voice of the speaker said the word *dog.* This test was given through earphones using audiotape with the same equipment as used for the PPST.

Dichotic Digits Test-Double Pairs (AV). The DDT-DP (Willeford, 1976) tested the ability to remember and repeat a series of four digits when presented dichotically. The task involved listening to a series of four digits, presented simultaneously, two digits in each ear. The child repeats the four digits. The test was administered on audiotape using the same equipment as was used for the PPST. A set of training items was presented before the test items were given.

Electrophysiological Tests

Visual Evoked Potentials (AV). A *Neuroscan* electrophysiological workstation (STIM acquisition and analysis software, *SynAmp* amplifiers) was used to acquire and analyze sensory evoked potential data on a total of 60 children. The STIM software package (Neuroscan Inc.) was used to generate checkerboard stimuli subtending visual angels of 15' and 30'.Pattern reversal visual evoked responses were acquired via independent stimulation of each eye with both check sizes at a rate of 1.7 reversals per second. Absolute and interside latencies of the N75, P100 and N145 components were determined. In addition, an examination of N75-P100 amplitude ratios will be undertaken to identify any possible asymmetries. Photopic luminance levels were studied. Scotopic levels required a longer testing time because of the need for dark adaptation, which were studied psychophysically.

Visual evoked potentials were discontinued on the advice of the External Advisory Committee in June 2001 after initial data from the 60 subjects suggested that it was not discriminating between risk categories.

Cognitive Event Related Potentials (C). Long-latency ERPs were recorded while the subject performs a visual Continuous Performance Task. The subject was seated 1 meter in front of a computer screen in a dimly lit room and presented with a series of five different letters in the center of the screen. Stimuli were presented at a rate of approximately one every 2 seconds. The subject was instructed to press the button on a response pad as quickly as possible whenever the letter "X" (target) appeared and to do nothing when a different letter (non-target) appeared. Averages were constructed from responses evoked from both correctly identified target and non-target stimuli. In addition to electrophysiological data, several behavioral measures were collected. Behavioral data included (1) total number correct responses, (2) percentage correct responses, (3) reaction time for correct responses, (4) total number errors of commission, (5) reaction time or errors of commission, and (6) total number errors of omission.

Data acquisition occurred with an Analog/Digital sampling rate of 500 Hz. Data were obtained from three scalp locations, frontal (Fz), central vertex (Cz), and parietal (Pz) referenced to linked earlobes. Vertical eye-movements (EOG) were recorded from electrodes attached above and below the left eye. Bandpass filters were set at 0.1 Hz to 30 Hz. A continuous data file was acquired containing the electrophysiological data and trigger pulses identifying each letter as either a target or non-target. Prior to off line construction of the average, all raw data were manually reviewed for excessive muscle or eye-movement artifacts. Any such data were marked and excluded from the subsequent average. An off line averaging routine was conducted to create an 1100 msec. epoch, with the first 100 msec. serving as a pre-stimulus baseline for determination of baseline-to-peak component amplitude.

Auditory Evoked Potentials (AV). Brainstem auditory evoked potentials were acquired unilaterally (from the right ear)[3] using standard clinical receiving parameters. Data were obtained using rarefaction clicks presented at levels of 60 and 80 dBnHL at rates of 19.9, 39.9, and 69.9 Hz. The waveform response peaks I, III, and V were identified, the absolute latencies (wave I, III and V) analyzed and interwave latencies (I-III, III-V and I-V) were calculated at each intensity and presentation rate. We then calculated a *latency shift* by subtracting the latency or interwave latency at a specific level (dBnHL for 19.9 Hz) from the corresponding component at a faster presentation rate (e.g., wave V absolute latency at 80 dBnHL for 20 Hz presentation rate minus the absolute wave V latency at 80 dBnHL for the 40 or 70 Hz presentation rate). This endpoint was chosen instead of absolute latencies because they are unaffected by other factors that affect absolute latencies, such as gender. The latency shift data were compared across groups to determine possible differences in the brainstem conduction between groups.

Otoacoustic Emissions (AV). Spontaneous, click evoked, and distortion product (DP) emission testing was completed for the right ear[3]. Subjects were comfortably seated and a small ear probe (microphone and speakers) placed in the outer portion of the ear canal. The number and level of spontaneous otoacoustic emissions were recorded and used for later analysis. Click evoked otoacoustic emissions were elicited using an 80 dB ($\pm$3 dB) broadband click stimulus. The signal-to-noise response was analyzed in 1/3-octave bands from 1000-4000 Hz. The amplitude of the response and the signal-to-noise of the response of the cubic difference (2F1-F2) distortion product emission were determined for F2 frequencies of 1000, 2000, 3000, 4000 and 6000 Hz. The distortion product emissions was determined using the level (L1) of the lower frequency (F1) evoking stimulus of 70 and 55 dB SPL while the level (L2) of the higher frequency (F2) evoking stimulus is (60 and 45 dB SPL). The ratio of the F2/F1 was 1.22 as has been shown to produce the most robust distortion product emission.

Experimental Tasks (Neuropsychological)

CANTAB (C, AV). Children were tested for motor abilities, attention, and memory using the *Cambridge Neuropsychological Test Automated Battery (CANTAB)*. This computerized battery used a touch screen monitor, reducing variation related to an individual subject's computer experience. The stimuli for all the tests were patterns of shapes and colors, not recognizable figures or symbols, allowing the battery to be used across cultures. CANTAB has been used extensively in populations diagnosed with Alzheimer's disease and two tests within CANTAB (paired associates learning and the graded naming test collectively) were recently described as a preclinical marker of the disease (Blackwell et al., 2004). The battery included five CANTAB sub-tests used successfully in a previous study of lead exposed children (Canfield, et al., 2004), and field tested in the pilot study: *Delayed Match to Sample, Paired Associate Learning, Big/Little Circle, Intra-Dimensional/Extra-Dimensional Shift, and Reaction Time.* The detailed procedures were as follows:

- *Delayed Match to Sample*: subjects are presented with a complex visual pattern for 4.5 seconds and then after a brief delay, four choice patterns. Delay intervals between sample presentation and choice were simultaneous or 0, 4 or 12 seconds. After 3 practice trials, there are 40 counterbalanced test trials, including 10 at each of the delay intervals (simultaneous, 0 4, or 12 seconds).

- *Paired Associated Learning*: Six boxes are initially drawn on the screen. All are opened in a randomized order, with one containing a pattern. After the last box has been closed, the pattern is shown in the middle of the screen for 3 seconds and the subject must point to the box where the pattern was located. If the choice is correct, the procedure is repeated with a single, new pattern. If incorrect, all the boxes are reopened, after which the subject must choose again. After three correct sets with a single pattern, the number of patterns is increased to two for two sets, to three for two sets, to six and then to eight for one set each. If any list has not been correctly completed within ten presentations, the test is terminated.

- *Big/Little Circle*: a series of pairs of circles, one large and one small are presented. Subject must first point to the smaller of the two and then after 20 trials to the larger for 20 trials.

- *ID/ED Shift*: Stage 1 involved learning a simple visual discrimination using color-filled nonsense shapes. At each stage, the child touched shapes and received automated feedback until meeting the criterion of 6 consecutive correct responses. For stage 2, the task contingencies were reversed such that the previously correct stimulus was now incorrect. The next two stages involved the addition of an irrelevant dimension, with the dimension adjacent to the familiar shapes in Stage 3 and overlapping the shapes in Stage 4. After reaching criterion, the correct shape was again reversed, but the irrelevant line patterns continued to be presented in Stage 5.

Stages 6 and 7 presented the intradimensional shift by introducing new shape and line configuration stimuli. After meeting criterion with the new stimuli (Stage 6), the correct stimulus was again reversed (stage 7). Stage 8 introduced the extradimensional shift which meant that the previous irrelevant dimension of the stimuli became the relevant dimension. Stage 9 imposed a reversal of the newly learned discrimination. If criterion was not reached in 50 trials for any stage, the test automatically terminated.

- *Reaction Time*: In the first condition, the subject simply had to press a yellow dot that appeared in the center of the screen. After achieving 5/6 correct, or a total of 18 trials, the choice reaction task was introduced where the dot could now appear in any of 5 locations. Criterion was again 5/6 correct or a maximum of 30 trials. If the subject failed to achieve criterion on either of the first two stages, the test is terminated. If successful, subjects were now required to a touch key and to hold down the touch key until the dot appeared in the center of the screen followed by a release of the key. Stage 4 was identical except that the subject had to touch the dot after releasing the touch key. In stage 5, the dot could then appear in any of 5 locations. In all stages, the subject was trained to a criterion of 5/6 correct or 18 trials for simple reaction time or 40 trials for choice reaction time.

FI Self Control (C). Attention deficit disorder is associated with increased response rates on fixed interval (FI) schedules of reinforcement, a schedule in which the first response occurring after a designated fixed interval of time has elapsed results in reinforcement, with responses occurring during the interval itself having no programmed consequences. Boys with ADHD have been shown to display increased response rates during both the FI and extinction components of the schedule, exhibiting shorter inter-response times (times between successive responses), defined by the authors as response bursts and *impulsiveness*. Impulsivity has been a hallmark of children with attention deficits and is defined experimentally (in self-control or delay of gratification paradigms) as a preferred choice of an immediate small reward relative to a larger but delayed reward. Thus, children with attention deficit-hyperactivity disorders exhibit both increased response rates on FI schedules and impulsivity (the choice of small but immediate rewards in self-control procedures).

Two studies of normal children by Darcheville and colleagues (1992; 1993) have shown that *increased FI response rates actually predict impulsivity***.** Specifically, children ranging in age from 3 months to 6 years exhibiting high response rates and short post-reinforcement pause times on a FI schedule also systematically chose the smaller immediate reinforcer rather than the longer but delayed reinforcers in a self-control procedure, i.e., children with increased FI rates were more impulsive. In contrast, children with very low response rates and long post-reinforcement pause times on the FI schedule (typically a single response per interval) preferred the longer but delayed rewards, i.e., they exhibited self-control. Thus *increased FI response rates were a surrogate for impulsivity*.

A multiple FI Self-Control procedure was used to obtain FI measures and percentage of self-control choices. Moreover, since each subject performed both tasks, correlations were calculated to quantify the relationship between FI and self-control performance. Software for this procedure was programmed on a PC using *LabView*.

FI Component. Subjects earned 1 point (reinforcer) for the first lever press occurring at least 30 sec after the prior reinforcer delivery. Responses during 30-second interval had no programmed consequences. This component lasted approximately five minutes, i.e. until the 30-sec interval in progress ended.

Self-Control Component. Similar procedures have been used in pigeons, rats, and humans. The paradigm, shown in Figure 3, consisted of a series of trials offering a choice of a response lever associated with 1 point after a 5 sec delay (Lever SC) or a lever associated with 5 points after a 20 sec delay (Lever I). Each session consisted of 3 blocks of 14 trials. The first four trials of each block consisted of forced trials to ensure continued experience with the contingencies associated with each lever, two with only Lever SC available, and the other two with only Lever I available with the order randomized so that neither lever was always presented for the first forced trial of a block. The remaining 10 trials of each block consisted of free choice trials with both levers available. On any trial, occurrence of a response results in a tone and a removal of one (forced trials) or both (free choice trials) operandum. After the designated delay elapsed (either 5 or 20 sec), reward was delivered (1 point after a 5 sec delay choice or 5 points after a 20 sec delay choice). Each trial lasted 45 sec; therefore, the time between the end of the reinforcement and the beginning of the next trial (post-reinforcement interval) varied depending on the subject's response time for that trial and the delay associated with the response choice.

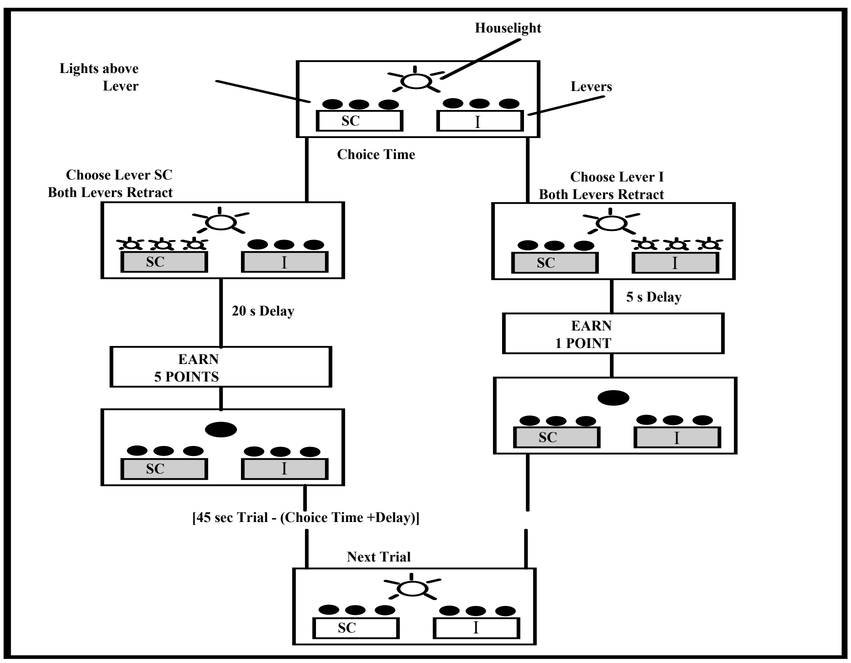

Fig. 3 A Schematic of the Self Control Paradigm.

Experimental Tasks (Sensory and Motor)

Fine Motor Control (FM). The system for assessing motor performance involved a task requiring the subject to make positioning responses in accordance with a computer display. As shown in Figure 4a, the subject gripped a handle attached to a rod; the rod acts as a lever whose fulcrum lies at the elbow. Rotation of the forearm in the vertical plane changed the angular position of a rotary variable displacement transducer located at the fulcrum. Transducer output was coupled to an analog-digital converter whose signal is transmitted to a digital computer. On each of 20 trials spaced 20 sec apart, a narrow band appeared on the computer display screen at 0º, then moved to a horizontal displacement of ±25º. The sequence was randomly selected. The subject's task was to rotate the lever so that a large dot, representing the angular position of the forearm, fell within the band and rested there for 8 seconds (see Fig 4b). At the end of the 8-second period, which is accompanied by a sound, and provided the target follower had remained stable, the band moved to a new position. This system allowed measurement of reaction time (latency to begin movement), movement time (duration of movement from its inception until the target follower fell within the target for 2 seconds), and forearm tremor (sampled at 1 kHz for 6 seconds). Tremor was quantified with a Fast Fourier Transform to yield a power spectrum plot of amplitude versus frequency.

Fig. 4a. The Fine Motor Control Apparatus. Picture shows the handgrip and the position of the subject's arm during testing.

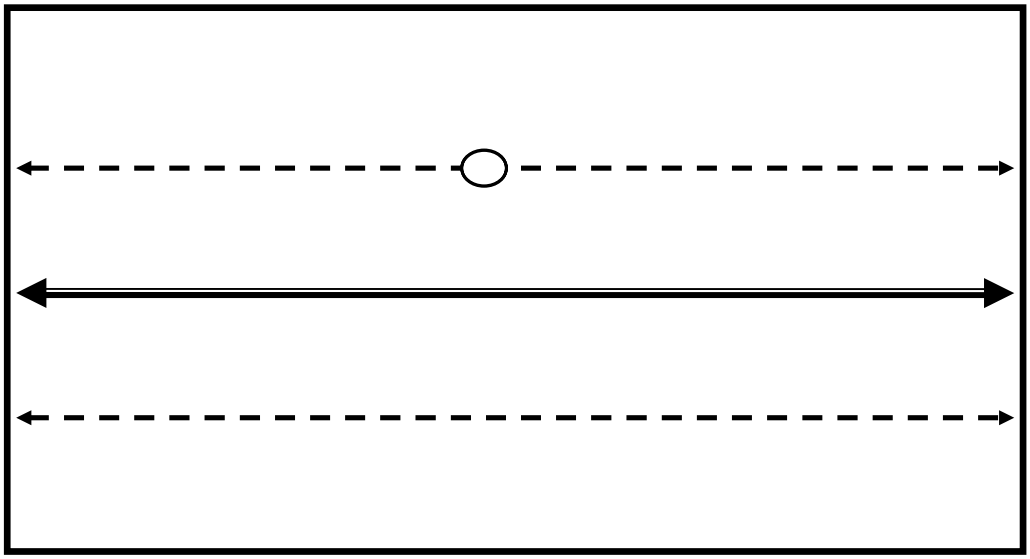

Fig. 4b. The Fine Motor Control Apparatus Display. The subject's task was to keep the oval icon in the center of the display screen.

Scotopic Visual Form Discrimination (SS). In a totally darkened room, the subject viewed a display with three light-emitting diodes whose elements represented a figure-8 configuration. A *LabView* digital-analog instrument was used to drive light-emitting diodes (LEDs). The subject rested his or her hands on a panel with three buttons. On each trial, two of the LEDs presented identical patterns selected from the seven display elements. The third was different, and appeared randomly in one of the three positions. The subject's task was to select the odd position by pressing the corresponding button. For this purpose, we converged to a series of thresholds with a PEST procedure (Parameter Estimation by Sequential Testing). We used a series of thresholds because, over the 20-min test period, the eye becomes increasingly sensitive (dark adaptation). The brightness of the LEDs was determined by the applied current, which was controlled by the computer. Because commercially available LEDs display intrinsic variations, we had to select units that closely match. The endpoints were the slope of the adaptation curve and the R^2 associated with that slope. The function is typically curvilinear so we needed a parameter that estimated the entire 20-minute period.

Visual Spatial Contrast Sensitivity (SS). We used a commercially available back lighted display that provided constant luminance independent of ambient lighting (*Vistech Corp*). It displayed six rows and nine columns of circular targets varying in both contrast and in spatial frequency. The children were asked to indicate, as they proceeded from left (most discriminable contrast) to right (less discriminable), the orientation of the display (left tilt, right tilt, or vertical) by the position of their hand. Each eye was tested separately.

Alertness and Coordination: Complex Perceptual-Motor Performance

Monitoring and Vigilance (PM). Alertness and coordination were measured using a system of complex controls and displays that simulated the demands of computer games that involve aircraft piloting or air traffic control. The system, depicted in Figure 5a, relied for the joint assessment of coordination and vigilance on a two-dimensional joystick control, accessory response buttons, and a computer-controlled video display. During testing, the subject responded by movements of a joystick and by activating detent buttons on the joystick assembly. The subject's job was to maintain pointers indicating attitude, velocity and fuel in nominal zones on each of three indicators, using a joystick and foot pedals (see Figure 5b). Instructions to the subject were given on the computer screen, with illustrations. The endpoints included percentages of total tracking time in which alarms, hazards and tracking errors occurred and alarm duration during the final five minutes of the task.

Fig. 5a. The Multitasking Apparatus used for Measuring Alertness and Coordination.

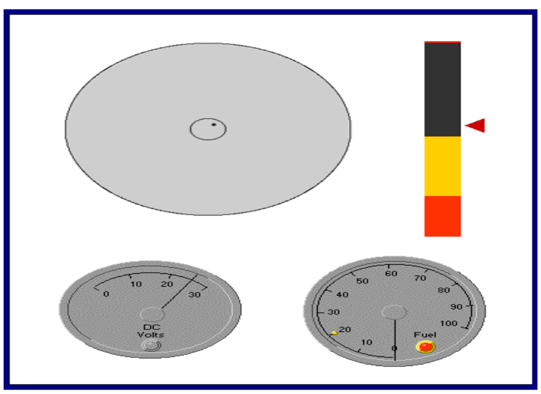

Fig. 5b. The Monitoring and Vigilance Display. The DC Volts Dial is to be kept between 10 and 20 by pressing joystick buttons. The indicator light on fuel gauge (show as red) must be maintained as green and is refueled by holding a button. The Velocity indicator drifts randomly and is kept in black area by foot pedal. The drifting cursor in the oval tracking display is maintained within inner circle by moving the joystick itself.

Testing Procedure

The test battery was divided into five groupings, each administered at a different station by a separate examiner:

1. Central Auditory Processing Tests (PPST and CEST)
2. Auditory Electrophysiology (Audiometry, tympanometry, OAE, BAER, DDT-DP)
3. Visual Electrophysiology (VEPs, CPT)
4. Neuropsychological Experimental Tasks (CANTAB and FI-Self Control)
5. Sensory and Motor Experimental Tasks

Station 1 required only a quiet room. Stations 2-4 each required a separate computer and different peripherals, and Station 5 also had to be capable of total darkness to perform the scotopic vision examination. Testing took place in a suite of rooms at the Golisano Children's Hospital at Strong or in the University of Rochester's General Clinical Research Center.

Administration of the tests at each station required about 45 minutes to one hour, for a total of four and one-half hours on site. All children were tested on all stations on the same day in two sessions separated by lunch. Order of testing stations was randomized.

The study team developed standardized instructions and methodology, based in part upon previous use of various tests with other child populations in Rochester or in the Seychelles pilot study. These instructions and methods were then followed by team members who administered the tests. Quality assurance was provided by investigators responsible for individual tests, although it was not feasible to obtain test administration reliability, i.e., simultaneous double scoring of test protocols by the regular administrator and a "gold standard".

Data Management and Statistical Analysis Plan

Analysis Plan. The primary analysis involved a total of 63 endpoints from the various tests and tasks. The analysis plan involved investigating the capacity of each test or task to independently predict each risk category separately. Primary endpoints represented those variables which experimenters judged to be most likely to predict risk status for one or more of the risk categories. The tests or tasks that demonstrated acceptable predictive capacity were then grouped into a final *battery.* Since the *battery* consists of endpoints chosen based on their independent predictive characteristics, an analysis of the predictive capacity of the battery as a whole based on the same data could incur a large bias. Validation of the battery will require further study and more data.

Distributions of residuals from multiple regression models were assessed using quantile-quantile (Q-Q) plots (Chambers, Cleveland, Kleiner and Tukey, 1983), and standard descriptive statistics, including coefficients of variation. To help establish external validity, tests of similar functions were compared and scores of the risk groups were contrasted. Test scores from all groups for those tests with norms were compared using Q-Q plots and other descriptive measures to scores on the normative samples.

Sensitivity and specificity were obtained from curves estimating receiver operating characteristics (ROC). This methodology has been used since the 1960's as a means for estimating the probability that a human (the receiver) can detect a signal from a background of noise (see Swets, Tanner & Birdsall, 1961; Swets, 1988). It has also been used successfully in studies to establish sensitivity and specificity of psychological endpoints (Etzioni, et al., 1999; Ting, et al., 1997; Tosteson, et al., 1994). The regression analysis of the receiver operating characteristic (ROC) curves examined the effects of covariates on the accuracy of using measured endpoints to predict the status of neonatal risk, learning disability, and IQ (low: <85, high: ≥85). This type of indirect modeling of ROC curves is described in Pepe (1998) and in Zhou, Obuchowski, & McClish (2002). Our analysis involved choosing a threshold c such that a measurement less (or greater, depending on the measurement) than c was classified as *healthy*. For a given value of c, the accuracy of such a decision rule was measured by sensitivity (true-positive rate) and specificity (true-negative rate). An ROC curve is a plot of sensitivity versus 1 - specificity for all possible choices of c. The area under the ROC curve was used as a summary index of overall diagnostic accuracy. An area of 0.5 indicates no discriminating ability, while an area of 1.0 indicates perfect discrimination.

Linear regression models were used to assess the effects of covariates on the slopes of ROC curves. Linear regression was chosen over logistic regression. Logistic regression models allow the assessment of covariate effects on the probability of risk but not on the accuracy of using endpoints to predict the status of each risk status category. Since the latter is the main objective, the chosen regression approach must be such that the endpoint is the dependent variable and the status variable is a covariate. The approach taken in the report satisfies this condition; a logistic regression approach does not.

Fitting a linear regression model with normal errors induces ROC curves of the form:

$$1 - \Phi[\Phi^{-1}(1 - p) - (\mu(1) - \mu(0))/\sigma] \text{ vs. } p,\ 0 \le p \le 1,$$

where $\mu(1)$ ($\mu(0)$) is the predicted response for those with (without) the condition, σ is the residual standard error, Φ (Φ^{-1}) is the standard normal distribution (quantile) function and p is dummy or argument variable. The induced area under the curve is given by $\Phi[(\mu(1) - \mu(0))/(\sigma\sqrt{2})]$. An ROC curve of this form will be affected by a covariate only if it interacts with the status variable (that is, only if its effect on the response differs for those with and without the condition). Therefore, testing for the effect of a covariate on an ROC curve is equivalent to testing for the first-order interaction between the covariate and the risk status variable. The sign and magnitude of estimates of these interactions indicate how a covariate affects the ROC curve. A concave ROC curve indicates that the test performs worse than chance for that subgroup. It can be made convex, e.g., Area Under the Curve (AUC) >0.5, by reversing the direction of the classification rule. Corresponding nonparametric estimates of the ROC curves were concave (AUC<0.5) whenever the parametric ones were also concave. Our analysis assessed the significance of the curve, whether convex or concave.

The ROC curves are empirical in the sense that they are based on data. However, the methods used are parametric, and so some assumptions about the data were made. These assumptions are those that are required for a multiple linear regression model and were verified in each case to ensure that the parametric approach was reasonable. In some cases, nonparametric ROC curves were constructed for various subgroups to confirm the adequacy of the parametric ROC curves.

Each test or task served as the dependent variable in each of three multiple linear regression models (one for each risk status variable). Regression models for each endpoint included the risk status variable, all covariates, and all first-order interactions between covariates and the risk status variable. Beginning with the full model, which included all predetermined main and first-order interaction effects, interactions were eliminated (one at a time) from the model until only significant interactions remained. Significance was defined as $p<0.05$ for a two-sided t-test. All main (non-interaction) effects were kept in the final model regardless of whether they were significant or not. Estimated ROC curves were plotted for all endpoints with discriminating ability significantly better than chance, or with covariates that significantly influenced predictiveness.

Calculations were performed using SAS and S-Plus. The fitting of the regression models and the plotting of the ROC curves was done in S-Plus. As described in the report, the ROC curves depend on several parameter estimates produced by the estimated regression models. In particular, estimates of the main effect of the status variable and any interactions involving it, and an estimate of residual standard error are utilized in the estimated ROC curves.

Covariates. Covariates included age at testing, gender, experience with computer manipulanda and experience with computer games (both ascertained via questionnaire), and hearing status on audiometry and tympanometry. Covariates in all analyses were first checked for multi-colinearity using multiple correlation (variance inflation factors), as well as pair-wise correlations and simple cross-tabulations. No multi-colinearity was indicated. Continuous covariates include age and IQ (except when IQ was the binary status variable). All other covariates were binary, and were coded in the standard "0/1" fashion. Each interaction of covariates was coded as the product of the values of each covariate, and is actually handled by the statistical software. No "re-scaling" was done. The mean age at testing for each level of each predictor and covariate is shown in Table 6.

Table 6
Mean(SD) Age at Testing (Years) of Subjects by Covariate Level

Variable	Yes or High		No or Low	
	Mean	SD	Mean	SD
Neonatal Risk	12.2	1.9	12.0	2.0
IQ	11.8	1.9	12.8	2.0
LD	12.5	2.0	11.9	1.9
Gender (male, female)	12.1	1.9	11.9	2.0
Handedness (left, right)	11.8	2.0	12.1	1.9
Hearing Status	13.0	1.1	12.0	2.0
Experience with Computer Manipulana	12.0	2.0	11.8	2.6
Experience Playing Video Games	12.0	2.0	12.0	2.0

We were unable to adjust for a number of other covariates that might have affected performance, including lack of sleep, diet and eating habits prior to testing, native language other than English, history of head trauma resulting in loss of consciousness, or ingestion of performance altering medications. The frequencies of occurrences were small and did not justify identification of any as separate sources of variance. Educational achievement was not measured. All could have affected results if they introduced bias; but the sample was probably large enough that these variables were probably randomized.

Data Management and Quality Control. Data forms were designed for direct keypunching. Completed forms were stored and sent in groups to the Study Coordinator, who logged the forms and recorded the IDs. The study coordinator reviewed forms individually for completeness and accuracy. The data were entered into an INGRES database after they were keypunched and verified. The database was periodically downloaded into SAS and listings prepared for checking. Range and logic checks were performed and covariates for use in future statistical analyses were checked completely.

Sample Size and Power

Power calculations for a particular study must be based on either a pilot study or some other previous study. This allows one to plan the current study, and it avoids the use of so-called "observed power," which has a one-to-one relationship with the p-value. It is difficult to calculate the power of tests for parameters in an ROC regression model. Since the AUC measures the difference in locations of the distributions of two groups, it may be reasonable to assume that the power calculations actually presented are appropriate for the analyses that were conducted.

The original sample size calculation was based on detection of a 0.33 SD shift on adjusted WISC-III IQ scores, with a mean of 100 and SD of 15. We re-computed power using an effect size estimation strategy developed by Cohen (1988). Cohen's d' is the number of standard deviations separating two group means and can be used as a measure of effect size. This measure is computed as follows:

$$d' = (M1\text{-}M2/SD)$$

where M1 is the mean for Group 1, M2 is the mean for Group 2, and SD is an estimate of the standard deviation taken from the analysis of variance summary table. For example, if the mean of Group 1 were 20 and the mean of Group 2 were 10 and the standard deviation were 5, then d' would be 2. The means are two standard deviations apart. Although there are no generally accepted criteria for determining whether a given d' is large enough to be important, Cohen recommends that d'=0.25 is a small effect, a d'=0.50 is a medium sized effect, and d'=0.75 is a large effect.

To compute effect size, we decided on criteria of p=0.05 and 80% power and the ability to detect a d' of 0.33, close to what Cohen calls a small effect. To reach d'=0.33, the design would require 73 subjects per group for a two-tailed test. This effect size would represent a 5-point IQ difference on the WISC-III. Table 7 shows the results of applying *d'*=0.33 to our endpoints from the fine motor control task, using data from the Seychelles field-test. This task involves three endpoints: Reaction time in msecs, Median frequency in Hz, and Tremor total power, a derived score reflecting displacement per unit time, equal to the average sum of the absolute value of each 4000-msec series and reported in cm/sec.

Table 7
Fine Motor Control Task Data

Measure	Mean	SD	Detectable Difference at 80% Power, *p*=0.05 and *d'*=0.33
Reaction Time (ms)	540	162	53 ms
Total power (cm/sec)	0.22	0.07	0.02 cm/sec
Median frequency (hz)	1.01	0.18	0.06 hz