FYI-1095-1260

American Petroleum Institute
1220 L Street, Northwest
Washington, D.C. 20005

Robert T. Drew, Ph.D.
Director, Health and
Environmental Sciences
(202) 682-8308
(202) 682-8270 (FAX)

Contains No CBI

FYI-95-001260
INIT 10/11/95

84960000002

OPPT Document Processing Center (7407)
ATTN: FYI Coordinator
U.S. EPA
401 M St. SW
Washington DC 20460

October 4, 1995

Dear FYI Coordinator:

In accordance with API's policy of providing the federal government with copies of research designed to determine whether any chemical substance or mixture manufactured, processed or distributed by API member companies may cause a risk of injury to health or the environment, we are enclosing a copy of the following draft final report:

(Identification no: FYI-not assigned) Draft Final Report, Human Neurobehavioral Research Methods: Subject Variables

Please note that this information is provided in accordance with the full disclosure policy of API and does not constitute a formal submission as required by a test rule. This document does not contain confidential information. If you have any questions, please communicate with me.

Sincerely,

Robert T. Drew, Ph.D.

RECEIVED
OPPT CBIC
95 OCT 11 PM 3:23

95 OCT 13 PM 2:08
OPPT NCIC

DRAFT

Contains No CBI

HUMAN NEUROBEHAVIORAL RESEARCH METHODS: SUBJECT VARIABLES

W. Kent Anger, PhD (Principal Investigator)*
O.J. Sizemore, PhD; Julie A. Glasser; Sandra J. Grossmann; Craig A. Kovera
Center for Research on Occupational and Environmental Toxicology
Oregon Health Sciences University (L606)
Portland, OR 97201

Richard Letz, PhD; Crystal Barnwell; Zack Moore; Deb Harris-Abbott
Emory University School of Medicine
Atlanta, GA

Rosemarie Bowler, PhD; Francisco Cuadros; Brigitte Johnson
San Francisco State University (SFSU)
San Francisco, CA

Final Report of a Research Project Conducted for
the American Petroleum Institute (API)

July, 1995

*W. Kent Anger bears exclusive responsibility for study analysis, report writeup, and conclusions, with significant contributions from other OHSU authors; authors from other institutions contributed significantly to the testing.

DRAFT

FOREWARD

[API to supply]

ABSTRACT

Behavioral tests are used to detect and characterize the effects of neurotoxic chemical exposures in human populations. These tests have been used extensively in worksite research, but little attention has been paid to the potentially large influence of subject variables on test performance. This project sought to evaluate the impact of two subject variables, education and cultural group, on widely-used tests of neurotoxic insult. Subjects aged 26-45 were recruited through media advertisements and employment agencies. Behavioral tests from the two consensus neurotoxicity test batteries (established by the World Health Organization and the US Agency for Toxic Substances and Disease Registry) were administered to 715 people with 0-18 years of education. The cultural groups studied were European-descent majority, Native American Indian, African-American, and Latin-American populations. Differences in educational level and locale (rural vs. urban) and gender were examined in the majority population.

Education, cultural group, age and gender all affected the outcome of the behavioral tests studied as revealed by ANOVA, MANOVA and multiple regression techniques. Education followed by cultural group explained the most variance in the tests studied. More importantly, years of educational and cultural group had 13-25% shared variance on the cognitive tests, suggesting that these factors should be controlled in the design of a study rather than attempting to "control" the factors in the statistical analysis. Failure to do so can lead to false conclusions about the presence or absence of neurotoxic effects.

Four critical confounding factors which can mimic neurotoxic effects, or obscure them, in workplace epidemiological investigations are defined by this study for 34 measures drawn from 22 frequently-used tests. Also established are key factors needed to plan and analyze a competent cross-sectional workplace study, statistical power analyses, the distribution for each test, and data transformations for increasing distribution normality.

ACKNOWLEDGEMENTS

THE FOLLOWING INDIVIDUALS ARE RECOGNIZED FOR THEIR CONTRIBUTIONS OF TIME AND EXPERTISE DURING THIS STUDY AND IN THE PREPARATION OF THIS REPORT:

API STAFF CONTACTS

David Mongillo

MEMBERS OF THE API REVIEW STAFF

To be named

DISCLAIMER

This report and its conclusions are the full responsibility of the Principal investigator and do not necessarily represent the views of the other contributors to this research project (named on the title page) or the views of the American Petroleum Institute.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Behavioral tests are used to detect and characterize effects of neurotoxic chemical exposures in people. Three batteries (collections) of behavioral tests which dominate cross-sectional epidemiological research at the worksite and in the community are: (a) the World Health Organization-recommended Neurobehavioral Core Test Battery (NCTB), (b) the Neurobehavioral Evaluation System (NES), and (c) the Agency for Toxic Substances and Disease Registry's (ATSDR) Adult Environmental Neurobehavioral Test Battery (AEnTB). The AENTB includes tests from the NCTB and NES. Both the NCTB and the AENTB are consensus batteries selected by designated experts for the respective organizations (Johnson *et al.*, 1987; Anger *et al.*, 1994; Amler *et al.*, 1994).

The tests in these batteries have been used extensively in field research (summarized in Anger, 1990), but the influence of subject variables on their outcomes has not been systematically examined. This study evaluated the impact of two potentially confounding factors, educational level and cultural background, on the tests in these consensus behavioral test batteries. In addition, information collected on educational locale (rural and urban), subject gender and age in the 26-45 year range provided the basis for investigating the impact of these variables on performance.

METHODS

Subjects aged 26-45 were recruited through media advertisements, employment agencies, and strategically placed flyers. Sixteen widely-used behavioral tests selected from the NCTB, NES, and AENTB were administered to people with 7-18 years of education from four major US cultural groups: European-descent majority (tested in Oregon), Native American Indian (in Oregon), African-American (in Atlanta and Oregon), and Latin-American (Mexican immigrants in San Francisco). The latter group included subjects with 0-6 years of education. Men and women from the majority population, the only group to include females, educated in rural and urban locales were sought to provide variety in the educational background. The tests are divided functionally into sensory, motor, and cognitive performance tests, and measures of affect and vocabulary. The test measures employed in this study are listed in Table 1, column 1.

FINDINGS

Distributions

The distributions of performance scores revealed that several tests are not normally distributed, an important factor considering the pervasive use of "parametric" statistical analytic techniques (e.g., Analysis of Variance or ANOVA) in this field of research. Transformations to increase normality in the distributions had limited impact and were not used in the analysis. Test measures with normal distributions and the most successful transformations to increase normality for the other measures are identified in Table 1, column 2.

Education, Cultural Group, Gender, Age

Education, cultural group (majority US population, Latin-American, African-American, and Native American Indian), gender, and age (in the narrow range of 26-45 years) produced statistically significant effects on performance in most behavioral tests studied. Hierarchical multiple regression was employed to determine which factors accounted for enough variance to have an important impact on behavioral testing. If a factor accounted for more than 5% of the variance associated with a particular measure, it is identified with a • in Table 1. This reveals that education and cultural group were the two most important factors explaining performance on motor and cognitive (including vocabulary) tests. **If ignored, these factors could mimic or obscure neurotoxic effects in population-based studies.** Therefore, worksite or community epidemiological research to detect neurotoxic effects of chemical exposures must adjust for these factors in the experimental design (preferably) or the statistical analyses, or both.

Table 1. Summary of Distribution and Subject Factors.[1]

MEASURE	Dist[2]	Sd/Mn Ratio[3]	Educ	Urban/ rural[5]	Cult Grp	Gend[5]	Age
Sensory							
Acuity/line compl'd.	log	0.23					
Contrast Sens. C	N	0.28					
Contrast Sens. D	N	0.26					
Lanthony Cl/Left [5]	-	0.22			[5]		
Lanthony Cl/Right[5]	-	0.22			[5]		
Vibratron/Index [5]	Sq. Rt.	0.36			[5]		
Vibratron/Small[5]	worse	0.41			[5]		
Motor							
Dynam./Strength [5]	N	0.19			[5]	•	
Dynam./Fatigue [5]	N	0.12			[5]		
SRT/mean RT	worse	0.16	•		•		
NES SRT/mn latency	log	0.16	•		•		
NES Tap/trial 1	N	0.27	•			•	
NES Tap/trial 2	N	0.18			[5]		
NES Tapping/alt [4]			•				
Santa Ana/Pref	N	0.15	•				
Santa Ana/Nonpref	SqRoot	0.19	•		•		
Aiming/correct	N	0.33	•		•		
Pegboard/Right	N	0.14				•	
Pegboard/Left	SqRoot	0.15					
Pegboard/Both	SqRoot	0.15	•		•		
Cognitive							
Benton/correct	SqRoot	0.13	•		•		
Digit Span/Forward	N	0.29	•		•		
Digit Span/Back	N	0.35	•		•		
Digit Symbol	N	0.20	•		•		
NES Sym Dig/best	log	0.26	•		•		
NES SymDig/Recall[5]	N	0.47			[5]		
NES Ser. Digit Learn	N	1.09	•		•		
Raven/errors [5]	N	0.51	•		[5]		
Affect							
NES Mood/Tension[6]	N	0.30					
NES Mood/Depress[6]	N	0.39					
NES Mood/Anger [6]	N	0.40					
NES Mood/Fatigue [6]	N	0.27					
NES Mood/Confus [6]	N	0.32					
Vocabulary							
NES Vocabulary [6]	log	0.28	•		•		
WAIS (scaled) [7]	N	0.25	•		•		[8]
Peabody [7]	N	0.17	•		•		[8]

[1] (•) = Accounts for more than 5% of variance on test performance.
[2] Skew, Kurtosis ≤1 = Normal (N); best transformation listed if not normal.
[3] Ratio from male majority subject data; ratios similar for other populations.
[4] Data from minority subjects only. [5] Data from majority subjects only.
[6] Test not administered to Latins. [7] Latins received Spanish-Language version.
[8] Test adjusted for age.

Performance on some behavioral tests is more variable than on others due to individual human variability. This affects the likelihood that a test will detect small differences between groups exposed to neurotoxic chemicals. The degree of variability in test performance is effectively summarized by the ratio of the standard deviation (variability) to the mean (average) of a group tested. This ratio is identified for each test measure in column 3 of Table 1 from males in the majority population. The <u>larger</u> the number in column 3, the <u>less able</u> is the test to detect small differences between the groups (i.e., for statistical significance).

The means and standard deviations are also the basis for power analyses, a calculation that identifies the number of subjects needed to detect a given effect size for a given level of confidence (i.e., the power). While the confidence level or power is not standardized in science, this report selects a 5% probability level (i.e., the significance level is $p = 0.05$) at 95% power. The selection of effect size is a matter of expectation or speculation. Effect size defines the amount of deficit anticipated or detectable in the exposed population, or the difference between the reference (presumably normal or control) group and the exposed group. A 20% deficit on a test is a substantial loss that may be conceptualized as a reduction in the number of items remembered in a memory test from 100 correct to 80; a 5% loss would be a reduction from 100 to 95. Power analyses ($p=0.05$ at 95% power) have been calculated for each test measure in this study for 5%, 10%, 15%, or 20% deficits (Table 2).

The measures in Table 2 have been ordered by the number of subjects (n) required to detect a significant difference. Tests at the bottom of the Table require a very large number of subjects to detect even large differences between groups and are thus of limited utility in worksite research. When the functions represented by these tests are important to test, nonparametric tests may be better choices for data analysis. From Table 2, the likelihood that a test/measure will detect a difference of a given magnitude in a study can be identified.

Table 2. Number of Subjects Required to Detect Effect Sizes of 5, 10, 15, or 20% Based on Power Analyses (5% significance at 95% power) for Males and Females in the Majority (Western European-derivative) Population.

Females (majority)					Males (majority)				
Measure	5%	10%	15%	20%	Measure	5%	10%	15%	20%
Pegboard/Right	71	18	8	4	Dynam/Fatigue	62	16	7	4
Dyn./Fatigue	80	20	9	5	Benton/correct	72	18	8	4
Santa Ana/Pref	82	20	9	5	Pegboard/Rt	88	22	10	6
Pegboard/Left	82	20	9	5	Pegboard/Left	97	24	11	6
Benton/correct	97	24	11	6	Pegboard/Both	98	25	11	6
Pegboard/Both	107	27	12	7	Santa Ana/Pref	98	25	11	6
Santa Ana/NPrf	110	27	12	7	NES SRT/mean	106	26	12	7
NES Tap/trial 2	122	30	14	8	SRT/Mean RT	113	28	13	7
SRT/mean RT	130	33	14	8	Peabody Vocab	121	30	13	8
NES SRT/mn lat	140	35	16	9	NES Tap/trial 2	141	35	16	9
Peabody Vocab	164	41	18	10	Sant Ana/NPrf	152	38	17	9
Digit Symbol	165	41	18	10	Dyn/Strength	163	41	18	10
Lanthony CI/Lt	185	46	21	12	Digit Symbol	166	41	18	10
Dyn/Strength	207	52	23	13	Lanthony CI/Lt	209	52	23	13
Lanthony CI/Rt	212	53	24	13	Lanthony CI/Rt	214	53	24	13
NES Sym Dig/bst	238	60	26	15	WAIS SCLD	263	66	29	16
WAIS SCLD	312	78	35	19	NES SymDig/bst	283	71	31	18
NES Tap/trial 1	314	79	35	20	NES Tap/trial 1	304	76	34	19
MOOD/Tension	353	88	39	22	MOOD/Fatigue	314	79	35	20
MOOD/Fatigue	354	89	39	22	NES Vocabulary	346	87	38	22
Digit Span Forw	389	97	43	24	Dig Span Forw	365	91	41	23
NES Vocabulary	390	98	[illegible]	24	MOOD/Tension	402	100	45	25
MOOD/Confusion	421	105	47	26	MOOD/Confus.	447	112	50	28
Digit Span/Back	528	132	59	33	Digit Span/Bck	525	131	58	33
MOOD/Anger	605	151	67	38	Vibr./Index	569	142	63	36
Vibr./Index	723	181	80	45	MOOD/Depress	650	162	72	41
MOOD/Depress.	728	182	81	46	MOOD/Anger	681	170	76	43
Vibr./Small	808	202	90	50	Vibr./Small	717	179	80	45
Sym Dig Recall	1379	345	153	86	Sym Dig Recall	947	237	10[illegible]	59
Raven/errors	1523	381	169	95	Raven/errors	1123	281	125	70
NES Ser Dig Lrn	5223	1306	580	326	NES Ser Dig Lrn	[illegible]	1288	573	322

The power calculations in Table 2 are based on US majority (Western European-derivative) subjects, although substantial differences can be expected given the cultural group findings seen here. When the same power analyses were applied to the Latin Americans tested in this study, most tests required more subjects to detect a given difference at a given power. Due to the differences in education between the majority and the Latin subjects in this study, the most appropriate comparisons are between subjects with 7 or more years of education. When the number of subjects required for significance (comparable to the N in the right columns of Table 2) was averaged across 17 measures from 12 tests, Latin Americans with 7-9 years of education required a mean of 21 more subjects per study (assuming the study involved only one exposed and one reference group), and Latin subjects with 10 or more years of education required a mean of 9 additional subjects per study. This simply reflects the greater individual variability or spread in test performance in the Latin-American subjects in this study.

Finally, it is important to note that behavioral studies almost never use a single measure of behavior. Adjustment of the significance level (alpha) or selection of an appropriate statistic is required to avoid the problem of falsely detecting differences which arise from multiple comparisons (sometimes called alpha inflation). One conservative approach, the Bonferroni correction, divides the level of significance equally across each behavioral measure. Thus, in a study with 10 comparisons (from 10 test measures) of the same subjects, this correction would require differences at the 0.005 level before accepting a difference as statistically significant (0.005 = 0.05/10).

<u>Questionnaire</u>
Analyses of answers to questions about previous institutionalization for drug or alcohol abuse and prior symptoms of numbness and tingling revealed notable performance impacts on many tests. The abuse findings were almost exclusively due to results from majority subjects. Conversely, the subjects answering yes to the question

"have you consumed more than one glass of any form of alcohol or any drug in the past 48 hours" performed about as well as those who answered no to the question. Previous institutionalization and evidence of numbness and tingling thus should be accounted for in a complete analysis of neurotoxicity studies.

CONCLUSIONS

Several conclusions can be drawn and recommendations offered:

- The influence of education and cultural group on cognitive test performance is so intertwined that they cannot be controlled statistically. If different cultural groups and a range of educational attainment can be anticipated in an epidemiologic study, at least one of these factors must be balanced in the experimental design and subject recruitment if cognitive tests are planned, although this is less important for motor, sensory, or affect tests. Inclusion of both education and cultural group in the data analysis is essential for obtaining accurate results.

- Gender had an impact on many tests studied here, although the variance accounted for by this factor was small except in the case of strength testing where male and female data must be treated separately. Gender balance in study designs and inclusion as a factor in the statistical analyses is essential if adverse effects are to be evaluated in both males and females.

- Age through the range of 26-45 years has a detectable impact on several behavioral tests. This factor should be included in subject recruitment or sampling plans and in data analyses, although the amount of variance accounted for by this factor was small.

- Many of the tests included here do not have normal distributions. A search for other established tests of the same functions but which have improved psychometric properties including more normal distributions, is recommended.

• For neurotoxicity research, tests with a standard deviation to mean ratio greater than 0.5 should not be used (including the Raven, NES Serial Digit Learning and NES Symbol Digit Recall), or improved protocols are needed.

• Subject reports of numbness and tingling were associated with performance declines on a number of tests. This suggests an underlying neurologic factor that has not been identified here. This and other questions in the NES pre-test questionnaire need careful reformulation to provide interpretable answers.

• People reporting previous institutionalization for drug or alcohol abuse may have degraded performance on some neurobehavioral tests, and this should be tracked by research protocols.

Section 1
INTRODUCTION

Through the 1940s, adverse effects of workplace chemicals on the nervous system were discovered in workers by alert clinicians or were identified through animal research conducted by industry. In the 1960s and 1970s, as exposure concentrations in industry steadily declined, objective behavioral tests were employed to identify early-stage or low-concentration adverse neurotoxic effects that were not obvious to the clinician. These non-invasive methods measured a broad range of motor, sensory, and cognitive capabilities in humans (Anger, 1990; Anger and Johnson, 1992).

DEVELOPMENT OF STANDARDIZED BATTERIES

From the mid-1960s to the early 1980s, some 60 different behavioral test methods were used in approximately 50 worksite epidemiological studies (primarily of lead-, mercury-, and carbon disulfide-exposed workers; Johnson and Anger, 1983). By the end of the decade, the number of unique behavioral tests had expanded to 250 and the number of cross-sectional studies had grown to 185 (Anger, 1990). The continued growth in the number of behavioral methods employed in this research has spawned a growing interest in developing standardized tests for this field.

Two major human test batteries were developed in the 1980s to investigate neurotoxic chemicals in worksite research. The first was developed in a 1983 meeting of field investigators using behavioral and neurological tests sponsored by the World Health Organization (WHO) and the US National Institute for Occupational Safety and Health (NIOSH). The assembled field investigators recommended seven field-proven behavioral tests with established sensitivity to neurotoxic chemicals. They named these tests the Neurobehavioral Core Test Battery (thus, the WHO NCTB), creating the first consensus behavioral test battery to assess neurotoxic chemicals. The goal was to build a data base of the neurotoxic effects of chemicals by using standardized tests (Johnson et al., 1987). This remains an unattained goal.

The second major battery, developed by Baker and Letz (Baker et al., 1985), implemented on a personal computer 22 behavioral tests used to evaluate nervous system function. The battery included adaptations of five of the seven WHO NCTB tests. The Baker and Letz battery of tests was named the Neurobehavioral Evaluation System, or NES (Letz, 1990). The major advantages of the NES are that it can be readily administered in a more reliable and efficient manner than the NCTB, which is administered by a trained examiner using "pencil and paper" tests. The major advantages of the NCTB are the inclusion of field-proven motor tests (not in the NES), economy of instrumentation, the ability to use it in non-industrialized countries (computers require electricity and service), and a greater potential for testing poorly educated subjects (the NES has screen-printed instructions while the NCTB is administered orally).

Pressure to evaluate hazardous waste sites has led to the development of another test battery, in this case to study US community groups near hazardous waste sites. The Agency for Toxic Substances and Disease Registry (ATSDR) convened a meeting in 1991 to propose such a test battery. The resultant battery includes tests from the NCTB and the NES, as well as other tests not in these batteries (Anger et al., 1994). ATSDR subsequently selected all but one of the tests proposed by the expert panel, naming it the Adult Environmental Neurobehavioral Test Battery (AENTB) (Hutchinson et al., 1992; Amler et al., 1994). This represents the second consensus test battery developed to assess in humans the effects of neurotoxic chemicals.

Tests from both the NES and NCTB have been widely used (Letz, 1990; Cassitto et al., 1990; Liang et al., 1990; Anger et al., 1991, 1993). The NES is the most extensively used computerized battery (Letz, 1990; Anger, 1990); the NCTB has the largest base of published control data from US subjects (Anger et al., 1993) and consists of those behavioral tests which have most consistently identified neurotoxic effects in worksite settings (Anger, 1990).

FIELD ASSESSMENTS OF NES, NCTB, AENTB

Baker, Letz, and others have used various NES tests (there is no standard configuration or set of NES tests) in diverse settings to study impaired or potentially impaired people. The on-screen instructions for NES tests have been translated into several languages and employed widely in industrialized countries, particularly the US and in Europe (Letz, 1990). The WHO NCTB tests continue to be among the most extensively used in non-computerized neurotoxicity assessments, although the NCTB has not been frequently employed as a battery per se. Recently, the AENTB has been used by ATSDR in three field studies of people living adjacent to hazardous waste sites (Amler et al., 1994).

Under the sponsorship of WHO and national occupational health institutes, a study to investigate the feasibility of using the NCTB in a wide range of cultural groups and languages was undertaken with the expectation that significant baseline data would be developed in those countries, including the US. Data from the NCTB feasibility study were collected from 2300 subjects in 10 countries distributed across three continents. The largest subject base was studied in the US, where over 900 male and female subjects between 16 and 65 years of age have been tested. Performance on some tests (Simple Reaction Time and Benton) was highly consistent across countries, while the other NCTB tests were less consistent across countries (Anger et al., 1993).

The performance data from one country, Nicaragua, was substantially inferior to the performance in other countries on all tests except the Santa Ana test of dexterity (requiring the subjects to manipulate pegs). Correlational analyses and anecdotal observations suggested that limited education was a significant factor in the Nicaraguan performance. While subjects from other countries had at least 8 years of education (13-15 years in the US) and lived in urban settings, the Nicaraguan subjects had a mean of only 3 years of education and lived in a rural setting (Anger et al., 1993). The relatively low level of education in many developing countries constrains the use of the NCTB for many potential test subjects. It could not be overlooked, however, that Nicaraguan subjects were also the only people from a Latin culture, raising questions about the feasibility of using the NCTB in Latin populations.

These questions are very relevant to the increasingly multi-national US industry. Sensitive medical tests are needed to detect the early stages of neurotoxic disorders prior to the development of serious, irreversible health effects, and for routine monitoring where there are demonstrated concerns. Tests with norms are needed to address the more difficult problem of dtermining that a medical problem doesn't exist in an individual, although this is exceedingly difficult where baseline data do not exist. A major limitation of the tests used in Behavioral Neurotoxicology is that they do not have norms or even a large normative database relevant to the working US population to: (a) use in lieu of unexposed reference groups; or, (b) assess the adequacy of performance in a reference group.

Interpreting the results of behavioral tests requires an understanding of the potential confounding factors that can affect their outcome. It is recognized that subject gender, education, age, and cultural background can confound or modify a study of neurotoxicity that employs behavioral tests, but the magnitude and potential interactive nature of these variables in the general population is not known.

PROJECT GOALS

This project assessed the impact of years of education and cultural group on performance on behavioral tests used inneurotoxicology research. In the majority population, possible performance differences resulting from urban vs. rural educational settings, and gender were also investigated. In addition, the influence of age in the narrow range of 26-45 years was analyzed. Finally, information was provided on the nature of the underlying distributions of each test and transformations to improve their normality when relevant.

The goal of this project was to provide a quantitative basis for understanding the relative impact of two potential confounding factors, education and cultural group, that can mimic or obscure neurotoxicity. The impact of these and other variables are described, providing the basis for conducting more definitive and interpretable studies of neurotoxic disorders.

Section 2
METHODS

SUBJECTS

This project sought subjects from four different cultural groups in the United States. These were male Mexican immigrants to California, male African Americans in Georgia and Oregon, male Native American Indians in Oregon, and male and female European-descent (i.e., US majority population) subjects in Oregon. Those with from 0-12 years of completed education were sought in each group, although individuals with more than 12 years of education were tested and are included in the results. Majority subjects with either rural or urban educational backgrounds were specifically targeted by testing in both rural and urban locations. Educational backgrounds were originally defined as urban (population of 50,000 or more) or rural (less than 50,000) on the basis of US census data from the decade in which each subject spent the majority of their school years.

Subjects were recruited in Atlanta, San Francisco, and in several cities in Oregon. Oregon subjects were recruited primarily by newspaper advertisements and posted flyers in Portland, Enterprise, Roseburg, Carver, Salem, and Springfield. African American subjects were primarily recruited in Atlanta through employment agencies, and Latin-Americans were recruited in San Francisco through employment agencies. Recruitment details are described in Appendix A, along with frequently-reported occupations by subjects and Native American Indian tribes represented in the study. A total of 715 people who were tested met the criteria for inclusion in the study (Table 3).

Table 3. Distribution of Tested Subjects by Years of Education and Cultural Group.

Cultural Group (state*)	Education (years) 0-5	6-9	10-12	13-16	Totals
Latin American male (CA)	33	81	22	5	141
African American male (GA/OR)	0	52	58	-	110
Native American male (OR)	0	10	36	6	52
Majority male (OR)	0	24	91	79	194
Majority female (OR)	0	26	105	87	218
Overall Totals:	33	193	312	177	715

* CA = California; GA = Georgia; OR = Oregon.

Subject statements as to their age, years of education and the location of their educational experience were accepted as accurate, although an identification card (e.g., driver's license) was requested and used to verify age (mismatches were rare) when available.

BEHAVIORAL TESTS

The project had two phases (funded several months apart) that began at different times but continued contemporaneously to the study's conclusion. The first phase was directed at testing minority (Latin-American, African-American, Native American Indian) populations; the second phase was directed at the majority (European-descent) population.

Eighteen tests were administered to minority subjects in this study (the untranslated NES vocabulary and Mood tests were not administered to the Latin-American subjects who spoke Spanish). Four additional tests from the ATSDR AENTB (ATSDR, 1992; Amler et al., 1994) that had not been finalized at the start of Phase I were administered to European-descent subjects. Table 4 lists the tests, functions they assess, and the test batteries in which they are included. The individual tests and the measures (dependent variables) analyzed from each test are described in Appendix B.

Table 4. Behavioral Tests by Battery from which the Tests Were Drawn.

Test	Function Assessed	AENTB	Nature of Data
NCTB			
Simple Reaction Time (SRT)	Response speed	Yes (altered)	Mean reaction time
Digit Span	Attention, memory		Number of spans recalled (forward, backward)
Santa Ana	Coordination, speed	Yes	Pegs turned by preferred, nonpreferred hand
Digit Symbol	Coding		Symbols completed
Benton Visual Retention	Visual memory		Correct choices
Aiming	Coordination, speed		Dots in circles (correct responses)
NES2 (selected)			
Serial Digit Learning	Learning, memory	Yes	Errors (NES2 formula)
Symbol Digit	Coding	Yes	Trial latencies
Symbol Digit Recall	Memory	Yes	Number recalled
SRT (extended)	Speed, attention	Yes	Mean reaction time
Tapping	Coordination	Yes	2 (majority) or 4 trials of number of taps
Mood Scale[2]	Affect	Yes	Scales: Tension, depression, anger, fatigue, confusion
Vocabulary[2]		Yes	Correct definitions
Other			
Peabody[3]	Vocabulary		Scaled score/correct
WAIS vocabulary[3]	Vocabulary		Scaled score/correct
Grooved Pegboard (Purdue)	Speed, coordination		Pegs inserted (right, left, both hands)
Dynamometer[1]	Strength, Fatigue	Yes	Mean of trials 1 to 3; trial 1 minus trial 5/trial 1
Contrast Sensitivity	Visual acuity	Yes	Correct choices C, D
Acuity	Visual acuity	Yes	Smallest all correct
Lanthony D-15[1]	Color vision	Yes	Errors; norm/abnorm
Vibratron[1]	Vibration sensitivity	Yes	Index, small finger threshold
Raven[1]	Logical reasoning	Yes	Correct choices; test time

[1] Tests given only to majority subjects in Oregon.
[2] Test not administered to Spanish-speaking Latin subjects in California.
[3] Spanish language version administered to Latin subjects.

0028

PROCEDURES

The Oregon Health Sciences University Human Research Committee approved the human subjects consent form and advertisement used in this study. A training manual developed for this study was the basis for training all Examiners who administered the tests. Protocols for the NCTB tests followed the NCTB Operational Guide, NES2 tests followed the protocols in the NES2 manual, and AENTB tests followed the protocols in the ATSDR AENTB Examiner's manual (Anger and Sizemore, 1993). Protocols for the remaining tests were developed with a neuropsychologist and pilot-tested in varying numbers of people. The battery was pilot-tested to assess subject acceptability, timing, and logistics; minor modifications were made to adjust for these factors prior to subject testing. The order of test presentation is listed in Table 5 for both phases.

Table 5. Order of test presentation in Phases I (Minority) and II (Majority).

Phase I (Minority)	Phase II (Majority)
Acuity (Optec)	Acuity (Optec)
Contrast Sensitivity (Optec)	Contrast Sensitivity (Optec)
NCTB SRT	Lanthony D-15 Color Vision (II only)
NCTB Digit Span	Vibrotactile Threshold (II only)
NCTB Santa Ana pegboard	NCTB SRT
NCTB Digit Symbol	NCTB Digit Span
NCTB Benton	NCTB Santa Ana
NCTB Aiming	NCTB Digit Symbol
Grooved Pegboard	NCTB Benton
NES2 Tapping (1 hand & alternating)	NCTB Aiming
NES2 Simple Reaction Time (SRT)	Grooved pegboard
NES2 Symbol Digit	Dynamometer (II only)
NES2 Symbol Digit Delayed Recall	Raven Progressive Matrices (II only)
NES2 Serial Digit Learning	NES2 Symbol-Digit
NES2 Mood Scale [1]	NES2 Tapping Test (one hand)
NES2 Vocabulary [1]	NES2 Simple Reaction Time (SRT)
WAIS-R Vocabulary Test [2]	NES2 Serial Digit Learning
Peabody Picture Vocabulary Test [2]	NES2 Symbol Digit Delayed Recall
	NES2 Vocabulary
	NES2 Mood Scale
[1] Not administed to Latin subjects	WAIS R Vocabulary Test
[2] Spanish version for Latin subjects	Peabody Picture Vocabulary Test

Test instructions placed in a looseleaf notebook were read to each subject after the subject signed the consent form. Instructions to all

Latin subjects were given in Spanish using the same instructions used in a previous study of Nicaraguan subjects (Anger et al., 1993). Instructions to all other subjects were given in English.

In Oregon, target towns or areas were visited to identify test locations and determine the best newspapers or alternative methods for advertising the study. A toll-free phone number was installed for subjects to schedule testing. Subjects were paid $30 for completing the tests.

EXAMINERS

A Master's degree candidate (FC) in Counseling at San Francisco State University (SFSU) and a Psychology undergraduate (BJ) at SFSU administered tests in Spanish to all subjects tested in California. Both were fluent in Spanish. To assure the fullest understanding of subject responses, the vocabulary tests were administered by the Examiner with the most extensive experience speaking Spanish (FC).

Two post-baccalaureate students (CB, ZM) at Emory University School of Medicine administered the tests in Atlanta to African American Subjects. One (CB) was an African American; the other was of European descent. Each administered half the test battery to any subject, alternating the tests administered regularly. An additional Examiner of European descent (DH-A) infrequently administered the entire battery in Atlanta.

One Examiner (JAG) administered the tests to 69% of the Oregon subjects. Two additional Examiners (SJG, CAK) tested the remaining Oregon subjects. Each Examiner had a Master's degree in psychology or equivalent coursework (degree pending).

00388

DIVERGENCE FROM PROPOSAL/CONTRACT

One test identified in the proposal was not administered as planned. The Rey Auditory Verbal Learning Test was not accepted by ATSDR as part of their AENTB, so it was eliminated from this study, as well. Three educational levels were originally sought (Table 6). They were 0-5 years; 6-9 years; and 10-12 years of education. As seen in Table 7, the cells for 0-5 years of education were only partially filled despite extensive recruitment efforts. However, the Wechsler Adult Intelligence Scale (WAIS) Vocabulary test revealed a broad spread of results suggesting a broad range of intelligence, thus approximating the intent of the study. The other incomplete cells were those of the Native American Indian population which could not be accessed directly on the reservations due to tribal decisions, despite contacts with all major tribes in Oregon (e.g., see Appendix A for a description of the 24 separate approaches made to representatives of Native American groups during recruitment).

Table 6. Subject Distribution Sought in Study.

Cultural Group	Gender	Years of Education 0-5	6-9	10-12	13-16	Totals
Latin American	male	50	50	50		150
African American	male	50	50	50		150
Native American Indian	male	50	50	50		150
Majority/urban	male	35	35	35	35	140
Majority/urban	female	35	35	35	35	140
Majority/rural	male		35	35		70
Majority/rural	female		35	35		70
Totals		220	290	290	70	870

Table 7. Distribution of Subjects Tested and Included in Analysis.

Cultural Groups	Gender	Years of Education 0-5	6-9	10-12	13-16	Percent of Plan
Latin American	male	33	81	22	5	94%
African American	male	0	52	58	-	73%
Native American Indian	male	0	10	36	6	35%
Majority/urban	male	0	9	45	42	69%
Majority/urban	female	0	16	46	34	69%
Majority/rural	male	-	15	41	35	130%
Majority/rural	female	-	12	55	45	160%
Totals		33	195	303	167	80%

Section 3
RESULTS

DISTRIBUTION OF EXAMINERS

Examiners can influence the results of behavioral tests, especially if they have a role in controlling the pace of testing. In this study, the tests judged most easily affected by variability in test administration are the WAIS vocabulary, Dynamometer grip strength and fatigue, Raven Progressive Matrices, Vibratron (index and small fingers), Digit Span and Santa Ana tests.

Subject test performance led to the discovery of differences in the implementation of the protocol between Oregon Examiners on the WAIS vocabulary test (one Examiner likely wrote down more responses and may have allotted the subject more time for responses on this test). Since Oregon Examiners CAK and SJG contributed unsystematically to the total N in the study (Figure 1), no data were excluded on the basis of potential Examiner differences.

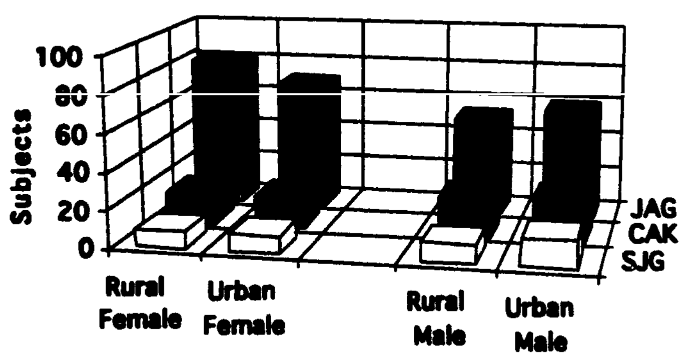


Figure 1. Number of majority male and female subjects educated in urban and rural schools tested by the three Oregon Examiners (JAG, CAK, SJG).

SUBJECT DEMOGRAPHICS
The number of subjects by cultural group is depicted in Figure 2. Subjects tended to fall close to levels associated with mandated educational requirements for basic education or a degree program.

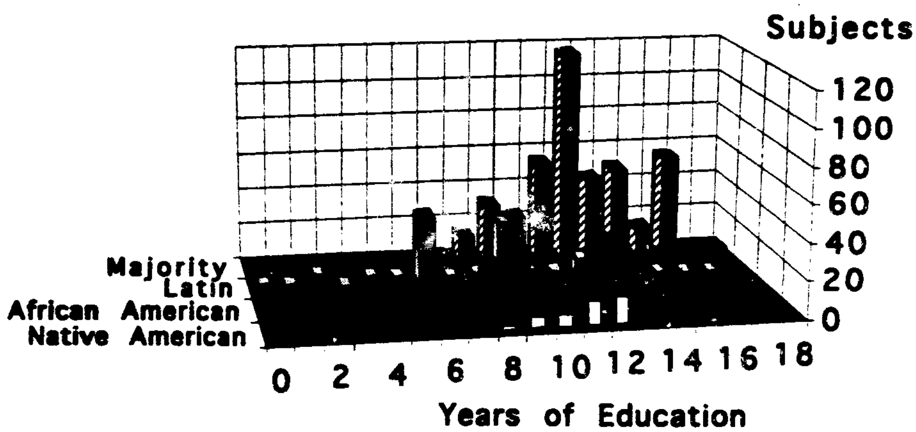


Figure 2. Distribution of years of education, by grade, in each cultural group.

A complete summary of the subject distributions (mean [and sd] age and years of education by cultural group and gender) is presented in Table 8.

Table 8. Number of Subjects, Mean (and SD) Age and Years of Education of Latin American, African American, Native American Indian, and Majority (Rural/Urban) Subject Groups, by Gender and Education subgroups.

Cultural Group	Gend	Educ. Range	N[1]	Age Mean	Age SD	Educ. Mean	Educ. SD
Latin American	M	0-18	146[2]	29.8	3.3	7.2	3.5
African American	M	0-12	110	32.0	2.9	9.9	1.8
American Indian	M	0-18	52	31.8	3.8	11.1	1.7
Majority/urban	M	0-18	91	36.2	5.7	12.4	2.2
Majority/urban	F	0-18	93	34.8	5.6	12.2	2.4
Majority/rural	M	0-18	74	35.8	5.3	12.3	2.3
Majority/rural	F	0-18	101	34.4	5.5	12.3	2.3
Detail							
Latin American	M	0-5	33	29.7	3.1	2.6	1.7
	M	6-9	48	29.6	3.2	7.4	1.4
	M	10-12	44	30.2	3.6	11.2	0.9
	M	13-18	17	32.2	3.7	14.8	1.6
African American	M	6-9	52	31.8	3.2	8.2	0.9
	M	10-12	57	32.2	2.6	11.3	0.8
	M	13-18	0				
American Indian	M	6-9	10	31.4	3.5	8.7	0.5
	M	10-12	32	31.1	3.3	11.2	0.8
	M	13-18	6	30.7	3.1	14.2	1.2
Majority/urban	M	6-9	9	36.4	5.8	8.6	0.53
	M	10-12	44	35.7	5.3	11.4	0.7
	M	13-18	38	36.7	6.1	14.5	1.3
Majority/urban	F	6-9	16	34.2	4.5	8.5	0.9
	F	10-12	42	35.1	6.0	11.4	0.6
	F	13-18	35	34.7	5.7	14.7	1.2
Majority/rural	M	6-9	12	38.3	4.4	8.6	0.51
	M	10-12	32	35.9	4.8	11.6	0.6
	M	13-18	30	34.6	5.9	14.4	1.4
Majority/rural	F	6-9	11	34.9	6.4	8.3	1.0
	F	10-12	51	34.5	5.4	11.4	0.8
	F	13-18	39	34.1	5.5	14.5	1.3

[1] This table includes data from subjects with education up to 18 years, including subjects excluded from the previous table.

[2] Age missing for 4 subjects.

ANALYTIC STRATEGY

Research in non-occupational populations has demonstrated the impact of education, gender, age, and cultural group on the behavioral tests used in neurotoxicology research (e.g., Lezak, 1995). The goal of this study and thus the analysis was to identify factors which influence test performance to an extent that would jeopardize conclusions about alternative, neurotoxic exposures. Therefore, traditional parametric statistical tests are employed across the behavioral measures without adjustments for the large number of multiple comparisons. The quantitative analyses described below reveal a wide array of statistically significant findings; statistically significant differences in this study did arise from very small actual performance differences between the groups due to the large number of subjects tested.

To convey the magnitude of the effects reported and their importance, additional analyses were conducted. A deductive analytic approach focused on effect consistency or replication was employed to convey the generality and the magnitude of the various results. Multiple hierarchical regression was employed to convey the importance of the individual findings. The order in which the analyses are presented is summarized next.

Distributions

The analysis begins with a depiction of the distributions of performance on each test and an evaluation of the degree of normality of the distribution. That the populations being studied are normally distributed is an assumption underlying the use of standard parametric techniques such as ANOVA, and violation of this assumption is thought to affect the outcome of the analyses in unpredictable ways. Several data transformations were applied to non-normal distributions (defined by kurtosis or skew exceeding 1.0), and the transformation which most improved normality is identified. The transformed data introduced only modest improvements in normality (as judged by changes in skew and kurtosis) and therefore were not used in the analyses reported here.

Overall Statistical Comparisons

An initial Multivariate ANOVA (MANOVA) was conducted with data from all subjects to determine if there were overall effects of the factors studied in this research. Each was significant by Wilks' Lambda (Table 9, top). A probability (P) of 0.05 or less was accepted as significant here and throughout the analyses. The MANOVA probabilities were adjusted for multiple comparisons (the factors). The substantial evidence of significance of the overall factors justified in-depth analyses of the individual factors, described in the remainder of the report.

Table 9. MANOVA Table for Main Study Factors.

Factor	F	Num, Den DF	P(robability)
Educational level	4.237	48, 978	0.0001
Cultural Group	5.232	48, 978	0.0001
Gender	7.589	24, 489	0.0001
Age level	2.969	24, 489	0.0001

Individual analyses were conducted on the factors of gender (majority subjects only), age (grouped for consistency with Anger et al.'s 1993 research in 10-year age ranges of 26-35 and 36-46), cultural group (males only), and education. To be included in the urban vs. rural education comparisons, 75% of total non-college education was required to be in a single city (to establish the population size). The definition of "rural" and "urban" proved complex and to a degree elusive. Only those subjects who spent 75% of their pre-college school in one of three categories of city size were eligible for the statistical analysis described above. These were subjects: (a) who went to school in cities with a population of less than 16,000 (using the US census figures from the decade in which they spent the largest portion of their time in school); (b) 16,000-349,999; and, (c) 350,000 and above. This division was selected to capture rural, suburban, and urban community distinctions and coincidentally yielded 135 (in a), 132 (b), and 131 (c) majority subjects in the respective groups. Other subjects were excluded from this analysis because they were educated in two or more cities from different population groupings.

Importance of Factors: Multiple Regression

Hierarchical multiple regression techniques were used to estimate the amount of variance accounted for by education, cultural group, gender, and age, test measure by test measure. Each hierarchical regression was begun by determining the percent of variance accounted for by a particular covariate (e.g., age), followed by the addition of a second factor (e.g., educational level) and recalculation of the equation. This series was repeated to identify unique and shared variance for each pair of factors. The unique variance identifies the degree to which a given factor can account for performance when the other factor is controlled. Shared variance describes factors which are linked and co-vary. When the percent variance explained (unique or shared) exceeds 10% it can be considered an important factor to be controlled in the design or analysis, and when it exceeds 25% it can be considered critical for evaluating test performance in neurotoxicity studies. When there is shared variance (variance which is common to two factors), the analysis in neurotoxicity studies becomes exceedingly complex, suggesting that one of the factors should be separated out into different analyses (virtually into different experiments).

The regressions were performed with all pairs of factors (education years with cultural group, with gender, with age, with urban/rural background; gender with age, with rural/urban background; and cultural group with age). Analyses of gender with other factors and age with rural/urban background revealed no shared variance and little unique variance. Therefore, these data are not reported by factor.

Consistency (Replication)

Next, effect consistency or within-study replication was addressed by inspection of mean test performance differences across common variables in different population groupings. For example, an effect of education should be seen in all cultural groups, or at least in cultural groups from similar educational systems. Differences between rural and urban education should be seen in both males and females. A difference between one cultural group and others should be seen in group members with different levels of educational attainment.

Effect consistency as employed here imposes the stringent requirement that differences must be seen across all or most comparisons, whereas a single comparison, if the effect size is very large, can produce significance in a purely statistical analysis. Consistency builds a high degree of confidence in the generality of conclusions as it reveals replications of findings in different sub-groups. Of course, since different factors can interact to cancel or enhance each other, complex relationships are not revealed by this type of analysis and must be detected statistically.

Data from the dynamometer strength test are illustrative. Predictably, males have greater grip strength in the dynamometer pull (more kilograms force) than females at each age range (Figure 3). Education would not be expected to play a significant role in grip strength, and indeed the means are similar across years of education (front to back on the graph) for both males and females. There are only minor differences between the two age ranges (26-35, 36-45) presented in Figure 3 (viewed against the back wall of the graph), as might be expected in this narrow age range. Thus, the only consistent difference of any magnitude seen in the graph is the strength difference between males and females. This is a finding of general significance. Men have greater grip strength than women regardless of education (and the host of variables correlated with this factor) or age (at least between 26 and 45 years).

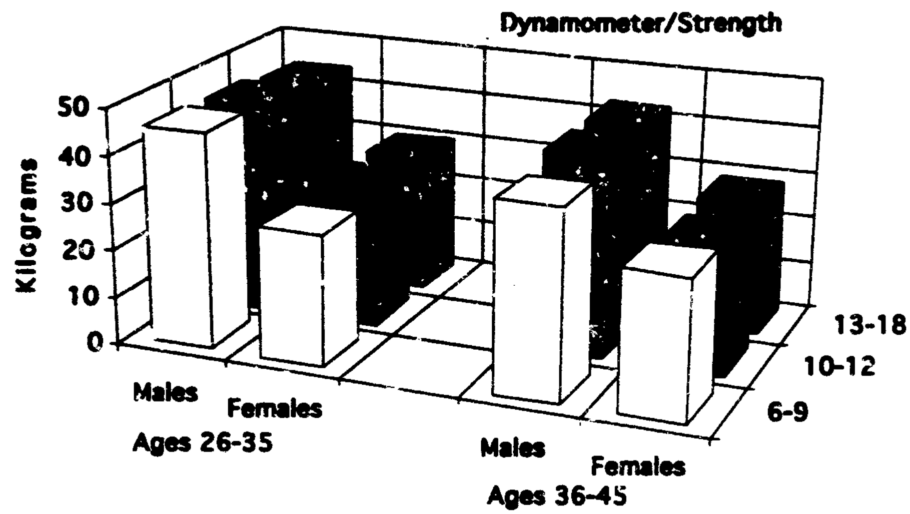


Figure 3. Male and female strength measurements (kilograms of force) in ages 26-35 and 36-45 across 3 educational ranges (6-9, 10-12, and 13-18 years of education).

Magnitude of the Effects

The magnitude of an effect can be conveyed by comparing group differences (i.e., mean performance differences) to the standard deviation (SD) of the differences. This is a convenient method of comparison that accurately parallels the results of simple ANOVA statistical techniques such as used here. A group difference in the range of 0.5-0.75 SD units is typically statistically significant with the size N tested here, and a difference of 1 SD unit or more will almost always be significant with an N in the range of 30-50 per group (a larger N is needed when multiple measures are involved).

This method of comparison can be illustrated with the data in Figure 3. It reveals that females have a mean dynamometer pull of approximately 27 Kilograms (kg), while males have a pull of 42-43 kg. Since the SD for performance on this test is approximately 6 for females and 8 for males (Tables 30 and 31, respectively), the difference between 27 and 43 is 16 kg or 2-2.5 SD units. This indicates a huge difference that is highly statistically significant. It is the largest difference (i.e., magnitude) encountered in the results of this study.

Correlations Between Tests

Performance on some tests correlates highly with performance on other tests, suggesting the possibility that they may be testing a single underlying factor. This can be an issue in data analysis where the analysis of related measures must be adjusted for multiple comparisons. These relationships are thus decribed at the end of each section.

Background Factors and Exclusions

Subjects were not excluded from the data analysis except for procedural errors, which were few in number. Due to their impact on the nervous system, some diseases and substance abuse can affect performance on the behavioral tests used in this study. During testing, each subject was asked a series of questions to obtain information on these background factors. They took the form of "has a medical doctor ever told you that you had a disease named . . . " or "have you ever been placed in an institution for chronic drug or alcohol abuse?" The basic thrust of each question is reflected in column 1 of Table 10; the percent of respondents answering yes, by educational grouping, is in the other columns.

Table 10. Number of Subjects and Percent Responses to Questions on Diseases, Abuse, Recent Drug/Alcohol Consumption, by Educational Categories.

Medical Problem (%)	Years of Education 0-5	6-9	10-12	13 and up
(Number of Subjects)	(33)	(193)	(312)	(181)
Carpal Tunnel	0	0	2	3
Tingling/Numbness (even once)	15	26	26	26
Lupus	0	0.5	0	0
Diabetes	0	0	0.6	0
Thyroid disease	0	0	1	0.6
Pernicious Anemia	0	0	0.3	0
Peripheral Neuropathy	0	0	0	0
Substance Abuse (Institutionalized)	9	17	13	4
1 drink/drug last 48 hours	12	23	20	22
Use of corrective lenses	9	19	35	52

There are clearly disproportional differences in symptom and abuse reports among the educational groups shown in Table 10. Similar distributional differences were also seen in the cultural groups studied. The effects of these variables were extensively explored; the results are described in Appendix D. Statistically significant differences were found in some subject subgroups reporting previous substance abuse or numbness or tingling. The performance differences seen on most tests were small (due to a small group of subjects in the majority male population), most lacked statistical significance (Appendix D), and they were judged to contribute minimally to error variance.

Subjects reporting medical problems, symptoms of numbness or tingling, or previous substance abuse were not excluded or otherwise controlled in the following statistical analyses. Nonetheless, the differences found in this study (Appendix D) confirm the need to query the existence of serious substance abuse problems to determine if the groups in an epidemiologic study are equal with respect to this factor for possible exclusion from the analysis.

Organization of Results

The following data presentation is divided into psychological domains of sensory, motor, cognitive, and affective tests. Vocabulary, a subdomain of cognitive function is addressed last due to its special

0041

status as a measure of intelligence and education. The purpose of this organization of the data presentation is to facilitate digestion by the reader.

SENSORY TESTS

The sensory tests assess visual acuity (Acuity, Contrast Sensitivity C and D), color vision (Lanthony), and vibration threshold (Vibratron index [finger] and small [finger]). Acuity was typically determined with the viewer wearing needed corrective lenses (if available) and was intended primarily to verify that the subjects could see the tests and the NES instructions on the computer screen. Less than 10 of the 715 subjects tested had scores below 3 on the acuity test, indicating that virtually all subjects had sufficient visual acuity to see the tests.

For the Lanthony, a color Confusion Index scale (Bowman, 1982) was calculated to quantify the degree of confusion between colors and is treated as measured on an interval scale. Vibration measures are in arbitrary amplitude units read from the test device.

Distributions (Sensory)

The distribution parameters for each sensory measure (all subjects) are detailed in Table 11, along with the best transformation to normalize the distribution, if needed (skew or kurtosis above 1). Acuity scores produce a skewed and peaked (kurtotic) distribution; a log transformation improved both skew and kurtosis, respectively, from -1.75 to 0.45 and 3.31 to -0.80. The square root transformation improved the Vibratron (index) skew from 0.88 to -0.06 and kurtosis from 2.99 to 1.32. Transformation of vibratron (small) distributions increased rather than decreased skew. Transformations are not appropriate for Lanthony data where "normal" scores are 1. The untransformed distributions are depicted in Appendix C.

Table 11. Mean, Median, Standard Deviation, Skew, Kurtosis, and the Best Transformation to Approximate Normality in Sensory Measures for All Subjects.

MEASURE	Mean	Mdn	SD	Skew	Kurt	Best Transf
Acuity/line completed	5.9	6.0	1.37	-1.75	3.31	log
Contrast Sensitivity C	5.5	6.0	1.56	-0.78	0.48	
Contrast Sensitivity D	5.4	6.0	1.41	-0.89	0.25	
Lanthony/Lt eye CI[1 2]	1.2	1.2	0.26	1.36	1.54	
Lanthony/Rt eye CI[1 2]	1.3	1.2	0.28	1.00	0.17	
Vibratron/Index[2]	1.4	1.5	0.61	0.88	2.99	Sq Root
Vibratron/Small[2]	1.5	1.5	0.56	0.88	2.35	(worse)

[1] Data from 231-233 subjects were used in this calculation. [2] Majority subjects.

Overall Statistical Comparisons (Sensory)

ANOVAs were calculated with education level, urban vs. rural education location, cultural group, gender, and age level as the independent (grouping) variable. The resulting (unadjusted) probabilities are listed for each sensory measure in Table 12. Education and age had a significant impact on color vision, and urban/rural differences affected Acuity, Contrast Sensitivity D, and color vision (right eye only). There was no impact of cultural group on sensory test measures.

Table 12. P(robability) of Effects of Main Study Factors on Sensory Tests.[1]

MEASURE	Educ Level	Urban vs. Rural[2,3]	Cultural Group	Gender[3]	Age Level
Acuity/line completed	ns	0.04	ns	0.04	ns
Contrast Sensitivity C	ns	0.02	ns	ns	ns
Contrast Sensitivity D	ns	ns	ns	ns	ns
Lanthony/Left CI[3]	0.04	ns	[3]	ns	0.01
Lanthony/Right CI[3]	0.0007	0.03	[3]	ns	0.01
Vibratron/Index[3]	ns	ns	[3]	ns	ns
Vibratron/Small[3]	ns	ns	[3]	ns	ns

[1] By ANOVA. [2] (city size groups: < 16,000; 16,000-349,999; 350,000 and up).
[3] Data collected from majority subjects only.

Importance of Factors: Multiple Regression (Sensory)

The probabilities in Table 12 suggest that the main factors studied here produce only marginal effects on sensory measures. This is borne out by the hierarchical multiple regression analyses which reveal that none of the factors examined in this research accounted for more than 5% of the variance on sensory test performance (Table 13).

Table 13. Percent Variance Accounted for by Main Study Factors on Sensory Tests using Hierarchical Multiple Regression.[1]

MEASURE	Age (Educ contr)	Educ (Age contr)	Shared var.	Educ (Cult Grp contr)	Cult Grp (Educ contr)	Shared var.
Acuity/line	0	2	0	0	0	1
Contrast Sens. C	0	2	0	1	0	1
Contrast Sens. D	0	4	0	2	1	2
Lanthony/Lt CI[2]	2	2	0	[2]	[2]	[2]
Lanthony/Rt CI[2]	1	5	0	[2]	[2]	[2]
Vibr./Index[2]	0	1	0	[2]	[2]	[2]
Vibr./Small[2]	1	1	0	[2]	[2]	[2]

MEASURE	Educ [3] (U/R contr)	U/R [3] (Educ contr)	Shared var.[3]	Age (Cult Grp contr)	Cult Grp (Age contr)	Shared var.
Acuity/line	1	0	0	1	2	0
Contrast Sens. C	1	0	0	0	1	0
Contrast Sens. D	4	0	0	0	3	0
Lanthony/Lt CI[2]	2	2	0	[2]	[2]	[2]
Lanthony/Rt CI[2]	5	3	0	[2]	[2]	[2]
Vibr./Index[2]	1	1	0	[2]	[2]	[2]
Vibr./Small[2]	1	0	0	[2]	[2]	[2]

[1] Header lists factor studied (factor controlled in parentheses). Middle columns list variance unique to factor. [2] Test administered only to majority subjects.
[3] Analysis conducted only on data from majority subjects. U/R = Urban/rural.

<u>Consistency (Replication) and Magnitude of Differences (Sensory)</u>

Higher levels of education were slightly associated with better vision on the acuity and contrast sensitivity C and D tests. Rural- and urban-educated subjects differed significantly on measures of Acuity and Contrast Sensitivity C. Majority and Native Americans had slightly better vision than Latin and African American subjects at most educational levels. The performance differences between the cultural groups are relatively small, on the order of a fraction of a standard deviation (using majority subject SD's in Table 11).

Males and females have comparable sensory capabilities and the test results reflected this, as only acuity of the sensory measures varied with respect to gender (Table 12). While not statistically significant, females had lower thresholds (i.e., better performance) on the tests of visual Contrast Sensitivity C and D and vibration sensitivity in both the index and small fingers. Vibration thresholds, presented in Figure 4, typify the magnitude of the gender differences seen in the sensory tests. Differences are on the order of 0.25 to 0.5 standard deviation units (Table 11) for each test.

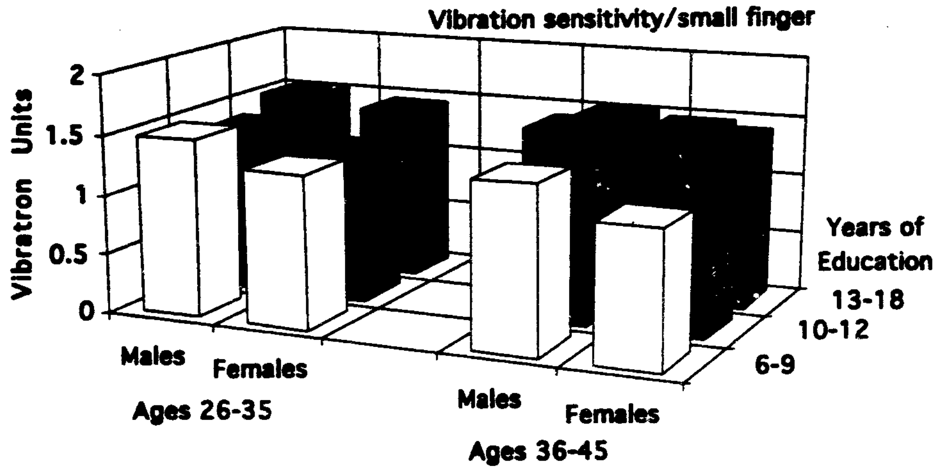


Figure 4. Male and female vibration thresholds in age ranges 26-35 and 36-45, across 6-9, 10-12, and 13-16 years of education.

<u>Correlations Between Tests (Sensory)</u>

Acuity and Contrast Sensitivity tests should be very highly correlated as they both assess visual resolution. Visual acuity correlated moderately well with Contrast Sensitivity C (r=0.39) and D (r=0.45), and Contrast Sensitivity C correlated highly with D (r=0.72). Vibration thresholds of the index and small finger also were highly correlated (r=0.72). All correlations are listed in Appendix E.

MOTOR TESTS

Tests that primarily assessed motor function included the Dynamometer (strength and fatigue), Simple Reaction Time (SRT), NES Simple Reaction Time, NES Tapping (trials 1 and 2 [minority subjects had four trials]), Santa Ana (preferred and nonpreferred hand), Aiming (dots in circle [correct]) and Purdue Pegboard (right, left, both hands simultaneously).

Distributions (Motor)

The parameters of the distributions for each motor measure are detailed in Table 14, along with the best transformation to normalize the distributions. The SRT, Santa Ana/nonpreferred hand, and Pegboard Left- and Both-hands measures are not normally distributed. The untransformed distributions are depicted in Appendix C.

Table 14. Mean, Median, Standard Deviation, Skew, Kurtosis, and the Best Transformation to Approximate Normality for Motor Measures for All Subjects.

MEASURE	Mean	Mdn	SD	Skew	Kurt	Best Transf	Trans Skew	Trans Kurt
Dyn./Strngth-M[1]	43	43	8.4	0.18	-0.01			
Dyn./Strngth-F[1]	29	29	6.3	0.35	0.54			
Dyn./Fatigue	0.85	0.86	0.11	-0.62	0.05			
SRT/mean RT	268	258	46.4	1.74	5.03	(worse)		
NES SRT/latency	281	270	48.5	1.68	5.15	log	0.31	0.85
NES Tap./trial 1 [2]	159	167	42.7	-0.38	0.56			
NES Tap./trial 2 [2]	169	172	30.6	0.14	0.84			
Santa Ana/Pref.	42.1	43.0	6.65	-0.23	0.76			
Santa Ana/Nonprf.	39.1	39.0	7.15	-0.91	4.03	Sq. Rt.	0.06	1.00
Aiming/correct	133	135	43.5	-0.25	-0.11			
Pegboard/Right	14.2	14.0	1.98	-0.17	0.96			
Pegboard/Left	13.7	14.0	2.03	-0.55	2.30	Sq. Rt.	-0.7	1.28
Pegboard/Both	22.4	22.0	5.36	3.60	57.57	Sq. Rt.	0.2	1.3

[1] Males and females separated due to large differences between them on this test (Figure 3). [2] First trial for all subjects; trial 2 only for minority subjects.

The transformations increased the normality of the distributions of NES SRT, Santa Ana Non-preferred, and Pegboard Left and Both hands (Table 14, rightmost columns). The "best" transformation (log) of the SRT/mean RT distribution actually increased both skew and kurtosis (noted as "worse" in column 7 of Table 14). The change in Pegboard Both for comparable data from all subjects tested is shown in Figure 5. The improvement is not immediately apparent, although the reduction in both skew and kurtosis is substantial (Table 14, bottom row).

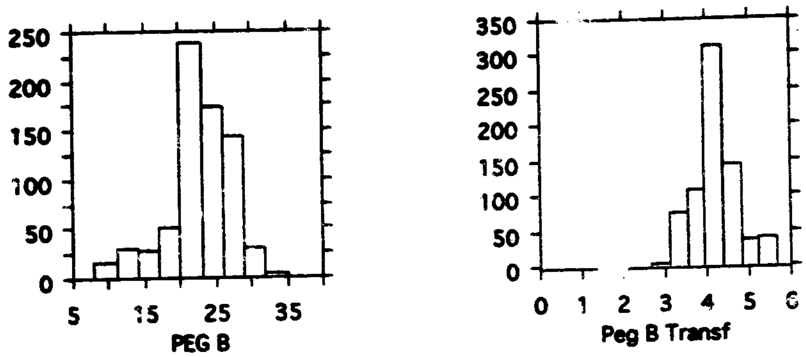


Figure 5. Original distribution of Pegboard Both hands (left side) and same distribution after a square root transformation was applied (right side).

Overall Statistical Comparisons (Motor)

ANOVAs were calculated with education level, urban vs. rural education location, cultural group, gender, and age level as the independent (grouping) variable. The resulting (unadjusted) probabilities are listed for each motor measure in Table 15. Education, cultural group, gender, and age each affect performance on virtually all of the motor tests, although the rural-urban split had a negligible impact.

Table 15. P(robability) of Effects of Main Study Factors on Motor Tests.[1]

MEASURE	Educ Level	Urban vs. Rural[2,3]	Cultural Group	Gender[3]	Age Level
Dynamom./Strength[3]	ns	ns	[3]	<0.0001	ns
Dynamometer/Fatigue[3]	ns	ns	[3]	0.009	ns
SRT/mean RT	<0.0001	ns	<0.0001	<0.0001	0.02
NES SRT/mean latency	0.0003	0.03	<0.0001	0.0007	0.03
NES Tapping/trial 1	<0.0001	ns	0.002	<0.0001	ns
NES Tapping/trial 2[3]	0.0278	0.03	[3]	<0.0001	0.02
Santa Ana/Pref.	<0.0001	ns	0.008	ns	ns
Santa Ana/Nonpref.	<0.0001	ns	ns	ns	ns
Aiming/correct	<0.0001	ns	<0.0001	0.0001	0.003
Pegboard/Right	<0.0001	ns	0.0007	<0.0001	0.04
Pegboard/Left	<0.0001	ns	0.0001	0.0008	ns
Pegboard/Both	<0.0001	ns	<0.0001	<0.0001	ns

[1] By ANOVA [2] City size groups: < 16,000; 16,000-349,999; 350,000 and up).
[3] Data collected from majority subjects only.

Importance of Factors: Multiple Regression (Motor)

Little variance is accounted for by cultural group in motor tests when controlling for education, but it explains a substantial percent of the variance (12%, 15%, 20%, respectively, on NES SRT, Pegboard both, and Aiming) when age is controlled (Table 16, bottom right). Education explains more than 5% of the variance for several motor tests when age (SRT Tapping, Santa Ana, Aiming), cultural group (NES SRT, pegboard, or urban/rural) educational background (Tapping, Santa Ana, Aiming) are controlled.

Table 16. Percent Variance Accounted for by Main Study Factors on Motor Tests using Hierarchical Multiple Regression. [1]

MEASURE	Age (Educ contr)	Educ (Age contr)	Shared var.	Educ (Cult Grp contr)	Cult Grp (Educ contr)	Shared var.
Dynam./Strength[2]	0	0	0	[3]	[3]	[3]
Dynam./Fatigue[2]	1	0	0	[3]	[3]	[3]
SRT/mean RT	0	7	1	3	2	6
NES SRT/latency	1	0	0	15	2	2
NES Tap./trial 1	1	7	1	5	0	2
NES Tap./trial 2 [2]	1	2	0	[2]	[2]	[2]
Santa Ana/Pref.	3	5	0	2	4	0
Santa Ana/NonPr.	3	8	0	1	6	1
Aiming/correct	0	17	0	5	7	12
Pegboard/Right	5	5	0	2	3	0
Pegboard/Left	2	5	0	1	2	2
Pegboard/Both	1	3	0	11	0	2

MEASURE	Educ [3] (U/R contr)	U/R [3] (Educ contr)	Shared var.[3]	Age (Cult Grp contr)	Cult Grp (Age contr)	Shared var.
Dynam./Strength[2]	0	0	0	[2]	[2]	[2]
Dynam./Fatigue[2]	0	0	0	[2]	[2]	[2]
SRT/mean RT	0	2	0	0	8	1
NES SRT/latency	1	2	0	0	12	1
NES Tap./trial 1	6	1	0	1	3	0
NES Tap./trial 2	2	0	0	[2]	[2]	[2]
Santa Ana/Pref.	7	0	0	3	3	0
Santa Ana/NonPr.	8	1	0	4	3	0
Aiming/correct	7	2	0	1	20	0
Pegboard/Right	4	0	0	5	5	0
Pegboard/Left	0	1	0	2	4	0
Pegboard/Both	3	0	0	2	15	0

[1] Header lists factor studied (factor controlled in parentheses). Middle columns list variance unique to factor. [2] Test administered only to majority subjects.
[3] Analysis conducted only on data from majority subjects. U/R = Urban/rural.

Consistency (Replication) and Magnitude of Differences (Motor)

Differences in motor performance were expected across educational levels, due to experience with keyboards and writing instruments in occupations demanding higher levels of education. There were isolated cases of better performance in groups with higher levels of education on the Tapping, Santa Ana, and Pegboard tests, all of which require hand speed and coordination, and on the Dynamometer strength test, which was not expected. These differences correlated with education were particularly striking in the Latin American subjects (well in excess of 1 SD in some cases), although this was particularly to the poor performance in subjects with less than six years of education.

Consistent differences between cultural groups were also seen sporadically on motor tests. Compared to all other cultural groups, Latin-Americans were faster on the NES SRT, majority subjects completed more taps on trial 1 of the Tapping test, and African Americans turned fewer pegs on the Santa Ana (preferred) and Pegboard (both hands) tests than the other cultural groups. Figure 6 depicts the larger number of taps associated with increasing levels of education in Latin-American subjects and the larger number of taps completed by Majority followed by Native American Indian subjects as compared to Latin and African Americans. This is typical of the magnitude of the differences seen in motor performance; they were not striking, except when comparing people with fewer than 6 years of education to those with more than six years of education.

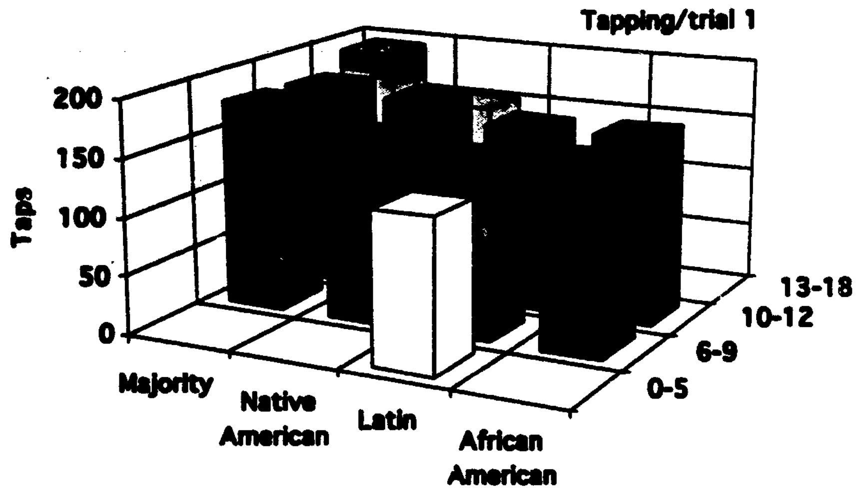


Figure 6. Number of taps by majority, American Indian, Latin, and African American subjects across educational levels 0-5, 6-9, 10-12, and 13-16.

Expected gender differences were seen on many tests involving strength, speed or coordination of the motor system. Majority males were stronger on the Dynamometer strength test than females (Figure 3, above), although females demonstrated consistently less fatigue (i.e., less decline between trials 1 and 5 of the Dynamometer test). Generally, however, majority males were faster than majority females on both reaction time tests and both tapping trials (e.g., Figure 7). The magnitude of the differences is generally 0.5 SD (Table 14). Other differences between males and females on motor tests were not consistent. For example, females moved more pins with their the left hand but not their right or both hands trials of the Pegboard test.

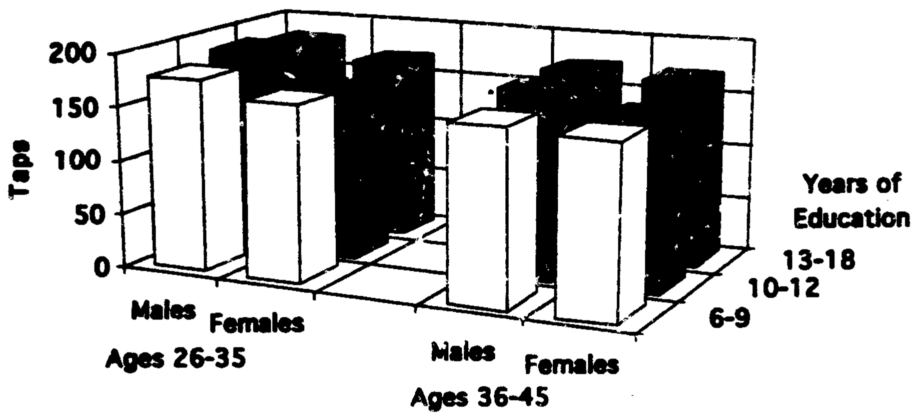


Figure 7. Tapping by males and females in age ranges 26-35 and 36-45 across education ranges 6-9, 10-12, and 13-16.

Correlations Between Tests

Several repeated measures of the same test were highly correlated. They are the two tapping trials (r=0.67), Santa Ana preferred with the non-preferred hand (r=0.66), Purdue Pegboard trials (r= 0.52 for the right- with the left-hand trials), and the NES and NCTB SRT measures (r=0.47). All correlations are listed in Appendix E.

COGNITIVE TESTS

The cognitive measures were the Benton (correct), Digit Span (forward and backward), Digit Symbol, NES Symbol Digit (latency, or time to complete trials), NES Symbol Digit Recall, NES Serial Digit Learning, and Raven Progressive Matrices (errors).

Distributions (Cognitive)

The parameters of the distributions for each cognitive measure are detailed in Table 17, along with the best transformation to normalize the distribution, where needed. The untransformed distributions are depicted in Appendix C.

Table 17. Mean, Median, Standard Deviation, Skew, Kurtosis, and the Best Transformation to Approximate Normality for Cognitive Measures in All Subjects.

MEASURE	Mean	Mdn	SD	Skew	Kurt	Best Transf	Transf Skew	Transf Kurt
Benton correct	8.3	9.0	1.60	-1.32	2.11	sq. root	0.6	0.1
Digit Span/Forw.	7.9	8.0	2.33	0.04	-0.05			
Digit Span/Back.	6.6	6.0	2.31	0.57	0.61			
Digit Symbol	53.1	54.0	14.6	-0.30	0.03			
NES Sym Dig/best	2.5	2.3	0.84	2.48	12.0	log	0.7	1.4
NES SymD/recall*	4.1	4.0	2.14	-0.11	-0.44			
NES Ser Dig Lrn	5.6	4.0	5.1	0.74	-0.74			
Raven/Errors*	9.5	2.0	5.3	0.77	0.40			

* Majority subjects only.

The Benton is skewed to the right, with almost one third of all subjects getting all 10 choices correct. The Serial Digit Learning task, which records errors, has a distribution that is a mirror image of the Benton, skewed to the left (Appendix C). The Benton is improved by the square root transformation, whereas the same transformation decreased skew but increased kurtosis in the Serial Digit Learning test (data not shown). The best overall impact of transformation on a cognitive measure occurred in the NES Symbol Digit Best distribution (Table 17 and Figure 8).

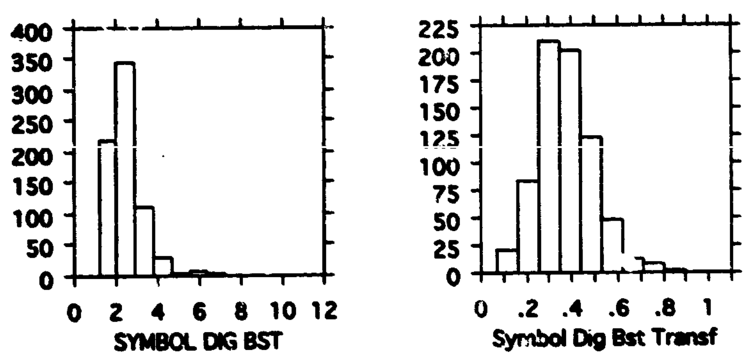


Figure 8. Original distribution of NES Symbol Digit latency (left side) and same distribution after a log transformation was applied (right side).

Overall Statistical Comparisons (Cognitive)

ANOVAs were calculated with educational level, urban vs. rural education locale, cultural group, gender, and age level as the independent (grouping) variable. The resulting (unadjusted) probabilities are listed for each cognitive measure in Table 18. Most factors produced differences that were highly significant, with the exception of the urban vs. rural location for education.

Table 18. P(robability) of Effects of Main Study Factors on Cognitive Tests.[1]

MEASURE	Educ Level	Urban vs. Rural[2]	Cultural Group	Gender[3]	Age Level
Benton correct	<0.0001	ns	<0.0001	ns	0.0002
Digit Span/Forward	<0.0001	0.04	<0.0001	ns	<0.0001
Digit Span/Backward	<0.0001	ns	<0.0001	0.002	<0.0001
Digit Symbol	<0.0001	ns	<0.0001	<0.0001	0.0003
NES Symbol Digit/best	<0.0001	ns	<0.0001	<0.0001	<0.0001
NES Symb. Digit/Recall[3]	ns	ns	[3]	0.002	0.0003
NES Serial Digit Learn.	<0.0001	ns	<0.0001	ns	0.0007
Raven/Errors[3]	<0.0001	ns	[3]	ns	ns

[1] By ANOVA. [2] (city size groups: < 16,000; 16,000-349,999; 350,000 and up).
[3] Data collected in majority subjects only.

Importance of Factors: Multiple Regression (Cognitive)

Hierarchical multiple regression revealed that the factor of age (in the limited span of 26-45 years) predicted less than 1% of the variance in all cases, but controlling it revealed the substantial contribution of education and cultural group to test performance (top left and bottom right panel of Table 19). When analyzed together, both cultural group and years of education contributed to the prediction of the cognitive measures, but most of the variance is shared (top right panel, Table 19). This suggests that the effects on these tests of education and culture cannot be separated and have profound effects on experimental design and data analysis in cross sectional studies using cognitive tests. Finally, controlling for urban/rural site for education (bottom left, Table 19) reveals different impacts on education than does controlling for cultural group (top right), but the range of variance attributable to these factors is roughly comparable.

Table 19. Percent Variance Accounted for by Main Study Factors on Cognitive Tests using Hierarchical Multiple Regression.[1]

MEASURE	Age (Educ contr)	Educ (Age contr)	Shared var.	Educ (Cult Grp contr)	Cult Grp (Educ contr)	Shared var.
Benton correct	0	15	2	4	4	13
Digit Span/Forw	1	17	2	10	3	16
Digit Span/Backw	0	14	4	5	4	14
Digit Symbol	0	36	1	4	4	13
NES Sym Dig/best	0	30	2	9	7	25
NES Sym Dig/Recall[2]	3	1	0	[2]	[2]	[2]
NES Serial Digit Lrn	0	17	2	13	2	18
Raven/Errors[2]	0	11	0	[2]	[2]	[2]

MEASURE	Educ[3] (U/R contr)	U/R[3] (Educ contr)	Shared var.[3]	Age (Cult Grp contr)	Cult Grp (Age contr)	Shared var.
Benton correct	10	0	2	0	15	1
Digit Span/Forw	5	0	0	0	24	2
Digit Span/Backw	8	0	0	0	15	4
Digit Symbol	15	0	0	2	32	0
NES Sym Dig/best	11	1	0	2	34	0
NES Sym Dig/Recall[2]	1	0	0	[2]	[2]	[2]
NES Serial Digit Lrn	5	0	0	1	29	1
Raven/Errors[2]	1	11	0	[2]	[2]	[2]

[1] Header lists factor studied (factor controlled in parentheses). Middle columns list variance unique to factor. [2] Test administered only to majority subjects.
[3] Analysis conducted only on data from majority subjects. U/R = Urban/rural.

<u>Consistency (Replication) and Magnitude of Differences (Cognitive)</u>

Both cultural group and years of education variables produced consistent and in some cases powerful effects on the cognitive tests. Increasing levels of education were associated with improved performance in the Benton, Digit Span (forward and reverse), Digit Symbol, and Symbol Digit (fastest completion), each finding seen in at least three of four cultural groups. Raven errors, too, were higher in people with less education. Differential performance in cultural groups was seen in the Digit Span (forward and backward) and Serial Digit Learning tests. Both educational and cultural group differences are reflected in Figure 9 (on this test the African Americans do not reflect the impact of education). The educational differences are on the order of 0.5 SD units (see Table 17) in all but the Latin-American subjects, in which they typically exceeded 1. No consistent differences were seen between the urban- and rural-educated (majority) subjects.

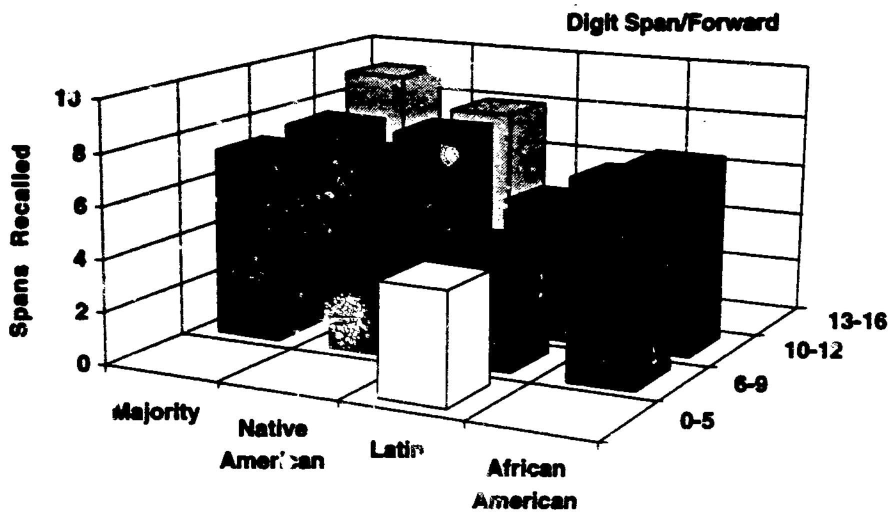


Figure 9. Number of spans recalled in the Digit Span (forward) test by Majority, Native American, Latin and African American subjects with 0-5, 6-9, 10-12, and 12-16 years of education.

The cultural group differences were seen primarily in the Latin-Americans who had considerably lower performance or a larger number of errors on the noted tests. African-American performance also revealed more errors on the Serial Digit Learning test and lower performance on the Digit Span reverse test. Differences in the range of 1-3 standard deviation units were seen (e.g., Figure 9) between the Latin population and the other populations tested here. However, the SDs generated by Latin subjects were also higher than in the other populations, indicating a greater variability in the Latin-Americans tested. The differences between the African American population and the other groups were in the 0.5-0.75 SD range.

Well-established gender differences were seen in the majority subjects in this study, most notably better performance by females on the Digit Symbol (completing more symbols) and NES Symbol Digit (faster completion) tests. Consistent findings across all comparisons also revealed that males recalled more digits (in both forward and backward order) in the Digit Span test. Digit repetition backward (i.e., in the reverse order that the digits are presented) is depicted in Figure 10. The differences are on the magnitude of 0.5 to 1 standard deviation units (cf. Table 17). The small, consistent effect of education, described above, can also be seen in Figure 10.

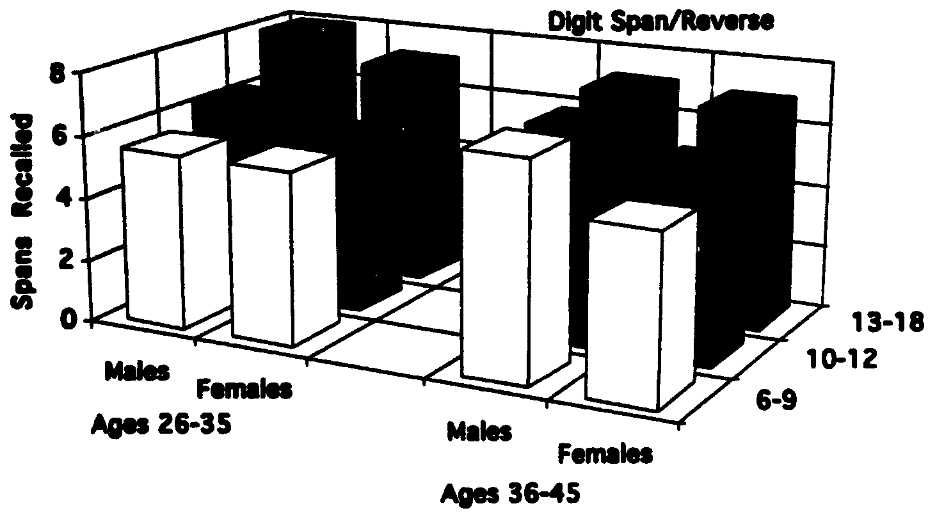


Figure 10. Digit spans (reverse) recalled by majority males and females, ages 26-35 and 36-45 with 6-9, 10-12, and 13-18 years of education.

Correlations Between Tests (Cognitive)

NES Symbol Digit test performance was highly correlated with performance on the analogous Digit Symbol (r=0.78) test from which the Symbol Digit was drawn. Correlations between cognitive measures were in the range of 0.40 and 0.51 (Benton total with Digit Symbol; Symbol Digit/Digit Span forward, Digit Span reverse/Benton total; Serial Digit Learning/Digit Span forward, Digit Span reverse, Digit Symbol, Raven errors/Symbol Digit). These tests tap related functions; relatively high correlations between these tests are expected. Correlations are in Appendix E.

MEASURES OF AFFECT

The measures of Affect studied here are the NES Mood test scales of Tension, Depression, Anger, Fatigue, and Confusion. This test was not administered to the Latin American subjects.

Distributions (Affect)

The parameters of the distributions for each measure of affect are detailed in Table 20. Depression, Anger and Confusion are skewed slightly to the right (high scores) (Appendix C), but Table 20 shows that skew and kurtosis are very low (under 1) for all affect measures. The untransformed distributions are depicted in Appendix C.

Table 20. Mean, Median, Standard Deviation, Skew, and Kurtosis for Measures of Affect in NES Mood Test (Latin-American Subjects Excluded).

MEASURE	Mean	Median	SD	Skew	Kurtosis
NES Mood/Tension	2.73	2.6	0.82	0.34	-0.31
NES Mood/Depression	2.18	2.0	0.87	0.99	0.44
NES Mood/Anger	2.03	1.8	0.83	0.88	0.69
NES Mood/Fatigue	2.77	2.8	0.82	0.35	-0.02
NES Mood/Confusion	2.40	2.2	0.77	0.53	-0.24

Overall Statistical Comparisons (Affect)

ANOVAs were calculated with education level, urban vs. rural education location, cultural group, gender, and age level as the independent (grouping) variable. The resulting (unadjusted) probabilities are listed for each measure of affect in Table 21. Differences seen on NES Mood measures were somewhat distributed.

Table 21. P(robability) of Effects of Main Study Factors on Affective Measures in NES Mood Test.[1]

MEASURE	Educ Level	Urban vs. Rural[2]	Cultural Group[3]	Gender[4]	Age Level
NES Mood/Tension	ns	ns	0.008	0.003	ns
NES Mood/Depression	ns	ns	ns	ns	ns
NES Mood/Anger	ns	ns	0.04	0.02	ns
NES Mood/Fatigue	0.003	ns	0.02	0.02	ns
NES Mood/Confusion	0.0004	ns	ns	ns	0.01

[1] By ANOVA. [2] City size groups: < 16,000; 16,000-349,999; 350,000 and up).
[3] Cultural group comparisons excluded Latin Americans (not given test).
[4] Majority subjects, only.

Importance of Factors: Multiple Regression (Affect)

None of the factors studied in this project explain any significant portion of the variance associated with the NES Mood measures (Table 22).

Table 22. Percent Variance Accounted for by Education and Cultural Groups on NES Mood Test of Affect using Hierarchical Multiple Regression.[1,2]

MEASURE	Age (Educ contr)	Educ (Age contr)	Shared var.	Educ (Cult Grp contr)	Cult Grp (Educ contr)	Shared var.
Mood/Tension	0	0	0	3	2	2
Mood/Depression	0	1	0	2	2	2
Mood/Anger	0	1	0	3	2	2
Mood/Fatigue	0	2	0	4	5	2
Mood/Confusion	1	3	0	0	4	0

MEASURE	Educ[3] (U/R contr)	U/R[3] (Educ contr)	Shared var.[3]	Age (Cult Grp contr)	Cult Grp (Age contr)	Shared var.
Mood/Tension	2	0	0	0	2	0
Mood/Depression	2	0	0	0	0	0
Mood/Anger	2	0	0	1	2	0
Mood/Fatigue	5	0	0	0	1	0
Mood/Confusion	3	0	0	1	0	0

[1] Header lists factor studied (factor controlled in parentheses). Middle columns list variance unique to factor. [2] Test not administered to Latin subjects.
[3] Analysis conducted only on data from majority subjects.

Consistency (Replication) and Magnitude of Differences (Affect)

There were no consistent differences between cultural groups or educational levels seen in comparisons between majority, Native American Indian and African American subjects who were given the

NES Mood test. In most NES Mood measures, majority females had higher scores (a greater intensity of the mood state) than majority male subjects, as exemplified in Figure 11. The differences on these measures were in the 0.5 SD range (Table 20).

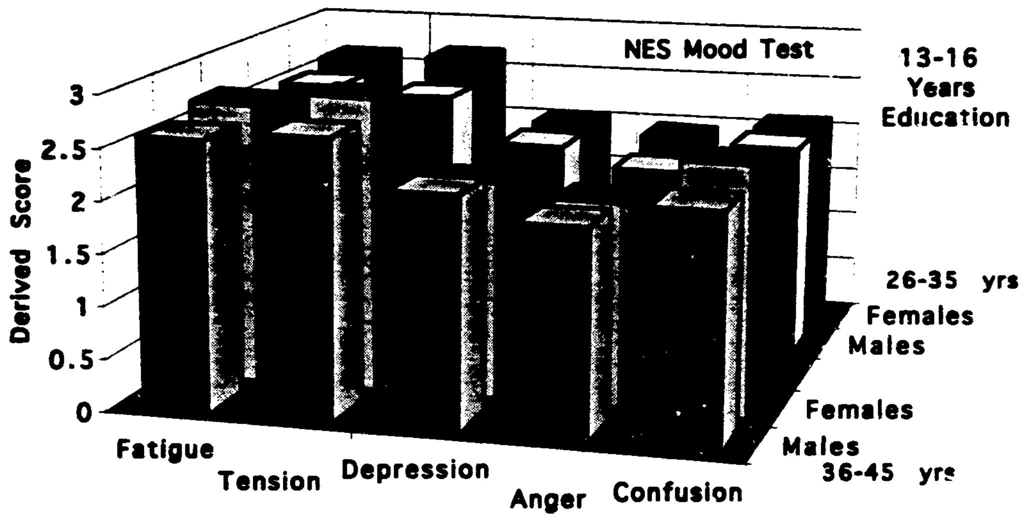


Figure 11. NES Mood scores (derived) in male and female majority subjects 26-35 and 36-45 years of age, with 13-16 years of education.

Correlations Between Tests (Affect)

Most of the Mood scales are moderately to highly correlated with each other (r=0.50 to 0.70). All correlations are listed in Appendix E.

MEASURES OF VOCABULARY

The NES vocabulary test, WAIS (scaled), and Peabody Picture vocabulary test provide the measures of vocabulary. The Spanish-language versions of the WAIS and Peabody vocabulary tests were administered to the Latin-American subjects, and norms unique to that population were used to score the tests. The NES was not administered to Latin subjects.

Distributions (Vocabulary)

The NES vocabulary test has a decidedly non-normal distribution, compared to the WAIS (scaled for age) and Peabody (scaled for age) scores (Figure 12).

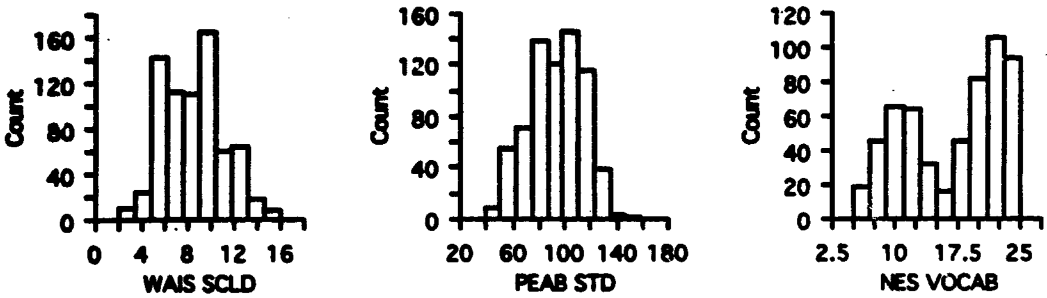


Figure 12. Distributions of scores on the WAIS, Peabody, and NES vocabulary tests (majority subjects only).

The parameters of the distributions for each measure of vocabulary (majority subjects only) are detailed in Table 23, along with the best transformation to normalize the NES Vocabulary test distribution.

Table 23. Mean, Median, Standard Deviation, Skew, Kurtosis, and the Best Transformation to Approximate Normality for the NES Vocabulary Test (Based on Majority Subjects Only).

MEASURE	Mean	Median	SD	Skew	Kurt.	Best Transform
NES Vocabulary	16.5	18.0	5.8	-0.32	-1.35	log
WAIS Vocabulary	8.4	8.0	2.61	0.36	-0.24	
Peabody Vocabulary	92.9	94.0	21.0	-0.17	0.49	

The NES Vocabulary test poses serious distribution problems for parametric analyses that are not adequately captured by measures of skew and kurtosis nor sufficiently improved by the log transformation (Figure 13). The distribution is U shaped and bimodal with or without the transformation. Figure 13 reveals the impact of the log transformation (which improved kurtosis from

-1.35 to -0.64). The Square root, reflected square root and inverse transformations were even less effective in improving the normality of the distribution.

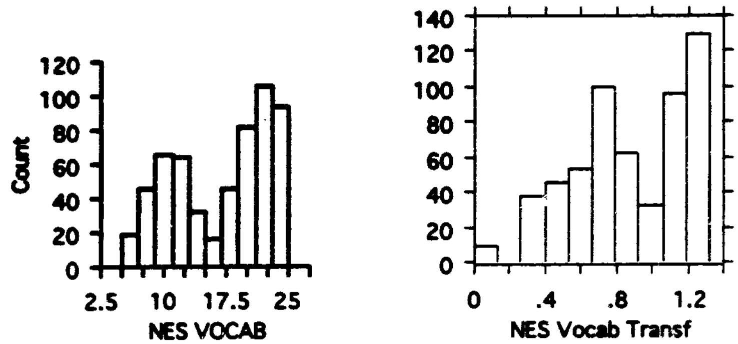


Figure 13. Original distribution of NES Vocabulary (left side) and the same distribution after a log transformation was applied (right side).

<u>Overall Statistical Comparisons (Vocabulary)</u>
Years of education (but not the urban vs. rural site of the education in majority subjects), cultural group membership, gender, and age all affected performance on the vocabulary tests (Table 24).

Table 24. <u>P(robability) of Effects of Main Study Factors on Vocabulary Tests.</u>[1]

MEASURE	Educ Level	Urban vs. Rural[2,3]	Cultural Group	Gender[3]	Age Level
NES Vocabulary[5]	< 0.0001	ns	<0.0001	ns	<0.0001
WAIS Vocabulary[6]	< 0.0001	ns	<0.0001	ns	-[4]
Peabody Vocabulary[6]	< 0.0001	ns	<0.0001	0.002	-[4]

[1] By ANOVA. [2] City size groups: < 16,000; 16,000-349,999; 350,000 and up).
[3] Data from majority subjects only. [4] Data are age-adjusted by test norms.
[5] Test was not administered to Latin subjects.
[6] Spanish-language versionadministered to Latin subjects

<u>Relation between Years of Education and Vocabulary Tests</u>
The overall differences in vocabulary across gross educational levels suggests that education does improve vocabulary scores. However, years of education was not equally effective for predicting the different vocabulary scores. Figure 14 presents regression plots across years of education in all subjects tested.

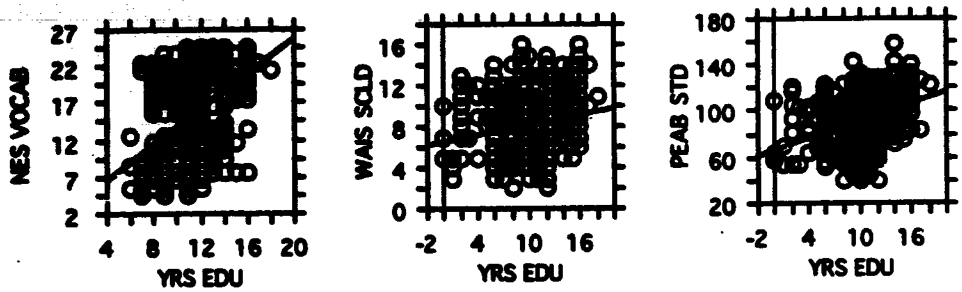


Figure 14. Simple regression plots of the NES, WAIS and Peabody vocabulary tests across years of education, for all subjects (Latin subjects excluded from NES and WAIS tests).

Inspection of regression plots for individual cultural groups reveals differential predictive value. Years of education is a better predictor of vocabulary in majority subjects (top panel of Figure 15) than in African American subjects (lower panel, Figure 15).

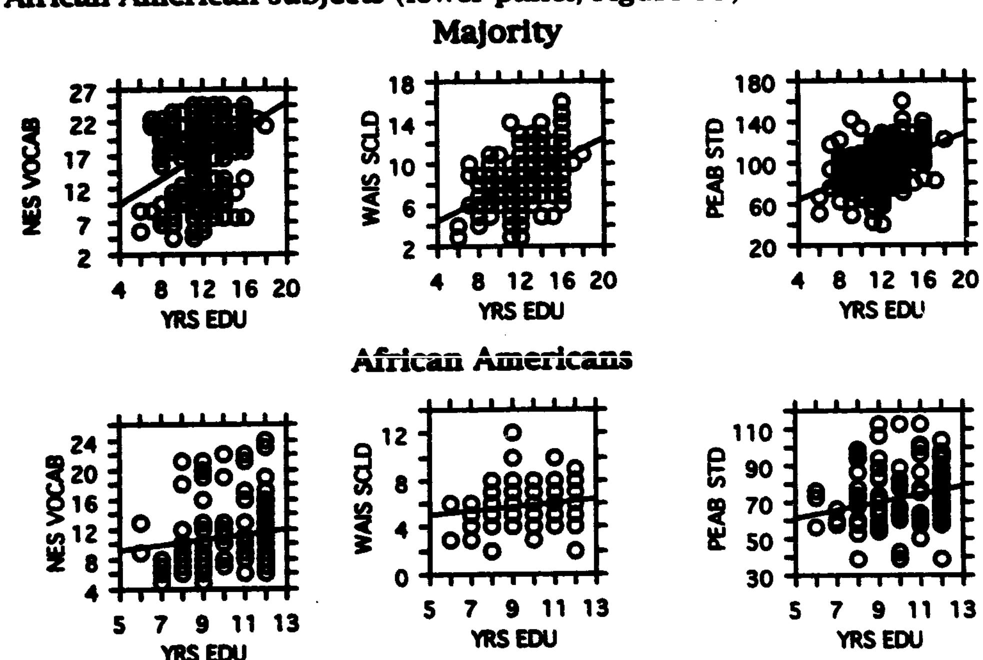


Figure 15. Regression plots of years of education in majority and African American subjects.

Importance of Factors: Multiple Regression (Vocabulary)

Multiple regression was used to gauge the impact of vocabulary on the individual motor and cognitive tests. The percent of variance explained by the WAIS vocabulary test (Table 25, left side) can be compared to the percent of variance explained by years of education (right side). Neither variable can explain more than a few percentage points of the variance on the motor tests (and less on sensory tests omitted from the Table), as would be expected.

In contrast, more variance in the cognitive tests is accounted for by the WAIS vocabulary test than by years of education for the people studied here. (Differences in the Latin population must be qualified by the Spanish-language version of the WAIS which only they received.) Clearly, vocabulary accounts for more variance in the Latin-American and majority subjects than in Native Americans and African Americans in this study, and years of education is dominant in the Latin-American subjects, less so in majority subjects and again not important in the African American and Native American populations on these tests (Table 25).

Table 25. Percent Variance Explained by WAIS Vocabulary Test (a Surrogate for Intelligence) and Years of Education across Cultural Groups.

MEASURE	WAIS Vocabulary Test				Years of Education			
	Afr	Lat	Maj	Nat	Afr	Lat	Maj	Nat
Motor								
SRT/mean RT	3	7	5	1	4	6	< 1	8
NES SRT/mean latency	2	6	7	-	2	6	1	8
NES Tapping/trial 1	0	12	12	0	0	10	6	0
NES Tapping/trial 2 [1]	-	-	3	-	-	-	2	-
Santa Ana/Preferred	7	5	12	-	1	6	7	1
Santa Ana/Nonpref.	5	3	9	-	1	8	8	1
Pegboard/Right	6	2	5	-	< 1	4	4	-
Pegboard/Left	6	7	3	-	4	6	-	2
Pegboard/Both	-	2	3	2	6	3	3	3
Cognitive								
Benton correct	6	17	5	5	1	16	3	-
Digit Span/Forward	6	10	19	3	-	7	5	< 1
Digit Span/Backward	9	15	22	2	2	12	8	-
Digit Symbol	13	23	17	-	3	32	15	< 1
NES Symbol Digit/Best	13	21	16	4	< 1	16	11	2
NES Sym Digit/Recall [1]	-	-	1	-	-	-	0	-
NES Serial Digit Learn	10	10	17	< 1	3	3	5	1
Raven/Errors [1]	-	-	17	-	-	-	11	-
Vocabulary								
NES [3]	48	-	54	52	2	-	17	2
WAIS (scaled) [2]	-	-	-	-	4	24	26	3
Peabody [2]	61	54	54	48	6	24	25	5

[1] Majority subjects only. [2] Spanish-language version administered to Latin-Americans. [3] Test not administered to Latin-Americans.

The strongest association between education and test performance is seen on the Digit Symbol test in the Latin American population (Figure 16) where years of education accounted for 32% of the variance (Table 25).

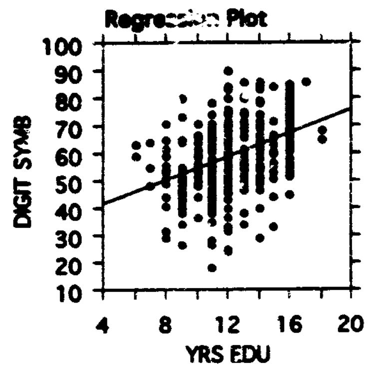


Figure 16. Scatterplot with regression line of Digit Symbol performance in Latin American subjects against years of education.

Multiple regression employing covariates reveals more of the complexity of the relationship. In Table 26, education, cultural group, and age are entered into the equations, with age, cultural group and education variables factored out of the analysis when relevant.

Table 26. Percent Variance Accounted for by Main Study Factors on Vocabulary Tests using Hierarchical Multiple Regression.[1]

MEASURE	Age, (Educ. cont)	Educ., (Age cont)	Shared Var.	Educ (Cult Grp cont)[2]	Cult Grp (Educ cont)[2]	Shared Var.[2]
NES Vocab.	3	21	2	10	10	13
WAIS Vocab. [2]	x	x	x	39	16	12
Peabody Vocab. [2]	x	x	x	23	16	2

[1] Hierarchical Multiple Regression (header lists factor studied followed by factor controlled). [2] Spanish-language version administered to Latins.
x = not applicable.

Since age is intentionally scaled out of the WAIS and Peabody scores, the lack of importance of this variable is virtually assured (column 2 of Table 26). The NES vocabulary test has more shared variance than that accounted for by education or cultural group alone, but the reverse is the case for both the WAIS and Peabody. Shared variance is a larger factor for the WAIS but essentially trivial for the Peabody.

Consistency (Replication) and Magnitude of Differences (Vocabulary)
Years of education had an unexpectedly small but detectable impact on absolute vocabulary scores, as seen in the data from females in Figure 17. The WAIS, Peabody, and NES tests revealed similarly proportioned improvements in vocabulary scores with increasing years of education, although this is somewhat obscured in Figure 17 due to the large numerical Peabody totals. Data from males revealed a similar but less consistent trend.

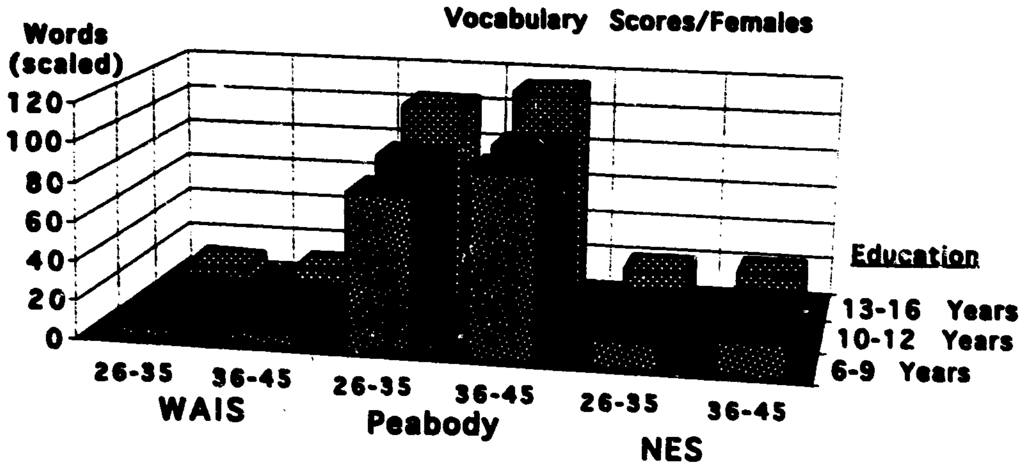


Figure 17. Differences on vocabulary test performance across three education levels in majority females.

A comparison of majority subjects educated in urban settings vs. rural settings reveals no substantial differences. Figure 1? compares rural and urban subjects for performance on the WAIS vocabulary test. The differences in years of education is apparent in people aged 26-35 and 36-45, although it is even more apparent that a high school education in the past 10-20 years (the 26- to 35-year olds) does not produce the level of vocabulary found in people educated 10 years earlier by a factor of nearly 1 SD (Table 23).

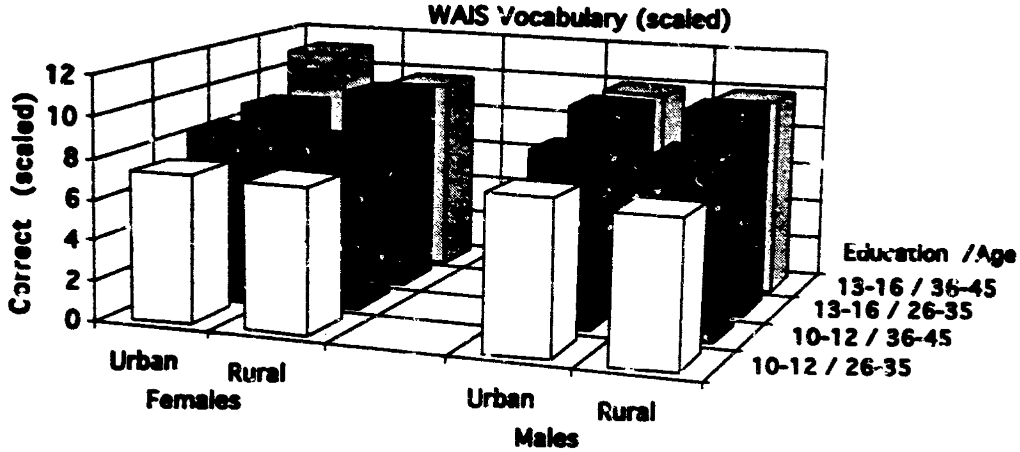


Figure 18. Urban/Rural WAIS vocabulary comparisons for majority females (left columns) and males (right columns) in educational (10-12, 13-16) and age (26-35, 36-45) groups.

Correlations Between Tests (Vocabulary)

The three measures of vocabulary and reported years of education correlated highly in the 388 male and female majority subjects who had 75% of their education in one city (Table 27).

Table 27. Correlations between reported years of education and WAIS, NES, and Peabody Vocabulary Tests in Majority Subjects.

	Years of Education	NES	WAIS	Peabody
Years of Education	1.000			
NES	.420	1.000		
WAIS	.512	.730	1.000	
Peabody	.508	.703	.729	1.000

In majority subjects, the highest correlation between years of education and a vocabulary test was with the WAIS (0.512), although the more objective Peabody test had virtually the same correlation (0.508). The NES vocabulary test had a high correlation with the WAIS (0.730) and the Peabody (0.703), although its correlation with years of education was lower (0.420) than the other vocabulary tests. This differs from the correlations between vocabulary tests and years of education in minority subjects (e.g., Figure 15).

Section 4
DISCUSSION

DISTRIBUTIONS (NORMALITY)

To employ sensitive but potentially labile parametric statistical analyses, a distribution must be normal. A priori expectations of distribution normality guide the selection of statistical tests and thus statistical power calculations. Appendix C reveals that data on many tests are normally distributed but a large number are not. The normal distributions, defined by skew and kurtosis measures under 1.0, are identified by an "N" in Table 28, along with the most appropriate transformation to reduce skew and kurtosis. Transformed data were not, however, used in the analyses reported here.

THE MAJOR FACTORS (SUMMARY)

This study assessed the impact of critical subject factors on the performance of behavioral tests used in neurotoxicity evaluations. Education, cultural group, gender, and age (range of 26-45 years only) all have an impact on some of the tests used in this field. Table 29 presents those factors found in this study to explain more than 5% of the variance. Dots therefore identify those variables which need to be balanced in the study design or statistically controlled in the data analysis. It behooves the investigator to consider such design and analytic strategies when using most behavioral tests in neurotoxicology research. This table reflects cases in which it is imperative.

Table 28. Summary of Distribution Characteristics of Tests

MEASURE	Skew	Kurtosis	Best Transf.	Transf. Skew	Transform Kurtosis
Sensory					
Acuity Category	-1.75	3.31	log	0.45	-0.80
Contrast Sensitivity C	-0.78	0.48	N		
Contrast Sensitivity D	-0.89	0.25	N		
Lanthony/Left eye CI[1]	1.36	1.54	-		
Lanthony/Right eye CI[1]	1.0	0.17	-		
Vibratron/Index[1]	0.88	2.40	Sq Root	-0.06	1.32
Vibratron/Small[1]	0.89	3.04	made worse		
Motor					
Dynamom./Strength-M[1]	0.18	-0,01	N		
Dynamom./Strength-F[1]	0.35	0.54	N		
Dynamometer/Fatigue[1]	0.11	-0.62	N		
SRT/mean RT	1.74	5.03	made worse		
NES SRT/mean latency	1.68	5.15	log	0.31	0.85
NES Tapping/trial 1	-0.38	0.56	N		
NES Tapping/trial 2[1]	0.14	0.84	N		
Santa Ana/Preferred	-0.23	0.76	N		
Santa Ana/Nonpref.	-0.91	4.03	Sq. Root	0.06	1.00
Aiming/correct	-0.25	-0.11	N		
Pegboard/Right	-0.17	0.96	N		
Pegboard/Left	-0.55	2.30	Sq. Root	-0.7	1.28
Pegboard/Both	3.60	57.57	Sq. Root	0.2	1.3
Cognitive					
Benton correct	-1.32	2.11	Sq. Root	0.6	0.1
Digit Span/Forward	0.04	-0.05	N		
Digit Span/Backward	0.57	0.61	N		
Digit Symbol	-0.30	0.03	N		
NES Symbol Digit/Best	2.48	12.0	log	0.7	1.4
NES Symbol Dig/Recall[1]	-0.11	-0.44	N		
NES Serial Digit Learn.	0.74	-0.74	N		
Raven/Errors[1]	0.77	0.40	N		
Affect					
NES Mood/Tension	0.34	-0.31	N		
NES Mood/Depression	0.99	0.44	N		
NES Mood/Anger	0.88	0.69	N		
NES Mood/Fatigue	0.35	-0.02	N		
NES Mood/Confusion	0.53	-0.24	N		
Vocabulary					
NES	-0.32	-1.35	log	-0.49	-0.64
WAIS (scaled)	0.36	-0.24	N		
Peabody	-0.17	0.49	N		

[1] Majority subjects only.

Table 29. Summary of Distribution and Subject Factors.[1]

MEASURE	Dist[2]	Sd/Mn Ratio[3]	Educ Level	Urban/ rural[5]	Cult Grp	Gend[5]	Age Level
Sensory							
Acuity/line compltd.	log	0.23					
Contrast Sens. C	N	0.28					
Contrast Sens. D	N	0.26					
Lanthony CI/Left[5]	-	0.22			[5]		
Lanthony CI/Right[5]	-	0.22			[5]		
Vibratron/Index[5]	Sq. Rt.	0.36			[5]		
Vibratron/Small[5]	worse	0.41			[5]		
Motor							
Dynam./Strength[5]	N	0.19			[5]	•	
Dynam./Fatigue[5]	N	0.12			[5]		
SRT/mean RT	worse	0.16	•		•		
NES SRT/mn latency	log	0.16	•		•		
NES Tap/trial 1	N	0.27	•			•	
NES Tap/trial 2	N	0.18			[5]		
NES Tapping/alt[4]			•				
Santa Ana/Pref	N	0.15	•				
Santa Ana/Nonpref	SqRoot	0.19	•		•		
Aiming/correct	N	0.33	•		•		
Pegboard/Right	N	0.14				•	
Pegboard/Left	SqRoot	0.15					
Pegboard/Both	SqRoot	0.15	•		•		
Cognitive							
Benton/correct	SqRoot	0.13	•		•		
Digit Span/Forward	N	0.29	•		•		
Digit Span/Back	N	0.35	•		•		
Digit Symbol	N	0.20	•		•		
NES Sym Dig/best	log	0.26	•		•		
NES SymDig/Recall[5]	N	0.47			[5]		
NES Ser. Digit Learn	N	1.09	•		•		
Raven/errors[5]	N	0.51	•		[5]		
Affect							
NES Mood/Tension[6]	N	0.30					
NES Mood/Depress[6]	N	0.39					
NES Mood/Anger[6]	N	0.40					
NES Mood/Fatigue[6]	N	0.27					
NES Mood/Confus[6]	N	0.32					
Vocabulary							
NES Vocabulary[6]	log	0.28	•		•		
WAIS (scaled)[7]	N	0.25	•		•		[8]
Peabody[7]	N	0.17	•		•		[8]

[1] (•) = Accounts for more than 5% of variance on test performance.
[2] Skew, kurtosis ≤1 = Normal (N); best transformation listed if not normal.
[3] Ratio from male majority subject data; ratios similar for other populations.
[4] Data from minority subjects only. [5] Data from majority subjects only.
[6] Test not administered to Latins. [7] Latins received Spanish-Language version.
[8] Test score adjusted for age.

When comparing or choosing tests for neurotoxicology research, the SD/mean ratio reveals the likelihood of detecting a difference if it exists in the population, or by extension how large a deficit would be detectable with a given number of subjects (Table 29, column 3). Ratios below 0.20-0.25 indicate tests that can detect 5-10% performance decrements with relatively small groups in cross-sectional studies. The information is included here to provide a single tabular reference for the test measures.

EDUCATION

The accuracy of the self-reported years of education measure in this study needs to be addressed. Schedulers were careful to not emphasize the years of education factor during the phone interview, in order not to steer responses to specific answers to questions about years of education. Until the last 100 subjects, no one was turned down because of education, rather study personnel targeted populations likely to have lower levels of education. People were accepted into the study even if they reported higher levels of education in order to not steer responses. Consequently, it is likely that responses to questions about years of education were accurate, although Examiners reported that many people with less than 12 years tended to be hesitant and appeared uncomfortable when responding to this question.

None of the vocabulary tests were ideal for studies of neurotoxicity. The Peabody is more objective and takes less time than the WAIS vocabulary test to administer and score, but both are susceptible to Examiner error. The NES Vocabulary test which correlates highly with both the WAIS and Peabody is not susceptible to Examiner error, but it has a non-normal distribution limits its use in parametric analyses.

Vocabulary tests correlate highly with intelligence (Lezak, 1995) and moderately with years of education. Vocabulary tests have been used as an index of intelligence or education during the analysis of data in neurotoxicology research. The psychological literature supports the relationship with intellience, although education is clearly positively correlated with intelligence. They are often considered a "negative control" test (a test affected only by very large concentration exposures to neurotoxic chemicals and therefore resistant to exposure concentrations encountered in modern industry), a proposition not addressed here. Other tests also correlate well with years of education, most notably the Digit Symbol (r=0.61) and Symbol Digit (r=0.56) which have consistently discriminated between groups exposed to neurotoxic chemicals and those not so exposed (Anger, 1990). It study provides strong evidence that education influences performance on cognitive tests. This suggests that years of education is an important variable to control in designing epidemiological studies, if possible, and is essential for analyzing study results involving cognitive tests. It is seldom reported in published reports on human neurotoxicology.

Urban vs. Rural Education (Majority Subjects)
People educated in urban schools performed as well as people educated in rural schools. To provide a more fine-grained view of rural/urban education, regression plots were generated for each test against the population in the town in which majority subjects were educated on test performance. Due to the wide range of population sizes, population data were log normally transformed (Figure 19). There was virtually no impact of the size of the town in which individuals were educated. Simple reaction time, Raven errors, and NES vocabulary test performance depict typical results.

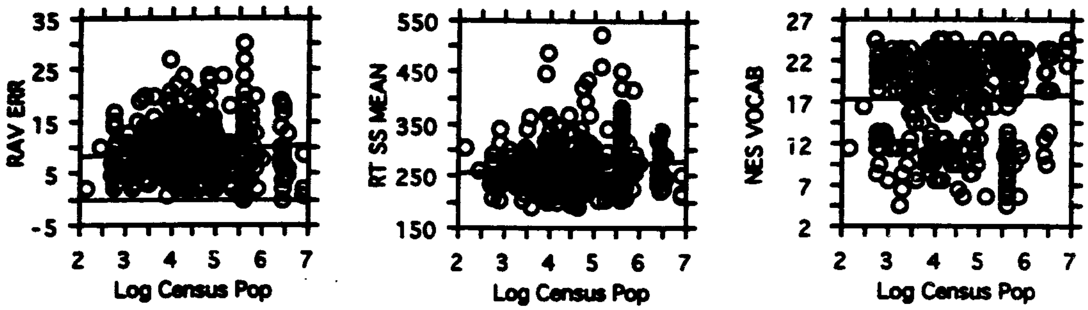


Figure 19. Regression plots of population in the town in which subjects were educated (log normally transformed) against Raven Errors, mean Simple Reaction Time, and NFS vocabulary test performance in majority subjects.

This fine-grained view points to even less of an effect than found in the three-group (< 16,000; 16,000-349,999; 350,000 and up) analysis (in the various sections above), and substitution of different urban/rural break points in the data analysis (not documented here) revealed that the differences were labile. These data suggest that a rurally-educated population in the US can be readily compared to an urban-educated US population in epidemiologic studies using the tests employed here, although caution is indicated for some tests.

CULTURAL GROUP

Substantial differences were seen between performance by the Latin subjects and other cultural groups in the cognitive tests, and sporadic differences were seen between other cultural groups studied. Cultural group is an important factor that must be balanced or, less preferably, statistically controlled in epidemiologic research. The complex relationship between years of education and cultural groups further supports the need to treat such groups differently. This must be qualified by the fact that the Latin American subjects in this study were educated in Mexico and test instructions were given in Spanish. Since the Latin data had a significant impact on the outcome of the analyses, the conclusions regarding unique and shared variance may be restricted to the types of backgrounds studied here. The term "cultural" as opposed to "ethnic" group is used here to emphasize both genetic heritage and experience.

Since the tests used in this study and thus in the field of Behavioral Neurotoxicology were developed and standardized on western-European derivative (majority US) populations, it is not surprising that the majority group had performance scores as high as or higher than minorities on average and especially minority subjects educated outside the US (i.e., the Latin Americans tested).

GENDER
Clear cut performance differences between majority males and females were seen on a some motor tests, notably the grip strength test, in this study. Consistent gender differences of much smaller magnitude were seen in the cognitive tests. Gender balance is thus desirable in a study sample, or appropriate statistical control (a less desirable alternative) is needed when motor tests are used.

AGE
The age range in this study was fairly narrow (26-45 years) and did not include the ages in which declines were likely to begin in the tests employed here. Nonetheless, there were performance declines on some motor, cognitive, and vocabulary tests over the age range studied. This is a variable to balance in the design (especially if the age range is over 20 years as in this study), if possible, or to record and enter into the statistical analysis.

VOCABULARY TESTS
The Vocabulary tests are a special category of test. They are intended to reflect native intelligence in some studies, a factor believed unlikely to be affected by low-level chemical exposures. Because the vocabulary tests account for substantial variance on the cognitive tests, their inclusion in cross-sectional studies is warranted and recommended where the exposed and referent subjects are drawn from different sources or job types with differing educational requirements (as is often the case). Unfortunately, none of the tests employed here appear to be ideal for that purpose, as noted below.

The WAIS and Peabody are highly correlated and appear equally suited for epidemiologic studies of chemical exposure effects. The WAIS requires a trained investigator and the potential for even inadvertent bias is high (noted above was the discovery that one Examiner gave more response time to the subjects on the WAIS, thus giving them more time to surface the correct response) and the time to administer the test is considerably longer than the Peabody. However, the Peabody does not assess people toward the higher end of the intelligence spectrum as well as the WAIS. The NES vocabulary test would seem a desirable alternative due to its high correlation with the other tests (the WAIS remains the gold standard of vocabulary tests) and its shorter administration time due in part to its computerization. However, the NES Vocabulary test's bimodal distribution is a serious impediment to use in parametric analyses (e.g., ANOVA, ANCOVA) that is largely unexplored and would be a matter of controversy and challenge. It, therefore, cannot be recommended until potential impacts are analyzed. A good test of vocabulary is needed for this field.

POWER ANALYSES

Power analyses provide the information regarding the number of subjects needed to detect a given level of effect in a planned study. The ratio of the standard deviation to the mean for a test is the main datum on which the analysis is based. The size of the effect to be detected can then be factored into the equation, and the number of subjects required to be able to detect that effect is calculated (Tables 30 and 31, for females and males, respectively.

The tests in Tables 30 and 31 are ordered on the basis of the N required to detect a given effect size. The correspondence between the resulting orderin the two Tables is striking. Other factors need to be considered, however, as in the case of the Benton test which appears near the top of both tables, but has a very nonnormal distribution because such a high percentage of subjects make few errors on the test. This test may detect only substantial performance deficits.

Table 30. Power Analyses for Female Majority Subjects.

Measure	Gen	Mean	SD	SD/Mn	5%	10%	15%	20%	30%
PEG R	F	14.88	1.90	0.13	71	18	8	4	2
DYNA FTG	F	0.83	0.11	0.14	80	20	9	5	2
SANTA PREF	F	42.60	5.85	0.14	82	20	9	5	2
PEG L	F	14.23	1.96	0.14	82	20	9	5	2
BNTN TOT	F	8.74	1.31	0.15	97	24	11	6	3
PEG B	F	24.20	3.80	0.16	107	27	12	7	3
SANTA N PREF	F	39.20	6.24	0.16	110	27	12	7	3
TAP TWO	F	162	27	0.17	122	30	14	8	3
RT SS MEAN	F	278	48	0.17	130	33	14	8	4
RT NES AVE	F	289	52	0.18	140	35	16	9	4
PEAB STD	F	95.65	18.61	0.19	164	41	18	10	5
DIGIT SYMB	F	62.76	12.25	0.20	165	41	18	10	5
LANTH CI L	F	1.21	0.25	0.21	185	46	21	12	5
DYNA STR	F	28.59	6.25	0.22	207	52	23	13	6
LANTH CI R	F	1.26	0.28	0.22	212	53	24	13	6
SYMBOL DIG BST	F	2.00	0.47	0.23	238	60	26	15	7
WAIS SCLD	F	8.39	2.25	0.27	312	78	35	19	9
TAP ONE	F	154	41	0.27	314	79	35	20	9
MOOD TENSION	F	2.90	0.83	0.29	353	88	39	22	10
MOOD FATGUE	F	2.91	0.83	0.29	354	89	39	22	10
DIG SPN F	F	7.72	2.32	0.30	389	97	43	24	11
NES VOCAB	F	17.93	5.38	0.30	390	98	43	24	11
MOOD CONFUS	F	2.42	0.76	0.31	421	105	47	26	12
DIG SPN R	F	6.22	2.17	0.35	528	132	59	33	15
MOOD ANGER	F	2.16	0.81	0.37	605	151	67	38	17
VIBRO INDEX	F	1.42	0.58	0.41	723	181	80	45	20
MOOD DEPRESS	F	2.24	0.92	0.41	728	182	81	46	20
VIBRO SMALL	F	1.40	0.61	0.43	808	202	90	50	22
SYMBOL DIG RCL	F	3.81	2.15	0.56	1379	345	153	86	38
RAV ERR	F	9.61	5.70	0.59	1523	381	169	95	42
DIGIT LEARN	F	3.73	4.10	1.10	5223	1306	580	326	145

Table 31. Power Analyses for Male Majority Subjects.

Measure	Gen	Mean	SD	SD/Mn	5%	10%	15%	20%	30%
DYNA FTG	M	0.86	0.10	0.12	62	16	7	4	2
BNTN TOT	M	8.86	1.14	0.13	72	18	8	4	2
PEG R	M	13.86	1.98	0.14	88	22	10	6	2
PEG L	M	13.58	2.03	0.15	97	24	11	6	3
PEG B	M	22.74	3.42	0.15	98	25	11	6	3
SANTA PREF	M	42.82	6.45	0.15	98	25	11	6	3
RT NES AVE	M	272	43	0.16	106	26	12	7	3
RT SS MEAN	M	257	42	0.16	113	28	13	7	3
PEAB STD	M	101.25	16.90	0.17	121	30	13	8	3
TAP TWO	M	178	32	0.18	141	35	16	9	4
SANTA N PREF	M	40.01	7.49	0.19	152	38	17	9	4
DYNA STR	M	43.27	8.38	0.19	163	41	18	10	5
DIGIT SYMB	M	55.76	10.91	0.20	166	41	18	10	5
LANTH CI L	M	1.24	0.27	0.22	209	52	23	13	6
LANTH CI R	M	1.30	0.29	0.22	214	53	24	13	6
WAIS SCLD	M	8.73	2.15	0.25	263	66	29	16	7
SYMBOL DIG BST	M	2.22	0.57	0.26	283	71	31	18	8
TAP ONE	M	173	46	0.26	304	76	34	19	8
MOOD FATGUE	M	2.73	0.74	0.27	314	79	35	20	9
NES VOCAB	M	18.23	5.15	0.28	346	87	38	22	10
DIG SPN F	M	8.06	2.34	0.29	365	91	41	23	10
MOOD TENSION	M	2.66	0.81	0.30	402	100	45	25	11
MOOD CONFUS	M	2.32	0.74	0.32	447	112	50	28	12
DIG SPN R	M	6.92	2.41	0.35	525	131	58	33	15
VIBRO INDEX	M	1.49	0.54	0.36	569	142	63	36	16
MOOD DEPRESS	M	2.20	0.85	0.39	650	162	72	41	18
MOOD ANGER	M	1.98	0.79	0.40	681	170	76	43	19
VIBRO SMALL	M	1.49	0.61	0.41	717	179	80	45	20
SYMBOL DIG RCL	M	4.46	2.08	0.47	947	237	105	59	26
RAV ERR	M	9.46	4.82	0.51	1123	281	125	70	31
DIGIT LEARN	M	3.37	3.67	1.09	5153	1288	573	322	143

CONCLUSIONS/RECOMMENDATIONS

Several conclusions can be drawn and recommendations offered:

- The influence of education and cultural group on cognitive test performance is so intertwined that they cannot be controlled statistically. If different cultural groups and a range of educational attainment can be anticipated in an epidemiologic study, at least one or preferably both of these factors must be balanced in the experimental design and subject recruitment if cognitive tests are planned. This is less important for motor, sensory, or affect tests. Inclusion of both education and cultural group as factors in the data analysis is essential for obtaining accurate results.

- Gender had an impact on many tests studied here, although the variance accounted for by this factor was small except in the case of strength testing where male and female data must be treated separately. Gender balance in study designs and inclusion as a factor in the statistical analyses is essential if adverse effects are to be evaluated in both males and females.

- Age through the range of 26-45 years has a detectable impact on several behavioral tests. This factor should be included in subject recruitment or sampling plans and in data analyses, although the amount of variance accounted for by this factor was small.

- Many of the tests included here do not have normal distributions. A search for other established tests of the same functions but which have improved psychometric properties including more normal distributions, is recommended.

- For neurotoxicity research, tests with a standard deviation to mean ratio greater than 0.5 should not be used (including the Raven, NES Serial Digit Learning and NES Symbol Digit Recall), or improved protocols are needed.

• Subject reports of numbness and tingling were associated with performance declines on a number of tests. This suggests an underlying neurologic factor that has not been identified here. This and other questions in the NES pre-test questionnaire need careful reformulation to provide interpretable answers.

• People reporting previous institutionalization for drug or alcohol abuse may have degraded performance on some neurobehavioral tests, and this should be tracked by research protocols.

Section 5
REFERENCES

Amler, R.W., J.A. Lybarger, W.K. Anger, B.L. Phifer, W. Chappell and L.J. Hutchinson. 1994. Adoption of an Adult Environmental Neurobehavioral Test Battery (AENTB). *Neurotoxicology and Teratology*. 16: 525-530.

Anger, WK. 1990. Worksite behavioral research: Results, sensitive methods, test batteries and the transition from laboratory data to human health. *NeuroToxicology*. 11: 629-720.

Anger, W.K. and M.G. Cassitto. 1993. Report on Workshop: WHO NCTB and other neurobehavioral test batteries. *Environmental Research*. 60: 84-86.

Anger, W.K., M.G. Cassitto, Y-x. Liang, R. Amador, J. Hooisma, D.W. Chrislip, D. Mergler, M. Keifer and J. Hortnagl. 1993. Comparison of performance from three continents on the WHO-recommended Neurobehavioral Core Test Battery (NCTB). *Environmental Research*. 62: 125-147.

Anger, W.K., R. Letz, D.W. Chrislip, H. Frumkin, K. Hudnell, J.M. Russo, W. Chappell, and L. Hutchinson. 1994. Neurobehavioral test methods for environmental health studies of adults. *Neurotoxicology and Teratology*. 16: 489-497.

Anger, W.K., and B.L. Johnson. 1992. Human behavioral neurotoxicology: Workplace and community assessments. In: W. Rom, ed. *Environmental and Occupational Medicine*, 2nd ed. Little, Brown, and Co., Boston, MA (USA). pp. 573-592.

Hutchinson L.J., R.W. Amler, J.A. Lybarger and W. Chappell. 1992. Neurobehvioral Test Batteries for Use in Environmental Health Field Studies. Atlanta, GA: Agency for Toxic Substances and Disease Registry; December.

Baker E.L., R. Letz and A. Fidler. 1985. A computer-administered neurobehavioral evaluation system for occupational and environmental epidemiology. Rationale, methodology, and pilot study results. *Journal of Occupational Medicine*. 27: 206-212.

Cassitto, M.G., D. Camerino, H. Hanninen and W.K. Anger. 1990. International collaboration to evaluate the WHO Neurobehavioral Core Test Battery. In: B.L. Johnson, W.K. Anger, A. Durao and C. Xintaras, eds. *Advances in Neurobehavioral Toxicology: Applications in Environmental and Occupational Health.* Lewis Publishers, Chelsea, MI (USA). pp. 203-224.

Johnson, B.L. and W.K. Anger. 1991. Behavioral toxicology. In: W.R. Rom, ed. *Environmental and Occupational Medicine*. Little, Brown, and Co., Boston, MA (USA). pp. 329-350.

Johnson, B.L., W.K. Anger, A. Durao and C. Xintaras. 1990. *Advances in Neurobehavioral Toxicology: Applications in Environmental and Occupational Health.* Lewis Publishers, Chelsea, MI (USA).

Johnson, B.L., E.L. Baker, M. El Batawi, R. Gilioli, H. Hanninen, A.M. Seppalainen and C. Xintaras. 1987. *Prevention of Neurotoxic Illness in Working Populations.* John Wiley and Sons, New York, NY (USA).

Letz, R. 1990. The neurobehavioral evaluation system: An international effort. In: B.L. Johnson, W.K. Anger, A. Durao and C. Xintaras, eds. *Advances in Neurobehavioral Toxicology: Applications in Environmental and Occupational Health.* Lewis Publishers, Chelsea, MI (USA). pp. 189-201.

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

Liang, Y-x., Z-q. Chen, R-k. Sun, Y-f. Fang and J-h. Yu. 1990. Application of the WHO Neurobehavioral Core Test Battery and Other Neurobehavioral Screening Methods. In: B.L. Johnson, W.K. Anger, A. Durao, and C. Xintaras, eds. *Advances in Neurobehavioral Toxicology: Applications in Environmental and Occupational Health.* Lewis Publishers, Chelsea, MI (USA). pp. 225-243.

Section 6
APPENDICES

Appendix A
SUBJECT OCCUPATIONS, NATIVE AMERICAN TRIBAL REPRESENTATION, AND RECRUITMENT.

OCCUPATIONS OF SUBJECTS

The occupations reported frequently by subjects were: Carpenter, cashier, nurse's aide, child care provider, clerk, construction (general), cook, custodian, driver, farmer, field hand, fisher, homemaker, laborer, landscaper, machinist, mechanic, mill worker, painter, roofer, sales, secretary, student, waiter/waitress, warehouse worker, welder. Approximately 30 subjects indicated they were unemployed and could not identify an occupation. There were 30 painters, although most of them were from the San Francisco population of day laborers who had emigrated from Mexico where they had been a painter at one time. Eight people identified their occupation as welder. The remaining occupations did not have obvious links with neurotoxic chemicals, although such exposures cannot be ruled out. The distribution of subjects employed full time vs. part time in these occupations is not known.

SUBJECT RECRUITMENT/TEST SITES

European-Descent Majority Subjects (Rural and Urban)

European-descent (Majority) subjects were sought in Oregon, which has a low percentage of minorities. Rural subjects were targeted in Oregon counties in which no town had a population of more than 5000 people, and urban subjects were targeted in Portland (population approximately 370,000 in the central urban area during the years subjects were attending school) and adjacent areas including Salem (population 89,000) and Springfield (42,000). Testing was advertised and conducted in rural Oregon counties with a large number of residents with less than 9 years of education or a high density of Native American Indians, including those with less than 9 years of education. The largest number of people in the study were tested at three sites in Portland, plus adjacent towns of Carver and Salem.

Enterprise (northeast Oregon) The study was advertised in the county (Wallowa Chieftan) and local (Enterprise) newspapers, and flyers were distributed to numerous stores. Testing was conducted in a 12' by 15' suite at a centrally-located motel.

Roseburg (southwest Oregon) Subjects were recruited by a display advertisement in the local (Roseburg News Review) newspaper, 60-second advertisements on a Roseburg country music station, and bulletins placed at the Roseburg Family Head Start program offices and at Umpqua Community College GED program. In addition, 70 letters were directed to Project Literacy tutors, although mailing was not confirmed and may have yielded no subjects. Testing was conducted in a 5' by 7' office at centrally-located Douglas Community Hospital.

Portland (northwest Oregon)/OHSU Advertisements were placed in the major local newspaper (Oregonian) to recruit subjects. Testing was conducted in a 12'x15' room and two 10' x 12' test rooms at Oregon Health Sciences University (OHSU).

Portland (northwest Oregon)/PCC Recruitment presentations were given in six Adult Literacy classes (approximately 10-12 people per class) at the Southeast campus of Portland Community College (PCC). This PCC campus is a major center for adult basic education and literacy programs. Testing was conducted in available classrooms and small offices at OHSU and PCC.

Carver, OR (northeastern Portland border) Contacts were made with six Portland Head Start offices (4 in Multnomah County and 2 in Clackamas County) to advertise the study. Contacts were made with Head Start programs (a program which includes weekly or monthly parent contact through home visits) at a group meeting, and flyers were distributed to approximately 21 home visitors (teachers), each of whom visits the homes of 20-30 children in any given month. Home visitors were asked to distribute the flyers in each household where they knew at least one household member had less than 10 years of education. Testing was conducted in a 4' x 8' hearing screening room at a centrally-located Head Start school in Carver, OR.

Salem (45 miles south of Portland). Residents and drop-ins to a Salvation Army building were recruited through flyers and suggestions from administrative personnel. Testing was conducted in a 10' by 12' room used by shelter personnel to eat lunch.

Springfield (west-central Oregon). Advertisements were placed in the local newspaper, the Springfield unemployment office, and a country music radio station to recruit subjects. The test site was a 6' by 10' office in a hospital located in the area of town inhabited by a large number of poorly educated people (according to census data).

African Americans

Most African American subjects were recruited and tested in Atlanta, GA, where employment agencies provided the main source of subject recruitment. Day employment or short-term employment agencies located adjacent to or in high density African-American communities arranged access to the subjects. In addition, migrant African-American watermelon harvesters near Atlanta were sought to locate subjects with lower levels of education.

Atlanta. Conference rooms at an employment agency, a county services building, and a nonprofit residential shelter for homeless and other populations were employed in Atlanta.

Portland (northwest Oregon)/PCC. A small African American group was tested at the Southeast Center of PCC. The test room could accommodate approximately 50 people, although only the Examiner and subject were present in the room during testing.

Latin (Mexican) Americans

Employment agencies were the main source of subjects in San Francisco. Day employment or short-term employment agencies close to the relevant communities supplied access to the subjects. Most subjects in San Francisco were tested in two 6' x 6' rooms at San Francisco State University. A small number were tested in examination rooms in residential clinic offices.

Native American Indians

Approaches to various Native American representative groups were carried out by study principals and consultants. The groups were: (1) Confederated Tribes of Oregon (primarily Paiutes and Umpquas) at the Warm Springs Reservation (by consultant Dr. Connie Hunt); (2) Klamath Tribe of Klamath, OR (by consultant Dr. Connie Hunt); (3) Celilo Falls, a rural Native American area near The Dalles, OR; (4) Madras (OR) where advertisements were placed in the newspaper and radio stations (testing was cancelled due to lack of response); (5) Siletz tribal offices in Siletz, OR (near the Pacific coast of central Oregon) where advertisements were placed in their newsletter and flyers were distributed to several establishments during a "pow wow" prior to scheduled testing (testing was cancelled due to a lack of response); (6) the administrative offices of the Cow Creek Band of the Umpqua tribe, leaving advertising flyers (prior to testing in nearby Roseburg); (7) Confederated Tribes of Siletz (Portland branch) adult education program to initiate advertisements through their offices (testing was conducted at OHSU); (8) Portland Indian Clinic (flyers posted); (9) North American Rehabilitation Association; (9) American Indian Association of Portland (promised newspaper advertisement was not placed); (10) Northwest Portland Indian Health Board; (11) Inter-Tribal Fish Commission (flyers sent); (12) Tahana White Crow of Salem (flyers distributed); (13) Salem Plasma Center; (14) Salem SOS Club; (15) Stayton Center for Referral (Salem area); (16) North Salem Adult and Family Services; (17) Salem branch of Confederated Tribes of Siletz; (18) Salem Veterans' Center; (19) Salem Salvation Army; (20) Public Health Service Indian Health Center (Chemawa area of Salem); (21) Grand Ronde Tribal Human

Services Division of Grand Ronde, OR (declined to participate); (22) Medicine Wheel Clinic of Confederated Tribes of Siletz (Siletz, OR); (23) Warm Springs Human Services Clinical Manager (declined participation due to lack of direct benefit to population); and (24) Yakima Indian Agency of Toppenish, WA.

The larger tribes, with the exception of the widely distributed Siletz group, declined participation in the study due to the lack of direct benefit to the tribe or their members and concerns that American Indians were often over-studied. The tribal contacts also refused to allow posting of advertisements in their reservation, including their publications, or any other form of contact with tribe members.

Native American Indians who were tested in this project responded to Oregonian advertisements (Portland's and the State's largest newspaper) and flyers posted in health clinics for Native American Indians or administrative offices for the Siletz and smaller tribes. Most Native American Indians tested in this study lived in cities rather than on reservations. Their reported primary tribal background represented a broad range of heritage or affiliation: Cherokee (2); Chippewa (1); Gabriolands (1); Gila River (1); Klamath (2); Klickitat (1); Mohawk (1); Northern Shian; (1); Red Lake (1); Siletz (3); Sioux (3); Sisseton (1); Warm Springs (1); Yakima (2); Not asked/not known (29).

Portland (northeast Oregon)/OHSU.Most subjects were tested at OHSU, described above, although a few Native American Indians were tested in the other locations around the state.

The Dalles (north-central OR) Testing was carried out in a 12' by 20' room at a church near the Celilo center (see 3, above).

Subjects with Less than 6 Years of Education

The only subjects with less than 6 years of education were Mexican immigrants in the Latin population described above. After an extensive search and contacts throughout Atlanta and neighboring parts of Georgia, and similar approaches in Oregon, it is clear that subjects in the under-6 years education group are not readily available or at least not willing to be tested. The most promising opportunity was at a prison where people with defined levels of education are incarcerated. However, the OHSU Human Research Committee has adopted, for all OHSU research, the Department of Health and Human Services policy of allowing testing of prisoners only if the hypothesis being studied has a direct benefit to the life of people who are prisoners. This was not the case, and we were unable to pursue this approach.

Appendix B
DATA DESCRIPTION FOR EACH TEST

ACUITY TEST (OPTEC model 1000CS)
This test measures visual acuity by presentation of successively smaller sets of letters. The line number on which the Participant correctly identifies every letter is the primary measure for this test. The further down on the scale (i.e., the smaller the letters or the larger the line number) the Subject can read, the better their vision. In conventional terms, perfect performance on the respective test rows is equivalent to: (1) 20/200; (2) 20/100; (3) 20/70; (4) 20/50; (5) 20/40; (6) 20/30; (7) 20/20.

The highest possible score is 7; larger numbers reflect better performance. This datum is treated as being measured on an interval scale.

CONTRAST SENSITIVITY TEST (OPTEC model 1000CS)
This test measures visual acuity by determining the threshold level at which the Subject can discriminate the lines within visual gratings. The lines vary in contrast and width in five sets of gratings (A through E). The two most sensitive measures of visual acuity are the thresholds on sets C and D. Those data were analyzed for this study.

An upper limit was established for each group of gratings. In most instances this was relatively straightforward since the Subjects typically answer correctly to the level of their ability and then label each subsequent circle as "blank." A single-digit score is developed for each set of gratings; this consists of the last (i.e., highest number) correct choice. In those cases in which the Subject made an incorrect response and it was followed by one correct response, the threshold was identified as the grating just preceding the incorrect response. If two consecutive correct responses followed the incorrect response, the threshold was established by the next error (viz., the last correct response).

Since the last circle in each group of gratings is blank, the highest possible score for any set of gratings is 8. Larger numbers (more gratings correctly identified) reflect greater visual acuity. These data are measured on an interval scale.

FARNSWORTH D-15 COLOR VISION TEST (DESATURATED)
This test measures color vision by having the Subject arrange colored chips in order of color similarity. Two data points were developed for each eye. A computerized scoring program was implemented by Dr. Sizemore to establish the "confusion index" (CI) regarding cap placement errors. Normal color vision and several types of color vision abnormality were also identified. Color defects are deuteranopia (green blindness), protanopia (red blindness), tritanopia (blue blindness), and abnormal trichromatic vision (various combinations of color vision loss).

Lower CI scores indicate better performance, with a 1.0 score indicating no visual errors or "confusion." The CI are continuous data; the color defect categorization is a nominal scale.

VIBROTACTILE THRESHOLD TEST
A modified Method of Limits procedure was used to determine the Subject's threshold. The subject was presented with vibration amplitudes of 6.0, 5.0, 4.0, 3.0, 2.5, 2.0, 1.5, 1.0, 0.5 until the Participant responded "no[t felt]" two times in a row. This was followed by an ascending trial using the same steps until two yes replies were made in a row. (The starting point was modified after trials 1 and 2 to be closer to the threshold level.)

Subjects received five trials for the index finger and five trials for the small finger. Trial 1 and the two most extreme of the remaining four trials were discarded, and the mean of the two remaining trials became the threshold score.

Smaller numbers reflect lower thresholds or better performance and thus greater sensitivity. These data were treated as measured on an interval scale.

NCTB SRT
The Subject presses a button as rapidly as possible after a light appeared for 64 trials (light presentations). The mean of the 64 response times (time between light onset and a button press) defines the mean reaction time.

Lower scores represent faster or better performance. The data are measured on an interval scale.

NCTB DIGIT SPAN
Digit strings are read to the Subject who responds by repeating the numbers in the order presented (forward) or in the reverse of that order (reverse). Strings of increasing length from 3 numbers long (2 for reverse trials) are presented to the subject until the subject incorrectly repeats two digit strings of the same length. Each digit string repeated without error is scored as 1 point.

Higher scores represent better performance or more digit strings remembered. The test is intended to continue until the subject fails, and the upper limit is 14 on both forward and reverse trials. These data are measured on an interval scale.

SANTA ANA
The subject rotates cylindrical pegs (with square bases located in square holes) 180 degrees as fast as they can for 4 30-second trials (two with the preferred and two with the nonpreferred hand). The total number of pegs turned 180 degrees on both trials from the preferred and nonpreferred constitute the score.

A larger number of pegs turned reflects greater coordination. The data are measured on an interval scale.

NCTB DIGIT SYMBOL
The subject writes, as quickly as possible, symbols below numbers on the basis of a code of digit-symbol pairings at the top of the page. The number of correct symbols written in 90 seconds is the datum for this test.

Larger numbers reflect more symbols and better performance. The data are measured on an interval scale.

NCTB BENTON VISUAL RETENTION TEST
The subject views a set of geometric symbols for 10 seconds and then selects the same set of symbols from four similar alternatives in a multiple choice format. There are 10 sets of symbols.

Larger numbers reflect more correct choices and thus better performance. Ten is the maximum score on this test. The data are measured on an interval scale.

NCTB AIMING II
Using a pencil, the subject places dots in 2-mm diameter circles as fast as possible consistent with accurate placement inside the line of the circles. Two 60-second trials are given. Total dots in the circles summed over the two trials and total dots touching the line or outside the circle are the data for this test. It is very sensitive to inter-rater differences. All Aiming data in this study were measured by one Examiner (JAG) during a 2-month period.

Larger numbers reflect more dots completed and therefore better performance. The data are measured on an interval scale.

GROOVED PEGBOARD
The subject removes 1-cm x 2-mm metal rods (termed pins) from a concave depression and places them in a row of 2-mm+ holes as fast as possible in 30 seconds. The first trial employed the right hand, the second trial the left hand, and the third trial was with both hands (simultaneously).

Larger numbers reflect more pins placed in the holes and therefore better performance. The data are measured on an interval scale.

DYNAMOMETER
The reading on the outer dial (KG) of the dynamometer served as the datum for each trial. The mean score from trials 1, 2, and 3 is the measure of strength. Trial 5 divided by Trial 1 serves as the measure of fatigue (fatigue ratios greater than 1.0 were scored as 1.0).

For the strength data, larger numbers reflect better performance (i.e., greater strength). For the fatigue ratio, larger numbers indicate less fatigue (with 1.0 indicating no fatigue and 0 indicating complete fatigue. These data are measured on an interval scale.

RAVEN PROGRESSIVE MATRICES
This test of conceptual reasoning presents a series of geometric forms with one missing form to be chosen from among several multiple choice options that complete the series. There are three sets of progresively difficult concepts. The sum of the errors from sets A, B, and C provide the primary datum for this test. There are 36 stimulus presentations, so the maximum number of errors is 36. The time to complete the 34 presentations after the first two "practice" trials is the second datum for this test.

Smaller numbers reflect fewer errors and less time to complete the test. These data are measured on an ordinal scale since all concepts are not equally difficult. However, an argument could be made that an interval scale is achieved, and the data were so treated here.

SYMBOL-DIGIT (NES)
This is the analogue of the Digit-Symbol test in which the subject must type numbers associated with symbols in a code of symbol-digit pairings. There are one practice and five test trials. From the test trials, the "Best Two Trials" in the NES summary program are selected as the mean latency (time to complete the tests) from the two fastest trials.

Smaller numbers reflect better (faster) performance. These data are measured on an interval scale.

TAPPING TEST (NES)
This test requests the subject to press one or two buttons as fast as possible for 30 seconds. Majority subjects received two trials with their preferred hand. Minority subjects received four trials: (1) trial 1 with their preferred hand; (2) trial 2 with their nonpreferred hand; (3) trial 3 with their preferred hand [approximately equivalent to trial 2 in the majority subjects]; and (4) trial 4 alternating between left and right hands. Total taps on trials 1 and 2 (Majority) or 1-4 (minority) are the data for this test.

Larger numbers reflect better (faster) performance. These data are measured on an interval scale.

SIMPLE REACTION TIME TEST (NES)
This is equivalent to the NCTB SRT test, but its duration is 10 minutes. The data are latencies (light onset to response). Blocks of 15 responses are recorded. Block 1 is disregarded; the mean of blocks 2-9 is a measure of response speed. The duration of the test for Majority subjects was slightly longer than for minority subjects in that the Majority (AENTB) protocol called for an additional block of 15 trials to reach the desired minimum test duration of 10 minutes. Responses slower than an upper limit of 1 second for minority and 3 seconds for majority Subjects were not included in the calculation of the mean. This is not believed to affect comparability of the data between majority and minority subjects.

Smaller numbers reflect better (faster) performance. The data are measured on an interval scale.

SERIAL DIGIT LEARNING (NES)
The subject is presented with a series of 8 digits and asked to type them into the computer in the order in which they saw them. The same series of digits is presented for 8 trials, or until the subject types them back correctly on two consecutive trials. The datum from this test consists of one "overall error score." It is proportional to the absolute number of errors on the 8 trials. The score can range from 0 to 16; 0 means the Participant remembered all 8 digits correctly on the first try (and on a repeat trial), and 16 means the Participant failed to accurately replicate the sequence of digits (which would be logged as 2 errors per trial) during the test.

Smaller numbers reflect less errors and better performance. This datum can be treated as if they were from an interval scale.

SYMBOL-DIGIT DELAYED RECALL TEST
An additional trial of the Symbol-Digit test is presented, but without the code of symbol-digit pairings, for the subject to type in as many of the digits associated with the symbols as they can recall. The number of correct digits served as the datum for this test. Nine is the largest number possible on this test.

Larger numbers reflect better recall (more symbol-digit combinations recalled). The data are measured on an interval scale.

VOCABULARY TEST (NES)
The subject is presented with 25 common English words (unless an error criterion is met) and asked to pick the correct response from among four definitions. The NES summary program provides a normalized score (the calculated score if the subject had answered all 25 words), the datum for this test.

Larger numbers reflect knowledge of the definition of more words (*i.e.*, better performance). Strictly speaking, this is measured on an ordinal scale (all words are not equally easy to define), but the measurement scale is frequently considered to qualify as interval or higher. It was so treated here.

MOOD TEST (NES)
Adjectives descriptive of various mood states are presented to the subject who is to select one of five alternatives (from not at all to extremely) that best indicates how the subject has been feeling recently. The score for each of five sub-scales is calculated by the NES summary program based on previous research on the original Profile of Mood States.

Higher scores indicate a greater degree of dysfunction. The data are measured on an ordinal scale, but can be treated as interval. They are treated as an interval measurement scale in the analysis.

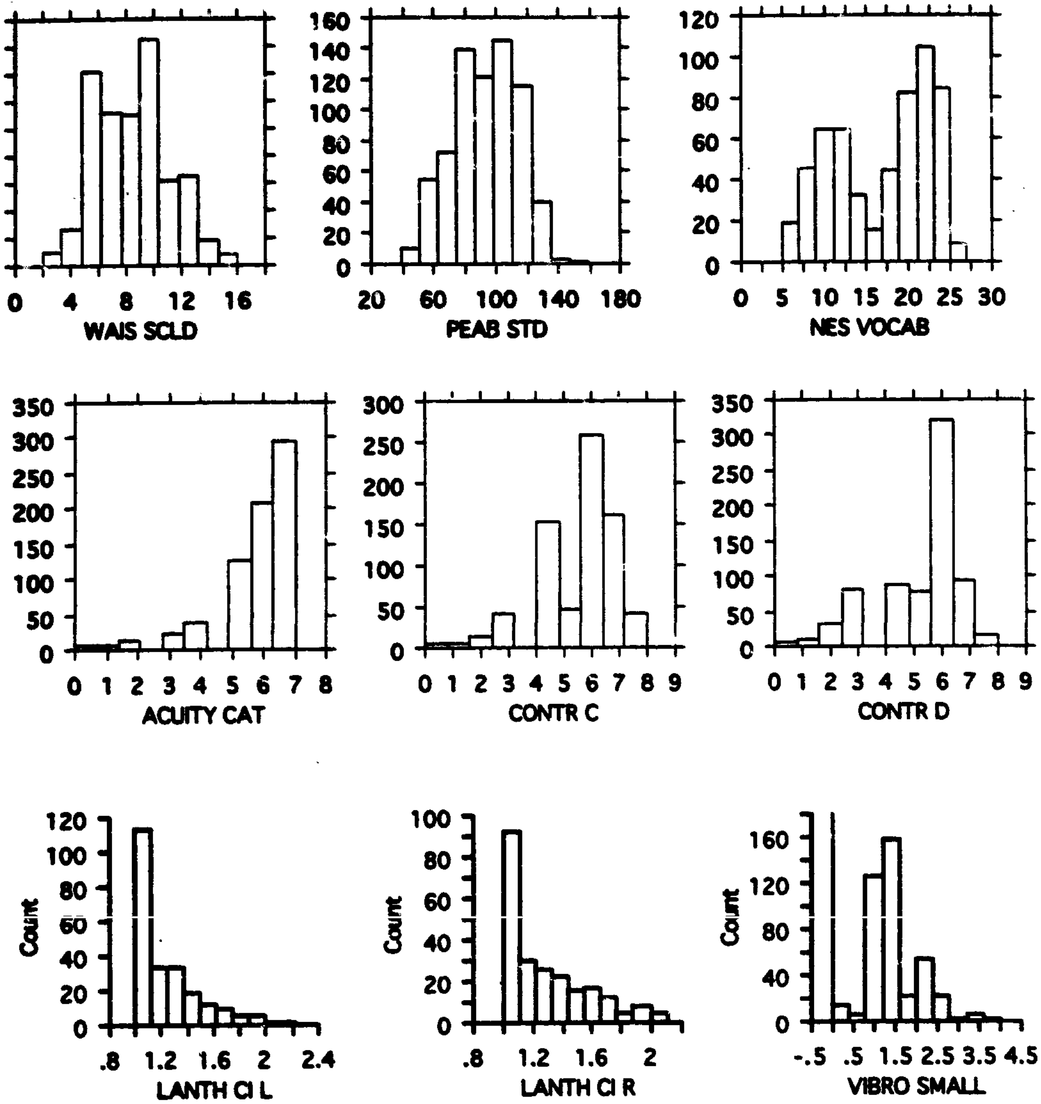
WAIS R VOCABULARY TEST
The subject is asked to define a series of words of increasing difficulty until a fixed number of errors occurs. The Examiner records the responses for subsequent evaluation against a standardized set of definition elements developed as part of the Wechsler Adult Intelligence Scale (WAIS). The Spanish language test differs from the English language test and is thus not strictly comparable. One Examiner (JG) evaluated all the English-language tests, while another Examiner (FC) evaluated all the Spanish-language tests (all Latin subjects). Each definition is scored on a scale of 0, 1 or 2 and the points are summed and scaled in accordance with the WAIS protocol, which included an adjustment for age.

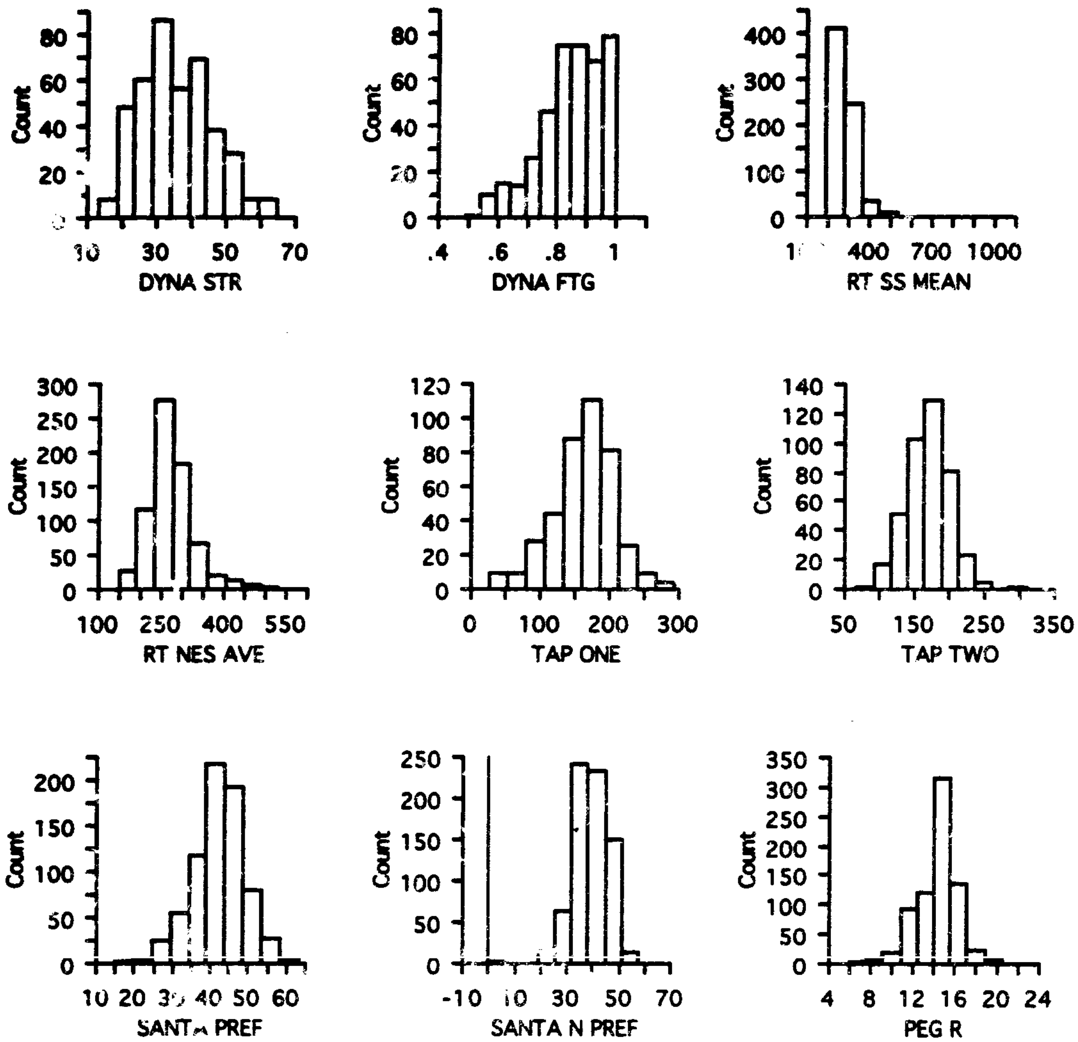
Larger scores reflect more correct words and are therefore indicative of a greater vocabulary. This scaled score is treated as if it were measured on an interval scale in the analysis.

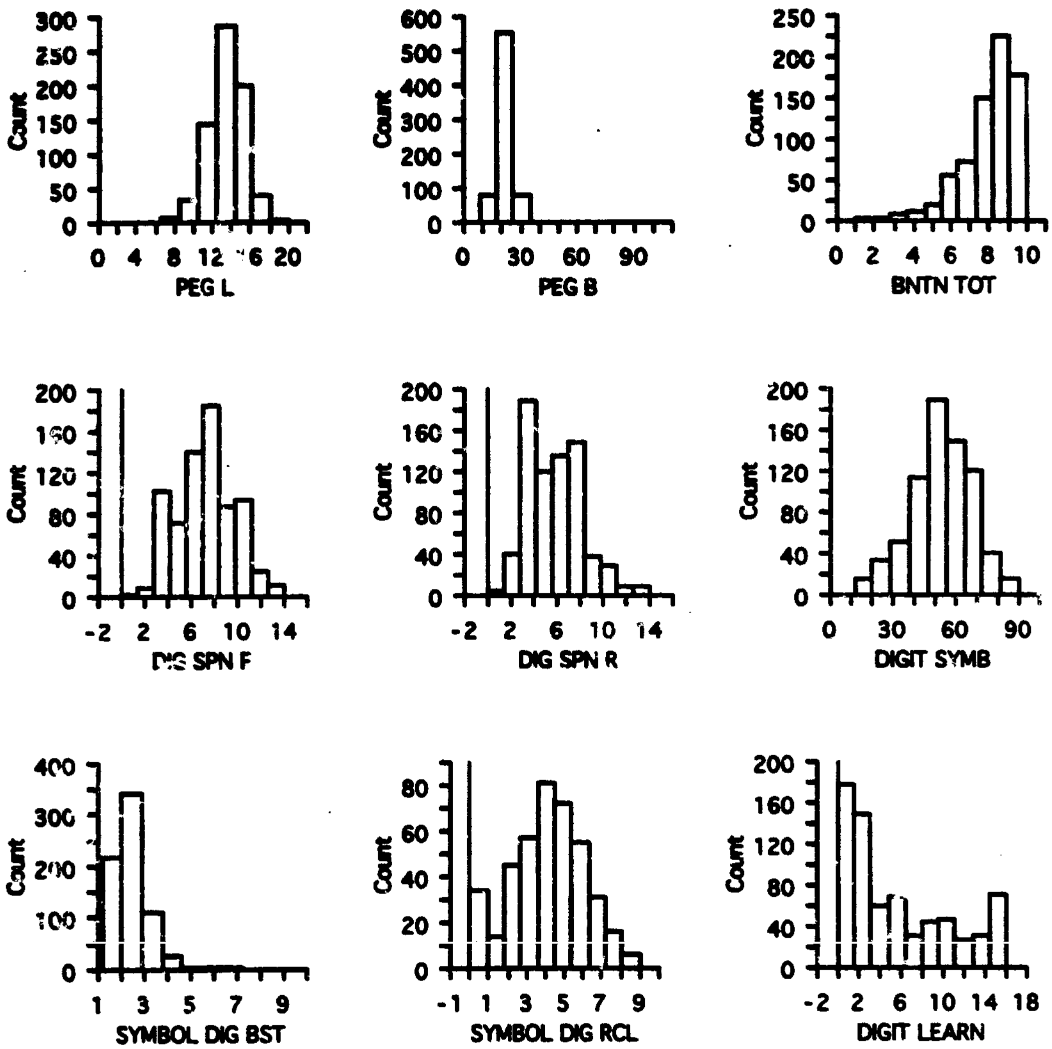
PEABODY PICTURE VOCABULARY TEST (PPVT)
An English (Spanish for Latin Americans) word is spoken to the Subject, who is asked to pick the one of four pictures that most closely represents the word. The Examiner first establishes a basal score (the lowest point of 8 consecutive answers), then a ceiling score (highest point of a series in which at least 6 of 8 are correct). All errors are subtracted from the ceiling score to produce the raw score. The raw scores are standardized in accordance with the PPVT protocol. The test relies on speech frequency to establish the hierarchical word list. Thus, Spanish speaking subjects saw a different set of pictures than English-speaking counterparts. The former is not as well as developed as the English version and resulted in a larger number of unuseable results. The data are adjusted for age.

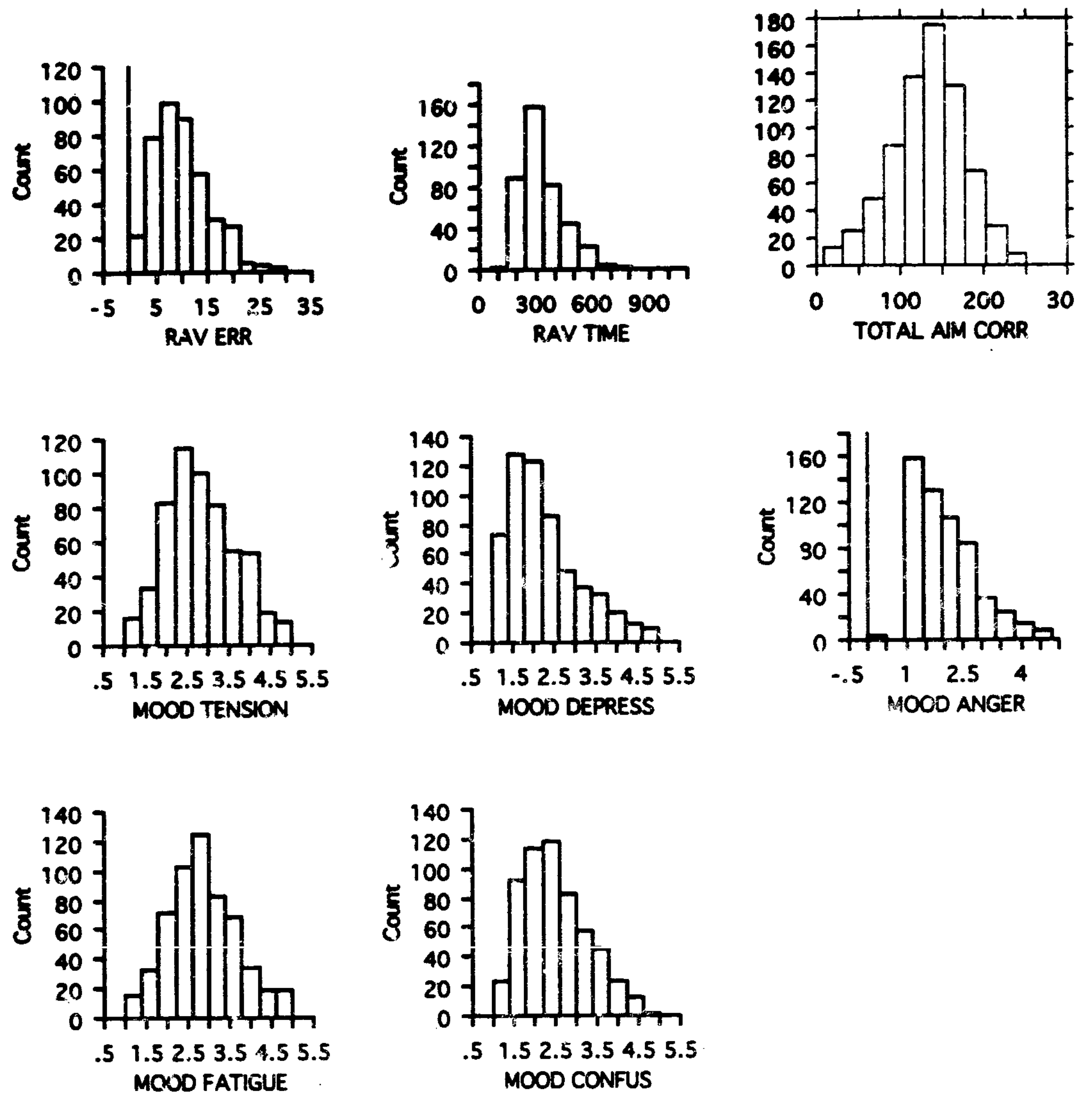
Larger scores represent more words correctly chosen and thus a better vocabulary. These data are treated as interval in the analysis.

Appendix C
DISTRIBUTIONS OF TEST RESULTS FROM ALL SUBJECTS TESTED




Count
80
60
40
20
30
50
70
DYNA STR
Count
80
60
40
20
0
.4
.6
.8
1
DYNA FTG
Count
400
300
200
100
0
400
700
1000
RT SS MEAN
Count
300
250
200
150
100
50
0
100
250
400
550
RT NES AVE
Count
100
80
60
40
20
0
0
100
200
300
TAP ONE
Count
140
120
100
80
60
40
20
0
50
150
250
350
TAP TWO
Count
200
150
100
50
0
20
40
50
60
PREF
Count
200
150
100
50
0
-10
30
50
70
SANTA N PREF
Count
350
300
250
200
150
100
50
0
4
8
12
16
20
24
PEG R


Count
300
250
200
150
100
50
0
0 4 8 12 16 20
PEG L
Count
600
500
400
300
200
100
0
0 30 60 90
PEG B
Count
250
200
150
100
50
0
0 2 4 6 8 10
BNTN TOT
Count
200
160
120
80
40
0
-2 2 6 10 14
DIG SPN F
Count
200
160
120
80
40
0
-2 2 6 10 14
DIG SPN R
Count
200
160
120
80
40
0
0 30 60 90
DIGIT SYMB
Count
400
300
200
100
0
1 3 5 7 9
SYMBOL DIG BST
Count
80
60
40
20
0
-1 1 3 5 7 9
SYMBOL DIG RCL
Count
200
160
120
80
40
0
-2 2 6 10 14 18
DIGIT LEARN


Count
RAV ERR
-5 5 15 25 35
0 20 40 60 80 100 120
RAV TIME
0 300 600 900
0 40 80 120 160
TOTAL AIM CORR
0 100 200 30
0 20 40 60 80 100 120 140 160 180
MOOD TENSION
.5 1.5 2.5 3.5 4.5 5.5
0 20 40 60 80 100 120
MOOD DEPRESS
.5 1.5 2.5 3.5 4.5 5.5
0 20 40 60 80 100 120 140
MOOD ANGER
-.5 1 2.5 4
0 40 80 120 160
MOOD FATIGUE
.5 1.5 2.5 3.5 4.5 5.5
0 20 40 60 80 100 120 140
MOOD CONFUS
.5 1.5 2.5 3.5 4.5 5.5
0 20 40 60 80 100 120 140

Appendix D.
BACKGROUND FACTORS

Distributions of the symptom questions by cultural group are shown in Table D-1. Few people in the study were affected by the diseases listed in Tables 10 and D-1; subjects reporting diseases were included in the data analyses. Of greater concern is the relatively higher percentage of Native American Indian (38%) and African American (22%) subjects reporting substance abuse problems, compared with 7% and 8% respectively in majority and Latin American subjects, respectively), and the relative higher percentage of tingling and numbness reported by majority and American Indian subjects (32% and 27% vs. 13%). Targeted recruitment in rehabilitation programs designed to reach minority subjects with low years of education likely explains the disproportionate distribution of past substance abuse.

Table D-1. Number of Subjects and Percent Responses to Questions on Diseases, Abuse, Recent Drug/Alcohol Consumption, by Cultural Group.

Medical Problem	Majority	African American	Latin	Native American
Number of Subjects*	415	110	142	52
Corrective Lenses (Yes)	43	27	11	35
Carpal Tunnel	3	0	0	0
Tingling/Numbness	32	13	13	27
Lupus	0.2	0	0	0
Diabetes	0.5	0	0	0
Thyroid disease	0.7	0	0	0
Pernicious Anemia	0.2	0	0	0
Peripheral Neuropathy	0	0	0	0
1 drink/drug last 48 hours	21	24	19	21
Substance Abuse (Institutionalized)	7	22	8	38

* Number of subjects responding to questions.

Subjects who reported that they had experienced numbness and tingling, substance abuse resulting in institutionalization, and alcohol or drug consumption in the past 48 hours were respectively compared to the remaining subjects for each factor using Analysis of Variance (ANOVA), with each test measure. The unadjusted probabilities are listed in Table D-2. For subjects reporting previous abuse institutionalization, all differences except Digit Span/Backward were in the direction of inferior performance compared to those not previously institutionalized. The abuser group had just over 1 year less education than those not reporting prior institutionalization. Performance differences associated with abuse in the total subject pool could be attributed primarily to the male majority subjects (² in Table D-2).

The disproportion between the cultural groups on percent subjects reporting prior institutionalization for substance abuse (Table D-1) places the generality of these findings in doubt. To further explore these differences, previously institutionalized majority subjects were matched for age and education with non-institutionalized majority subjects, and the analyses were repeated on equal-sized groups (18 vs. 19 subjects). Differences on measures of mood (tension, depression, anger, confusion), Santa Ana (preferred and nonpreferred hand), and tapping (trials 1 and 2) were significant, although the magnitude of the differences was less than 1 standard deviation.

Table D-2. Probability of an Effect on Test Performance (ANOVA) of Reports of Numbness or Tingling, Institutionalization for Substance Abuse, Alcohol Consumption in last 48 hours, and in All Subjects.

MEASURE	Tingling or Numbness	Substance Abuse	Drink/Drug in Last 48 Hrs
Sensory			
Acuity Category	ns	ns	ns
Contrast Sensitivity C	ns	ns	ns
Contrast Sensitivity D	ns	ns	ns
Lanthony/Left eye CI [1]	ns	ns	ns
Lanthony/Right eye CI [1]	ns	0.01	ns
Vibratron/Index [1]	ns	ns	ns
Vibratron/Small [1]	ns	ns	ns
Motor			
Dynamometer/Strength [1]	ns	0.01 [2]	0.03
Dynamometer/Fatigue [1]	ns	ns	0.03
SRT/mean RT	ns	ns	ns
NES SRT/mean latency	ns	0.02 [2]	ns
NES Tapping/trial 1 [1]	0.02	0.0001 [2]	ns
NES Tapping/trial 2 [1]	0.04	0.0003 [2]	ns
Aiming/correct	ns	ns	ns
Santa Ana/Preferred	ns	0.0002 [2]	ns
Santa Ana/Nonpreferred	ns	0.0005 [2]	ns
Pegboard/Right	ns	0.003 [2]	ns
Pegboard/Left	ns	ns	ns
Pegboard/Both	0.05	0.03	ns
Cognitive			
Benton correct	0.003	ns	ns
Digit Span/Forward	ns	ns [2]	ns
Digit Span/Backward	0.007	0.01 [2]	ns
Digit Symbol	0.02	0.001 [2]	ns
NES Symbol Digit/Best	0.005	< 0.0001 [2]	ns
NES Symbol Digit/Recall	ns	ns	ns
NES Serial Digit Learning	0.006	0.002 [2]	ns
Raven/Errors [1]	ns	0.003 [2]	ns
Affect			
NES Mood/Tension	0.0006	0.03 [2]	ns
NES Mood/Depression	0.04	0.005 [2]	ns
NES Mood/Anger	0.003	0.003 [2]	ns
NES Mood/Fatigue	0.002	ns [2]	ns
NES Mood/Confusion	ns	0.0002 [2]	ns
Vocabulary			
WAIS (scaled)	ns	< 0.0001 [2]	ns
Peabody	ns	< 0.0001 [2]	ns
NES Vocabulary	ns	< 0.0001 [2]	ns

[1] Majority subjects only. [2] Significant differences in male majority subjects.

01 02

The differences are broken down in Table D-3 by cultural group. There were virtually no test score differences on the factor of institutionalization in the other cultural groups, raising further doubt about the generality and even the accuracy of the differences between such small subgroups and the much larger remaining subject pool.

Subjects reporting numbness and tingling performed more poorly than subjects who did not on every test except Digit Span/Backward, and many of the differences on tests of cognition and affect were statistically significant (Table D-3). This question was inserted to detect possible cases of carpal tunnel or peripheral neuropathy. The only differences expected in such cases were on the vibratron (early sensory losses), tapping, and dynamometer (later strength losses) tests. Of the expected differences, all showed deficits, although only the tapping differences were significant. Unexpected significant differences on such diverse measures as Serial Digit Learning and Mood/tension suggest Central Nervous System (CNS) differences and bear further investigation. However, these findings do not present a consistent picture and must therefore be regarded with some skepticism.

Table D-3. Probability (P) of Performance Differences among Cultural Groups Previously Institutionalized for Abuse vs. Never-institutionalized Subjects[1]

MEASURE	Majority	Latin Americans[2]	American Indians[3]	African Americans
Sensory				
Acuity Category	ns	ns	ns	ns
Contrast Sensitivity C	ns	ns	ns	ns
Contrast Sensitivity D	ns	ns	ns	ns
Lanthony/Left eye CI [4]	ns	[4]	[4]	[4]
Lanthony/Right eye CI [4]	0.01	[4]	[4]	[4]
Vibratron/Index [4]	ns	[4]	[4]	[4]
Vibratron/Small [4]	ns	[4]	[4]	[4]
Motor				
Dynamometer/Strength [4]	ns	[4]	[4]	[4]
Dynamometer/Fatigue [4]	ns	[4]	[4]	[4]
SRT/mean RT	ns	ns	ns	ns
NES SRT/mean latency	ns	ns	ns	ns
NES Tapping/trial 1 [4]	< 0.0001	[4]	[4]	[4]
NES Tapping/trial 2	0.008			
Aiming/correct				
Santa Ana/Preferred	0.01	ns	0.005	ns
Santa Ana/Nonpreferred	ns	ns	0.003	ns
Pegboard/Right	0.002	ns	ns	ns
Pegboard/Left	ns	ns	ns	ns
Pegboard/Both	ns	ns	ns	ns
Cognitive				
Benton correct	ns	ns	ns	ns
Digit Span/Forward	ns	ns	ns	ns
Digit Span/Backward	ns	ns	ns	ns
Digit Symbol	ns	ns	ns	ns
NES Symbol Digit/Best	ns	ns	0.004	0.04
NES Symbol Digit/Recall [4]	ns	[4]	[4]	[4]
NES Serial Digit Learning	ns	ns	0.04	ns
Raven/Errors [4]	0.003	[4]	[4]	[4]
Affect				
NES Mood/Tension [5]	0.0007	[5]	ns	ns
NES Mood/Depression [5]	0.001	[5]	ns	ns
NES Mood/Anger [5]	0.0008	[5]	ns	ns
NES Mood/Fatigue [5]	0.04	[5]	ns	ns
NES Mood/Confusion [5]	0.0003	[5]	ns	ns
Vocabulary				
WAIS	0.02	0.01	0.05	ns
Peabody	ns	ns	ns	0.04
NES Vocabulary [5]	ns	[5]	ns	0.005

[1] ANOVA p levels (unadjusted); for 24 comparisons (measures); a p = 0.002 would be required for significance at p = 0.05.
[2] Only 12 of 146 people answered "yes," limiting the potential for significance.
[3] Sample of only 52 people. [4] Majority subjects only.
[5] Latin subjects not included.

Appendix E.
TEST CORRELATIONS

MEASURE	Acu	Ctr C	CtrD	Ln/L[1]	Ln/R[1]	Vib/I[1]	Vibs[1]					
Sensory												
Acuity/line	1.00											
Contr Sens. C	.390	1.00										
Contr Sens. D	.449	.721	1.00									
Lanthony/Lt[1]	-.097	-.117	-.227	1.00								
Lanthony/Rt[1]	-.089	-.102	-.171	.641	1.00							
Vibr./Index[1]	.025	.034	-.029	-.045	.056	1.00						
Vibr./Small[1]	.036	.062	.035	-.047	.025	.719	1.00					
Motor	Dy/S[1]	Dy/F[1]	SRT	N srt	Ntap1	Ntp2[1]	Aim	SA/P	SA/N	PegR	PegL	PegB
Dyn./Strngth[1]	1.00											
Dyn./Fatigue[1]	.167	1.00										
SRT	-.309	-.086	1.00									
NES SRT	-.272	-.130	.470	1.00								
NES Tap 1	.340	.129	-.338	-.182	1.00							
NES Tap 2[1]	.351	.161	-.333	-.266	.669	1.00						
Aiming	.024	.112	-.249	-.128	.237	.231	1.00					
Santa Ana/P.	-.017	.029	-.305	-.224	.261	-.078	.416	1.00				
Santa Ana/N	.031	.017	-.283	-.265	.276	-.028	.419	.663	1.00			
Pegboard/R	.022	.010	-.165	-.171	.187	.050	.378	.430	.348	1.00		
Pegboard/L	.002	.021	-.240	-.163	.202	-.009	.359	.385	.449	.552	1.00	
Pegboard/B	-.057	.041	-.150	-.109	.148	.042	.395	.289	.251	.382	.392	1.00
Cognitive	Bent	DS/F	DS/B	Dsym	Nsd/b	Nsd-r[1]	Nsdl	Rav[1]				
Benton	1.00											
Digit Span/F	.394	1.00										
Digit Span/B	.386	.618	1.00									
Digit Symbol	.447	.487	.440	1.00								
NES Sym Dig	-.507	-.454	-.442	-.776	1.00							
NES SymD/R[1]	-.144	-.126	-.101	-.228	.330	1.00						
NES SerD Lrn	-.491	-.561	-.547	-.561	.566	.183	1.00					
Raven/Er[1]	-.366	-.350	-.319	-.348	.347	.204	.445	1.00				
Affect	Tens[2]	Dpr[2]	Ang[2]	Fatig[2]	Conf[2]							
NES/Tens[2]	1.00											
NES/Depr[2]	.701	1.00										
NES/Ang[2]	.641	.704	1.00									
NES/Fatig[2]	.549	.588	.437	1.00								
NES/Conf[2]	.636	.625	.496	.554	1.00							
Vocabulary	Yeduc	Nvoc	WAIS	Peab								
Yrs Education	1.00											
NES Vocab [2]	.420	1.00										
WAIS	.512	.730	1.00									
Peabody [2]	.508	.703	.729	1.00								

[1] Majority subjects only. [2] Test not administered to Latin subjects.

American Petroleum Institute
1220 L Street, Northwest
Washington, D.C. 20005

interoffice memo

DATE: October 4, 1995

FROM: Re'Naye Re'Naye

TO: HESD Staff

SUBJ: Vacation Schedules

Please indicate on the attached form what your pending/tentative vacation plans are and return your responses to me by Tuesday, October 17th.

Thank you.

0 1 0 6

HESD STAFF - PENDING VACATION SCHEDULE

NAME	OCTOBER 95'	NOVEMBER 95'	DECEMBER 95'	JANUARY 96'
J. Vail				
W. Ollison				
K. Talley				
H. Hopkins				
A. Shanks				
Y. Kobayashi				
R. Williams				
R. Drew				
P. Martino				
B. Bauman				
R. Claff				
A. Steen				
J. Shaw				
M. Sellouk				
H. Feldman				
D. Lax				
M. Paxton				
C. Sharp				
D. Mongillo				
R. Rhoden				
R. Barter				
C. Gosnell				
K. Vaucrosson				
R. Landry				
E. Barbaza				
J. Stovall				
P. Green				

(T) Tentative vacation date(s)

ONCE YOU'VE DECIDED ON YOUR VACATION DAY(S), PLEASE WRITE THEM IN.