



American
Petroleum
Institute

STD-API/PETRO PUBL 4648-ENGL 1996 0732290 0563499 7T7



HUMAN NEUROBEHAVIORAL STUDY METHODS: EFFECTS OF SUBJECT VARIABLES ON RESEARCH RESULTS

Health and Environmental Sciences Department
Publication Number 4648
December 1996



Documents
Document Sharing Hub



One of the most significant long-term trends affecting the future vitality of the petroleum industry is the public's concerns about the environment, health and safety. Recognizing this trend, API member companies have developed a positive, forward-looking strategy called STEP: Strategies for Today's Environmental Partnership. This initiative aims to build understanding and credibility with stakeholders by continually improving our industry's environmental, health and safety performance; documenting performance; and communicating with the public.

API ENVIRONMENTAL MISSION AND GUIDING ENVIRONMENTAL PRINCIPLES

The members of the American Petroleum Institute are dedicated to continuous efforts to improve the compatibility of our operations with the environment while economically developing energy resources and supplying high quality products and services to consumers. We recognize our responsibility to work with the public, the government, and others to develop and to use natural resources in an environmentally sound manner while protecting the health and safety of our employees and the public. To meet these responsibilities, API members pledge to manage our businesses according to the following principles using sound science to prioritize risks and to implement cost-effective management practices:

- ❖ To recognize and to respond to community concerns about our raw materials, products and operations.
- ❖ To operate our plants and facilities, and to handle our raw materials and products in a manner that protects the environment, and the safety and health of our employees and the public.
- ❖ To make safety, health and environmental considerations a priority in our planning, and our development of new products and processes.
- ❖ To advise promptly, appropriate officials, employees, customers and the public of information on significant industry-related safety, health and environmental hazards, and to recommend protective measures.
- ❖ To counsel customers, transporters and others in the safe use, transportation and disposal of our raw materials, products and waste materials.
- ❖ To economically develop and produce natural resources and to conserve those resources by using energy efficiently.
- ❖ To extend knowledge by conducting or supporting research on the safety, health and environmental effects of our raw materials, products, processes and waste materials.
- ❖ To commit to reduce overall emission and waste generation.
- ❖ To work with others to resolve problems created by handling and disposal of hazardous substances from our operations.
- ❖ To participate with government and others in creating responsible laws, regulations and standards to safeguard the community, workplace and environment.
- ❖ To promote these principles and practices by sharing experiences and offering assistance to others who produce, handle, use, transport or dispose of similar raw materials, petroleum products and wastes.

Human Neurobehavioral Study Methods: Effects of Subject Variables on Research Results

Health and Environmental Sciences Department

API PUBLICATION NUMBER 4648

PREPARED UNDER CONTRACT BY:

W. KENT ANGER, PhD*, O.J. SIZEMORE, SANDRA J. GROSSMANN,
JULIE A. GLASSER, AND CRAIG A. KOVERA
CENTER FOR RESEARCH ON OCCUPATIONAL AND ENVIRONMENTAL TOXICOLOGY
OREGON HEALTH SCIENCES UNIVERSITY (L606)
PORTLAND, OREGON 97201

DECEMBER 1996

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



FOREWORD

API PUBLICATIONS NECESSARILY ADDRESS PROBLEMS OF A GENERAL NATURE. WITH RESPECT TO PARTICULAR CIRCUMSTANCES, LOCAL, STATE, AND FEDERAL LAWS AND REGULATIONS SHOULD BE REVIEWED.

API IS NOT UNDERTAKING TO MEET THE DUTIES OF EMPLOYERS, MANUFACTURERS, OR SUPPLIERS TO WARN AND PROPERLY TRAIN AND EQUIP THEIR EMPLOYEES, AND OTHERS EXPOSED, CONCERNING HEALTH AND SAFETY RISKS AND PRECAUTIONS, NOR UNDERTAKING THEIR OBLIGATIONS UNDER LOCAL, STATE, OR FEDERAL LAWS.

NOTHING CONTAINED IN ANY API PUBLICATION IS TO BE CONSTRUED AS GRANTING ANY RIGHT, BY IMPLICATION OR OTHERWISE, FOR THE MANUFACTURE, SALE, OR USE OF ANY METHOD, APPARATUS, OR PRODUCT COVERED BY LETTERS PATENT. NEITHER SHOULD ANYTHING CONTAINED IN THE PUBLICATION BE CONSTRUED AS INSURING ANYONE AGAINST LIABILITY FOR INFRINGEMENT OF LETTERS PATENT.

Copyright © 1996 American Petroleum Institute

iii

ACKNOWLEDGMENTS

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

API STAFF CONTACT

David Mongillo, Health and Environmental Sciences Department

MEMBERS OF THE NEUROTOXICOLOGY TASK FORCE

Wayne Daughtrey, Exxon Biomedical Sciences, Inc

David Logan, Mobil Oil Corporation

Charles Ross, Shell Oil Company

Ceinwen Schreiner, Mobil Business Resources Corporation

Christopher Skisak, Pennzoil Company

CONTRACTOR'S ACKNOWLEDGMENTS

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

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

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 a range of advertising. 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 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, and the distributions for each test.

TABLE OF CONTENTS

<u>Section</u>	<u>Page</u>
EXECUTIVE SUMMARY	ES-1
1. INTRODUCTION	1-1
PROJECT GOALS	1-1
DEVELOPMENT OF STANDARDIZED BATTERIES.....	1-2
FIELD ASSESSMENTS OF NES, NCTB, AENTB	1-3
2. METHODS	2-1
SUBJECTS	2-1
BEHAVIORAL TESTS.....	2-2
PROCEDURES.....	2-2
EXAMINERS	2-4
DIVERGENCE FROM PROPOSAL/CONTRACT.....	2-5
3. RESULTS	3-1
DISTRIBUTION OF EXAMINERS.....	3-1
SUBJECT DEMOGRAPHICS.....	3-2
ANALYTIC STRATEGY	3-2
SENSORY TESTS	3-10
MOTOR TESTS	3-13
COGNITIVE TESTS.....	3-19
MEASURES OF AFFECT	3-24
MEASURES OF VOCABULARY	3-26
4. DISCUSSION.....	4-1
DISTRIBUTIONS (NORMALITY)	4-1
THE MAJOR FACTORS (SUMMARY).....	4-1
EDUCATION	4-4
CULTURAL GROUP	4-6
GENDER.....	4-6



Documents

Document Sharing Hub

TABLE OF CONTENTS

<u>Section</u>	<u>Page</u>
4. DISCUSSION (Continued)	
AGE.....	4-7
VOCABULARY TESTS	4-7
POWER ANALYSES	4-8
CONCLUSIONS/RECOMMENDATIONS	4-10
REFERENCES	R-1
APPENDICES	
APPENDIX A.....	A-1
APPENDIX B.....	B-1
APPENDIX C.....	C-1
APPENDIX D.....	D-1
APPENDIX E.....	E-1

**Documents**

Document Sharing Hub

LIST OF FIGURES

<u>Figure</u>	<u>Page</u>
1. Number of majority male and female subjects educated in urban and rural schools tested by the three Oregon Examiners.....	3-1
2. Distribution of years of education, by grade, in each cultural group	3-2
3. Male and female strength measurements in ages 26-35 and 36-45 across 3 educational ranges.....	3-7
4. Male and female vibration thresholds in ages 26-35 and 36-45, across 6-9, 10-12, and 13-16 years of education	3-13
5. Original distribution of Pegboard Both hands and same distribution after a square root transformation was applied.....	3-15
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.....	3-17
7. Tapping by males and females in age ranges 26-35 and 36-45 across education ranges 6-9, 10-12, and 13-16.....	3-18
8. Original distribution of NES Symbol Digit latency and same distribution after a log transformation was applied	3-20
9. Number of spans recalled in the Digit Span test by Majority, Native American, Latin and African American subjects with 0-5, 6-9, 10-12, and 12-16 years of education	3-22
10. Digit spans recalled by majority males and females, ages 26-35 and 36-45 with 6-9, 10-12, and 13-18 years of education	3-23
11. NES Mood scores in male and female majority subjects 26-35 and 36-45 years of age, with 13-16 years of education	3-26
12. Distributions of scores on the WAIS, Peabody, and NES vocabulary tests	3-27
13. Original distribution of NES Vocabulary and the same distribution after a log transformation was applied	3-28
14. Simple regression plots of the NES, WAIS and Peabody vocabulary tests across years of education, for all subjects	3-29
15. Regression plots of years of education in majority subjects and African American subjects	3-30

LIST OF FIGURES (Continued)

<u>Figure</u>	<u>Page</u>
16. Scatterplot with regression line of Digit Symbol performance in Latin American subjects against years of education.....	3-32
17. Differences on vocabulary test performance across three education levels in majority females.....	3-33
18. Urban/Rural WAIS vocabulary comparisons for majority females and males in educational and age groups	3-34
19. Regression plots of population in the town in which subjects were educated against Raven Errors, mean Simple Reaction Time, and NES vocabulary test performance in majority subjects.....	4-5


Documents

Document Sharing Hub

LIST OF TABLES

<u>Table</u>	<u>Page</u>
1. Summary of Distribution and Subject Factors	ES-3
2. Number of Subjects Required to Detect Effect Sizes of 5, 10, 15 or 20% Based on Power Analyses for Males and Females in the Majority Population	ES-5
3. Distribution of Tested Subjects by Years of Education and Cultural Group	2-2
4. Behavioral Tests by Battery from Which the Tests Were Drawn	2-3
5. Order of Test Presentation in Phases I (Minority) and II (Majority)	2-4
6. Subject Distribution Sought in Study	2-6
7. Distribution of Subjects Tested and Included in Analysis	2-6
8. Number of Subjects, Mean Age and Years of Education of Latin American, African American, Native American Indian, and Majority (Rural/Urban) Subject Groups, by Gender and Education Subgroups	3-3
9. MANOVA Table for Main Study Factors	3-5
10. Number of Subjects and Percent Responses to Questions on Diseases, Abuse, Recent Drug/Alcohol Consumption, by Educational Categories	3-9
11. Mean, Median, Standard Deviation, Skew, Kurtosis, and the Best Transformation to Approximate Normality in Sensory Measures for All Subjects	3-11
12. Probability of Effects of Main Study Factors on Sensory Tests	3-11
13. Percent Variance Accounted for by Main Study Factors on Sensory Tests Using Hierarchical Multiple Regression	3-12
14. Mean, Median, Standard Deviation, Skew, Kurtosis, and the Best Transformation to Approximate Normality for Motor Measures for all Subjects, and Skew and Kurtosis of Transformed Data	3-14
15. Probability of Effects of Main Study Factors on Motor Tests	3-15
16. Percent Variance Accounted for by Main Study Factors on Motor Tests Using Hierarchical Multiple Regression	3-16



Documents

Document Sharing Hub

LIST OF TABLES (Continued)

<u>Table</u>	<u>Page</u>
17. Mean, Median, Standard Deviation, Skew, Kurtosis, and the Best Transformation to Approximate Normality for Cognitive Measures in All Subjects.....	3-19
18. Probability of Effects of Main Study Factors on Cognitive Tests	3-20
19. Percent Variance Accounted for by Main Study Factors on Cognitive Tests Using Hierarchical Multiple Regression	3-21
20. Mean, Median, Standard Deviation, Skew, and Kurtosis for Measures of Affect in NES Mood Test	3-24
21. Probability of Effects of Main Study Factors on Affective Measures in NES Mood Test	3-25
22. Percent Variance Accounted for by Education and Cultural Groups on NES Mood Test of Affect Using Hierarchical Multiple Regression	3-25
23. Mean, Median, Standard Deviation, Skew, Kurtosis, and the Best Transformation to Approximate Normality for the NES Vocabulary Test	3-27
24. Probability of Effects of Main Study Factors on Vocabulary Tests	3-29
25. Percent Variance Explained by WAIS Vocabulary Test and Years of Education across Cultural Groups.....	3-31
26. Percent Variance Accounted for by Main Study Factors on Vocabulary Tests Using Hierarchical Multiple Regression	3-32
27. Correlations between Reported Years of Education and WAIS, NES, and Peabody Vocabulary Tests in Majority Subjects	3-34
28. Summary of Distribution Characteristics of Tests.....	4-2
29. Summary of Distribution and Subject Factors	4-3
30. Power Analyses for Female Majority Subjects	4-9
31. Power Analyses for Male Majority Subjects.....	4-10
D-1. Number of Subjects and Percent Responses to Questions on Diseases, Abuse, Recent Drug/Alcohol Consumption, by Cultural Group	D-1

LIST OF TABLES (Continued)

<u>Table</u>	<u>Page</u>
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.....	D-4
D-3. Probability of Performance Differences among Cultural Groups Previously Institutionalized for Abuse vs. Never-Institutionalized Subjects.....	D-5



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 was collected on educational locale (rural and urban), subject gender and age in the 26-45 year range to provide 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. Eighteen widely used behavioral tests selected from the NCTB, NES, and AENTB were administered to subjects 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 first three groups had an educational range of 7-18 years, whereas the educational range of the Latin American

group was 0-18 years. Men and women from the majority population (the only group to include females) with either rural or urban educational locales were specifically 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 yield results that 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. Factors which accounted for more than 5% of a measure's variance are 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	Distribution ²	SD/Mn Ratio ³	Edu- cation	Urban/ Rural ⁵	Cultural Group	Gender ⁵	Age
Sensory							
Acuity/line completed	log	0.23					
Contrast Sensitivity C	N	0.28					
Contrast Sensitivity D	N	0.26					
Lanthony CI/Left ⁵	-	0.22			5		
Lanthony CI/Right ⁵	-	0.22			5		
Vibratron/Index ⁵	Square Root	0.36			5		
Vibratron/Small ⁵	worse	0.41			5		
Motor							
Dynamometer/Strength ⁵	N	0.19			5	.	
Dynamometer/Fatigue ⁵	N	0.12			5		
SRT/Mean RT	worse	0.16	.		.		
NES SRT/Mean latency	log	0.16	.		.		
NES Tapping/Trial 1	N	0.27	.			.	
NES Tapping /Trial 2 ⁵	N	0.18			5		
NES Tapping/Alternating ⁴			.				
Santa Ana/Preferred	N	0.15	.				
Santa Ana/Nonpreferred	Square Root	0.19	.		.		
Aiming/Correct	N	0.33	.		.		
Pegboard/Right	N	0.14				.	
Pegboard/Left	Square Root	0.15					
Pegboard/Both	Square Root	0.15	.		.		
Cognitive							
Benton/Correct	Square Root	0.13	.		.		
Digit Span/Forward	N	0.29	.		.		
Digit Span/Backward	N	0.35	.		.		
Digit Symbol	N	0.20	.		.		
NES Symbol Digit/Best	log	0.26	.		.		
NES Symbol Digit/Recall ⁵	N	0.47			5		
NES Serial 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/Confusion 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 Distribution: Skew, kurtosis ≤ 1 = Normal (N); best transformation listed if not normal. ³ Standard Deviation (SD)/Mean(Mn) 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 Adjusted for age.

ES-3

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 control groups and those 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 larger the number in column 3, the less able the test is to detect small differences between the groups (i.e., for statistical significance).

The means and standard deviations are also the basis for power analysis, a calculation that identifies the number of subjects needed to detect a given effect size (i.e. 5, 10, 15%, etc.) 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 statistical techniques 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.

ES-4

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	Dynamometer/Fatigue	62	16	7	4
Dynamometer/Fatigue	80	20	9	5	Benton/Correct	72	18	8	4
Santa Ana/Preferred	82	20	9	5	Pegboard/Right	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/Preferred	98	25	11	6
Santa Ana/Nonpreferred	110	27	12	7	NES SRT/mean	106	26	12	7
NES Tapping/Trial 2	122	30	14	8	SRT/Mean RT	113	28	13	7
SRT/Mean RT	130	33	14	8	Peabody Vocabulary	121	30	13	8
NES SRT/Mean latency	140	35	16	9	NES Tapping/trial 2	141	35	16	9
Peabody Vocabulary	164	41	18	10	Santa Ana/Nonpreferred	152	38	17	9
Digit Symbol	165	41	18	10	Dynamometer/Strength	163	41	18	10
Lanthy CI/Left	185	46	21	12	Digit Symbol	166	41	18	10
Dynamometer/Strength	207	52	23	13	Lanthy CI/Left	209	52	23	13
Lanthy CI/Right	212	53	24	13	Lanthy CI/Right	214	53	24	13
NES Symbol Digit/Best	238	60	26	15	WAIS Scaled	263	66	29	16
WAIS Scaled	312	78	35	19	NES Symbol Digit/Best	283	71	31	18
NES Tapping/Trial 1	314	79	35	20	NES Tapping/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/Forward	389	97	43	24	Digit Span/Forward	365	91	41	23
NES Vocabulary	390	98	43	24	Mood/Tension	402	100	45	25
Mood/Confusion	421	105	47	26	Mood/Confusion	447	112	50	28
Digit Span/Backward	528	132	59	33	Digit Span/Backward	525	131	58	33
Mood/Anger	605	151	67	38	Vibratron/Index	569	142	63	36
Vibratron/ Index	723	181	80	45	Mood/Depression	650	162	72	41
Mood/Depression	728	182	81	46	Mood/Anger	681	170	76	43
Vibratron/ Small	808	202	90	50	Vibratron/Small	717	179	80	45
Symbol Digit/Recall	1379	345	153	86	Symbol Digit/Recall	947	237	105	59
Raven/Errors	1523	381	169	95	Raven/Errors	1123	281	125	70
NES Serial Digit Learn	5223	1306	580	326	NES Serial Digit Learn	5153	1288	573	322

The power calculations in Table 2 are based on US majority (Western European-derivative) subjects, although substantial differences can be expected in other cultural groups given the 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 among 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 seen 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 ($0.005 = 0.05/10$).

Questionnaire

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 these factors 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; however, inclusion of both education and cultural group as factors in the data analysis is still 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 of gender as a factor in the statistical analyses are 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. Older ages could have been expected to have a larger impact on test performance.
- Many of the tests included here do not produce results with 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 resulting in 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). Improved protocols are needed for these tests.

- 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 in 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 chemical 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).

PROJECT GOALS

This project assessed the impact of years of education and cultural group on performance on behavioral tests used in neurotoxicology 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, when relevant, transformations to improve their normality.

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.

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 advantage of the NES is 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 living 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, Rohlman and Sizemore, 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 United States 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 United States. 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 United States, where more than 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 (e.g., 13-15 years in the United States) 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, which raises questions about the feasibility of using the NCTB in Latin populations.

These feasibility questions are very relevant to the increasingly multi-national U.S. 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 determining that a medical problem does not 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 U.S. 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 are not known.

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 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 his or her 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).

Subject statements about age, years of education and the location of 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.

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

Cultural Group (state*)	Education (years)				Totals
	0-5	6-9	10-12	13-16	
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.

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.

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

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; normal/abnormal
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.

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 and 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
	WAIS R Vocabulary Test
	Peabody Picture Vocabulary Test

¹ Not administered to Latin subjects. ² Spanish version for Latin subjects.

Test instructions placed in a loose-leaf 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

Two Examiners, both fluent in Spanish, from San Francisco State University (SFSU) administered tests in Spanish to all subjects tested in California. One of the Examiners (BJ) was an undergraduate in psychology; the second (FC) was a Masters degree

candidate in counseling. 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 (ZM) was of European descent. Each administered half the test battery to any subject, regularly alternating the tests administered. 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 Masters degree in psychology or equivalent coursework (degree pending).

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 its AENTB; therefore, 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				Totals
		0-5	6-9	10-12	13-16	
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				Percent of Plan
		0-5	6-9	10-12	13-16	
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

The results of behavioral tests can be influenced by the Examiners, 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 were 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 among 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.

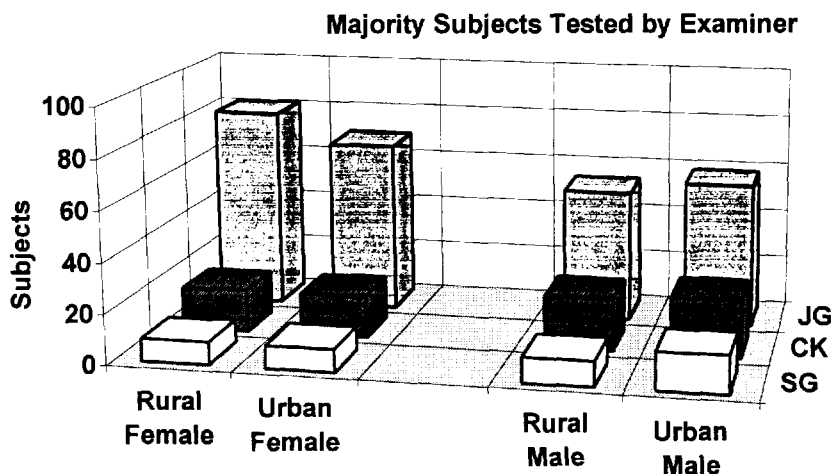


Figure 1. Number of majority male and female subjects educated in urban and rural schools tested by the three Oregon Examiners (JAG, CAK, SJG).