OXFORD JOURNALS
OXFORD UNIVERSITY PRESS

Archives of Clinical Neuropsychology 24 (2009) 31–46

Archives of CLINICAL NEUROPSYCHOLOGY

To Err is Human: "Abnormal" Neuropsychological Scores and Variability are Common in Healthy Adults

Laurence M. Binder[a,*], Grant L. Iverson[b,c], Brian L. Brooks[b]

[a]*Department of Neurology, Oregon Health & Science University, Beaverton, Oregon, USA*
[b]*British Columbia Mental Health & Addiction Services, Vancouver, BC, Canada*
[c]*University of British Columbia, Vancouver, BC, Canada*

Accepted 19 November 2008

Abstract

Normative studies of variability in performance by healthy adults on neuropsychological batteries are reviewed. Regarding test score scatter, normative participants often have large discrepancies between best and worst scores. When "abnormality" was defined as a score more than one standard deviation below the mean, in test batteries with at least 20 measures, the great majority of normative participants had one or more abnormalities. Restricting samples to participants with above average IQ or educational levels and using more conservative definitions of abnormality, such as two standard deviations below the mean did not eliminate the presence of abnormal scores. We conclude that *abnormal* performance on some proportion of neuropsychological tests in a battery is psychometrically *normal*. Abnormalities do not necessarily signify the presence of acquired brain dysfunction because low scores and large intraindividual variability often are characteristic of healthy adults. We recommend that test battery developers provide data on the amount of variability in normal samples and also provide base rate tables with false positive rates that can be used clinically when interpreting test performance.

Keywords: Normal variability; Normal cognitive function; Misdiagnosis; Prevalence of abnormal scores; Base rates

Introduction

The definition of abnormality remains unsettled in clinical neuropsychological practice. There are multiple definitions of abnormality for individual scores. In one widely used normative data set, individual scores are defined as abnormal if they are more than one standard deviation below the mean (Heaton, Grant, & Matthews, 1991; Heaton, Miller, Taylor, & Grant, 2004), but other cutoff scores have been established by test developers. For example, Wechsler tests have traditionally classified scores below the 10th percentile as "borderline" and scores below the second percentile as "extremely low" (Wechsler, 1997a, 1997b). Benton defined scores at or below the fifth percentile as "defective" (Benton, Van Allen, Hamsher, & Levin, 1978; Benton, Varney, & Hamsher, 1977). There also is ambiguity surrounding the clinical and research use of the terms "abnormal" and "abnormality." In this paper, these terms are used synonymously with "low" when referring to the level of neuropsychological performance. We do not imply that "low" or "abnormal" scores are pathological or always indicative of brain dysfunction. Indeed, this paper summarizes data from several sources indicating that low or abnormal scores are common in normative samples.

Regardless of how abnormality on a single test is defined, clinicians must also consider how many low scores are obtained by healthy adults when using a large battery of measures. Almost invariably, clinicians use multiple tests, and then may need to evaluate the clinical significance of a few abnormal scores. Test batteries may lead to dozens of scores. For example, a battery limited to the Wechsler Adult Intelligence Scale-III (WAIS-III) and the Wechsler Memory Scale-III (WMS-III) might often include, not counting any optional or supplemental measures, 13 WAIS-III primary subtest scores and 9 primary WMS-III

* Corresponding author at: 4900 SW Griffith Drive, Suite 244, Beaverton, OR 97005, USA. Tel: +1-503-626-5246.
E-mail address: pdxlarry@aol.com (L.M. Binder).

doi:10.1093/arclin/acn001 Advance Access publication on 6 March 2009

subtest scores. The WAIS-III includes three IQ scores and four index scores, and the WMS-III has eight index scores. The resulting battery of WAIS-III and WMS-III tests includes a total of 22 subtest scores and 15 IQ and Index scores, exclusive of any supplemental measures.

If multiple measures are used, then the probability of obtaining a single abnormal score will increase as the number of measures increases. Statistical models have been used to estimate the probability that abnormal results would be obtained in a test battery. Using binomial probability distributions and a cutoff of one standard deviation below the mean, Ingraham and Aiken (1996) estimated that in a battery of only six tests, more than 20% of the normative sample would be classified as having an abnormal performance on at least two measures. Schretlen, Testa, Winicki, Pearlson, and Gordon (2008) reported that the statistical model of Ingraham and Aiken accurately estimated the observed frequency of scores one standard deviation below the mean in a large test battery administered to a normative sample. However, at more stringent, conservative definitions of abnormality, the statistical model overestimated the frequency of abnormalities. Axelrod and Wall (2007) found that the binomial probability distributions underestimated the prevalence of low scores compared to actual base rates in a college sample that was administered five tests comprising seven scores from the Halstead–Reitan Neuropsychological Test Battery (Reitan & Wolfson, 1993).

Crawford, Garthwaite, and Gault (2007) used Monte Carlo simulation to estimate the number of low scores that would be expected in a test battery. The Monte Carlo simulation method takes into account the intercorrelations between measures in a battery. This method appears to predict the number of low scores in a normative sample more accurately than a method based on binomial probability (Ingraham & Aiken, 1996), which assumes independence of measures and thus ignores intercorrelations. On the WMS-III (The Psychological Corporation, 1997), with abnormality defined as a score more than one standard deviation below the mean, fewer than 16% of the normative sample of 1,250 participants had an abnormal score on any one index when that index was considered in isolation, a finding consistent with a normal distribution of scores. However, the probability that a participant had one or more abnormal index scores, when considering performance across all indexes simultaneously, was much higher than 16%. According to statistical software (Crawford et al., 2007), at least one abnormal score was obtained by 44.8% of the sample. Moreover, at least two abnormal scores were obtained by 27.8%, three or more by 19.6%, four or more by 14.4%, and six or more by 6%. Crawford and colleagues also noted that the binomial prediction method underestimated the number of two or more abnormal scores.

Schretlen and colleagues (2008) examined performance in 327 healthy adults and provided a comparison between actual base rates for batteries of varying length, the binomial prediction method (Ingraham & Aiken, 1996), and the Monte Carlo method (Crawford et al., 2007). Several important findings were reported. First, regardless of the method employed, healthy adults usually obtained some low test scores. Second, adjusting the test scores for demographic factors and predicted intelligence using regression equations increased the number of low scores. Third, the binomial prediction method tended to overestimate the percentage of people by obtaining two or more low scores when using less stringent cutoffs, when examining scores that were not adjusted for demographic factors or predicted intelligence, and when examining longer batteries (i.e., more intercorrelated test scores). Finally, the Monte Carlo method of estimating the percentage of people obtaining two or more low scores was roughly consistent with the actual base rates of low scores (i.e., most differences between estimated and observed data represented less than 3% of the sample, with the highest difference representing a 9% discrepancy).

Statistical methods for predicting the frequency of low scores (Crawford et al., 2007) are useful when base rate data are unavailable. However, now there are several published sources of base rate data on normative sample abnormalities and discrepancies between scores. Citing existing normative studies of neuropsychological performance on large test batteries showing a high frequency of abnormal scores, Sherman, Slick, Strauss, and Spreen (2006) called for additional investigation of normal variability and the prevalence of low scores.

The purpose of this paper is to review literature regarding the base rates of large discrepancy scores between measures, to review studies on the prevalence of low scores across neuropsychological test batteries in adult normative populations, and to discuss the implications of these findings for future test development, research, and interpretation of data obtained in clinical evaluations. This review concerns the specificity of neuropsychological assessment. Data regarding the probability of normal findings in patients with cerebral dysfunction and the sensitivity of neuropsychological tests or test batteries also are important but are beyond the scope of this review.

Test Score Scatter and Discrepancy Analyses

Test Score Scatter in the Wechsler Intelligence Scales

Test manuals for measures such as the Wechsler scales provide information on the statistical significance of differences between scores. The level of statistical significance of a difference between two scores in a battery of tests, however, provides

no information on the *frequency* of such differences. Matarazzo and Herman (1985) and Matarazzo and Prifitera (1989) distinguished between statistical and clinical significance by showing the frequency of such differences. In the Wechsler Adult Intelligence Scale-Revised (WAIS-R) standardization sample ($N = 1{,}880$), many normal individuals had large deviations between Verbal and Performance IQ scores and between highest and lowest subtest scores. Although an absolute difference between Verbal and Performance IQ of 15 points or more was statistically significant (i.e., $p < .05$), it occurred in 17.7% of the normative sample (Matarazzo & Herman, 1985). Absolute differences of 10 or more points occurred in 37.8% of the WAIS-R normative sample. The mean difference between the highest and lowest WAIS-R subtest scaled scores in the normative sample was 6.7 points ($SD = 2.1$), a difference equivalent to more than two standard deviations (Matarazzo & Herman, 1985). Large, statistically reliable differences between subtest scaled scores were also common in the WAIS-R standardization sample. Approximately 69% had differences of six or more scaled score points, nearly 49% had differences of seven or more points, 32% had differences of eight or more points, and 18% had differences of nine or more points (Matarazzo & Prifitera, 1989). Intraindividual subtest differences as great as eight points, for example, between a high score of 14 at the 91st percentile and a low score of 6 at the 9th percentile, were found in nearly one of every three adults in the WAIS-R standardization sample. Replication of the findings using the WAIS-III (Wechsler, 1997) and WAIS-IV texts (Wechsler, 2008) are discussed below.

Schinka, Vanderploeg, and Curtiss (1994) examined the WAIS-R subtest score variability in the normative sample as a function of the highest subtest scaled score. Subtest scaled scores have means of 10 and standard deviations of 3. The frequency of large discrepancies between highest and lowest scores was greater among adults with higher maximum scores and higher Full Scale IQ. The correlation between Full Scale IQ and score range was $r = .33$, indicating that participants with higher intelligence had more variability. The correlation between scatter range and maximum score was $r = .65$, indicating more variability in participants with better than average best scores than in participants with average or below best scores. For example, in the 113 adults with a high subtest scaled score of 16 (98th percentile), the *mean* lowest subtest scaled score was 7.6. Moreover, 64.6% had a lowest subtest scaled score of eight or below and 12.4% had a lowest scaled score of five or below. Among the 205 adults with a high subtest scaled score of 12, the mean lowest subtest score was 5.7, 98.1% had a lowest scaled score of eight or below, and 9.1% had a lowest scaled score of three or below. Among the 85 adults with a high subtest scaled score of 9 (37th percentile), the mean lowest subtest scaled score was 3.9, 90.6% had a lowest scaled score of five or below, and 12.9% had a lowest scaled score of two or below. Clearly, a high degree of variability across even a small battery of tests, such as the WAIS-R, is the rule. It would be a serious mistake to assume, for example, that high functioning adults with high average or superior IQs should have uniformly high test scores across a comprehensive battery of tests.

The amount of test scatter in healthy adults has been presented for the WAIS-III (see Table B.5 in the Administration and Scoring manual; Wechsler, 1997a). If the six WAIS-III Verbal subtests are given, 90.3% of the population will have a one standard deviation difference of at least three points and 31.3% will have a two standard deviation difference of at least six points between highest and lowest scaled scores. The wide Verbal subtest scatter occurs despite the intercorrelations of the subtests; Pearson's *r* between any two Verbal subtests range from .39 to .77.

Similarly, if five WAIS-III Performance subtests are given, 37.5% of the normal population show a two standard deviation split between the highest and the lowest subtest scaled scores. The more tests administered, the greater the proportion of the normal population who show large discrepancies between scores. For example, with administration of the 11 subtests that comprise the four index scores, 71.4% will show a two standard deviation or greater difference between highest and lowest subtest scores, and 17.6% will show a three standard deviation or greater difference. Research on the effect of IQ and highest scaled score on WAIS-III subtest score variability comparable to prior research with the WAIS-R (Schinka et al., 1994) has not been published to our knowledge. Clearly, the WAIS-R findings are applicable to clinical interpretation of the WAIS-III and WAIS-IV.

Normal variability across subtests is illustrated further using the WAIS-III in Table 1. As seen in this table, the likelihood of having large differences between test scores increases with the number of measures. With 14 subtests, including the optional Object Assembly subtest, nearly one in four normal subjects (23.9%) will have a three standard deviation difference of nine points between highest and lowest scaled scores. Examples of 9-point spreads in subtest scores include: 3–12 (1st percentile to 75th percentile), 5–14 (5th percentile to 91st percentile), and 7–16 (16th percentile to 98th percentile). Therefore, wide variability across test scores is normal and expected, even among intercorrelated tests such as the Verbal subtests of the WAIS-III.

Discrepancies between Wechsler Intelligence and Memory Scales

In clinical practice, discrepancies between IQ and memory scores have been regarded as useful markers of a clinically significant memory deficit. However, researchers have reported only modest clinical usefulness of some discrepancy scores. Bornstein, Chelune, and Prifitera (1989) examined discrepancies between the WAIS-R and Wechsler Memory

Table 1. Variability across WAIS-III subtests in the normative sample

	Six verbal subtests	Five performance subtests	Seven verbal subtests	Seven performance subtests	11 Index subtests	14 Subtests
Mean difference	4.7	5.0	5.2	5.7	6.7	7.3
Median difference	5	5	5	6	6	7
Amount of scatter						
Frequency of three or more scaled points, % (1 *SD*)	90.3	90.0	95.0	96.8	99.5	100
Frequency of six or more scaled points, % (2 *SD*)	31.3	37.5	41.4	50.7	71.4	82.2
Frequency of nine or more scaled points, % (3 *SD*)	3.7	5.4	5.1	8.9	17.6	23.9

Notes: The average score for a WAIS-III (Wechsler, 1997) subtest is 10 with a standard deviation of three points. This table was adapted from Table B.5 on page 211 of the WAIS-III Administration and Scoring Manual. Index subtests do not include comprehension, picture arrangement, or object assembly.

Scale-Revised (WMS-R) in 110 normative participants and 192 patients with brain dysfunction, of whom 88.5% had diagnoses either of epilepsy or dementia. Control subjects had IQ and memory standard scores averaging about 100, close to the means for the standardization samples. Bornstein and colleagues reported that the discrepancies based on scales measuring immediate verbal or visual memory were not clinically useful. For example, 24% of control subjects and 29% of patients had Verbal IQ minus verbal immediate memory discrepancies $\geq$12 points. Discrepancies between Full Scale IQ and the WMS-R Delayed Memory Index (based on delayed verbal and visual memory) were more sensitive to brain dysfunction than discrepancies between IQ and immediate verbal or visual memory. A discrepancy $\geq$12 points between Full Scale IQ and Delayed Memory was found in 18.2% of control participants and 36.6% of patients, and a discrepancy $\geq$22 occurred in 5.5% of controls and 18.8% of patients. Hawkins and Tulsky (2003) provided extensive information on discrepancy analyses relating to the WAIS-III and WMS-III. They cautioned that more research was needed to clearly establish that discrepancy analyses can be used to improve diagnostic validity over examining individual index scores alone.

Base rate analyses for WAIS-III FSIQ and index scores also showed that large discrepancies were common in the normative sample of 2,450 participants (Tulsky, Rolfhus, & Zhu, 2000). Moreover, the frequency and magnitude of discrepancy scores varied by level of intelligence. Tables D.1–D.5 in the WAIS-III/WMS-III Technical Manual (The Psychological Corporation, 1997) provide the base rates of index difference scores at different FSIQ levels. Information from these five tables was distilled to create Table 2. As seen in this table, as the level of FSIQ increases, the likelihood of large differences between index scores increases. For example, a 20-point difference between the Verbal Comprehension Index and the Processing Speed Index occurs in only 8.1% of healthy subjects with FSIQ $\leq$79, but a difference of this magnitude occurs in 28.6% of subjects of superior or better intelligence with FSIQ $\geq$120. Similarly, a 20-point spread between the Perceptual Organization Index and the Processing Speed Index occurs in 10.2% of subjects with low average FSIQs of 80–89 and in 21.4% of subjects with high average FSIQs of 110–119. Therefore, the higher the FSIQ, the greater the probability that he or she will show large variability across the WAIS-III index scores.

IQ and index scores are summary scores comprised of multiple subtests. Summary scores eliminate some of the variability across the subtests through aggregation. There is more variability among individual subtests than summary scores.

Recently, the WAIS-IV was published. Table 3 provides the base rates for discrepancies between WAIS-IV index scores at various levels of FSIQ, based on data distilled from Table B.2 of the *Administration and Scoring Manual* (Wechsler, 2008). The data in Table 3 largely agreed with the WAIS-III data in Table 2, except that in the WAIS-IV there are no Verbal or Performance IQ scores (unlike its predecessor). Like the WAIS-III, the WAIS-IV normative data show higher frequencies of large discrepancies between index scores with higher FSIQs than with lower FSIQs. Across all FSIQ levels, the frequency of large discrepancies between pairs of index scores including the Processing Speed Index are greater than the frequency of large discrepancies not including the Processing Speed Index. The Processing Speed Index has a larger standard error of measurement and lower reliability than the other three index scores (Wechsler, Coalson, & Raiford, 2008), leading to greater discrepancies among index score pairs that include Processing Speed.

Dori and Chelune (2004) provided base rates of standardization sample discrepancy scores from the WAIS-III and WMS-III, two test batteries that were co-normed on 1,250 persons. Base rates were provided separately for four levels of education. Large discrepancies, favoring verbal intellect over memory scores, occurred in a sizeable minority of subjects in the normative sample, especially among those with more education. For example, in the group with 13–15 years of education, 15% had Verbal IQ minus General Memory (delayed memory) discrepancies $\geq$15 points. In the group with 16 or more years of education, 21% had Verbal IQ minus General Memory (delayed memory) discrepancies $\geq$15 points. Discrepancies favoring nonverbal intellect over memory also occurred in a sizeable minority of healthy adults, but these discrepancies appeared unrelated to educational level. In the group with 16 or more years of education, 18% had Performance IQ minus General Memory

Table 2. Percentages of normative participants with 10 or more, 15 or more, 20 or more, and 25 or more point discrepancies between WAIS-III Full Scale IQ and Index scores by IQ level

Amount of discrepancy	Verbal IQ – performance IQ	Verbal comprehension–perceptual organization	Verbal comprehension–working memory	Perceptual organization–processing speed	Verbal comprehension–processing speed	Perceptual organization–working memory	Working memory–processing speed
Full scale IQ ≤79							
10 points	20.6	31.1	26.3	31.1	37.8	34.3	45.5
15 points	7.2	11.0	13.1	14.4	16.7	15.2	21.2
20 points	1.4	5.3	5.1	5.7	8.1	7.1	12.1
25 points	0	1.4	3.0	1.0	1.9	3.0	4.0
Full scale IQ 80–89							
10 points	32.3	39.5	36.3	38.1	43.9	36.8	44.8
15 points	13.3	18.2	20.4	23.2	25.4	24.9	25.9
20 points	4.7	9.1	12.4	10.2	15.2	12.9	12.4
25 points	1.4	4.4	4.5	4.7	6.6	7.5	5.0
Full scale IQ 90–109							
10 points	37.2	42.5	46.2	48.5	53.3	44.7	45.8
15 points	17.5	23.1	24.3	29.4	32.1	24.3	30.4
20 points	7.0	11.3	12.1	16.6	18.3	12.9	16.3
25 points	2.4	5.1	5.3	7.8	9.0	6.3	10.4
Full scale IQ 110–119							
10 points	43.8	45.7	48.7	54.3	59.1	50.4	52.1
15 points	22.6	29.2	32.1	36.3	40.4	24.8	32.9
20 points	9.2	16.5	9.8	21.4	22.4	14.5	20.9
25 points	4.6	7.8	5.6	11.2	12.4	10.3	11.5
Full scale IQ ≥120							
10 points	48.1	47.7	59.1	54.4	67.2	48.8	55.9
15 points	24.5	29.9	37.0	33.6	44.8	30.7	38.6
20 points	14.5	15.4	22.0	22.8	28.6	15.7	26.0
25 points	4.6	7.1	8.7	12.0	17.0	7.1	16.5

Notes: Tables D.1–D. 5 in the WAIS-III/WMS-III Technical Manual (The Psychological Corporation, 1997) provide the base rates of index difference scores at different IQ ability levels. Information from these five tables was distilled to create this table. Frequencies are bidirectional, for example, the frequency of scores for verbal comprehension minus working memory is added to the frequency of scores for working memory minus verbal comprehension. The average IQ or Index score is 100 with a standard deviation of 15. An example of a 20-point split is 80–100 (9th–50th percentile).

discrepancies ≥15 points, compared with 15% of the group with 13–15 years of education, 15% of the group with 12 years, and 14% of the group with fewer than 12 years (Dori & Chelune, 2004).

Variability in the Aging, Brain Imaging, and Cognition Study

Schretlen, Munro, Anthony, and Pearlson (2003) illustrated the extent of normal variability in the Aging, Brain Imaging, and Cognition (ABC) study of healthy adults ranging in age from 20 to 92 years with a mean age of about 55 years and approximately 14 years of education. After screening for dementia with the Mini-Mental Status Exam and excluding participants with other relevant neurological history, the 197 remaining participants received a neuropsychological battery of 15 tests yielding 32 measures. Raw test scores were transformed to standardized scores based on data from this research sample, and each participant's maximum discrepancy between highest and lowest score was computed. The mean maximum discrepancy was 3.4 standard deviation units. No participant had a maximum discrepancy of less than 1.6 standard deviation units; only four participants had a maximum discrepancy of less than 2.0 standard deviation units and the largest maximum discrepancy score was 6.1. Sixty-five percent of the sample had maximum discrepancy scores of 3.0 or more and 20% had scores of 4.0 or more units.

To determine how outlying scores inflated the variability, Schretlen and colleagues (2003) then excluded each individual's highest and lowest score, reducing the range in the entire sample to 1.4 to 5.1 standard deviation units from a range prior to excluding highest and lowest scores of 1.6 to 6.1 and reducing the mean variability to 2.7 from 3.4 units. After restricting the variability in this manner, 27% of the participants had maximum discrepancy scores of 3.0 standard deviation units or more and no participant had a score of 1.0 standard deviation units or less.

Table 3. Percentages of normative participants with 10 or more, 15 or more, 20 or more, and 25 or more point discrepancies between WAIS-IV Index scores by IQ level

Amount of discrepancy	Verbal comprehension–perceptual reasoning	Verbal comprehension–working memory	Perceptual reasoning–processing speed	Verbal comprehension–processing speed	Perceptual reasoning–working memory	Working memory–processing speed
Full scale IQ ≤79						
10 points	40.8	36.5	37.1	45.0	36.0	41.3
15 points	19.0	13.8	21.2	28.0	11.6	25.9
20 points	7.4	6.4	9.0	14.8	6.8	9.6
25 points	3.7	1.0	4.2	8.5	2.1	5.3
Full scale IQ 80–89						
10 points	45.9	45.9	49.5	55.9	42.0	40.1
15 points	21.9	20.3	30.7	38.6	22.5	31.6
20 points	10.1	6.7	15.2	22.5	11.6	19.5
25 points	3.6	3.0	7.6	13.6	4.5	10.0
Full scale IQ 90–109						
10 points	45.9	42.4	51.6	53.0	46.6	49.2
15 points	26.2	25.0	30.8	33.8	25.8	29.2
20 points	15.1	12.5	18.8	20.1	15.1	17.2
25 points	6.7	6.0	9.5	11.4	6.4	10.7
Full scale IQ 110–119						
10 points	45.7	50.8	56.9	56.9	55.9	54.0
15 points	31.2	27.9	38.8	35.9	30.6	37.5
20 points	17.3	15.1	24.2	24.2	15.9	24.5
25 points	8.8	8.0	13.0	15.4	8.8	15.4
Full scale IQ ≥120						
10 points	53.5	54.0	59.5	59.0	52.5	42.0
15 points	35.5	32.5	39.5	37.5	31.0	35.5
20 points	18.5	16.0	28.0	26.5	16.0	22.0
25 points	9.5	9.5	16.0	16.0	5.5	15.5

Notes: Table B.2 in the WAIS-IV Administration and Scoring Manual (Wechsler, 2008) provide the base rates of index difference scores at different IQ ability levels. Frequencies are bidirectional, for example, the frequency of scores for verbal comprehension minus working memory is added to the frequency of scores for working memory minus verbal comprehension.

Schretlen and colleagues (2003) suggested that clinicians engage in a reasoning process that includes an estimate of premorbid intellectual level and that clinicians are more concerned with low scores than with high scores. To address how a clinician might interpret individual patients' performances, they computed the discrepancy between the National Adult Reading Test-Revised scores (NART-R; Blair & Spreen, 1989) and the lowest of the other 31 scores. The NART-R is a measure of single word reading/decoding that was designed to estimate premorbid IQ. The *average* lowest score was 1.9 standard deviation units below estimated premorbid IQ. That is, on average, healthy subjects had a neuropsychological test score that was two standard deviations lower than their NART-R score. Eighteen percent of the participants had a lowest score more than 3.0 standard deviations below the estimated IQ score. The investigators also pointed out the problem, assuming that reading level was at least as high as other neuropsychological scores. The highest neuropsychological score exceeded the reading-level-based estimated IQ score by a mean of 1.5 standard deviation units.

Base Rate of Low Scores

It makes intuitive sense that if healthy adults have a substantial amount of test score scatter, and large discrepancy scores are relatively common, then it is likely that healthy adults will obtain a certain number of low test scores. However, intuitive sense does not easily translate into clinical practice, particularly without psychometric evidence of how often low test scores are obtained in healthy people. In this section, we explore the frequency of low test scores in healthy people.

There are several key psychometric principles that neuropsychologists must consider when interpreting multiple test scores. This section provides data from the literature in support of these principles. Some of the more important, yet often overlooked or poorly understood principles, include: (a) obtaining some low scores from a battery of tests is the rule, not the exception,

(b) the more tests that are given, the more likely the person is to have a large spread between high and low scores, (c) people with fewer years of education and/or lower levels of intellectual abilities are expected to have more low scores compared with those with more years of education and/or higher intellectual abilities, and (d) people with more years of education and/or higher intelligence obtain some low scores. An isolated low test score, when administering and interpreting a battery of measures, should be considered common in healthy people. To properly understand what constitutes abnormal performance on neurocognitive testing, we must first appreciate what is considered to be normal performance on the testing. A small number of research groups have examined the profiles of healthy adults across a battery of tests. The work of these groups is summarized below.

Base Rate of Low Scores in the Halstead–Reitan Battery

On the Halstead–Reitan Battery, using an older scoring system, normal examinees were not expected to pass every measure (Reitan & Wolfson, 1985). The Halstead Impairment Index (HII) of the Halstead–Reitan Battery is the proportion of abnormal scores, ranging from 0 (better than the cutoffs on all seven measures) to 1.0 (worse than the cutoffs on all measures). Scores on the HII of 0.1 (one score worse than the cutoff) to 0.4 (three scores worse than the cutoffs) were considered normal. The more recently developed Generalized Neuropsychological Deficit Scale for the battery also yields a total score in the normal range despite some abnormal scores (Reitan & Wolfson, 1993).

Heaton and colleagues (1991) studied normal variability in an expanded Halstead–Reitan Battery. Initially, they reported on 40 measures in 455 healthy adults. More recently, they reported on 25 measures in 1,189 people (2004). In each study, the number of participants varied somewhat for each measure, and scores were demographically adjusted for age, education, and gender. In the second study (Heaton et al., 2004), norms for Caucasians and African-Americans were reported separately. The number of low scores obtained by their normative samples is shown in Table 4. When using a cutoff score of more than one standard deviation below the mean, (a) in the initial study with 40 measures, only 10% of the sample had zero low scores and (b) in the second study with 25 measures, only 13% had no low scores. With 40 measures, 12% had more than 11 low scores, and with 25 measures 13% had more than eight low scores. With more conservative cutoff scores of more than 1.5 or 2.0 standard deviations below the mean, abnormalities in a battery of 25 measures remained common (Robert K. Heaton, personal communication, December 27, 2007). Using the cutoff score of T< 35 (more than 1.5 *SD* below the mean), the majority

Table 4. Prevalence of low scores on the expanded Halstead–Reitan Neuropsychological Battery

Number of low scores	T score <40		T score <35		T score <30		Number of low scores
	%	Cum%	%	Cum%	%	Cum%	
21	0.1	0.1	–	–	–	–	21
20	0.2	0.3	–	–	–	–	20
19	0.1	0.3	–	–	–	–	19
18	0.3	0.6	–	–	–	–	18
17	0.6	1.2	–	–	–	–	17
16	0.5	1.7	–	–	–	–	16
15	0.7	2.3	–	–	–	–	15
14	0.8	3.2	–	–	–	–	14
13	1.0	4.2	0.3	0.3	–	–	13
12	1.2	5.4	0.2	0.4	–	–	12
11	1.9	7.3	0.2	0.6	–	–	11
10	2.2	9.5	0.3	0.8	0.1	0.1	10
9	3.5	12.9	0.6	1.4	0.0	0.1	9
8	4.4	17.3	1.4	2.9	0.0	0.1	8
7	4.1	21.4	2.3	5.1	0.3	0.4	7
6	7.0	28.4	1.9	7.1	0.1	0.5	6
5	7.7	36.1	3.1	10.2	0.6	1.1	5
4	9.3	45.3	4.7	14.9	1.3	2.4	4
3	13.1	58.4	9.6	24.5	2.4	4.7	3
2	13.3	71.7	12.0	36.5	5.6	10.3	2
1	15.1	86.8	22.9	59.4	17.5	27.8	1
0	13.2	100	40.6	100	72.2	100	0

Notes: Based on 25 scores derived from the Expanded Halstead–Reitan Neuropsychological Battery (Heaton et al., 2004). Additional data provided by R. Heaton (personal communication, December 27, 2007). Cum%, cumulative percentage. For example, for T < 40, 9.3% of the sample had exactly four scores in this range, 45.3% had four or more scores in this range, 4.4% had exactly eight scores in this range, and 17.3% had eight or more low scores in this range.

(59%) had at least one low score, 15% had more than three, 10% had more than four, and 5% had more than six low scores. Abnormalities with the cutoff score of more than 2.0 standard deviations below the mean ($T < 30$) were as follows: 28% had one or more, 10% had two or more, and 5% had three or more.

Base Rate of Low Scores in a Flexible Battery

In another study of the prevalence of low scores (Palmer, Boone, Lesser, & Wohl, 1998), performance was examined across a flexible battery of neuropsychological tests in 132 neurologically and psychiatrically healthy older adults between the ages of 50 and 79 ($M = 63.8$ years, $SD = 7.7$). The battery consisted of measures of global cognitive abilities (i.e., Mini-Mental State Exam), attention and processing speed (i.e., WAIS Digit Span, WAIS Digit Symbol Coding, and Stroop Word Reading and Color Naming), confrontation naming (i.e., Boston Naming Test), learning and memory (i.e., WMS-R Logical Memory, WMS-R Visual Reproduction, Rey Osterrieth Complex Figure, and Warrington's Recognition Memory Test for Words and for Faces), visual constructional abilities (i.e., Rey Osterrieth Complex Figure copy), and working memory/executive functions (i.e., Controlled Oral Word Association Test, Stroop Interference, Auditory Consonant Trigrams, and Wisconsin Card Sorting Test). Twenty-six age-corrected test scores were considered simultaneously, and low scores were defined as less than or equal to the 10th percentile and less than or equal to two standard deviations below the mean. Most (73%) of the healthy older adults had one or more scores at or below the 10th percentile and 37% had one or more scores at or below two standard deviations from the mean (Palmer et al., 1998).

Base Rate of Low Scores in the ABC Study

Schretlen and colleagues (Diaz-Asper, Schretlen, & Pearlson, 2004; Testa & Schretlen, 2006; Schretlen, Munro, Anthony, & Pearlson, 2003; Schretlen, Testa, Winicki, Pearlson, & Gordon, 2008) have published several papers related to normative performance on a battery of measures used in the ABC study of normal aging. The ABC publications varied in terms of numbers of participants and measures, because participants and a few measures were added during the study, and also because the investigators addressed research questions related to the number of measures. Diaz-Asper, Schretlen, and Pearlson (2004) examined 16 tests yielding 28 scores in 221 healthy, non-demented participants. Based on FSIQ from the seven subtest short forms of the WAIS-R, the sample was divided into groups with above average (>109), average (90–109), and below average (<90) FSIQ scores. Standard scores for each variable were derived from this research sample. A Cognitive Impairment Index was computed for each participant. The Cognitive Impairment Index was the proportion of tests that were two or more standard deviations below the mean after scores were statistically adjusted for age. The low IQ group mean Cognitive Impairment Index of 0.115 indicated impairment on almost 12% of the measures, the average IQ group mean Index of 0.028 showed impairment on 3% of the measures, and the above average IQ group mean Index of 0.008 showed impairment on almost 1% of the measures. IQ was inversely related to the Cognitive Impairment Index. Comparison of the below average and average IQ groups revealed a mean Cohen's *d* effect size of 0.73, and comparison of the average and above average IQ groups revealed a mean effect size of 0.41. Diaz-Asper and colleagues (2004) concluded that IQ and neuropsychological performance are related, but that the relationship was stronger for people with low IQ than for people with high IQ.

In a related study, Testa and Schretlen (2006) examined 25 scores after demographic correction. Participants were screened for health problems and averaged 52 years of age and 14 years of education with a mean WAIS-R/WAIS-III prorated IQ of 106. Three cutoff scores for abnormality were examined. As shown in Table 5, three or more abnormal scores were produced by 53% of the sample using a cutoff of more than one standard deviation below the mean. The same threshold was reached by 16% of the sample using a cutoff of more than 1.5 standard deviations below the mean and 3% using a cutoff of more than two standard deviations below the mean.

Base Rate of Low Scores in the Neuropsychological Assessment Battery

Base rates of low scores have recently been studied extensively in the large normative sample used for the standardization of the Neuropsychological Assessment Battery (NAB; Stern & White, 2003). The NAB is comprised of 24 co-normed tests across five modules (i.e., Attention, Language, Memory, Spatial, and Executive Functions), which yield 36 demographically adjusted T scores that contribute to five domain Index scores and a Total Index score. In addition to evaluating five cognitive domains, the NAB was also co-normed with a measure of intellectual abilities, the Reynolds Intellectual Screening Test (RIST; Reynolds & Kamphaus, 2003).

Iverson, Brooks, White, and Stern (2008) presented the prevalence of low Index and primary scores on the NAB in 1,269 healthy adults across the lifespan. Using a cutoff of more than one standard deviation below the demographically adjusted

Table 5. Frequency of low, demographically adjusted test scores in various neurocognitive test batteries

	Battery used	Number of test scores	Sample size	>1.0 *SD* below mean	>1.5 *SD* below mean	>2.0 *SD* below mean
Heaton et al. (1991)	E-HRNB	40	455	90% > 0, Median = 4 27% > 8 19% > 9 12% > 11	64% > 0	32% > 0
Heaton et al. (2004)[a,b]	E-HRNB	25	1,189	87% > 0 Median = 3 28% > 5 21% > 6 17% > 7 13% > 8 10% > 9 5% > 11	59% > 0 Median = 1 24% > 2 15% > 3 10% > 4 5% > 6	28% > 0 10% > 1 5% > 2
Diaz-Asper et al. (2004)	ABC study	28	221			11.5% low Average IQ 2.8% average IQ 0.8% above Average IQ
Testa and Schretlen (2006)	ABC study	25	269	53% > 2	16% > 2	3% > 2
Schretlen et al. (2008)	ABC study	10	220–327	35.7% > 1 Mean = 1.49 (*SD* = 1.93)	15.0% > 1 Mean = 0.63 (*SD* = 1.21)	3.4% > 1 Mean = 0.23 (*SD* = 0.62)
Schretlen et al. (2008)	ABC study	25	220–327	75.2% > 1 Mean = 3.63 (*SD* = 4.43)	40.1% > 1 Mean = 1.61 (*SD* = 2.70)	14.4% > 1 Mean = 0.54 (*SD* = 1.28)
Schretlen et al. (2008)	ABC study	43	220–327	91.1% > 1 Mean = 6.23 (*SD* = 7.00)	56.8% > 1 Mean = 2.70 (*SD* = 4.16)	23.8% > 1 Mean = 0.91 (*SD* = 1.94)
Iverson et al. (2008)	NAB	36	1,269	91.5% > 0 43.6% > 4 18.7% > 8 15.5% > 9 9.3% > 12	70.2% > 0 48.5% > 1 15.6% > 4 9.0% > 6	44.3% > 0 21.8% > 1 12.2% > 2
Iverson et al. (2008)	WAIS-III/WMS-III	20	1,250	77.5% > 0 Median = 3 33.7% > 4 23.0% > 6 15.7% > 8 10.4% > 10	43.3% > 0 Median = 0 28.0% > 1 17.7% > 2 11.9% > 3 8.3% > 4	26.6% > 0 Median = 0 14.0% > 1 5.7% > 2 3.5% > 3

Note: E-HRNB, Expanded Halstead–Reitan Neuropsychological Battery; ABC, Aging, Brain Imaging, and Cognition.
[a]R. Heaton (personal communication, December 27, 2007) provided the distributions of abnormalities to supplement published information. E-HRNB, Expanded Halstead–Reitan Neuropsychological Battery; ABC, Aging, Brain Imaging, and Cognition.
[b]See Table 4 for additional data from the E-HRNB.

mean, 36.9% of the normative sample had an abnormality on at least one of the five index scores and 19.3% had an abnormality on at least two index scores. With a more conservative cutoff score of more than two standard deviations below the demographically adjusted mean, 6.3% had at least one index score that was abnormal.

Low index scores were more common among members of the NAB normative sample with low average RIST intelligence scores than in those with higher RIST scores. For those with low average intelligence (i.e., RIST scores between 80 and 89), 82.1% had at least one score below one standard deviation from the mean. In fact, 27.6% had at least four, and 10.4% were abnormal on all five index scores. The low average intelligence group often had index scores more than two standard deviations below the mean (standard scores <70); 19.4% had at least one such score and 9.0% had at least two. The base rates of low

index scores was related to IQ level, except that there was little difference between rates of abnormality between the high average and superior intelligence groups. In the superior group, with RIST scores >119, 17.2% had at least one index score more than one standard deviation below the mean. In the high average group, 16.7% had at least one such low index score, and in the average group, 39.1% had at least one and 16.6% had at least two low scores.

Iverson and colleagues (2008) found frequent low scores when simultaneously considering the 36 demographically adjusted NAB scores in the standardization sample. Five or more low scores (T < 40, more than one standard deviation below the mean) were found in 43.6% of the total sample. Ten or more low scores more than one standard deviation below the mean were found in 15.5% of the sample. One or more frankly abnormal scores (i.e., more than two standard deviations below the mean) were found in 44.3% of the normative sample, and two or more frankly abnormal scores were found in 21.8% of adults. As was true of the index scores, intelligence measured by the RIST was strongly associated with the extent of low scores on the NAB.

Base Rate of Low Scores in the WAIS-III and WMS-III

The presence of low test scores in healthy people has also been examined on the WAIS-III/WMS-III co-normed batteries. Of the 2,450 participants in the WAIS-III normative sample, half also received the WMS-III. Iverson, Brooks, and Holdnack (2008) examined the prevalence of low subtest scores in 1,250 healthy adults when the 20 primary WAIS-III/WMS-III subtests, which are used to derive the 12 Index scores, are examined simultaneously. In the total sample, 77.5% of healthy adults had at least one subtest at or below one standard deviation, 51.1% had three or more, and 28.3% had six or more low scores. Low scores across the WAIS-III/WMS-III battery are common. In fact, a healthy person needs to obtain 12 or more low primary scores across the WAIS-III/WMS-III battery to have a profile defined as "uncommon" (found in fewer than 10% of healthy adults). Consistent with other published data, the number of low scores increases in those people with lower intellectual abilities, although those with higher levels of intelligence still have some low scores across a battery of tests.

Base Rate of Low Memory Scores

The presence of some low scores in healthy people has been illustrated in abbreviated batteries of tests too. For example, there have been a few studies documenting the presence of low scores, and in particular the normality of having some low *memory* scores, across a battery of memory measures in healthy older adults. There is an obvious impact for clinicians and researchers who conduct cognitive assessments with older adults for the purpose of identifying early dementia—most diagnostic criteria are based on the presence of one low memory score.

In one study, the base rates of low memory scores across a flexible battery of memory tests was examined in the 132 healthy older adults (Palmer et al., 1998). The five memory measures included: story learning (WMS-R Logical Memory; immediate and delayed recall, percent retention), recall of simple geometric designs (WMS-R Visual Reproduction; immediate and delayed recall, percent retention), recall of a complex figure (Rey Osterrieth Complex Figure; 3-min delayed recall, percent retention), word recognition (Warrington's Recognition Memory Test—Words), and face recognition (Warrington's Recognition Memory Test—Faces). When performance on the five memory measures (i.e., 10 age-adjusted normative scores) was examined collectively, nearly 40% had one or more low test scores and nearly 17% had two or more low test scores (i.e., ≤1.3 standard deviation below the mean or ≤10th percentile). Notably, 13% of the healthy older adults had one or more memory tests with a score in the frankly impaired range (i.e., ≤2 standard deviations below the mean; ≤2nd percentile). Despite rigorous exclusion criteria to ensure a healthy normative sample, Palmer and colleagues illustrated that low memory scores are common in healthy older adults when multiple tests are administered.

Brooks, Iverson, and White (2007) examined the prevalence of low NAB memory scores in 742 healthy older adults between the ages of 55 and 79 years. These older adults, who were part of the normative sample for the NAB, were screened for any neurological and psychiatric history that could contribute to cognitive impairment. The NAB Memory module consists of four tests (i.e., List Learning, Shape Learning, Story Learning, and Daily Living Memory) that yield 10 demographically adjusted (i.e., age, education, and gender) T scores.

Table 6 presents the number of low scores, below various cutoff scores, that would be considered common (i.e., broadly normal), uncommon (i.e., found in <20%), or unusual (i.e., found in <10%) when considering the 10 individual NAB memory tests simultaneously (this information was based on the tables presented in Brooks et al., 2007). For example, in the total standardization sample, it is unusual to have five or more scores of T < 40, three or more scores of T < 35, and two or more scores of T < 30. It is recommended, however, that the number of low scores be interpreted for different levels of intellectual abilities. When considering the cutoff of T < 35, it is unusual to have five or more low scores in those with low average intellectual abilities. However, for those with high average or superior intellectual abilities, it is unusual to have two or more low scores.

Table 6. Number of low memory scores on the NAB Memory Module and WMS-III that is common, uncommon, or unusual in healthy older adults

	Cutoff	Number of low memory scores		
		Common	Uncommon (<20%)	Unusual (<10%)
NAB memory module				
Total sample ($N = 742$)	<1 *SD*	0–2	3–4	5+
	<1.5 *SD*	0–1	2	3+
	<2 *SD*	0	1	2+
Low average intellectual abilities ($n = 85$)	<1 *SD*	0–6	7	8+
	<1.5 *SD*	0–3	4	5+
	<2 *SD*	0–1	2	3+
Average intellectual abilities ($n = 382$)	<1 *SD*	0–2	3–4	5+
	<1.5 *SD*	0–1	2	3+
	<2 *SD*	0	1	2+
High average intellectual abilities ($n = 166$)	<1 *SD*	0–1	2	3+
	<1.5 *SD*	0	1	2+
	<2 *SD*	0	–	1+
Superior intellectual abilities ($n = 100$)	<1 *SD*	0–1	2	3+
	<1.5 *SD*	0	1	2+
	<2 *SD*	0	1	2+
WMS-III primary memory measures				
Total sample ($N = 550$)	≤1 *SD*	0–3	4–5	6+
	<1.5 *SD*	0–1	2	3+
	≤2 *SD*	0	1	2+
Borderline intellectual abilities ($n = 32$)	≤1 *SD*	0–6	7	8
	<1.5 *SD*	0–3	–	4+
	≤2 *SD*	0–1	–	2+
Low average intellectual abilities ($n = 74$)	≤1 *SD*	0–5	6	7+
	<1.5 *SD*	0–2	–	3+
	≤2 *SD*	0–1	–	2+
Average intellectual abilities ($n = 312$)	≤1 *SD*	0–3	4	5+
	<1.5 *SD*	0–1	2	3+
	≤2 *SD*	0	1	2+
High average intellectual abilities ($n = 98$)	≤1 *SD*	0–2	3	4+
	<1.5 *SD*	0–1	2	3+
	≤2 *SD*	0	–	1+
Superior intellectual abilities ($n = 20$)	≤1 *SD*	0–1	2	3+
	<1.5 *SD*	0	–	1+
	≤2 *SD*	0	–	1+

Notes: NAB, Neuropsychological Assessment Battery (Stern & White, 2003). Ten test scores were considered *simultaneously* for the NAB analyses. WMS-III, Wechsler Memory Scale-III (Wechsler, 1997b). Eight subtest scores were considered *simultaneously* for the WMS-III analyses. "Common" refers to the number of low scores found in 20% or more of the standardization sample, "uncommon" refers to the number of low scores found in approximately <20%, and "unusual" refers to the number of low scores found in approximately <10%. Intellectual abilities for the NAB sample are based on Reynolds Intellectual Scales Test (RIST; Reynolds & Kamphaus, 2003). Estimated intellectual abilities for the WMS-III sample are derived from Wechsler Test of Adult Reading (WTAR; The Psychological Corporation, 2001).

Brooks, Iverson, Holdnack, and Feldman (2008) presented the prevalence of low WMS-III scores in 550 healthy older adults between 55 and 87 years from the standardization sample. The WMS-III consists of four co-normed memory tests (Logical Memory, Faces, Verbal Paired Associates, and Family Pictures) that, including immediate and delayed memory, yield eight age-adjusted subtest scores (Auditory Recognition and Working Memory subtests not included). The number of low scores, below various cutoff scores, that would be considered common (broadly normal), uncommon (found in <20%), or unusual (found in <10%) when considering the eight primary WMS-III memory scores simultaneously are presented in Table 6 (this information is derived from the tables presented in Brooks et al., 2008). When the eight age-adjusted subtest scores were examined simultaneously, it was unusual to have six or more scores more than one standard deviation below the mean ($T < 40$) and unusual to have two or more scores more than two standard deviations below the mean ($T < 30$).

When the analyses were stratified by levels of predicted intellectual abilities estimated by the Wechsler Test of Adult Reading (The Psychological Corporation, 2001), an unusual number of WMS-III scores of $T < 35$ were as follows: (a) four or more for those with predicted borderline intellectual abilities; (b) three or more for those with predicted low average, average, or high average intellectual abilities; and (c) one or more for those with predicted superior or very superior intellectual abilities.

Discussion

Data summarized here show that low or abnormal scores were common in various normative samples. As shown in Table 5, when defining abnormality as a score more than one standard deviation below the mean, test batteries with at least 20 measures yielded at least two abnormalities in most normal participants, and the median number of abnormalities typically was 10%–15% of the total number of test scores in the batteries (Heaton et al., 1991, 2004; Iverson et al., 2008). Abnormalities more than one standard deviation below the mean occurred in 25% or more of the measures in more than one-sixth of the healthy normal subjects (Heaton et al., 1991, 2004; Iverson et al., 2008). More stringent definitions of abnormality, for example, more than two standard deviations below the mean, did not eliminate all abnormalities in normal persons. Furthermore, the larger the test battery, the larger the number of abnormal test scores in the average normative participant.

Variability in test performance and the presence of some low test scores were not specific to any particular battery of tests or to a certain normative sample. Data on the frequency of abnormalities (i.e., prevalence of low scores) are now known for some normative data sets and test batteries, including an expanded Halstead–Reitan Neuropsychological Battery with 40 scores (Heaton et al., 1991), an expanded Halstead–Reitan Neuropsychological Battery with 25 scores (Heaton et al., 2004), the NAB (Stern & White, 2003) with 36 primary scores (Iverson et al., 2008), the WAIS-III/WMS-III battery (Wechsler, 1997a, 1997b) with 20 primary subtest scores (Iverson et al., 2008), Palmer's battery with 26 test scores (Palmer et al., 1998), and Schretlen's battery from the ABC studies (Schretlen et al., 2003), which has varied in size. Despite using different cognitive measures and different normative participants, all of these data sets showed similar frequencies of abnormalities.

There are many reasons why members of normative samples obtain some low scores in a large test battery. Explanations might include measurement error (broadly defined), longstanding weaknesses in certain areas, fluctuations in motivation and effort, psychological interference, and other situational factors such as inattentiveness, fatigue, or minor illness (Mitrushina, Boone, & D'Elia, 1999). Particularly in older normative groups, some individuals may have undiagnosed mild cognitive impairment that can only be demonstrated years later with longitudinal study (De Santi et al., 2008). The probability of obtaining abnormal scores is also related to demographic variables and inversely related to intelligence (Iverson et al., 2008; Schretlen et al., 2008). It is important that people with less than high school education, those with below average intelligence, and individuals from diverse ethnic or cultural backgrounds are more likely to get more low scores, and are thus at greater risk for being misdiagnosed with cognitive impairment (Iverson & Brooks, in press; Iverson et al., 2008).

The normative studies reviewed here used different criteria for exclusion of participants. In the ABC study, participants were excluded if they were institutionalized; were unable to communicate via telephone; if they had a Mini-Mental State Exam score of less than 24/30 (Diaz-Asper et al., 2004); if they had a condition such as current substance dependence, current major depression, or a neurological history likely to have affected cognitive function; or if they showed signs of brain dysfunction based on neurological and psychiatric screening (Schretlen et al., 2008). The normative study of Heaton and colleagues (2004) used history provided by the participants to exclude various conditions affecting condition including various forms of major psychiatric illness, substance abuse, brain disease or injury, and learning disabilities. The standardization of the NAB (Stern & White, 2003) utilized a procedure similar to Heaton and colleagues to exclude persons with neurological disease, acquired injury, psychiatric illness, treatment/medication, or physical impairment (i.e., color blindness, visual loss, hearing impairment, or physical disability) that would negatively impact test performance. The standardization of the WAIS-III/WMS-III (The Psychological Corporation, 1997) and the WAIS-IV (Wechsler et al., 2008) also used history provided by the potential participants and excluded person who reported a history of any condition likely to affect cognition including major psychiatric disorders, learning disorders, and brain disease or injury. Particularly in the studies that used history provided by potential participants for screening, it is possible that persons were included who had a history of cognitive dysfunction, especially learning disorders, despite the screening methods; some people with such histories might not be accurate historians. Moreover, it is a truism that participants at the low end of the normal curve intellectually will perform poorly on other cognitive tests.

None of the normative studies used scores on tests of effort as part of their exclusionary criteria. In our experience with recruited research participants, who are paid a small remuneration for taking a battery of cognitive tests, failure on effort tests is rare (e.g., Storzbach, Rohlman, Anger, Binder, & Campbell, 2001). We doubt that the findings summarized in this review can be explained on the basis of insufficient effort by participants in normative studies.

Participants included in the normative studies reviewed here are likely representative of the normative population. The ABC study focused on what the investigators called "*normal* as opposed to *optimal* aging" (Diaz-Asper et al., 2004, p. 83). We

believe the other studies reviewed here generally studied normative as opposed to optimum neuropsychological functioning. Clearly, the degree of variability in a normative sample is affected by the success of procedures designed to exclude participants at risk for cognitive dysfunction.

For developers of test batteries, the obvious implication of the studies reviewed here is that data on the extent of normal variability, the base rates of low scores, and the presence of large discrepancy scores should be provided to test users in the test manual and scoring software at the time of publication. Data on the frequency of low scores should be provided for both subtest and index/summary scores for different levels of abnormality, such as more than one standard deviation below the mean ($T < 40$; below the 16th percentile), at or below the fifth percentile, and more than two standard deviations below the mean ($T < 30$; below the 2nd percentile). Peer-reviewed publications containing these data that follow test publication both limit and delay the dissemination of critical information necessary for correct clinical interpretation of data. No journal article will be as widely disseminated or known to clinicians as information contained in the original test manual.

The widest possible dissemination of base rate data is essential to prevent misinterpretations of clinical data. For example, we continue to see misinterpretations by clinicians of discrepancies between scores on the WAIS-III and WMS-III and over-interpretation of small numbers of low test scores as we are writing this paper, 11 years after publication of these test batteries. Although inclusion of base rate information in the original test manual and scoring software could not have prevented all misinterpretations of WAIS-III/WMS-III data by clinicians, we believe that its inclusion would have reduced the frequency of these misinterpretations.

There are many clinical implications to the data reviewed here. One implication involves limits on the accuracy of the estimation of premorbid mental abilities using current cognitive test performance. A few scores much higher than other scores do not mean that premorbid mental abilities were at the same level as the highest scores. The "best performance method" of estimating premorbid mental abilities involves taking the highest obtained scores and assuming that those scores are representative of all premorbid abilities (Lezak, Howieson, & Loring, 2004). The best performance method will lead to frequent overestimates of premorbid abilities. Similarly, the use of tests of single word reading often will lead to overestimates of premorbid neuropsychological functioning. In the ABC study, the National Adult Reading Test-Revised (NART-R; Blair & Spreen, 1989) had a 95th percentile confidence interval of ± 15.4 points for predicting Full Scale IQ. In comparison, the NART-R was a weaker predictor of neuropsychological performance other than IQ, yielding 95% confidence intervals of $\pm 25.3-29.4$ points (Schretlen, Buffington, Meyer, & Pearlson, 2005).

Cautions against over-interpreting isolated low scores have been present in the literature for over two decades. Matarazzo and Herman (1985) and Matarazzo and Prifitera (1989) emphasized the extent of normal variability in the WAIS-R and distinguished between statistically and clinically significant test score differences. Heaton and colleagues (1991) wrote, "it is a serious mistake to assume that one or more test scores beyond the accepted cutoff scores always indicate of presence of an acquired cerebral disorder." (p. 36). Schinka and colleagues (1994) warned clinicians by stating, "Even very marked subtest scatter is likely to reflect the state of normal individual differences in many, or most cases." (p. 367). Neuropsychological data must be interpreted with reference to base rates of expected differences and abnormalities, taking into account the number of measures in the battery.

Previously published data show that large discrepancies occur fairly commonly between IQ, index, and subtest scores of the WAIS-R, WMS-R, WAIS-III, and WMS-III. Importantly, to our knowledge, no researchers or test publishers have provided information regarding the base rates of large discrepancy scores when *all* discrepancies are considered *simultaneously* (as is done in clinical practice). That is, we know how common it is to find a 15-point split between the WAIS-III Verbal Comprehension Index and the Processing Speed Index but we do not have empirical data showing how common it is to find a 15-point split between any pair of WAIS-WMS index scores when *all pairwise combinations* are reviewed in a single patient (as is done when reviewing the scoring printout). Fortunately, clinicians can obtain software useful for statistically estimating such findings when the intercorrelation between scores is known (Crawford et al., 2007). One or more "uncommon" discrepancy scores (e.g., occurring in fewer than 10% of healthy adults) is actually common when considering all possible combinations of discrepancy scores. To our knowledge, these data analyses have never been presented for the WAIS-WMS battery and should be the focus of future research. Crawford and colleagues *estimated* that 30.5% of the normal population would be expected to exhibit one or more WAIS-III index score abnormalities, with abnormality defined as a discrepancy between the mean index score and lowest index score exhibited by less than 10% of the population. Similar estimates can be made with his software for any test battery if the correlation matrix of the subtests in the battery is known. Clinicians should be cautious about attributing a discrepancy between two indexes as an indicator of acquired cognitive difficulty unless diagnostic accuracy statistics for that discrepancy in the clinical population of interest are known (Ivnik et al., 2001). Longitudinal research suggesting that the likelihood of developing dementia could be predicted by the degree of intraindividual variability does not yet have clinical utility (Holtzer, Verghese, Wang, Hall, & Lipton, 2008).

The degree of variability in test batteries increases as test reliability decreases because there is more measurement error in a score with low reliability than in a score with high reliability. Consequently, when using a test battery that has lower reliability and no base rate data on normal variability, one should be very cautious about interpreting large intraindividual discrepancies or a small percentage of low scores as abnormal. For example, the Delis Kaplan Executive Function System (Delis, Kaplan, & Kramer, 2001) has nine subtests with several conditions for most subtests. Some conditions of some subtests were rated as having low ($<.60$) reliability and others were rated as having marginal (.60 to .89) reliability (Strauss, Sherman, & Spreen, 2006). This caution regarding tests with lower reliabilities applies to a set of tests originally developed as a battery and also to an individualized test battery developed by a clinician for a particular patient or patient population in her or his practice.

Related to the issue of reliability is the value of composite scores, such as WAIS-III or WAIS-IV index scores. Reliabilities are higher for composite scores than for the subtests that comprise those same composite scores (The Psychological Corporation, 1997). A study cited above showed large variability between highest and lowest WAIS-R subtests (Schinka et al., 1994). It is worth remembering that a set of composite scores will show less intraindividual variability than less reliable subtest scores.

Normative data are relevant to the specificity of abnormal findings. Normative data do not provide statistics on the sensitivity of measures. One can improve specificity at the expense of sensitivity, but we must achieve balance between false-positive and -negative diagnostic errors. The challenge is to develop measures that are both reasonably sensitive to the presence of brain dysfunction and reasonably specific to brain dysfunction (Heaton et al., 2004).

The forensic implications of the data reviewed here are clear. Most civil forensic neuropsychological evaluations of adults are performed to determine if there is acquired cognitive impairment. A full appreciation of the extent of the base rates of low scores and large discrepancies in normal persons should decrease the frequency of misdiagnosis of cognitive impairment in forensic evaluations. Neuropsychological test findings that clinicians have attributed to various conditions such as neurotoxic exposure or mild traumatic brain injuries, for example, often are typical of normal persons (e.g., two to four low scores across a battery of tests). Thus, clinicians should guard against over-interpreting isolated low scores and adopt a more scientific approach to evaluating test results (cf. Franklin, 2004; Larrabee, 2005).

The implications of low scores as a normal occurrence in daily clinical practice are also quite evident in the assessment of older adults with suspected early cognitive decline, because the criterion for identifying cognitive impairment and, in particular, mild cognitive impairment sometimes has been erroneously based on the presence of *one* low memory score. Blackford and LaRue (1989) noted that "... in a memory battery with many measures, the chances are substantial that at least one score will fall into the impaired range" (p. 303). Other data show that many normal persons do not perform normally on an entire battery of memory tests (Brooks et al., 2007, 2008; de Rotrou et al., 2005; Loewenstein et al., 2006; Palmer et al., 1998). Moreover, there is an accumulating amount of literature that illustrates how a universal cutoff score will be more likely to misdiagnose those healthy older adults who have fewer years of education or lesser intelligence and miss a diagnosis in those who have more education and higher intelligence. By understanding how often healthy people obtain low memory scores across a battery of measures, neuropsychologists are better able to determine what an unusual number of low memory scores is and, subsequently, develop guidelines for determining memory impairment based on separating normal from abnormal performance.

Future research should address the question of whether certain domains of functioning or tests are more likely to yield low scores than others. We do not know if measures thought to be relatively sensitive to brain dysfunction such as list learning memory tests are more likely to show abnormality in normative samples than measures generally considered less sensitive to brain dysfunction. If all measures are equally likely to show abnormalities in normative samples, then the findings reviewed here are more valid for a neuropsychological battery that has not been studied for its extent of normal variability and abnormality. If, on the other hand, some measures are more likely to be abnormal than others, then the extent of abnormality of a battery without normative data can only be estimated from published data on other test batteries.

In clinical practice, it can be very difficult to determine when a particular patient is an exception to base rates and normal frequencies. The base rates summarized in this review provide probabilities; the base rates do not provide diagnostic certainty in most cases. Hence, diagnostic classifications based on base rates will not always be accurate, although the clinician is more likely to be correct when relying on sophisticated base rate information than when determining that a particular case is an exception to the base rates.

Classification of a patient as cognitively impaired when the frequency of abnormal scores is within normal limits is more likely to be correct when the abnormal scores are consistent with neurodiagnostic data. For example, if a right-handed patient with a partial complex seizure disorder has a right temporal epileptiform abnormality, then one or two abnormalities on tests of visual memory may have clinical significance. Such isolated neuropsychological abnormalities should be considered equivocal neuropsychological evidence of acquired brain dysfunction, and the neuropsychological report should state that the equivocal abnormality is given some weight because it is consistent with neurodiagnostic data or a clearly present neurological disease. On the other hand, clinicians should exercise much greater caution when interpreting isolated neuropsychological

abnormalities when there are no neurodiagnostic data consistent with brain dysfunction and when the neuropsychological abnormalities are the only data consistent with a neurological diagnosis.

We conclude that *abnormal* performance on some proportion of neuropsychological tests in a battery is psychometrically *normal*. Although this is especially true with a more liberal definition of abnormality such as a score more than one standard deviation below the mean, it also is true with more stringent definitions of abnormalities including a score more than two standard deviations below the mean. Statistically normal performance on all measures in a large test battery is not necessary in order to classify the overall result as normal. Several abnormal test scores do not necessarily imply the presence of acquired brain dysfunction in adults. Although people with higher IQ scores tend to have fewer low scores than people with lower IQ scores, normal persons of high intelligence often have some low test scores. We further conclude that large variability between highest and lowest scores is psychometrically normal, the degree of normal variability is greater in those people with higher IQ scores, and highly intelligent normal people sometimes show considerable contrast between neuropsychological strengths and weaknesses. As Schretlen and colleagues (2003) stated, "the findings reported here underscore the importance of basing clinical neuropsychological inferences about cerebral dysfunction on clinically recognizable patterns of performance in the context of other historical, behavioral, and diagnostic information, *rather than on psychometric variability alone*." (p. 869, emphasis added)

Conflict of Interest

Dr. Iverson has received past research funding from Psychological Assessment Resources, Inc., the company that publishes the Neuropsychological Assessment Battery. This study, however, was unfunded.

Acknowledgement

We thank Robert K. Heaton and the anonymous reviewers for their contributions.

References

Axelrod, B. N., & Wall, J. R. (2007). Expectancy of impaired neuropsychological test scores in a non-clinical sample. *International Journal of Neuroscience*, *117*, 1591–1602.

Benton, A. L., Van Allen, M. W., Hamsher, K. S., & Levin, H. S. (1978). *Test of facial recognition, form SL*. Iowa City, IA: Department of Neurology, University of Iowa Hospitals and Clinics.

Benton, A. L., Varney, N. R., & Hamsher, K. S. (1977). *Judgment of line orientation manual*. Iowa City, IA: Department of Neurology, University of Iowa Hospitals and Clinics.

Blackford, R. C., & LaRue, A. (1989). Criteria for diagnosing age-associated memory impairment: Proposed improvement from the field. *Developmental Neuropsychology*, *5*, 295–306.

Blair, J. R., & Spreen, O. (1989). Predicting premorbid IQ: A revision of the National Adult Reading Test. *The Clinical Neuropsychologist*, *3*, 129–136.

Bornstein, R. A., Chelune, G. J., & Prifitera, A. (1989). IQ-memory discrepancies in normal and clinical samples. *Psychological Assessment*, *1*, 203–206.

Brooks, B. L., Iverson, G. L., Holdnack, J. A., & Feldman, H. H. (2008). The potential for misclassification of mild cognitive impairment: A study of memory scores on the Wechsler Memory Scale-III in healthy older adults. *Journal of the International Neuropsychological Society*, *14*, 463–478.

Brooks, B. L., Iverson, G. L., & White, T. (2007). Substantial risk of "Accidental MCI" in healthy older adults: Base rates of low memory scores in neuropsychological assessment. *Journal of the International Neuropsychological Society*, *13*, 490–500.

Crawford, J. R., Garthwaite, P. H., & Gault, C. B. (2007). Estimating the percentage of the population with abnormally low scores (or abnormally large score differences) on standardized neuropsychological test batteries: A generic method with applications. *Neuropsychology*, *21*, 419–430. Test Software http://www.abdn.ac.uk/~psy086/dept/PercentAbnormKtests.htm.

de Rotrou, J., Wenisch, E., Chausson, C., Dray, F., Faucounau, V., & Rigaud, A. S. (2005). Accidental MCI in healthy subjects: A prospective longitudinal study. *European Journal of Neurology*, *12*, 879–885.

De Santi, S., Pirraglia, E., Barr, W., Babb, J., Williams, S., & Rogers, K., et al. (2008). Robust and conventional neuropsychological norms: Diagnosis and prediction of age-related cognitive decline. *Neuropsychology*, *22*, 469–484.

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *The Delis Kaplan executive function system: Technical manual*. San Antonio, TX: The Psychological Corporation.

Diaz-Asper, C. M., Schretlen, D. J., & Pearlson, G. D. (2004). How well does IQ predict neuropsychological test performance in normal adults? *Journal of the International Neuropsychological Society*, *10*, 82–90.

Dori, G. A., & Chelune, G. J. (2004). Education-stratified base-rate information on discrepancy scores within and between the Wechsler Adult Intelligence Scale: Third Edition and the Wechsler Memory Scale: Third Edition. *Psychological Assessment*, *16*, 146–154.

Franklin, R. D. (2004). Neuropsychological hypothesis testing. In R. D. Franklin (Ed.), *Prediction in forensic and neuropsychology: Sound statistical practices* (pp. 29–64). Mahwah, NJ: Lawrence Erlbaum Associates.

Hawkins, K. A., & Tulsky, D. S. (2003). WAIS-III WMS-III discrepancy analysis: Six-factor model index discrepancy base rates, implication, and preliminary consideration of utility. In D. S. Tulsky, D. H. Saklofske, G. J. Chelune, R. K. Heaton, R. J. Ivnik, R. Bornstein, A. Prifitera, & M. Ledbetter (Eds.), *Clinical interpretation of the WAIS-III and WMS-III* (pp. 211–272). Amsterdam: Academic Press.

Heaton, R. K., Grant, I., & Matthews, C. G. (1991). *Comprehensive norms for an extended Halstead–Reitan Battery: Demographic corrections, research findings, and clinical applications*. Odessa, FL: Psychological Assessment Resources.

Heaton, R. K., Miller, S. W., Taylor, M. J., & Grant, I. (2004). *Revised comprehensive norms for an expanded Halstead–Reitan Battery: Demographically adjusted neuropsychological norms for African American and Caucasian adults professional manual*. Lutz, FL: Psychological Assessment Resources.

Holtzer, R., Verghese, J., Wang, C., Hall, C. B., & Lipton, R. B. (2008). Within-person across-neuropsychological test variability and incident dementia. *Journal of the American Medical Association*, *300*, 823–830.

Ingraham, L. J., & Aiken, C. B. (1996). An empirical approach to determining criteria for abnormality in test batteries with multiple measures. *Neuropsychology*, *10*, 120–124.

Iverson, G. L., & Brooks, B. L. (in press). Improving accuracy for identifying cognitive impairment. In M. R. Schoenberg, & J. G. Scott (Eds.), *The black book of neuropsychology: A syndrome-based approach*. New York: Springer.

Iverson, G. L., Brooks, B. L., & Holdnack, J. A. (2008). Misdiagnosis of cognitive impairment in forensic neuropsychology. In R. L. Heilbronner (Ed.), *Neuropsychology in the courtroom: Expert analysis of reports and testimony* (pp. 243–266). New York: Guilford Press.

Iverson, G. L., Brooks, B. L., White, T., & Stern, R. A. (2008). Neuropsychological Assessment Battery (NAB): Introduction and advanced interpretation. In, A. M. Horton Jr. & D. Wedding (Eds.), *The neuropsychology handbook* (3rd ed., pp. 279–343). New York: Springer.

Ivnik, R. J., Smith, G. E., Cerhan, J. H., Boeve, B. F., Tangalos, E. G., & Petersen, R. C. (2001). Understanding the diagnostic capabilities of cognitive tests. *The Clinical Neuropsychologist*, *15*, 114–124.

Larrabee, G. J. (2005). A scientific approach to forensic neuropsychology. In G. J. Larrabee (Ed.), *Forensic neuropsychology: A scientific approach* (pp. 3–28). New York: Oxford University Press.

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological assessment* (4th ed.). New York: Oxford University Press.

Loewenstein, D. A., Acevedo, A., Ownby, R., Agron, J., Barker, W. W., & Isaacson, R., et al. (2006). Using different memory cutoffs to assess mild cognitive impairment. *American Journal of Geriatric Psychiatry*, *14*, 911–919.

Matarazzo, J. D., & Herman, D. O. (1985). Clinical uses of the WAIS-R: Base rates of differences between VIQ and PIQ in the WAIS-R standardization sample. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements, and applications* (pp. 899–932). New York: Wiley.

Matarazzo, J. D., & Prifitera, A. (1989). Subtest scatter and pre-morbid intelligence: Lessons from the WAIS-R standardization sample. *Psychological Assessment*, *1*, 186–191.

Mitrushina, M. N., Boone, K. B., & D'Elia, L. F. (1999). *Handbook of normative data for neuropsychological assessment*. New York: Oxford University Press.

Palmer, B. W., Boone, K. B., Lesser, I. M., & Wohl, M. A. (1998). Base rates of "impaired" neuropsychological test performance among healthy older adults. *Archives of Clinical Neuropsychology*, *13*, 503–511.

Reitan, R. M., & Wolfson, D. (1985). *The Halstead–Reitan Neuropsychological Test Battery: Theory and clinical interpretation*. Tucson, AZ: Neuropsychology Press.

Reitan, R. M., & Wolfson, D. (1993). *The Halstead–Reitan Neuropsychological Test Battery: Theory and clinical interpretation* (2nd ed.). Tucson, AZ: Neuropsychology Press.

Reynolds, C. R., & Kamphaus, R. W. (2003). *Reynolds Intellectual Assessment Scales and Reynolds Intellectual Screening Test professional manual*. Lutz, FL: Psychological Assessment Resources.

Schinka, J. A., Vanderploeg, R. D., & Curtiss, G. (1994). Wechsler Adult Intelligence Scale-Revised subtest scatter as a function of maximum subtest scaled score. *Psychological Assessment*, *6*, 364–367.

Schretlen, D. J., Buffington, A. L., Meyer, S. M., & Pearlson, G. D. (2005). The use of word-reading to estimate "premorbid" ability in cognitive domains other than intelligence. *Journal of the International Neuropsychological Society*, *11*, 784–787.

Schretlen, D. J., Munro, C. A., Anthony, J. C., & Pearlson, G. D. (2003). Examining the range of normal intraindividual variability in neuropsychological test performance. *Journal of the International Neuropsychological Society*, *9*, 864–870.

Schretlen, D. J., Testa, S. M., Winicki, J. M., Pearlson, G. D., & Gordon, B. (2008). Frequency and bases of abnormal performance by healthy adults on neuropsychological testing. *Journal of the International Neuropsychological Society*, *14*, 436–445.

Sherman, E., Slick, D. J., Strauss, E., & Spreen, O. (2006). Psychometrics in neuropsychological assessment. In E. Strauss, E. Sherman, & O. Spreen (Eds.), *A compendium of neurospsychological tests: Administration, norms and commentary* (3rd ed., pp. 3–43). New York: Oxford University Press.

Stern, R. A., & White, T. (2003). *Neuropsychological Assessment Battery*. Lutz, FL: Psychological Assessment Resources.

Storzbach, D., Rohlman, D. S., Anger, W. K., Binder, L. M., & Campbell, K. A. (2001). Neurobehavioral deficits in Persian Gulf veterans: additional evidence from a population-based study. *Environmental Research*, *85*, 1–13.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary* (3rd ed.). New York: Oxford University Press.

Testa, S. M., & Schretlen, D. J. (2006). *The frequency of abnormal neuropsychological test scores following demographic adjustment in healthy adults*. Paper presented at the American Academy of Clinical Neuropsychology, Philadelphia.

The Psychological Corporation. (1997). *WAIS-III/WMS-III technical manual*. San Antonio, TX: Author.

The Psychological Corporation. (2001). *Wechsler test of adult reading manual*. San Antonio, TX: Author.

Tulsky, D. S., Rolfhus, E. L., & Zhu, J. (2000). Two-tailed versus one-tailed base rates of discrepancy scores in the WAIS-III. *The Clinical Neuropsychologist*, *14*, 451–460.

Wechsler, D. (1997a). *Wechsler Adult Intelligence Scale: Third Edition*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997b). *Wechsler Memory Scale: Third Edition*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2008). *WAIS-IV Wechsler Adult Intelligence Scale: Fourth Edition: administration and scoring manual*. San Antonio, TX: NCS Pearson.

Wechsler, D., Coalson, D. L., & Raiford, S. E. (2008). *WAIS-IV. Wechsler Adult Intelligence Scale: Fourth Edition. Technical and interpretative manual*. San Antonio, TX: NCS Pearson.