Journal of the International Neuropsychological Society (2008), **14**, 436–445.
Copyright © 2008 INS. Published by Cambridge University Press. Printed in the USA.
doi: 10.1017/S1355617708080387

Frequency and bases of abnormal performance by healthy adults on neuropsychological testing

DAVID J. SCHRETLEN,[1,2] S. MARC TESTA,[1] JESSICA M. WINICKI,[1]
GODFREY D. PEARLSON,[1,3,4] AND BARRY GORDON,[5,6]

[1]Department of Psychiatry and Behavioral Sciences, The Johns Hopkins University School of Medicine, Baltimore, Maryland
[2]Russell H. Morgan Department of Radiology and Radiological Science, The Johns Hopkins University School of Medicine, Baltimore, Maryland
[3]Olin Neuropsychiatry Research Center, Hartford Hospital/Institute of Living, Hartford, Connecticut
[4]Department of Psychiatry, Yale University School of Medicine, New Haven, Connecticut
[5]Department of Neurology, The Johns Hopkins University School of Medicine, Baltimore, Maryland
[6]Cognitive Science Department, The Johns Hopkins University, Baltimore, Maryland

(RECEIVED October 29, 2007; FINAL REVISION December 6, 2007; ACCEPTED December 7, 2007)

Abstract

The frequency and determinants of abnormal test performance by normal individuals are critically important to clinical inference. Here we compare two approaches to predicting rates of abnormal test performance among healthy individuals with the rates actually shown by 327 neurologically normal adults aged 18–92 years. We counted how many participants produced abnormal scores, defined by three different cutoffs with test batteries of varied length, and the number of abnormal scores they produced. Observed rates generally were closer to predictions based on a series of Monte Carlo simulations than on the binomial model. They increased with the number of tests administered, decreased as more stringent cutoffs were used to identify abnormality, varied with the degree of correlation among test scores, and depended on individual differences in age, education, race, sex, and estimated premorbid IQ. Adjusting scores for demographic variables and premorbid IQ did not reduce rates of abnormal performance. However, it eliminated the contribution of these variables to rates of abnormal test performance. These findings raise fundamental questions about the nature and interpretation of abnormal test performance by normal, healthy adults. (*JINS*, 2008, *14*, 436–445.)

Keywords: Neuropsychology, Cognition disorders, Diagnostic errors, Monte Carlo method, Models, Binomial, Classification

INTRODUCTION

Abnormal cognitive test performance is variably defined. Scores falling more than 2 standard deviations (*SD*s) below the population mean are almost universally viewed as abnormal, but cutoffs as lenient as >1 *SD* are used by some. In a normal distribution, these cutoffs include from 2.3% to 15.9% of healthy persons. When a test battery includes multiple measures, the number of healthy persons who produce abnormal test scores increases, as do the number of abnormal scores they produce. Because the extent and bases of this increase are unclear, how to interpret the results of multiple tests remains unclear.

In 1189 neurologically normal adults who completed 25 or more tests from an expanded Halstead-Reitan Neuropsychological Battery (HRB), Heaton et al. (2004) used T-scores below 40 (i.e., >1 *SD* below the mean) to define "impaired" performance. Based on this, 87% of the participants produced at least one and 34% produced five or more "impaired" scores. When the T-score cutoff was lowered to <30 (i.e., >2 *SD*s below the mean), 28% still produced at least one "impaired" score. Because they adjusted the T-scores for age, sex, race, and education, these factors probably did not affect the likelihood of demonstrating "impairment." Rather, this likelihood probably depended primarily on how many tests were administered and the cutoff used to define "impaired" performance. Other factors might have contributed, but these presumably did *not* include disease or injury because the participants were neurologically normal. While

Correspondence and reprint requests to: David J. Schretlen, Ph.D., The Johns Hopkins Hospital, 600 N. Wolfe Street, Meyer 218, Baltimore, MD 21287–7218. E-mail: dschret@jhmi.edu

patients are expected to produce abnormal test scores, the causes and meaning of such performance by *healthy* adults remain unclear.

When raw score cutoffs are used to define "impaired" performance, many factors might contribute. For example, age and education clearly affect performance (Crum et al., 1993) on the Mini-Mental State Exam (Folstein et al., 1975). When a person's raw score is compared with those of healthy age peers, as is common for IQ and other tests, this minimizes age effects, but it leaves other factors in addition to disease or injury as potential causes of abnormal performance. Thus, it remains unclear how adjusting test scores for demographic factors alters the frequency and determinants of "impaired" performance by normal adults.

When one test is administered, the percentage of normal healthy adults who obtain an abnormal score depends on the cutoff used to define abnormality. If one defines a score that falls >2 *SD*s below the mean as abnormal, then approximately 2.3% of individuals who comprise a Gaussian distribution will obtain abnormal scores. Likewise, 6.7% or 15.9% of normal healthy individuals will produce an abnormal score if the more lenient cutoffs of >1.5 or >1 *SD*s below the mean, respectively, are applied. It is a mathematical truism that taking multiple tests inflates the odds of obtaining at least one abnormal score. Thus, the questions of how to identify "abnormality" based on a single test and how to interpret the results of multiple tests performed on the same individual must account for this mathematical truism.

Ingraham and Aiken (1996) noted that, when multiple uncorrelated tests are used, the binomial probability distribution can be used to predict how many normal persons will produce one or more abnormal scores. They tested this against data reported for HIV-1 seronegative men and HIV-1 seropositive men with and without signs or symptoms of infection (Janssen et al., 1989). The binomial model predicted that 49%, 14%, and 2% of healthy participants would produce 2 or more abnormal scores (out of 10 measures) by chance using cutoffs of ≥1, 1.5, and 2 *SD*s below the mean, respectively, to define abnormal performance. As hypothesized, the seronegative and asymptomatic HIV-positive groups showed rates of abnormal performance that were consistent with predictions, whereas symptomatic HIV-positive patients showed higher than predicted rates of abnormal performance using cutoffs of 1.5 *SD*s (31%) and 2 *SD*s (12%) below the mean. This approach assumes that healthy subjects produce abnormal scores by chance. Ingraham and Aiken (1996) were unable to test whether demographic or other characteristics correlated with the observed rates of abnormal performance by apparently healthy participants.

A limitation of the binomial model is that it assumes independence among measures that comprise a test battery, and this is virtually never the case. Ingraham and Aiken (1996) noted that any correlation among the measures tends to decrease rates of abnormal performance, but the effects of such correlation might be more complicated. Crawford et al. (2007) argued that the binomial model tends to *overestimate* how many individuals will produce one or more abnormal test scores, and *underestimate* how many will produce two or more abnormal scores. This can be best understood using an example. Suppose one administers 10 tests with a mean correlation of 0.5 to healthy adults, and someone earns an average score on test X. Because the measures are correlated, the odds that this person will perform abnormally on the other tests decrease slightly. Because most individuals perform normally on text X, the binomial model tends to overestimate how many will earn one or more abnormal scores. Conversely, suppose someone else produces an abnormal score on test X. Because the tests are correlated, that person's chances of performing abnormally on the other tests increase slightly. In this way, the binomial model might underestimate how many people obtain two or more abnormal scores when the tests are correlated.

In response to this limitation, Crawford et al. (2007) proposed using a Monte Carlo (MC) simulation method to estimate how many normal individuals will produce abnormal test scores. They described a generic MC simulation that requires only R, the $k \times k$ matrix of correlations among the k measures comprising a test battery. The method begins by obtaining the Choleski (also spelled Cholesky) decomposition of R, which can be seen as the square root of R (step 1). It then generates one million random vectors of k independent standard normal variates (step 2). Finally, it post-multiplies each vector by the Choleski decomposition matrix to produce one observation per vector from the desired multivariate normal distribution (step 3), resulting in a million "observations." Crawford et al. used these MC simulation results to estimate how many "individuals" (i.e., observations) would obtain one or more abnormal Wechsler Adult Intelligence Scale, 3rd Edition (WAIS-III) Index scores given the correlations among them reported in the test manual. The investigators designated scores below the 5th percentile (>1.67 *SD*s below the mean) as abnormal. Lacking raw data, Crawford et al. could not empirically test the accuracy of their MC predictions against actual observed rates of abnormal WAIS-III Index performance. However, while the binomial distribution predicts that 18.5% of normal adults would produce one or more abnormal scores (assuming correlations of zero among the WAIS-III Indices), the MC simulation found that 13.2% of normal adults would produce one or more abnormal Index scores (using correlations among measures reported in the WAIS-III manual). Furthermore, the MC simulation found that 4.6% of the one million individuals produced two or more abnormal scores, compared with 1.4% predicted by the binomial model. These findings suggest that the MC simulation method predicts different rates of abnormal performance than the binomial model. However, this has never been tested empirically. In addition, the effects of adjusting test performance for demographic characteristics and premorbid IQ on observed rates of abnormal performance have never been examined, despite the increasing practice of adjusting neuropsychological test performance for demographic variables.

https://doi.org/10.1017/S1355617708080387 Published online by Cambridge University Press

The aims of this study were threefold: First, we sought to compare rates of abnormal test performance predicted by the binomial model with those predicted by the Monte Carlo simulation method for batteries of varied length, using three different cutoffs to define abnormality. We also compared predicted rates of abnormal performance with those actually shown by 327 reasonably healthy adults. Second, we repeated these analyses after adjusting cognitive test performance for age, sex, race, years of education, and estimated premorbid IQ. Third, we examined the effects of demographic variables and estimated premorbid IQ on predicted rates of abnormal test performance using both unadjusted and adjusted T-scores.

METHOD

Participants

Data used for this analysis were drawn from a community sample of 394 adults recruited from the Baltimore, Maryland, and Hartford, Connecticut, areas to participate in the Aging, Brain Imaging, and Cognition (ABC) study. Participants were recruited by means of random digit dialing, written invitation to Medicare beneficiaries aged 65 and older, and telephone calls to listings selected in pseudorandom manner from residential directories. The ABC study was conducted in two phases. Participants ($n = 215$) who entered the study during phase 1 (1995–1998) were recruited from Baltimore. Those who entered during phase 2 (1999–2005) were recruited from Baltimore ($n = 86$) and Hartford ($n = 93$). Also, 110 phase 1 participants returned in phase 2, but they contributed scores only on tests that were added during phase 2. Thus, no participant was counted twice in the sample. All participants gave written informed consent, and the study was approved by the Johns Hopkins Medicine and Hartford Hospital Institutional Review Boards. Each participant underwent a physical and neurological examination, psychiatric interview, laboratory blood tests, brain magnetic resonance imaging scan, and cognitive testing over 1–2 days. We excluded 26 participants with Parkinson or Alzheimer disease, multiple sclerosis, or epilepsy, prior history of stroke or traumatic brain injury (with >1 hour loss of consciousness), a life-threatening illness such as pancreatic cancer, or a combination of diseases or conditions that together could be life-threatening, such as poorly controlled hypertension with coronary artery disease and a prior myocardial infarction. We also excluded 41 participants who had a history of major depression, bipolar disorder, schizophrenia, or substance dependence (a remote history of depression or substance use disorder was allowed). No person who contributed data to this analysis showed signs of brain dysfunction based on these screening procedures. Finally, we excluded one person who did not complete any cognitive tests. After excluding these participants (17% of the sample), the remaining 327 adults included 185 (57%) women and 142 (43%) men who ranged from 18 to 92 years of age ($M = 54.8$; $SD = 18.8$). They completed from 3 to 20 years of schooling ($M = 14.2$; $SD = 3.0$), and they included 262 (80%) whites, 59 (18%) blacks, and 6 (2%) persons of "other" racial/ethnic background. Overall, 245 participants (74.9%) were from Baltimore and 82 (25.1%) were from Hartford.

Neuropsychological Tests and Measures

Most participants completed 24 cognitive tests from which 43 measures were derived. Obviously, tests added during phase 2 were not administered to the entire sample. Data obtained from every participant who completed a measure and met inclusion criteria were used for these analyses. The tests, measures, and mean (±*SD*) raw scores are shown in Table 1. While all 43 measures were used for the largest battery, check marks (✓) show which measures were used for the 10- and 25-measure batteries. Because Hartford subjects entered the study during phase 2, they were administered a seven-subtest (Ward, 1990) version of the WAIS-III (Wechsler, 1997), rather than the seven-subtest Wechsler Adult Intelligence Scale, Revised (WAIS-R; Wechsler, 1981) given to participants from Baltimore. They also completed the second edition of Conners' Continuous Performance Test (CPT-II; Conners, 2000) rather than the original version administered to Baltimore participants (CPT; Conners, 1995). Most of the other tests are widely known and readily available, but a few are not. One less widely used instrument is a 30-item naming test whose items were drawn from the original, 85-item, experimental version of the Boston Naming Test (BNT; Kaplan et al., 1976). We also developed a recency discrimination task (RDT; Manning et al., 2007) based on this, as follows: Ten minutes after completing this BNT, all 30 items were presented again 2 at a time, and participants were asked which one of each pair they saw most recently. Scores for the RDT can range from 0 to 15 correct. Another less well-known instrument is the Design Fluency Test (DFT) that was first developed by Jones-Gotman and Milner (1977). We recorded the number of novel designs produced in 4 minutes, as described elsewhere (Kingery et al., 2006). We also administered a Verbal Fluency Test (VFT) to assess oral word list generation in response to letter (*S* & *P*) and semantic category (animals, supermarket items) cues during consecutive 1-min trials. Another test that is unique to our battery is the Prospective Memory Test (PMT). On this single-item instrument, participants were instructed to ask the examiner to return a borrowed item at the end of testing and were provided successively more explicit cues about the object and its location if they failed to ask for it back. This PMT was modeled on a similar item from the Rivermead Behavioural Memory Test (Wilson et al., 1985), and scores can range from 0 (best) to 4 (worst), depending on how many cues an examinee requires (Bakker et al., 2002). Two other tests that are rarely used in clinical practice include the Career Abilities Placement Survey (CAPS) Spatial Relations Test (Knapp et al., 1992) and the Perceptual Comparison Test (PCT; Salthouse, 1991). In the CAPS Spatial Relations Test, respon-

Table 1. Cognitive tests, measures included in batteries of 10, 20, and 43 measures, number of participants who completed each test, and raw score means ± standard deviations

Test name	Measure	B_{10}	B_{25}	*N*	Mean ± *SD*	Reference
Mini-Mental State Exam (MMSE)	Total correct	✓	✓	327	28.1 ± 1.7	(Folstein et al., 1975)
Wechsler Adult Intelligence Scale (WAIS-R/WAIS-III)[a]	Information			327	19.8 ± 5.5	(Wechsler, 1981, 1997)
	Digit Span			327	14.7 ± 3.9	
	Arithmetic			327	11.4 ± 3.9	
	Similarities			327	19.2 ± 4.8	
	Picture Completion			327	14.3 ± 3.2	
	Block Design			327	25.8 ± 10.4	
	Digit Symbol/Coding			327	47.5 ± 13.0	
Wechsler Abbreviated Scale of Intelligence (WASI)[b]	Matrix Reasoning			245	22.4 ± 7.3	(Wechsler, 1999)
Shipley Institute of Living Scale (SILS)[b]	Abstraction			231	12.9 ± 4.9	(Zachary, 1986)
Grooved Pegboard Test (GPT)	Dominant hand (sec)		✓	302	80.4 ± 28.1	(Klove, 1963)
	Non-dominant hand (sec)		✓	301	90.5 ± 34.7	
Perceptual Comparison Test (PCT)	Total correct		✓	326	64.5 ± 16.4	(Salthouse, 1991)
Trail Making Test (TMT)	Part A (sec)	✓	✓	327	34.9 ± 17.0	(Reitan, 1958)
	Part B (sec)	✓	✓	324	95.0 ± 69.4	
Conners' Continuous Performance Test (CPT/CPT-II)[c]	Hit reaction time (msec)			306	439.8 ± 68.0	(Conners, 1995, 2000)
	Hit RT standard error			305	6.9 ± 2.5	
	d-prime			305	3.4 ± 0.9	
Brief Test of Attention (BTA)	Total correct	✓	✓	321	15.4 ± 3.7	(Schretlen, 1997)
Wisconsin Card Sorting Test (mWCST)	Correct category sorts		✓	323	5.3 ± 1.3	(Nelson, 1976)
	Perseverative errors		✓	323	2.5 ± 3.9	
Iowa Gambling Task (IGT)[b]	Advantageous draws			229	56.0 ± 14.6	(Bechara, 2007)
Cognitive Estimation Test (CET)	Executive functioning			321	4.6 ± 2.4	(Axelrod & Millis, 1994)
Verbal Fluency (VFT)	Letter		✓	327	28.2 ± 9.2	(Schretlen et al., 2003)
	Category	✓	✓	326	44.8 ± 11.4	
Boston Naming Test (BNT-30)	Correct without cues	✓	✓	325	28.2 ± 2.6	(Kaplan et al., 1976)
Benton Facial Recognition (BFRT)	Total correct (short form)		✓	326	22.4 ± 2.3	(Benton et al., 1994)
Career Abilities Placement Survey (CAPS)[b]	Spatial Relations			229	7.6 ± 3.3	(Knapp et al., 1992)
Rey Complex Figure (RCFT)	Copy trial		✓	327	31.3 ± 4.3	(Rey, 1941)
Clock Drawing (CDT)[b]	Command + copy (sum)	✓	✓	220	9.5 ± 0.8	(Lu et al., 2005)
Design Fluency Test (DFT)	Total novel designs		✓	319	14.2 ± 7.2	(Kingery et al., 2006)
Wechsler Memory Scale (WMS-R)	Logical Memory I		✓	327	26.3 ± 6.9	(Wechsler, 1987)
	Logical Memory II		✓	327	22.4 ± 7.5	
	Visual Reproduction I			327	32.7 ± 6.1	
	Visual Reproduction II			327	22.7 ± 10.6	
Hopkins Verbal Learning Test (HVLT-R)	Total learning (trials 1–3)	✓	✓	327	24.6 ± 4.8	(Brandt & Benedict, 2001)
	Delayed free recall	✓	✓	327	8.7 ± 2.6	
	Delayed recognition	✓	✓	326	10.4 ± 1.6	
Brief Visuospatial Memory Test (BVMT-R)	Total learning (trials 1–3)		✓	327	22.2 ± 7.5	(Benedict, 1997)
	Delayed free recall		✓	327	8.7 ± 2.7	
	Delayed recognition		✓	327	5.6 ± 0.7	
Prospective Memory Test (PMT)	Cues required		✓	298	0.6 ± 0.7	(Bakker et al., 2002)
Recency Discrimination (RDT)	Total correct			327	11.2 ± 1.8	(Manning et al., 2007)

Note. Columns B_{10} and B_{25} show the tests and measures included in batteries composed of 10 and 25 measures, respectively. All 43 measures shown in the table were included in the complete battery.

[a] The WAIS-R was completed by 245 participants recruited from Baltimore, MD; the WAIS-III was completed by 82 participants recruited from Hartford, CT. Values shown in the table are for the WAIS-R. Corresponding raw score means (±*SD*) for the WAIS-III are as follows: Information, 19.8 ± 4.6; Digit Span, 17.7 ± 4.0; Arithmetic, 15.0 ± 3.3; Similarities, 24.9 ± 4.5; Picture Completion, 19.4 ± 3.8; Block Design, 40.4 ± 12.6; and Digit Symbol, 75.5 ± 15.8.

[b] Fewer participants completed his test because it was added to the protocol during phase 2 of the study.

[c] The CPT was completed by 233 participants recruited from Baltimore, MD; the CPT-II was completed by 73 participants recruited from Hartford, CT. Values shown in the table are for the CPT. Corresponding raw score means (±*SD*) for the CPT-II are as follows: Hit RT, 392.8 ± 54.2 ms; Hit RT standard error, 5.2 ± 1.7; and d-prime, 2.5 ± 1.4.

https://doi.org/10.1017/S1355617708080387 Published online by Cambridge University Press

dents are shown three-dimensional geometric figures that are "unfolded" and asked which of five solid figures match the "unfolded" stimuli. We recorded the total number of correct choices made in 5 min. The PCT measures speed of simple information processing based on four timed tasks, each of 30-s duration. Two tasks require the respondent to compare pairs of three- or six-letter strings and mark whether each pair is the same or different. The other two tasks require the respondent to decide whether pairs of three- or six-line designs are the same or different. We used the sum of correct comparisons over all four tasks for analysis. Finally, we used a modified 48-card version of the Wisconsin Card Sorting Test (mWCST; Nelson, 1976) from which we recorded the number of category sorts completed (out of six possible) and the number of perseverative errors.

Data Analysis

We first identified raw scores that fell closest to the 15.87, 6.68, and 2.28 percentiles of the distribution for each measure to approximate T-scores of 40, 35, and 30, respectively. The signs of all scores expressed in seconds or errors were reversed so that high T-scores always reflect better performance than low T-scores. We then counted the participants who fell in each of four T-score intervals: The first included T-scores of 40 or greater. These participants were classified as "normal." The other three T-score intervals (<30, 30–34.99, and 35–39.99) were classified as "abnormal." We also computed nine Cognitive Impairment Index (CII) scores for each participant. These represent the number of tests on which each person obtained "abnormal" scores defined by three cutoffs (<40, <35, and <30) for batteries of 10, 25, and 43 measures (i.e., 3 cutoffs × 3 batteries = 9 CII scores). We correlated the CII scores with age, sex, race, years of education, and estimated premorbid IQ based on the National Adult Reading Test (NART-R; Blair & Spreen, 1989). We also regressed CII scores on these variables using a stepwise procedure.

We next repeated these analyses after adjusting the T-scores for demographic variables and estimated premorbid IQ. General descriptions of the methods used for these adjustments have been reported elsewhere (Heaton et al., 2004; Ivnik et al., 1992). We first converted raw scores to scaled scores ($M = 10$; $SD = 3$) based on the observed distribution of each measure. We then regressed the scaled scores on age, sex, self-reported race (black *vs*. non-black), years of education, estimated premorbid IQ, and squared terms for age, education, and estimated IQ. We then converted the standardized residuals to T-scores. These adjusted T-scores were used to compute rates of abnormal performance based on the same cutoffs used for unadjusted T-scores.

Finally, we calculated predicted rates of abnormal performance using the binomial formula supplied by Ingraham and Aiken (1996) for three test batteries using T-score cutoffs of <40, <35, and <30. We also conducted a series of one million Monte Carlo simulations following Crawford et al. (2007). We derived *R* matrices from batteries of 10, 25, and 43 measures selected from the ABC study assessment. Monte Carlo simulations were conducted using both unadjusted and adjusted T-score cutoffs (<40, <35, and <30) to define abnormal performance.

RESULTS

In a Gaussian distribution, 15.9% of T-scores fall below 40. Consistent with this, 8.4 to 19.2% of our participants scored in this range across all 43 measures. For example, 11.0% scored in this range on the Facial Recognition Test, and 18.3% scored in this range on the delayed recall trial of the Hopkins Verbal Learning Test. Theoretically, 6.7% of T-scores fall below 35 in a normal distribution, and 4.3 to 8.6% of the scores produced by our study participants fell in this range across all 43 measures. Finally, 2.1% of T-scores should fall below 30, and 0.9 to 3.1% of our participants scored in this range. These findings demonstrate that converting raw scores to T-scores based on area transformations of the observed distributions ensured that the expected numbers of participants produced scores falling below the three specified T-score cutoffs.

We then conducted the multiple regression analyses to obtain demographic- and premorbid IQ-adjusted T-scores. Over the 43 measures, 13.2 to 19% of participants produced adjusted T-scores below 40, while 4.9 to 11.0% obtained adjusted T-scores below 35, and 0.9 to 5.6% produced T-scores below 30. Altogether, we conducted 258 χ^2 analyses (43 measures × 3 cutoff scores for both unadjusted and adjusted T-scores). The proportions of subjects whose scores fell below specified cutoffs differed from theoretical expectation for only three measures, each of which included fewer than the expected number of abnormal scores.

Model Predictions of Frequency of Abnormal Performance

We next estimated how many normal healthy adults would produce 2 or more abnormal scores out of 10, 25, and 43 measures when T-score cutoffs of <40, <35, and <30 were used to define abnormality. The percentages of participants predicted to obtain two or more abnormal scores by the binomial model are shown in Figure 1 (the black bar in each grouping, labeled BN_{pre}). Monte Carlo predictions vary, depending on the strength of correlation among measures included in the test battery. Consequently, we conducted separate MC simulations for unadjusted and adjusted test scores because they showed different degrees of correlation: The mean *r*'s were .35 for unadjusted scores and .16 for adjusted scores. Monte Carlo predictions for the unadjusted and adjusted T-scores are depicted by the second and fourth bars of each grouping, labeled "MC_{pre} unadj" and "MC_{pre} adj," respectively. As shown, the BN model usually predicted higher rates of abnormal performance than the MC method, but it predicted marginally lower rates for

https://doi.org/10.1017/S1355617708080387 Published online by Cambridge University Press

the 10- and 25-measure batteries using a T-score cutoff of <30 to define abnormality and for the 10-measure battery using a cutoff of <35 (see top and middle panels of Figure 1).

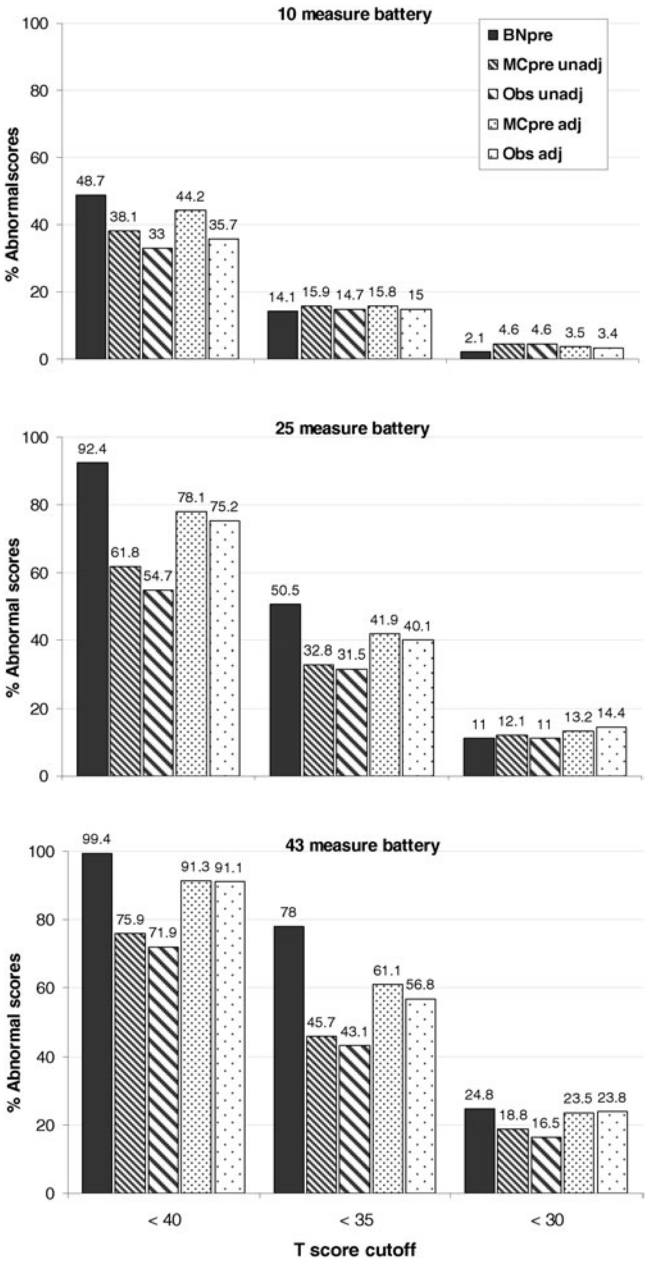

Fig. 1. Predicted and observed percentages of participants who produced two or more abnormal test scores (*y* axis) as defined by three different cutoffs (<40, <35, and <30 T-score points) on test batteries of varied length. The top panel depicts rates of abnormal test performance on a 10-measure battery. The middle and bottom panels show these rates for 25- and 43-measure batteries, respectively. Rates of abnormal performance predicted by the binomial distribution (BNpre) are shown by the first (black) bar in each grouping. Rates of abnormal performance predicted by Monte Carlo simulations are shown for both unadjusted and adjusted T-scores by the second (MCpre unadj) and fourth (MCpre adj) bars in each grouping, respectively. Actual observed rates of abnormal performance are shown for unadjusted and adjusted T-scores by the third (Obs unadj) and fifth (Obs adj) bars in each grouping, respectively.

Predicted *Versus* Observed Rates of Abnormal Performance

For this analysis, we computed the percentages of participants who actually produced 2 or more abnormal scores on batteries of 10, 25, and 43 measures using the three cutoffs. These are shown by the third and fifth bars for unadjusted and adjusted scores, respectively, of each grouping in Figure 1. As expected, the rates increased with the number of tests administered and decreased as the T-score cutoff used to define abnormality was lowered from <40 to <30. The observed rates of abnormal performance were lower than those predicted by the binomial model in most instances. For example, the binomial model predicts that 48.7% of normal individuals will produce 2 or more T-scores below 40 out of 10 measures. In fact, 33% of our participants earned two or more *unadjusted* and 35.7% earned two or more *adjusted* T-scores below 40 [$\chi^2_{(1)} \geq 21.8$; $p < .0001$ for both comparisons]. This is shown in Figure 1, top panel, first group of bars. In many instances, the rates of actual abnormal performance did not differ significantly from BN predictions. For example, the binomial model predicts that 14.1% of normal individuals will obtain 2 or more abnormal scores out of 10 when T-scores below 35 are used to define abnormality, and 14.7 of our participants obtained unadjusted T-scores in this range (Figure 1, top panel, second group of bars). In only one case, the BN model predicted that fewer (2.1%) than the actual (4.6%) number of participants would earn two or more scores [$\chi^2_{(1)} = 9.84$; $p < .002$]. This involved the 10-measure battery and an unadjusted T-score cutoff of <30 (Figure 1, top panel, third group of bars).

We next compared 18 MC simulation predictions with the actual percentages of participants who produced two or more abnormal scores using the three cutoffs for each test battery based on adjusted and unadjusted test scores. Only two of these revealed significant differences: First, the MC method predicted that 44.2% of the sample would produce at least two abnormal adjusted T-scores out of 10 tests using a cutoff of <40, whereas only 35.7% of participants did [$\chi^2_{(1)} = 9.40$; $p < .002$]. Second, the MC method predicted that 61.8% of the sample would obtain at least two abnormal unadjusted T-scores out of 25 tests using the same cutoff, whereas 54.7% did [$\chi^2_{(1)} = 7.11$; $p < .008$]. In every other comparison, the MC predicted and observed rates of abnormal performance differed by 4% or less, and in several cases they were identical.

Correlates of CII Scores

We computed nine CII values (one for each combination of battery length and cutoff used) based on the unadjusted

Table 2. Spearman (ρ) correlations between Cognitive Impairment Index (CII) scores based on unadjusted T-scores and age, sex, race, years of education, and estimated premorbid IQ using cognitive test batteries of varied length and three cutoff scores

Battery length	T-score cutoff	Mean (*SD*)	Min–Max[a]	Age	Sex[b]	Race[c]	Education[d]	NART-R IQ
10 measures	<40	1.49 (1.93)	0–9	.515***	−.046	.240***	−.263***	−.302***
10 measures	<35	0.63 (1.21)	0–7	.404***	−.070	.176*	−.247***	−.277***
10 measures	<30	0.23 (0.62)	0–5	.310***	−.066	.158*	−.254***	−.300***
25 measures	<40	3.63 (4.43)	0–22	.573***	−.029	.215***	−.327***	−.360***
25 measures	<35	1.61 (2.7)	0–18	.528***	−.039	.186**	−.325***	−.354***
25 measures	<30	0.54 (1.28)	0–9	.409***	−.066	.176*	−.312***	−.318***
43 measures	<40	6.23 (7.0)	0–34	.535***	.029	.247***	−.393***	−.417***
43 measures	<35	2.7 (4.16)	0–24	.482***	.023	.219***	−.357***	−.387***
43 measures	<30	0.91 (1.94)	0–15	.359***	−.014	.258***	−.344***	−.369***

Note. * = $p < 0.01$; ** = $p < 0.001$; *** = $p < 0.0001$.
[a] Min = minimum number of abnormal scores produced by any participant; Max = maximum number of abnormal scores produced by any participant.
[b] Coded as 1 = male; 2 = female.
[c] Coded as 1 = non-black; 2 = black.
[d] Highest grade completed.

scores obtained by each participant. As shown in Table 2, the number of abnormal scores each person produced increased with longer batteries and as more lenient cutoffs were used to define abnormal performance. The CII distributions were markedly skewed. Consequently, we used Spearman's rho to test zero-order correlations between CII values and demographic characteristics and estimated premorbid IQ (Table 2). These correlations likely reflect variance in the CII values that is shared by several predictors. To estimate the total variance that can be explained, we regressed each CII rate on age, sex, race, years of education, and estimated premorbid IQ using a stepwise method of variable entry. These analyses all yielded significant models (p's < 0.0001). The resulting R^2 estimates ranged from .211 to .538, indicating that the predictors accounted for 21.1% to 53.8% of the variance in CII. Eight models included three predictors; the ninth included four. Age was the first variable to enter seven of the nine models, and yielded R^2 changes of .137 to .308, indicating that age explained 13.7% to 30.8% of the variance in CII scores in these models. Age entered as the second best predictor (preceded by NART-R scores) in the other two models, where it produced R^2 changes of .105 to .146. In all seven models where age entered first, either NART-R scores or years of education emerged as the second best predictor and explained an additional 7.7% to 25.6% of the variance in CII scores. Race entered as the third best predictor in eight models and education entered third in the ninth model. Race (and education, when it entered third) yielded small R^2 changes (.008 to .048), accounting for less than 2% of the variance in most models. Finally, a term for sex entered as the fourth predictor in one model, accounting for an additional 1.1% of the variance in CII scores. These findings show that normal adults do not produce abnormal test results by chance alone. The strongest predictors of abnormal test performance are age and estimated premorbid IQ or education. Race and sex make modest contributions in some cases.

Effects of Adjusting Test Performance

We next investigated how adjusting test performance for demographic variables and estimated premorbid IQ affects the rates and correlates of abnormal performance. To address this question, we re-computed CII values using demographically and premorbid IQ-adjusted T-scores. Because the CII distributions were highly skewed, we compared matched pairs of adjusted and unadjusted CII values using the nonparametric Wilcoxon signed ranks test. These analyses revealed that unadjusted and adjusted CII values did not differ for the 10-measure battery using any of the three cutoff scores (all Z's < 1.4; p's $> .17$). However, participants produced higher adjusted than unadjusted CII scores for the 25- and 43-measure batteries using all three cutoffs (all Z's > 2.1; p's $< .05$). We also sought to determine whether using adjusted T-scores to compute CII would eliminate the dependence of abnormal test performance on demographic variables and estimated premorbid IQ. As hypothesized, Spearman correlation analyses revealed no significant associations between CII values based on adjusted T-scores and age, sex, race, education, or estimated premorbid IQ. We also repeated the multiple regression analyses using the CII values derived from adjusted T-scores. These analyses did not produce a single significant model, as no predictor variable met entry criteria. Likewise, forcing all five predictors *en bloc* into multiple regression equations failed to yield a single significant model (all p's $> .24$).

DISCUSSION

Four main findings emerged from this study. First, some neurologically normal individuals show abnormal performance on neuropsychological testing. The percentage who do and the number of abnormal scores they produce depend on how many tests are included in a battery and the cutoff score used to define abnormality. Second, healthy adults do not obtain abnormal scores purely by chance. The likelihood varies with age, sex, race, education, and estimated premorbid IQ. Third, adjusting test scores for these characteristics does not reduce the proportion of adults who obtain abnormal scores or the number of such scores they produce, but it eliminates the association of abnormal test performance with demography and premorbid IQ. Fourth, Monte Carlo simulations predict rates of abnormal performance more accurately than the binomial model.

The finding that some normal healthy individuals show abnormal neuropsychological test performance is not new. Using the binomial distribution, Ingraham and Aiken (1996) predicted that 49% of healthy individuals would score >1 *SD* below the mean on at least 2 of 10 cognitive measures, and found that 33% of healthy men reported in another study actually did. Likewise, 35.7% of our participants obtained 2 or more abnormal adjusted T-scores out of 10 using this cutoff. Heaton et al. (2004) reported that 72% of healthy adults scored below 40 on 2 or more of 25 demographically adjusted T-scores derived from an expanded HRB. We found that 75.2% of healthy adults scored below 40 on 2 or more adjusted T-scores derived from a very different set of 25 measures. Axelrod and Wall (2007) found that 29% of healthy young adults scored in the "impaired" range on 3 or more of 7 measures derived from the HRB based on raw score cutoffs recommended by Reitan and Wolfson (1985). Even when we lowered the T-score cutoff to below 30 (i.e., >2 *SD*s below the mean), 11–24% of our participants produced 2 or more abnormal scores on batteries of 25 or 43 measures. In short, the findings that healthy adults often produce abnormal cognitive test scores, and that the likelihood of doing so depends on the cutoff used to define abnormality and number of tests administered, are consistent with previous research.

Perhaps a less expected finding is that abnormal test performance by healthy adults is not due purely to chance, at least not when unadjusted scores are used. Both the binomial model and MC method conceptualize abnormal test performance by healthy adults as due to error. But this does not mean they are *random* events. In our sample, the number of participants who obtained abnormal unadjusted test scores and their CII values correlated with demographic variables and estimated premorbid IQ. Optimal models explained 21% to 54% of the total variance in CII scores. Age accounted for the most variance, followed by estimated premorbid IQ or years of education. Race accounted for relatively little unique variance, and sex entered only a single model. Given that estimated premorbid IQ (Schretlen et al., 2005), current IQ (Diaz-Asper et al., 2004; Horton, 1999; Tremont et al., 1998), and demographic variables (Heaton et al., 2004) all correlate with neuropsychological test performance by healthy adults, it is not surprising that they also predict concurrent rates of abnormal performance. Using demographically homogeneous samples likely minimizes these effects, whereas the demographic diversity of our sample amplifies them.

While demographic characteristics and premorbid IQ are strongly associated with abnormal neuropsychological test performance when unadjusted scores are considered, adjusting test scores for these factors eliminates the association. After adjusting performance using methods described previously (Schretlen et al., 2007), neither the number of adults who produced abnormal scores nor their CII values correlated significantly with age, sex, race, education, or estimated premorbid IQ. In fact, we could not explain significant variance in CII values based on adjusted test scores with any combination of predictor variables. This was not because the adjustments eliminated abnormal test performance. In fact, participants produced slightly more abnormal adjusted than unadjusted scores, although these differences shrank as the stringency of cutoffs used to define abnormality was increased. Rather, adjusting test scores eliminates only the effects of demographic characteristics and premorbid IQ on abnormal performance.

Deviations between predicted and observed rates of abnormal performance are interesting. Ingraham and Aiken (1996) noted that the binomial model overestimates rates of abnormal test performance when measures are correlated. This was true of our study, especially when we used unadjusted scores, lenient cutoffs, and larger test batteries. Demographically adjusting test scores brought actual rates of abnormal performance into closer conformity with BN predictions, likely because it reduced the size of correlations among test measures. The MC method of Crawford et al. (2007) predicted actual rates of abnormal performance more accurately than the binomial model, likely because the MC approach accounts for correlations among the measures, whereas the binomial model assumes that they are uncorrelated. Rates of abnormal performance predicted by MC simulations differed from those shown by actual study participants by less than 9% in all comparisons and by less than 3% in most.

It remains unclear whether abnormal scores represent incidental findings of impairment, anomalous but insignificant poor performance (measurement error), a statistical artifact of the procedure used to define abnormal performance, or some combination of these. In healthy individuals, abnormal demographically adjusted test scores likely represent chance findings, as conceptualized by the binomial and MC simulation methods. However, using adjusted scores does not guarantee that abnormal performance by healthy individuals is clinically meaningless. Adjusted T-scores do not locate a person in the distribution of raw test scores. They locate a person in the theoretical distribution of individuals with identical demographic background and estimated premorbid IQ. It is still possible that abnormal adjusted test scores reflect cerebral dysfunction due to transient changes in cerebral blood flow or metabolism, neurotransmitter availability, or some other aspect of brain functioning. Nor can we exclude the possibility that individual differences in wakefulness, effort, rapport with the examiner, personality, and myriad other factors contribute as well. In other words, this study does not clarify whether—or under what

https://doi.org/10.1017/S1355617708080387 Published online by Cambridge University Press

circumstances—abnormal test performance reflects cerebral dysfunction. We can neither confirm nor reject the possibility that some normal, healthy individuals produce abnormal adjusted test scores due to transient or circumscribed cerebral dysfunction. The present findings permit us to conclude only that such abnormal adjusted test scores are *not* due to individual differences in age, sex, race, education, or premorbid IQ. Ultimately, our findings underscore the distinction between "abnormal" test performance and "impaired" functioning. A test score can be abnormal for many reasons; impaired functioning is but one. The reverse is also true: Individuals with unambiguously impaired brain function can produce normal cognitive test scores. Determining whether abnormal test performance reflects impaired brain function requires inferential reasoning by the clinician. It is not a property of test scores. Indeed, our findings argue against the presence of a one-to-one relationship between abnormal test performance and cerebral dysfunction. Consequently, while this study demonstrates the frequency of abnormal performance by healthy adults, it also underscores the examiner's responsibility for interpreting abnormal findings.

The main limitation of this study, like all normative studies, is that some participants might have had unrecognized health conditions, and these could have accounted for their abnormal performance. However, 17% of the initial participants were excluded due to medical or psychiatric illness, and the health ratings of the remaining participants did not correlate with their CII values. Moreover, even if every member of a given sample enjoys *perfect* health, the distribution of their test scores presumably would still be Gaussian. In this case, similar numbers of subjects would still score >1.0, 1.5, and 2.0 *SD*s below the sample means (which might be shifted up), and the same percentages of individuals should still produce two or more abnormal scores on test batteries of comparable length. Thus, it seems unlikely that health problems account for the observed rates of abnormal test performance. Another weakness is that the sample is smaller than optimal for estimating rates of rare events, such as abnormal test performance defined by the tails of a score distribution. However, the actual rates of abnormal performance were closest to predictions when abnormality was defined by more stringent cutoffs, suggesting that the sample size was large enough to provide robust estimates of abnormal performance in the general population.

Three final implications of this study merit comment. First, using MC simulations to estimate rates of abnormal test performance could be clinically useful. Finding that a patient produces more abnormal scores than expected could strengthen confidence that cerebral dysfunction is responsible. Conversely, finding fewer abnormal scores than expected could help the clinician avoid over-interpreting a few anomalous performances. Second, if abnormal demographically adjusted test scores are more likely than abnormal raw scores to represent chance events, then the use of such adjustments might facilitate efforts to distinguish between benign and pathological patterns of abnormal test performance. We currently are exploring this possibility (Testa & Schretlen, 2006). Finally, the logic and findings reported here likely apply to other diagnostic procedures in psychology and medicine. Whenever multiple correlated measures are obtained, whether these involve psychiatric symptom ratings or laboratory blood tests, clinicians must be alert to the possibility that seemingly "abnormal" findings can occur by chance.

ACKNOWLEDGMENTS

This research was supported by NIMH grant MH60504, The Therapeutic Cognitive Neuroscience Fund, and the Benjamin and Adith Miller Family Endowment on Aging, Alzheimer's, and Autism Research. Dr. Schretlen receives royalties from sales of the Brief Test of Attention. No other coauthor has any conflict of interest related to the research reported in this article. The authors gratefully acknowledge Dr. John R. Crawford for his advice concerning the Monte Carlo simulation, as well as Dr. John Langfitt and Dr. Scott Millis for their advice regarding statistical analyses contemplated for this article. The authors also thank Dr. Yaser Mushtaq for his help with the health ratings.

REFERENCES

Axelrod, B.N. & Millis, S.R. (1994). Preliminary standardization of the Cognitive Estimation Test. *Assessment*, *1*, 269–274.

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.

Bakker, A., Schretlen, D.J., & Brandt, J. (2002). Testing prospective memory: Does the value of a borrowed item help people remember to get it back? *The Clinical Neuropsychologist*, *16*, 64–66.

Bechara, A. (2007). *Iowa Gambling Task Professional Manual*. Odessa, FL: Psychological Assessment Resources, Inc.

Benedict, H.R.B. (1997). *Brief Visuospatial Memory Test–Revised Professional Manual*. Odessa, FL: Psychological Assessment Resources, Inc.

Benton, A.L., Sivan, A.B., deS Hamsher, K., Varney, N.R., & Spreen, O. (1994). *Contributions to Neuropsychological Assessment: A Clinical Manual* (2nd ed.). New York: Oxford University Press.

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

Brandt, J. & Benedict, H.R.B. (2001). *Hopkins Verbal Learning Test–Revised Professional Manual*. Odessa, FL: Psychological Assessment Resources, Inc.

Conners, C.K. (1995). *Conners' Continuous Performance Test*. Toronto, Canada: Multi-Health Systems, Inc.

Conners, C.K. (2000). *Conners' CPT-II, Continuous Performance Test II*. North Tonawanda, NY: Multi-Health Systems Inc.

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.

Crum, R.M., Anthony, J.C., Bassett, S.S., & Folstein, M.F. (1993). Population-based norms for the mini-mental state examination by age and educational level. *JAMA*, *269*, 2386–2391.

https://doi.org/10.1017/S1355617708080387 Published online by Cambridge University Press

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.

Folstein, M.F., Folstein, S.E., & McHugh, P.R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*, 189–198.

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*. Lutz, FL: Psychological Assessment Resources, Inc.

Horton, A.M., Jr. (1999). Above-average intelligence and neuropsychological test score performance. *International Journal of Neuroscience*, *99*, 221–231.

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.

Ivnik, R.J., Malec, J.F., Smith, G.E., & Tangalos, E.G. (1992). Mayo's older americans normative studies: WAIS–R norms for ages 56 to 97. *The Clinical Neuropsychologist*, *6*, 1–30.

Janssen, R.S., Saykin, A.J., Cannon, L., Campbell, J., Pinsky, P.F., Hessol, N.A., O'Malley, P.M., Lifson, A.R., Doll, L.S., Rutherford, G.W., & Kaplan, J.E. (1989). Neurological and neuropsychological manifestations of HIV-1 infection: Association with AIDS-related complex but not asymptomatic HIV-1 infection. *Annals of Neurology*, *26*, 592–600.

Jones-Gotman, M. & Milner, B. (1977). Design fluency: The invention of nonsense drawings after focal cortical lesions. *Neuropsychologia*, *15*, 653–674.

Kaplan, E., Goodglass, H., & Weintraub, S. (1976). *Boston Naming Test. Experimental Edition*. Boston: Aphasia Research Center, Boston University.

Kingery, L.R., Schretlen, D.J., Sateri, S., Langley, L.K., Marano, N.C., & Meyer, S.M. (2006). Interrater and test-retest reliability of a fixed condition design fluency test. *The Clinical Neuropsychologist*, *20*, 729–740.

Klove, H. (1963). Clinical neuropsychology. In F.M. Forster (Ed.), *Medical Clinics of North America*. New York: Saunders.

Knapp, L., Knapp, R.R., & Knapp-Lee, L. (1992). *Career Ability Placement Survey: CAPS Technical Manual*. San Diego, CA: EdITS.

Lu, L., Yun, J., Meyer, S.M., & Schretlen, D.J. (2005). *Interrater reliability, construct validity, and normative data for the clock drawings of normal adults*. Paper presented at the International Neuropsychological Society, 33rd Annual Meeting, St. Louis, MO, p. 80.

Manning, K.J., Gordon, B., Pearlson, G.D., & Schretlen, D.J. (2007). The relationship of recency discrimination to explicit memory and executive functioning. *Journal of the International Neuropsychological Society*, *13*, 710–715.

Nelson, H.E. (1976). A modified card sorting test sensitive to frontal lobe defects. *Cortex*, *11*, 918–932.

Reitan, R.M. (1958). Validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and Motor Skills*, *8*, 271–276.

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

Rey, A. (1941). L'examen psychologique dans les cas d'encephalopathie traumatique. (Les problems). *Archives de Psychologie*, *28*, 215–285.

Salthouse, T.A. (1991). Mediation of adult age differences in cognition by reductions in working memory and speed of processing. *Psychological Science*, *2*, 179–183.

Schretlen, D.J. (1997). *Brief Test of Attention Professional Manual*. Odessa, FL: Psychological Assessment Resources, Inc.

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., Cascella, N.G., Meyer, S.M., Kingery, L.R., Testa, S.M., Munro, C.A., Pulver, A.E., Rivkin, P., Rao, V.A., Diaz-Asper, C.M., Dickerson, F.B., Yolken, R.H., & Pearlson, G.D. (2007). Neuropsychological functioning in bipolar disorder and schizophrenia. *Biological Psychiatry*, *62*, 179–186.

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.

Testa, S.M. & Schretlen, D.J. (2006). Diagnostic utility of regression based norms in schizophrenia. *The Clinical Neuropsychologist*, *20*, 206.

Tremont, G., Hoffman, R.G., Scott, J.G., & Adams, R.L. (1998). Effect of intellectual level on neuropsychological test performance: A response to Dodrill (1997). *The Clinical Neuropsychologist*, *12*, 560–567.

Ward, L.C. (1990). Prediction of verbal, performance, and full scale IQs from seven subtests of the WAIS-R. *Journal of Clinical Psychology*, *46*, 436–440.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale–Revised*. New York: Psychological Corporation.

Wechsler, D. (1987). *Wechsler Memory Scale–Revised*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale–Third Edition*. San Antonio, TX: The Psychological Corporation Harcourt Brace & Company.

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence*. San Antonio, TX: Harcourt Assessment, Inc.

Wilson, B.A., Cockburn, J., & Baddeley, A.D. (1985). *The Rivermead Behavioural Memory Test*. Bury St. Edmunds, UK: Thames Valley Test Company.

Zachary, R.A. (1986). *Shipley Institute of Living Scale Revised Manual*. Los Angeles: Western Psychological Services.

https://doi.org/10.1017/S1355617708080387 Published online by Cambridge University Press