Journal of the International Neuropsychological Society (2003), **9**, 864–870.

DOI: 10.1017/S1355617703960061

Examining the range of normal intraindividual variability in neuropsychological test performance

DAVID J. SCHRETLEN,[1] CYNTHIA A. MUNRO,[1] JAMES C. ANTHONY,[1,2] AND GODFREY D. PEARLSON[1,2]

[1]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland
[2]Department of Mental Hygiene, Johns Hopkins University School of Public Health, Baltimore, Maryland

(RECEIVED March 28, 2002; REVISED November 6, 2002; ACCEPTED November 13, 2002)

Abstract

Neuropsychologists often diagnose cerebral dysfunction based, in part, on marked variation in an individual's cognitive test performance. However, little is known about what constitutes the normal range of intraindividual variation. In this study, after excluding 54 individuals with significant health problems, we derived 32 *z*-transformed scores from 15 tests administered to 197 adult participants in a study of normal aging. The difference between each person's highest and lowest scores was computed to assess his or her maximum discrepancy (MD). The resulting MD values ranged from 1.6 to 6.1 meaning that the *smallest* MD shown by any person was 1.6 standard deviations (*SD*s) and the *largest* MD shown by any person was 6.1 *SD*s. Sixty-six percent of participants produced MD values that exceeded 3 *SD*s. Eliminating each person's highest and lowest test scores decreased their MDs, but 27% of the participants still produced MD values exceeding 3. Although MD values appeared to increase with age, adjusting test scores for age, which is standard in clinical practice, did not correct for this. These data reveal that marked intraindividual variability is very common in normal adults, and underscore the need to base diagnostic inferences on clinically recognizable patterns rather than psychometric variability alone. (*JINS*, 2003, *9*, 864–870.)

Keywords: Neuropsychology, Clinical inference, Intraindividual variability

INTRODUCTION

A "statistically significant" difference between any given pair of cognitive test scores means that the probability of obtaining such a discrepancy by chance or measurement error is low (e.g., $p < .05$) if the "true" difference between the scores is zero (Matarazzo & Herman, 1985). The underlying assumption of this approach to the assessment of intraindividual variability is that the person's "true" abilities (a_1 and a_2), as measured by the test score pair, are identical. The null hypothesis in this case could be expressed as H_0: $a_1 = a_2$. By extension, when a battery of neuropsychological tests measuring i abilities is administered, and all possible pairs of scores are compared, the null hypothesis is that all of the "true" scores are identical (i.e., H_0: $a_1 = a_2 = \cdots a_i$). However, even among normal, healthy persons there is little reason to expect that the null hypothesis is correct. Both the complexity of the human central nervous system and individual differences in the organization of neural circuits on which various mental abilities depend argue against the likelihood that any individual will be endowed with identical levels of ability across all domains of cognitive functioning. Empirical evidence also has shown that this null hypothesis is incorrect for most normal people, even over the restricted range of abilities assessed by the Wechsler Adult Intelligence Scale–Revised (WAIS–R; Wechsler, 1981) or Wechsler Adult Intelligence Scale–Third Edition (WAIS–III; Wechsler, 1997).

In a series of studies, Matarazzo and colleagues examined the distributions of Verbal minus Performance IQ score differences (VIQ–PIQ) and inter-subtest "scatter" produced by participants in the WAIS–R standardization sample (Matarazzo & Herman, 1985; Matarazzo et al., 1988; Matarazzo & Prifitera, 1989). These analyses revealed that "statistically significant" differences between each person's highest and lowest WAIS–R subtest scores were the rule rather than the exception. For example, 86% of partici-

Reprint requests to: David J. Schretlen, Ph.D., Johns Hopkins Hospital, 600 N. Wolfe St., Meyer 218, Baltimore, MD 21287-7218. E-mail: dschret@jhmi.edu

pants in the WAIS–R standardization sample and 87% of those in the WAIS–III standardization sample produced differences of ≥ 5 scaled score points (i.e., $p < .05$) between their highest and lowest subtest scores (Matarazzo & Prifitera, 1989; Wechsler, 1997). In fact, 18.1% of those in the WAIS–R standardization sample and 17.4% of persons in the WAIS–III standardization sample produced discrepancies of ≥ 9 scaled score points (i.e., 3 *SD*s) between their highest and lowest subtest scores. Given that the WAIS–R and WAIS–III assess the single construct of intelligence (with some fractionation of verbal, visual-spatial, processing speed, and working memory abilities), it is possible that normal persons will show even greater intraindividual variability across the broader range of abilities measured by a comprehensive neuropsychological test battery.

These findings should concern neuropsychologists, who often base clinical inferences about the presence of cerebral dysfunction, at least in part, on marked variation in a patient's level of cognitive test performance. Citing Silverstein (1982), for example, Lezak (1995) wrote: "The basic element of test score analysis is a significant discrepancy between any two or more scores," and added that "marked quantitative discrepancies in a person's performance suggest that some abnormal condition is interfering with that person's overall ability to perform at the characteristic level of cognitive functioning" (p. 165).

The aim of this study was to determine how much intraindividual variability healthy adults would show on a reasonably comprehensive battery of neuropsychological tests.

METHOD

Participants

During Phase 1 of the Johns Hopkins Aging, Brain Imaging, and Cognition (ABC) study, 214 adults were recruited primarily through random-digit dialing from households in the Baltimore metropolitan area. Another 37 participants were recruited during Phase 2 of the same ongoing study. Because the ABC study concerns "usual" rather than "optimal" aging, a substantial number (54) of the 251 participants had significant health problems and were excluded from further analyses. The remaining 197 participants ranged from 20 years to 92 years of age ($M = 55.1$; $SD = 19.1$). They included more women (56.9%) than men (43.1%) and more non-Hispanic Caucasians (79.7%) than African Americans (18.8%) and persons of other racial/ethnic backgrounds (1.5%). The participants completed from 3 years to 20 years of schooling ($M = 13.9$; $SD = 3.1$). They produced a mean Mini Mental State Examination (MMSE) score of 28.5 ($SD = 1.3$), a mean WAIS–R Full Scale IQ of 105.9 ($SD = 14.7$), and a mean New Adult Reading Test–Revised (NART–R; Blair & Spreen, 1989) estimated Full Scale IQ of 103.5 ($SD = 10.9$). In short, the final sample was broadly representative of normal community-dwelling adults.

Procedure

After giving written informed consent to participate in this study, which was approved by the Johns Hopkins University IRB, each participant provided a health history and underwent physical and neurological examinations, a structured psychiatric interview (Schedule for Clinical Assessment in Neuropsychiatry; Wing et al., 1996), brain magnetic resonance imaging, and neuropsychological testing. The entire assessment required approximately 6 hr to complete. On completion of the protocol, each person was paid a stipend for his or her participation. Although Phase 2 of the ABC study includes both cross-sectional and longitudinal components, only cross-sectional data were used for the present analyses.

Neuropsychological assessment

Each participant was administered a battery of 15 tests from which 32 measures were derived. The tests and measures used for the present analyses included Ward's (1990) seven-subtest short form of the WAIS–R from which raw scores from the Information, Digit Span, Arithmetic, Similarities, Picture Completion, Block Design, and Digit Symbol subtests were used. To avoid the inclusion of overlapping measures, WAIS–R Verbal, Performance and Full Scale IQ scores were not used. Language screening included a 30-item version of the Boston Naming Test (BNT; Goodglass & Kaplan, 1983) from which the total number of spontaneously named pictures was used, as well as two tests of Verbal Fluency, in which the total numbers of acceptable words beginning with the letters S and P (Letter Fluency) and animal names and supermarket items (Semantic Fluency) during consecutive 1-min trials, also were used. The times required to complete Parts A and B of the Trail Making Test (Reitan, 1958) were recorded, as were hit reaction times on the Conners' Continuous Performance Test (CPT) (Conners, 1995). From Nelson's (1976) modification of the Wisconsin Card Sorting Test (WCST), we recorded the number of category sorts completed and perseverative errors committed. We computed total deviation scores for the Cognitive Estimation Test (Axelrod & Millis, 1994), total correct on the Brief Test of Attention (Schretlen, 1997), and mean times to complete the Grooved Pegboard Test (Kløve, 1963) on two trials with each hand. From the Wechsler Memory Scale–Revised (WMS–R; Wechsler, 1987), we derived four measures, including immediate and delayed recall on the Logical Memory and Visual Reproduction subtests. In addition, we used measures of total learning over trials, delayed free recall, and delayed recognition ("hits" minus "false positive" errors) from both the Hopkins Verbal Learning Test–Revised (HVLT–R; Brandt & Benedict, 2001) and the Brief Visuospatial Memory Test–Revised (BVMT–R; Benedict, 1997). The number of correctly identified faces was recorded from the Facial Recognition Test (Benton et al., 1994), as were the total number of acceptable drawings produced in 4 min for the Design Fluency Test (Jones-

https://doi.org/10.1017/S1355617703960061 Published online by Cambridge University Press

Table 1. Means and standard deviations of raw scores for each neuropsychological test measure

Test measure[a]	*n*	*M* ± *SD*	Test measure[a]	*n*	*M* ± *SD*
WAIS–R Info[b]	197	20.1 ± 5.4	BNT–30	196	28.4 ± 2.4
WAIS–R DSp[b]	197	15.0 ± 3.8	VFT (letter)	197	27.3 ± 8.4
WAIS–R Arith[b]	197	11.7 ± 3.6	VFT (category)	197	42.9 ± 10.4
WAIS–R Sim[b]	197	19.2 ± 4.5	Design Fluency	195	15.1 ± 7.6
WAIS–R PC[b]	197	14.6 ± 2.9	Facial Recognition	196	22.1 ± 2.3
WAIS–R BD[b]	197	26.8 ± 9.6	Rey CFT (copy)	197	30.8 ± 4.3
WAIS–R DSym[b]	197	48.5 ± 12.5	WMS–R LM–I	197	26.7 ± 6.9
GPT Dom (s)	169	81.3 ± 27.2	WMS–R LM–D	197	22.6 ± 7.6
GPT N-Dom (s)	168	92.6 ± 36.4	WMS–R VR–I	197	32.9 ± 5.8
Trail Making A (s)	197	35.3 ± 15.0	WMS–R VR–D	197	23.6 ± 10.3
Trail Making B (s)	196	92.5 ± 57.1	HVLT–R (trials 1–3)	197	24.5 ± 4.8
Brief Test of Attention	197	14.9 ± 3.8	HVLT–R (delay)	197	8.8 ± 2.6
mWCST (cat.)	196	5.3 ± 1.2	HVLT–R (recog.)	197	10.4 ± 1.4
mWCST (per. err.)	195	2.4 ± 3.3	BVMT–R (trials 1–3)	197	22.7 ± 7.2
CPT Hit RT (ms)	190	436 ± 66	BVMT–R (delay)	197	8.9 ± 2.5
Cognitive Estimation	193	4.8 ± 2.5	BVMT–R (recog.)	197	5.6 ± 0.7

[a]WAIS–R Info, DSp, Arith, PC, BD, and DSym = Information, Digit Span, Arithmetic, Picture Completion, Block Design, and Digit Symbol, respectively; GPT Dom and N-Dom = Grooved Pegboard Test dominant and nondominant hands, respectively; mWCST cat. and per. err. = modified Wisconsin Card Sorting Test category sorts and perseverative errors, respectively; CPT Hit RT = "hit" reaction time; VFT letter and category = Verbal Fluency Test for letter (S & P) and semantic category (animal names & supermarket items) cues, respectively; WMS–R LM–I, LM–D, VR–I and VR–D = Wechsler Memory Scale–Revised Logical Memory (immediate and delayed) and Visual Reproduction (immediate and delayed), respectively; HVLT–R and BVMT–R trials 1–3, delay, and recognition = learning over trials, delayed recall, and delayed recognition ("hits" minus "false positive" errors), respectively, for the Hopkins Verbal Learning Test–Revised and Brief Visuospatial Memory Test–Revised.
[b]For ease of interpretation, mean (± *SD*) age-corrected subtest scores for the WAIS–R were: Information = 11.0 ± 3.2, Digit Span = 10.5 ± 2.6, Arithmetic = 10.3 ± 2.9, Similarities = 11.0 ± 2.6, Picture Completion = 11.1 ± 2.6, Block Design = 10.9 ± 2.7, and Digit Symbol = 10.7 ± 2.4.

Gotman & Milner, 1977). Finally, each person's copy accuracy (Meyers & Meyers, 1995) was recorded for the Rey-Osterrieth Complex Figure Test (Rey, 1993).

Data Analysis

First, the data from 54 participants whose examinations or health histories revealed moderate or severe health problems, such as dementia, Parkinson's disease, multiple sclerosis, current major depression, a history of stroke or traumatic brain injury with > 1 hr loss of consciousness, alcohol or drug dependence, and other conditions thought to involve the central nervous system, or who earned MMSE (Folstein et al., 1975) scores below 24/30, were excluded. Then, the 197 participants' scores on each of the 32 neuropsychological measures described above were *z*-transformed using SPSS for Windows Release 10.0.5 (SPSS, Inc.). Tests for which higher scores denote worse performance were multiplied by −1 to ensure that positive values always reflected better performance than negative values. Finally, the maximum discrepancy (MD) between each person's highest and lowest scores was derived as follows: Initially, each participant was assigned a single row of data, with the columns consisting of his or her *z*-transformed test scores. To derive MD values, the columns and rows were transposed, so that each person was assigned to a *column*, and each neuropsychological test variable was assigned to a *row*. Thereafter, each person's lowest score was subtracted from his or her highest score, yielding an MD value in standard deviation units for each participant.

RESULTS

For descriptive purposes, the means and standard deviations of performance on each neuropsychological test measure are shown in Table 1.

The MD values ranged from 1.6 to 6.1, with a mean of 3.4 (*SD* = 0.8), indicating that the smallest MD shown by any participant was 1.6 *SD*s, and the largest MD value shown by any participant was 6.1 *SD*s. As shown in Figure 1 (black bars), 65% of the participants produced MD values of 3.0 or greater, meaning that their best and worst cognitive test performances differed by at least 3 *SD*s (i.e., 45 standard score points), and 20% of the participants produced MD values of 4.0 or greater, meaning that their best and worst test scores differed by at least 4 *SD*s (i.e., 60 standard score points). To assess whether a few test score "outliers" inflated the MD distribution, we also computed MD values after removing each person's single highest and lowest *z*-scores.* As intended, doing so attenuated the range of *z*-transformed test scores. This shifted the distribution of

*Due to changes in the study protocol and technical problems, 37 (18.8%) participants were missing 1 or 2 test scores, and another 5 (2.5%) were missing 3 or 4 test scores.

https://doi.org/10.1017/S1355617703960061 Published online by Cambridge University Press

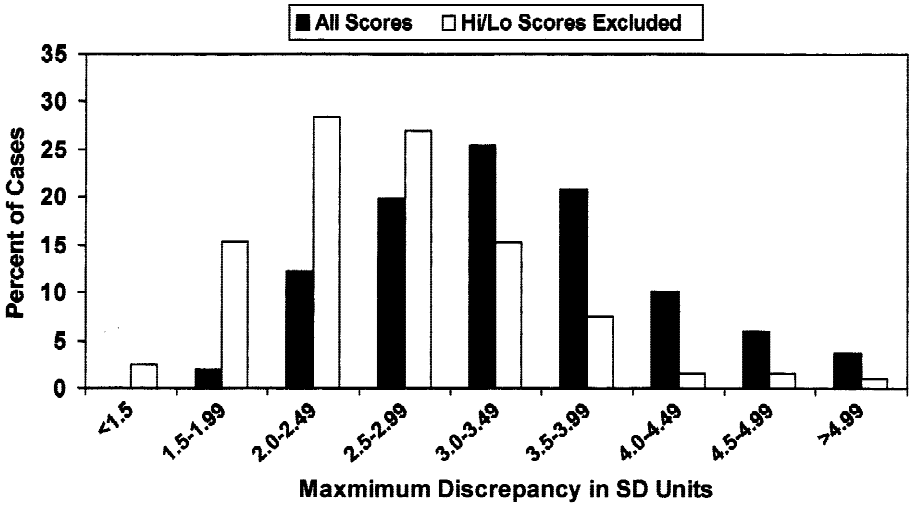

Fig. 1. Frequency distribution showing the percent of participants who produced MD values within specified ranges (expressed in *SD* units). The black bars depict the distribution of MD values based on 100% of each person's non-age-corrected cognitive test *z*-scores. The white bars depict the distribution of MD values after each person's highest and lowest test scores were excluded (Hi/Lo exclusion).

MD values downward, so that they then ranged from 1.4 *SD*s to 5.1 *SD*s, with a mean of 2.7 (Figure 1, white bars). Nevertheless, 27% of the participants still produced MD values of 3 or more standard deviations based on their remaining test scores. Not a single person showed "consistent" performance, when consistency was defined by an MD of 1 *SD* or less.

It could be argued that clinicians are more concerned with test scores that are very low than very high, and that many practitioners base their expectations of a given patient's neuropsychological test performance on his or her estimated "premorbid" ability. For this reason, we next converted each participant's NART–R estimated Full Scale IQ to a *z*-score, and then subtracted the lowest of his or her *z*-scores on the other 32 test measures from that. The resulting discrepancy scores ranged from −.6 to 5.1 ($M = 1.9$; $SD = 1.1$). This indicates that the average person's lowest neuropsychological test score fell 1.9 *SD*s below his or her estimated "premorbid" IQ. In fact, the lowest *z*-score fell more than 2 *SD*s below the estimated IQs for 86 (44%) participants, and more than 3 *SD*s below the estimated IQs for 35 (18%) participants. For purposes of comparison, each person's *z*-transformed IQ estimate also was subtracted from the highest of his or her other *z*-scores. The resulting differences ranged from −.5 to 3.6 ($M = 1.5$; $SD = .9$). This indicates that the average person's highest neuropsychological test score exceeded his or her estimated "premorbid" IQ by 1.5 *SD*s. In fact, the highest *z*-score exceeded estimated IQ by at least 2 *SD*s for 63 (32%) participants and by at least 3 *SD*s for 14 (7%) participants. Seven individuals produced NART–R estimated IQs that were lower than their *z*-scores on all of the other test measures, and eight individuals produced NART–R estimated IQs that exceeded their *z*-scores on all of the other cognitive measures.

Because our participants ranged from 20 years to 92 years of age, and some cognitive tests are more affected by age than others, we hypothesized that age-effects might have inflated the MD values. Correlation of the MD values with age yielded a Pearson *r* of .22 ($p < .002$), indicating that advancing age was associated with greater intraindividual variability in cognitive test performance. To further assess this, we regressed each cognitive measure on age, saved the residuals as *z*-transformed scores, and again computed MD values following the procedures described above. Contrary to our hypothesis, this did *not* reduce the resulting MD values, which then ranged from 1.7 to 6.2, with a mean of 3.5. Taken together, these findings suggest that intraindividual variability increases with age, and that age-adjusting the test scores does not appear to correct for this.

In an effort to determine whether a few specific tests with limited variance or highly skewed distributions, such as the BNT, might account for most of the extreme MD values, we reviewed the test scores of all 39 individuals whose MD values were 4.0 or greater. Altogether, 27 of the 32 test measures appeared as the highest or lowest *z*-score of at least one of these participants. Only five measures appeared as the highest or lowest score in more than four extreme MD score pairs. These included the HVLT–R (delayed recognition), BVMT-R (delayed recognition), Verbal Fluency (letters S & P), Cognitive Estimation, and CPT "hit" reaction time. Although Verbal Fluency invariably appeared as the high score in extreme MD score pairs, the other four measures appeared about equally often as the highest or lowest test score in extreme MD score pairs. Twenty-two test measures appeared as the highest or lowest score in 1–4 extreme MD score pairs. The only test variables that never appeared as the highest or lowest score of participants with extreme intraindividual variability included the BVMT–R (learning over trials), Block Design and Digit Symbol from the WAIS–R, Verbal Fluency (semantic), and the WMS–R Visual Reproduction subtest (immediate recall). Finally, contrary to expectation, four of the test measures that appeared most frequently in extreme MD score pairs showed less skewness (−.91 to .11) and kurtosis (−.27 to .74) than most of the other 27 test measures, whose skewness and kurtosis estimates ranged from −2.40 to .71 and from −.86 to 7.30, respectively. The BVMT–R delayed recognition measure appeared in six extreme MD score pairs

https://doi.org/10.1017/S1355617703960061 Published online by Cambridge University Press

(as highest in 3 and lowest in 3), and was found to be both negatively skewed (−2.07) and lepokurtic (4.25). Apart from this single measure, however, the present results suggest that even extreme MD values were not artifacts of just a few tests with peculiar distributions. Rather, 27 of the 32 measures that comprised our cognitive test battery appeared as the highest or lowest score for at least one of the 39 participants with extreme intraindividual variability. Only five of these test measures appeared in more than four extreme MD score pairs.

Finally, because increasing IQ has been associated with greater inter-subtest scatter and larger VIQ–PIQ discrepancies on the WAIS–R and WAIS–III, we examined the relationship between MD and IQ in the present sample. Surprisingly, MD values showed weakly negative Pearson correlations with WAIS–R Full Scale IQ scores ($r = -.14$; $p = .043$) and NART–R IQ estimates ($r = -.17$; $p = .014$). These mean that intraindividual variability did not increase with better performance on the WAIS–R or NART–R and, if anything, actually declined. Further, the correlation between MD and the mean of each participant's cognitive test *z*-scores was −.51 ($p < .0001$), indicating that intraindividual variability actually increased with *poorer* overall cognitive test performance, as shown in Figure 2.

DISCUSSION

In this study, 197 healthy adults demonstrated marked intraindividual variability in their performances on 32 measures derived from 15 neuropsychological tests. Only four participants (2% of the sample) produced discrepancies of *less than* 2 *SD*s between their highest and lowest test scores, while 130 (66% of the sample) showed discrepancies of 3 or more standard deviations between their highest and lowest test scores. Excluding each person's highest and lowest test scores attenuated the range of intraindividual variability. However, 53 individuals (27% of the sample) still showed discrepancies of at least 3 *SD*s between the highest and lowest of their remaining test scores. If one defines "consistent" neuropsychological test performance by scores that all fall within ±1 *SD* of each other, no participant demonstrated "consistent" performance across the test measurements included in this battery.

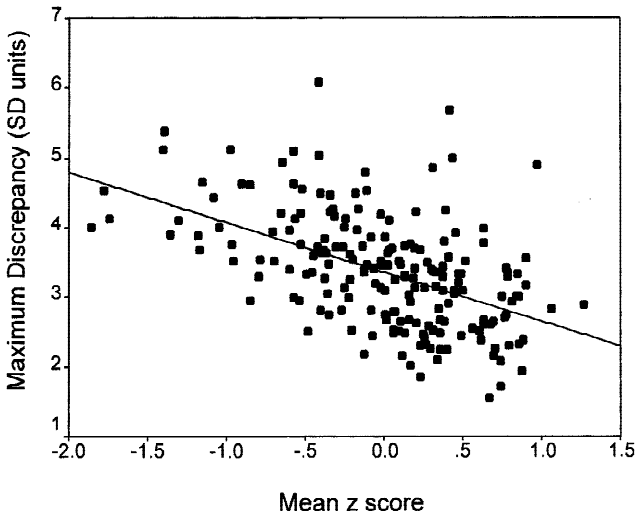

Fig. 2. Scatterplot depicting each participant's MD (in *SD* units) as a function of the mean of his or her *z* scores on all 32 cognitive test measures. The Pearson *r* was −.51, indicating that intraindividual variability decreased with better overall cognitive test performance.

Considered in the context of "premorbid" ability, each person's lowest *z*-score on the 32 test measures fell a mean of 1.9 *SD*s below his or her NART–R Full Scale IQ estimate. Conversely, each person's highest *z*-score on the 32 measures exceeded his or her NART–R IQ estimate by a mean of 1.5 *SD*s. These findings suggest that marked "downward" discrepancies between a person's estimated IQ and his or her poorest cognitive test performance are slightly more common than "upward" discrepancies. In any case, these findings clearly refute the notion that most individuals are endowed with equal ability across the spectrum of cognitive functions.

At one level, these findings are hardly surprising: Both the complexity of the central nervous system and minor variations in cerebral architecture make it highly unlikely that the "null hypothesis" of equal endowment in all abilities would hold true for most persons. Indeed, statistically significant differences between various WAIS–R and WAIS–III score pairs have been shown to characterize most normal individuals, despite the fact that these test batteries assess a relatively narrow range of abilities (Matarazzo & Herman, 1985; Matarazzo & Prifitera, 1989; Wechsler, 1997). More surprising was the magnitude of intraindividual variability that emerged and its persistence after eliminating each person's highest and lowest test scores. Perhaps most surprising, however, was the fact that many tests—rather than just a few—contributed extreme scores to the profiles of those who showed the largest intraindividual variations in performance. In fact, only five tests appeared in more than four of 39 score pairs with extreme MD values, and the distributions of most of these showed less skewness and kurtosis than many other tests (including some WAIS–R subtests). Thus, the intraindividual variability shown by this study cannot be attributed to a small number of tests with peculiar psychometric characteristics.

Interindividual variability in cognitive test performance has been found to increase with advancing age (Christensen et al., 1999; Schaie, 1994). Although the impact of age on intraindividual variability has received less research attention, Hultsch et al. (2002) recently reported that older adults showed greater intraindividual variability in reaction times than younger adults. Consistent with this, the present study revealed that MD values increased with age. However, the correlation was modest, as age accounted for only about 5% of the variance in MD values. Further, age-correcting the cognitive test scores, as is usually done in clinical practice, did *not* reduce intraindividual variability.

https://doi.org/10.1017/S1355617703960061 Published online by Cambridge University Press

The finding that intraindividual variability showed only a modest inverse correlation with IQ is interesting. It contrasts with previous demonstrations that within-person intersubtest "scatter" and VIQ–PIQ discrepancies increase with better overall performance on the WAIS–R and WAIS–III (Matarazzo & Herman, 1985; Matarazzo & Prifitera, 1989; Wechsler, 1997). In the present study, MD values not only failed to correlate positively with IQ, they actually showed a significant negative correlation with mean performance on the 32 measures that comprised our neuropsychological test battery. One possible explanation of this apparent contradiction is that the WAIS–R and WAIS–III were designed to measure abilities rather than deficits. As such, they generally yield score distributions that are more Gaussian than those of most neuropsychological measures. Because mean scores on several neuropsychological tests (e.g., Boston Naming, Wisconsin Card Sorting, Trail Making, and Grooved Pegboard) are relatively close to their ceilings, there might be less "opportunity" for normal persons to demonstrate marked intraindividual variability at the upper ends of the distributions on these tasks. This could explain why the average lowest *z*-score for those 39 individuals with extreme MD values was −3.1, whereas the average highest *z*-score for the same individuals was 1.5.

Though perhaps not surprising in themselves, the present findings have important implications for clinical neuropsychological inference. As in any medical diagnostic workup, the process of reaching a diagnostic neuropsychological formulation begins with the question of whether a given patient's examination is normal or abnormal. To answer this question, the neuropsychologist will consider the patient's presenting complaints, other diagnostic test results, history, appearance and behavior, and approach to cognitive tasks. In addition, a critical component of clinical neuropsychological inference involves the determination of whether a patient demonstrates "marked quantitative discrepancies" among his test scores (Lezak, 1995). To the extent that neuropsychologists approach the process of clinical inference in this fashion, the present findings should concern every practitioner because they show that many healthy adults demonstrate "marked quantitative discrepancies" among their cognitive test performances. Indeed, such discrepancies appear to be the rule rather than the exception.

Anticipating the question of whether some characteristic of our sample or some aspect of our approach to the data analysis might have inflated the degree of intraindividual variability observed here, we took steps to *minimize* variability. First, we excluded 54 participants with significant health problems in order to reduce the possibility that persons with cognitive deficits might inflate the range of intraindividual variability. Second, we derived all of the measures used for statistical analyses from the distributions of test scores produced by the sample, rather than from published norms, to avoid spuriously inflating MD values due to differences among the samples used for the standardization of each test. For both of these reasons, we believe that the maximum discrepancy values reported here are no larger than one would find with a similar cognitive test battery using tabled values to compute the age-adjusted scores.

Although all of the participants in this study lived independently in the community and were recruited for a study of normal aging, we cannot exclude the possibility that a few suffered from undiagnosed cognitive disorders that inflated the magnitude of observed intraindividual variability in cognitive test performance. However, each participant provided a health history and received thorough physical, neurological, and psychiatric examinations, on the basis of which we excluded 22% of the participants due to medical, neurological, or psychiatric illness in order to reduce this risk. Furthermore, our sample produced a mean prorated WAIS–R Full Scale IQ of 105.9 ($SD = 14.7$), which is quite close to what one would expect from a broadly representative sample of normal, healthy adults. Thus, we doubt that undiagnosed cognitive disorders significantly increased the intraindividual variability observed in this study.

It could be argued that because short forms of tests generally are associated with lower reliability than long forms, our use of several short forms might have inflated the obtained estimates of intraindividual variability. However, this is unlikely for several reasons. First, only five of the 32 cognitive test measures were based on short forms. These included the BNT (based on 30 items), WCST category sorts and perseverative errors (based on 48 cards), Letter Word Fluency (based on the letters S & P rather than F, A, & S), and Facial Recognition Test (based on 27 items). This concern about the use of short forms does not apply to the WAIS–R because all seven subtests were administered in their entirety and IQ scores were not included with the 32 *z*-scores used to compute MD values. Further, only one of the short form tests appeared in more than four extreme MD score pairs. Finally, most short form measures, except for those derived from the modified WCST, showed very little skewness or kurtosis.

Ultimately, 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.

ACKNOWLEDGMENTS

Portions of the findings reported here were presented at the Thirtieth Annual International Neuropsychological Society Conference, February 13, 2002, Toronto, Ontario, Canada. This research was supported, in part, by grants 2 R01 AG11859-04 from the National Institute on Aging and 5 R01 MH43775-11 from the National Institute of Mental Health, NIH.

REFERENCES

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

Benedict, H.R.B. (1997). *Brief Visuospatial Memory Test–Revised professional manual*. Odessa, Florida: Psychological Assessment Resources, Inc.

Benton, A.L., Sivan, A.B., Hamsher, K. deS., 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. *The Clinical Neuropsychologist, 3*, 129–136.

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

Christensen, H., Mackinnon, A.J., Korten, A.E., Jorm, A.F., Henderson, A.S., Jacomb, P., & Rodgers, B. (1999). An analysis of diversity in the cognitive performance of elderly community dwellers: Individual differences in change scores as a function of age. *Psychological Aging, 14*, 365–379.

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

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.

Goodglass, H. & Kaplan, E. (1983). *Boston Diagnostic Aphasia Examination (BDAE)*. Philadelphia, Pennsylvania: Lea and Febiger.

Hultsch, D.F., MacDonald, S.W., & Dixon, R.A. (2002). Variability in reaction time performance of younger and older adults. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 57*, P101–115.

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

Kløve, H. (1963). Clinical neuropsychology. In F.M. Forster (Vol. Ed.), *The medical clinics of North America, Vol. 47* (pp. 1647–1658). New York: Saunders.

Lezak, M.D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.

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* (pp. 899–932). New York: J. Wiley & Sons.

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

Matarazzo, J.D., Daniel, M.H., Prifitera, A., & Herman, D.O. (1988). Inter-subtest scatter in the WAIS–R standardization sample. *Journal of Clinical Psychology, 44*, 940–950.

Meyers, J.E. & Meyers, K.R. (1995). *Rey Complex Figure Test and Recognition Trial professional manual*. Odessa, Florida: Psychological Assessment Resources, Inc.

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.

Rey, A. (1993). Psychological examination of traumatic encephalopathy. J. Corwin & F.W. Bylsma (Trans.). *Clinical Neuropsychologist, 7*, 3–21. (Original work published 1941).

Schaie, K.W. (1994). The course of adult intellectual development. *American Psychologist, 49*, 304–313.

Schretlen, D. (1997). *Brief Test of Attention professional manual*. Odessa, Florida: Psychological Assessment Resources, Inc.

Silverstein, A.B. (1982). Pattern analysis as simultaneous statistical inference. *Journal of Consulting and Clinical Psychology, 50*, 234–249.

SPSS for Windows (Release 10.0.5). [Computer software]. (1999). Chicago, Illinois: SPSS, Inc.

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 manual*. San Antonio, Texas: The Psychological Corporation.

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

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale–Third Edition administration and scoring manual*. San Antonio, Texas: The Psychological Corporation.

Wing, J.K., Sartorius, N., & Üstün, T.B. (1996). *Schedules for Clinical Assessment in Neuropsychiatry (SCAN) version 2.1*. Geneva: World Health Organization.