OXFORD
UNIVERSITY PRESS

Archives of Clinical Neuropsychology 30 (2015) 217–227

Archives of CLINICAL NEUROPSYCHOLOGY

The Neuropsychological Assessment of Cognitive Deficits Considering Measures of Performance Variability

Céline Tanner-Eggen[1,2,*], Christian Balzer[2], Walter J. Perrig[1], Klemens Gutbrod[3]

[1]*Department of Psychology, University of Bern, Bern CH-3012, Switzerland*
[2]*Neurological Rehabilitation Centre, Reha Rheinfelden, Rheinfelden CH-4310, Switzerland*
[3]*Department of Neurology, Division of Cognitive and Restorative Neurology, Inselspital, Bern University Hospital, and University of Bern, Bern CH-3010, Switzerland*

*Corresponding author at: Reha Rheinfelden, Salinenstrasse 98, CH-4310 Rheinfelden, Switzerland. Tel.: +41 61 836 52 93.
E-mail address: tannerceline@icloud.com (C. Tanner-Eggen).

Accepted 17 February 2015

Abstract

Neuropsychologists often face interpretational difficulties when assessing cognitive deficits, particularly in cases of unclear cerebral etiology. How can we be sure whether a single test score below the population average is indicative of a pathological brain condition or normal? In the past few years, the topic of intra-individual performance variability has gained great interest. On the basis of a large normative sample, two measures of performance variability and their importance for neuropsychological interpretation will be presented in this paper: the number of low scores and the level of dispersion. We conclude that low scores are common in healthy individuals. On the other hand, the level of dispersion is relatively small. Here, base rate information about abnormally low scores and abnormally high dispersion across cognitive abilities are provided to improve the awareness of normal variability and to serve clinicians as additional interpretive measures in the diagnostic process.

Keywords: Neuropsychological assessment; Normal variability; Base rates; Low scores; Dispersion

In a neuropsychological examination, not only single test scores but also entire performance profiles should be considered to identify cognitive deficits. An unusual low performance on a single test might be interpreted as being reflective of acquired neurocognitive impairment if there is a correlation with a known brain lesion (e.g., isolated verbal memory deficit in a right-handed patient with a left hippocampal lesion). However, the psychometric principles associated with single-score distribution (i.e., Gaussian normal distribution; for an overview, see Slick, 2006) should not be applied to multiple-score distribution because, as more tests are administered, the chances of having abnormally low scores increase (Balzer, Moeller, Willmes, Gutbrod, & Eggen, 2011; Brooks & Iverson, 2010; Brooks, Strauss, Sherman, Iverson, & Slick, 2009; Brooks, Sherman, Iverson, Slick, & Strauss, 2011; Iverson & Brooks, 2011). As stated by Ingraham and Aiken (1996), when examining the results of multiple tests, the clinician is confronted with the problem of determining how many abnormal test scores are necessary to diagnose a profile as pathological. According to Binder, Iverson, and Brooks (2009), there is no agreement among neuropsychologists about the definition of abnormality.

For the interpretation of single test scores, Heaton, Grant, and Matthews (1991) and Heaton, Miller, Taylor, and Grant (2004) set the cutoff for low scores at >1 *SD* below the mean ($<$16th percentile). On the other hand, Wechsler tests traditionally classify test scores below the 10th percentile as "borderline" and scores below the 2nd percentile as "extremely low" (Wechsler, 1997a, 1997b). In this paper, we refer to the Heaton definition of abnormality. For the interpretation of entire performance profiles, no such recommendation concerning cutoff can be provided, because the number of low scores depends on the number of tests administered.

When assessing neurocognitive deficits, neuropsychologists usually focus on differences between individual test performances and the mean of the normative sample (i.e., inter-individual comparisons). However, an emphasis on these differences without considering intra-individual performance variability may lead to inaccurate inferences (e.g., Holtzer, Verghese, Wang, Hall, & Lipton, 2008; Nesselroade, 2002, in MacDonald, Li, & Bäckman, 2009). In the past few years, performance variability within

doi:10.1093/arclin/acv008 Advance Access publication on 15 March 2015

individuals (i.e., intra-individual comparisons) has been of great interest. It has been studied in different ways, using various definitions of variability (e.g., Binder et al., 2009; Crawford, Garthwaite, & Gault, 2007; Hilborn, Strauss, Hultsch, & Hunter, 2009; Schretlen, Munro, Anthony, & Pearlson, 2003). In the following, two types of performance variability and their importance for neuropsychological assessment are presented.

In recent years, a great deal of research has been done on the extent of low test performances of healthy individuals. Overall, these studies emphasize that clinicians need to be aware of the existence of low scores in the healthy population (Axelrod & Wall, 2007; Binder et al., 2009; Brooks & Iverson, 2010; Crawford et al., 2007; Iverson & Brooks, 2011; Palmer, Boone, Lesser, & Wohl, 1998; Schretlen, Testa, Winicki, Pearlson, & Gordon, 2008). But more than that neuropsychologists should also *use* information about the occurrence of low scores in the general population for a more accurate interpretation of a cognitive performance profile. In a recent review, Binder and colleagues (2009) summarized the work of several research groups examining the low scores of healthy adults across a battery of tests. These data showed that low scores were common in normative samples. The authors recommended that all test battery developers should provide information about the prevalence of variability in the general population. Using base rate tables of low scores supplements clinical interpretation and can help reduce the likelihood of misdiagnosis (also see Brooks, Iverson, Lanting, Horton, & Reynolds, 2012; Iverson, Holdnack, Brooks, & Lange, 2011). In this paper, information about low scores in a large normative sample of a comprehensive neuropsychological test battery will be provided. Since base rate data of low scores are not completely novel, this measure of performance variability will only be discussed briefly.

According to Brooks and colleagues (2009), the presence of low scores is due to the intra-individual variability (IIV) in the cognitive abilities of healthy people. There is growing evidence that considerable variation is prevalent across test performance within healthy individuals (Hilborn et al., 2009; Hultsch, Strauss, Hunter, & MacDonald, 2008; Schretlen et al., 2003). This seems to disagree with the general view in the variability literature that increased cognitive variability can be associated with the presence of central nervous system pathology (Hill, Rohling, Boettcher, & Meyers, 2013). The general term 'intra-individual variability (IIV)' has been defined in multiple ways (e.g., Hilborn et al., 2009; Hultsch, MacDonald, & Dixon, 2002; Hultsch & MacDonald, 2004; Tractenberg & Pietrzak, 2011), bringing forth two major aspects: inconsistency and dispersion. Inconsistency is the variability observed in a person's performance on a single task over a period of time (i.e., fluctuation). Dispersion refers to the variability of a person's performance across different tasks (i.e., profile scatter). In this study, the latter description of IIV was investigated. Most of the existing studies assessed IIV by measuring variability in reaction times (i.e., inconsistency), or examined variability in old or very old individuals (Bielak, Hultsch, Strauss, MacDonald, & Hunter, 2010; Bielak, Cherbuin, Bunce, & Anstey, 2014; Christensen et al., 1999; Hilborn et al., 2009; Hultsch et al., 2002; Rapp, Schnaider-Beeri, Sano, Silverman, & Haroutunian, 2005). So far, only few studies have examined dispersion of cognitive functioning across multiple tests (Hill et al., 2013; Holtzer et al., 2008; Kliegel & Sliwinski, 2004; Rabinowitz & Arnett, 2013). Hill and colleagues (2013) investigated IIV, using the concept of dispersion, in a large sample of individuals with traumatic brain injuries (TBIs). Holtzer and colleagues (2008) and Kliegel and Sliwinski (2004) examined dispersion as a predictor of cognitive decline in old age. A recent study of Rabinowitz and Arnett (2013) explored dispersion in college athletes before and after sports-related concussion. In their exploratory study, they found significant intra-individual variation across tests in the normative cognitive performance. The finding that cognitively healthy individuals display cognitive dispersion has important clinical implications that make it necessary to identify the characteristics of normal dispersion (Hilborn et al., 2009). Therefore, a second goal of the present paper is to provide base rate data about the level of abnormally high dispersion across cognitive abilities for a normative sample of healthy adults. Knowing this base rate in a healthy sample helps to decide, if further analysis is indicated to detect possible pathological performance profiles.

The number of low scores and the level of dispersion are two different measures of variability neuropsychologists should be aware of when interpreting neuropsychological performance profiles. Using the example of a comprehensive neuropsychological test battery, the main purpose of this paper is not only to improve the awareness of normal variability but also to provide information about these two measures of variability in healthy adults as additional interpretive methods in neuropsychological assessment.

Methods

Normative Sample

The normative sample consisted of 569 healthy adults aged 16–65 years (mean 38.6 years). The mean age of the 292 women was 38.8 years ($SD = 13.6$); the mean age of the 277 men was 38.4 years ($SD = 13.2$). The level of education was assessed according to the UNESCO's International Standard Classification of Education (ISCED; http://www.uis.unesco.org/Education/Pages/international-standard-classification-of-education.aspx). There are six different levels of ISCED: ISCED 0 = preschool (1–3 years), ISCED 1 = primary education (4, 5, or 6 years), ISCED 2 = lower secondary education (compulsory education; 9 years), ISCED 3 = upper secondary education (European Baccalaureate, vocational education; 12–13 years), ISCED 4 = post-

secondary non-tertiary education (adult education, programmes giving access to higher education), ISCED 5 = first stage of tertiary education (university education, higher vocational qualification), and ISCED 6 = second stage of tertiary education (post-graduate studies, doctorate). Instead of this more qualitative classification, the sum of years in school and occupational training was taken as a quantitative measure of education. The mean education was 13.8 years ($SD = 3.2$), with men having a slightly higher mean of education (14.6 years, $SD = 3.1$) than women (13.0 years, $SD = 3.1$). To ensure that all participants were healthy, exclusion criteria were formulated. These criteria were prepared as a list of questions that were asked as a standardized interview to every participant prior to examination. Excluded were persons experiencing any kind of accidents or illnesses with involvement of the central nervous system (e.g., TBI, cerebrovascular diseases, encephalitis, meningitis, dementia, Parkinson's disease, multiple sclerosis, epilepsy, brain tumours, and suffered hypoxia); serious physical or psychiatric illnesses (e.g., HIV-infection/ AIDS, whiplash-associated disorders, ADHD, sleep apnea syndrome, chronic lung diseases, diabetes mellitus, hypo- or hyperthyroidism, cancer, exposition to solvents, pesticides, or metals, illnesses affecting heart, lung, liver, kidney, pancreas, or pituitary gland, major depression, schizophrenia); former or actual alcohol abuse/drug consumption; current consumption of any kind of medication having an effect on cognitive performance; chronic or acute pain (e.g., migraine during examination); and limited vision or hearing. On the examination day, it was assured that a participant did not suffer from any physical or mental condition like the flu, indisposition, dizziness, acute mental imbalance, or other illnesses affecting test performance. Furthermore, individuals with a lack of fluency in German were excluded. The quota sample is representative for the population of Switzerland according to the statistical yearbook in the given age range.

Hence, in adherence to the strict exclusion criteria, all efforts were undertaken to prevent effects on findings by a pathological subgroup.

Materials

In this paper, the normative sample of a test battery called "Materials and Norms for the Neuropsychological Diagnostics" (MNND; Balzer, Berger, et al., 2011) was examined. These authors adapted and partly modified frequently used neuropsychological tests to subsume them in a test battery, which was standardized on one and the same large normative sample. This allows neuropsychologists to use procedures of the psychometric single case analysis to statistically analyze neuropsychological test profiles.

Since there is a large amount of test parameters in MNND, and not every test parameter is of similar importance, only the 20 most relevant (i.e., most often used in our clinical practice) test scores of 13 neuropsychological tests were chosen for analysis. Only test parameters with good psychometric properties were selected. The tests can be summarized in four cognitive domains: memory, attention, executive functions, and visuospatial functions. The tests and the 20 test parameters are listed in Table 1.

Analysis of low Scores

The number of low scores was calculated by considering performance on the 20 test parameters of Table 1 simultaneously. In a first step, the percentile rank of each raw score was determined for each test parameter separately. Mathematically, the percentile rank is defined as

$$PR = \left[\frac{cfi + .5(fi)}{N}\right] \times 100\%,$$

where *cfi* is the cumulative frequency for all scores lower than the score of interest, *fi* is the frequency of scores in the interval of interest, and *N* is the sample size.

According to Brooks and colleagues (2009), it is important to interpret test performance and the number of low scores within the context of a person's demographic characteristics. That is why regression-based normative data were used, adjusting the test scores of MNND by the impact of age, gender, education, and test version (forms A and B). Due to the education correction, no differences between various educational groups were expected in the present examination of low scores. However, individuals with less education (i.e., <12 years) differed significantly from individuals with more education (i.e., ≥ 12 years) in regard to the number of low scores (e.g., $p < .009$ at cutoff 16th percentile). Therefore, instead of regression-based adjustment for education, base rates of low scores were stratified by years of education, as has been done by Brooks and colleagues (2012).

Because the majority of the participants younger than 20 years old had not finished their education at the time of investigation, it could not be decided whether to put them in the low education group (i.e., <12 years) or in the high education group (i.e., ≥ 12 years). Therefore, all participants aged <20 years ($n = 39$) had to be excluded from the analysis of low scores. For the base rate analysis of low scores, the same procedure was used as in computing percentile ranks for a raw score distribution, with correction of

Table 1. Description of the neuropsychological assessment

Subtest	Adapted from	Description	Test parameter (*z*)
Verbal learning and memory test[a]	RAVLT (Rey, 1958, 1964)	Word list learning	**1.** Sum of learning (1–5) **2.** Delayed recall **3.** recognition
Figural learning and memory test[a]	RVDLT (Rey, 1964; Spreen & Strauss, 1991)	Figure list learning	**4.** Sum of learning (1–5) **5.** Delayed recall **6.** Recognition
Non-verbal learning and memory test[a]	RULIT (Ruff & Allen, 1999)	Route learning	**7.** Sum of learning (1–5) **8.** Delayed recall
Text memory[a]	WMS-R/WMS-III (Wechsler, 1987, 1997c)	Text recall	**9.** Immediate recall **10.** Delayed recall
Rey complex figure test[a]	RCFT (Rey, 1941; Taylor, 1969)	Figural memory	**11.** Delayed recall
Verbal memory span[a]	Wechsler adult intelligence test (German; von Aster, Neubauer, & Horn, 2006)	Digit span	**12.** Correct digits
Visual memory span[a]	Block tapping test (Milner, 1971; Schellig, 1997)	Block span	**13.** Correct blocks
Test des Deux-Barrages[b]	T2B (Zazzo & Stambak, 1964)	Selective attention	**14.** Correct items/min **15.** Errors in %
Stroop test[c]	Victoria stroop test (Regard, 1981)	Interference control	**16.** Stroop time
Word fluency test[c]	Word fluency/COWA (Benton, Hamsher, & Sivan, 1994; Thurstone, 1938)	Letter fluency (S)	**17.** Correct words
Design fluency test[c]	Five-point test (Regard, Strauss, & Knapp, 1982)	Figural fluency	**18.** Correct digits
Kramer categorization test[c]	Kramer intelligence test (German; Kramer, 1972)	Categorization task	**19.** Correct categories
Spatial test[d]	Test of primary mental abilities (Thurstone & Thurstone, 1941)	Mental rotation	**20.** Correct items

Notes: Numbers in bold show the 20 test parameters chosen for analysis in this study.
[a]Memory, [b] = attention, [c] = executive functions, and [d] = visuospatial functions.

cumulative frequencies to the center of the interval, to ensure the exact threshold of a cutoff. In this way, low scores were then examined for two different levels of education in the present paper: <12 ($n = 95$) and ≥ 12 years ($n = 435$).

The base rates of low scores were analyzed using three cutoff scores: first, >1 *SD* below the mean (<16th percentile); second, below the 7th percentile (1.5 *SD*); finally, >2 *SD* below the mean (<2nd percentile). In the theoretical frame of single test score interpretation, these classifications correspond to "mild impairment," "mild to moderate impairment," and "moderate impairment," respectively (Heaton et al., 1991, 2004; Schellig, Drechsler, Heinemann, & Sturm, 2009). In clinical practice, the 16th percentile is most commonly used as cutoff score for abnormality. Hence, to keep the data presentation manageable, only the results for the <16th percentile will be presented graphically. The results for all other cutoffs are listed in Table 2 below.

Analysis of Dispersion

Dispersion was examined using a similar procedure to Morgan, Woods, Delano-Wood, Bondi, and Grant (2011). Since no differences between various educational groups were found, the base rates of abnormally high dispersion levels were not stratified separately by education. Therefore, the total sample was included in the analysis. A regression-based adjustment for age, gender, education, and test version was conducted. To be able to perform statistical analyses, percentiles were converted into *z*-scores as standard equivalents and an intra-individual standard deviation (ISD) was computed across these selected *z*-scores for each participant. Dispersion, in this case, is characterized as a normal *SD*. A dispersion level (ISD-score) of 0 means that all *z*-scores are equal, whereas an ISD-score of 1 or more implies that the *z*-scores differ considerably. The procedure for the base rate analysis of dispersion was the same as described above.

The ISD-scores were not adjusted for level of performance in this paper, because such correction might complicate interpretation of results (Schmiedek, Lovden, & Lindenberger, 2009, in Morgan et al., 2011; Morgan, Woods, Grant, & HNRP, 2012). To analyze a possible relation between the level of dispersion and the level of performance, correlational analyses were calculated.

Results

Base Rates of Low Scores

The prevalence of low scores on the neuropsychological test battery MNND are presented in Table 2. Base rates are listed separately for the total sample and for two different educational levels: <12 years and ≥ 12 years. The cutoff for low scores had a

Table 2. Base rates of low scores stratified by years of education

Number of low scores	Total sample ($n = 530$)		Educational level			
			<12 years ($n = 95$)		≥12 years ($n = 435$)	
	Cum%	%	Cum%	%	Cum%	%
<16th percentile						
Zero	100	27.0	100	11.6	100	23.9
1	73.0	19.8	88.4	8.4	76.1	23.9
2	53.2	12.3	80.0	7.4	52.2	14.5
3	40.9	9.6	72.6	15.8	37.7	9.7
4	31.3	8.3	56.8	8.4	28.0	8.7
5	23.0	6.4	48.4	11.6	19.3	5.5
6	16.6	4.9	36.8	10.5	13.8	4.8
7	11.7	2.1	26.3	4.2	9.0	1.6
8	9.6	2.8	22.1	6.3	7.4	2.3
9	6.8	1.3	15.8	2.1	5.1	1.1
10	5.5	2.1	13.7	6.3	3.9	1.4
11	3.4	1.1	7.4	1.1	2.5	1.1
12	2.3	—	6.3	—	1.4	—
13	2.3	0.9	6.3	1.1	1.4	0.9
14	1.3	0.4	5.3	2.1	0.5	—
15	0.9	0.4	3.2	1.1	0.5	0.2
16	≤0.6	0.4	≤2.1	2.1	≤0.2	—
<7th percentile						
Zero	100	53.8	100	33.7	100	54.0
1	46.2	20.0	66.3	17.9	46.0	22.3
2	26.2	11.3	48.4	14.7	23.7	12.0
3	14.9	3.6	33.7	5.3	11.7	3.7
4	11.3	4.2	28.4	7.4	8.0	3.9
5	7.2	2.1	21.1	5.3	4.1	1.4
6	5.1	2.1	15.8	7.4	2.8	0.9
7	3.0	1.3	8.4	4.2	1.8	0.7
8	1.7	0.6	4.2	—	1.1	0.7
9	1.1	0.2	4.2	1.1	0.5	—
10	0.9	0.6	3.2	2.1	0.5	0.2
11	<0.5	0.2	≤1.1	—	≤0.2	0.2
<2nd percentile						
Zero	100	82.8	100	67.4	100	85.1
1	17.2	11.9	32.6	18.9	14.9	11.3
2	5.3	3.0	13.7	6.3	3.7	2.3
3	2.3	1.3	7.4	5.3	1.4	0.7
4	0.9	0.6	2.1	1.1	0.7	0.5
5	<0.5	—	≤1.1	—	≤0.2	—

Notes: There are slight variations due to rounding. Analyses are based on 20 age and gender adjusted z-scores derived from the test battery Materials and Norms for the Neuropsychological Diagnostics (Balzer et al., 2011). Cum% = cumulative percentage. For example, for a cutoff below the 16th percentile, 12.3% of the total sample had exactly two low scores, 53.2% had two or more low scores.

considerable impact on the frequency in the neurologically healthy population. Obtaining one or more low scores below the 16th percentile occurred in 73.0% of the total sample (cumulative percentage). At more conservative cutoffs, the number of low scores decreased. For example, in 46.2% one or more low scores below the 7th percentile were observed and 17.2% had one or more low scores below the 2nd percentile.

When considering the prevalence of low scores across educational level, there were significant differences for the three cutoffs. For example, having three or more low scores below the 16th percentile was found in 72.6% of the healthy adults with less education compared with 37.7% of those with more years of education ($\chi^2(11) = 49.723, p < .001$). Having three or more low scores below the 7th percentile was found in 33.7% of the less educated people compared with 11.8% of those with higher education ($\chi^2(7) = 42.627, p < .001$). And having three or more low scores below the 2nd percentile was found in 7.4% of the lower educated group compared with 1.4% of the higher educated group ($\chi^2(3) = 22.190, p < .001$). Statistically meaningful age or gender differences regarding the number of low scores were not present.

Fig. 1 provides a visual representation of the base rate data for low scores below the 16th percentile. Three different curves represent the likelihood that a particular individual would show a number of low scores. One curve shows the base rate data of the total sample and the other two curves show the data for the two different educational levels. As shown, healthy individuals with <12 years of education had significantly more low scores than more educated individuals.

One could argue that the more extreme ages have skewed the data because of the wide age range. To identify possible outliers in our data, the boxplot criterion was administered. No abnormalities were found, which is why it was assumed that basic statistical assumptions for analyses were met.

Base Rates of Dispersion

Table 3 shows the frequencies of different dispersion levels in the healthy population, measured by the intra-ISD. ISD-scores indicate the degree of performance variability, with higher scores representing scattered profiles with greater variability across measures, whereas lower values reflect flatter, more consistent profiles with little variability. The ISD-scores in this paper ranged from 0.400 to 1.275. The first line in the table marks the cutoff for abnormal dispersion at PR < 16. The exact ISD-score falling below the 16th percentile cutoff is >0.91. The second line in the table indicates the percentage of individuals with dispersion >1 *SD* of their own mean. Only 6.2% of the healthy sample showed a dispersion >1 *SD*.

In terms of the univariate analyses, there was a highly significant age difference for the level of dispersion, $F(2, 557) = 16.56$, $p < .001$, and a significant gender effect, $F(1, 557) = 4.34$, $p = .038$. Dispersion enhanced with increasing age. In *post hoc* tests, older adults showed a significantly higher amount of dispersion than younger ones (Tukey-HSD; $p < .001$). Men showed only a minimally higher amount of dispersion than women. The mean of the two groups was almost identical and it can be assumed that the significant gender effect was due to the large sample size. Neither an education effect nor any interactions were found.

Correlational analyses between dispersion and the level of performance did show a significant ($r = -.13$; $p = .002$), but very small correlation. Based on other studies (Hill et al., 2013; Rabinowitz & Arnett, 2013; Schinka, Vanderploeg, & Curtiss, 1994; Schretlen et al., 2003), the missing relationship between the overall test performance and the dispersion was perplexing for us and the reviewers. We thought that the regression-based adjustment for age, gender, education, and test version might be the reason for this result. Therefore, we calculated a control analysis with scores that were not adjusted for age, gender, education, and test version. This analysis did not reveal another result. The correlation between the overall test performance and the dispersion was still negligible ($r = -.16$; $p < .001$).

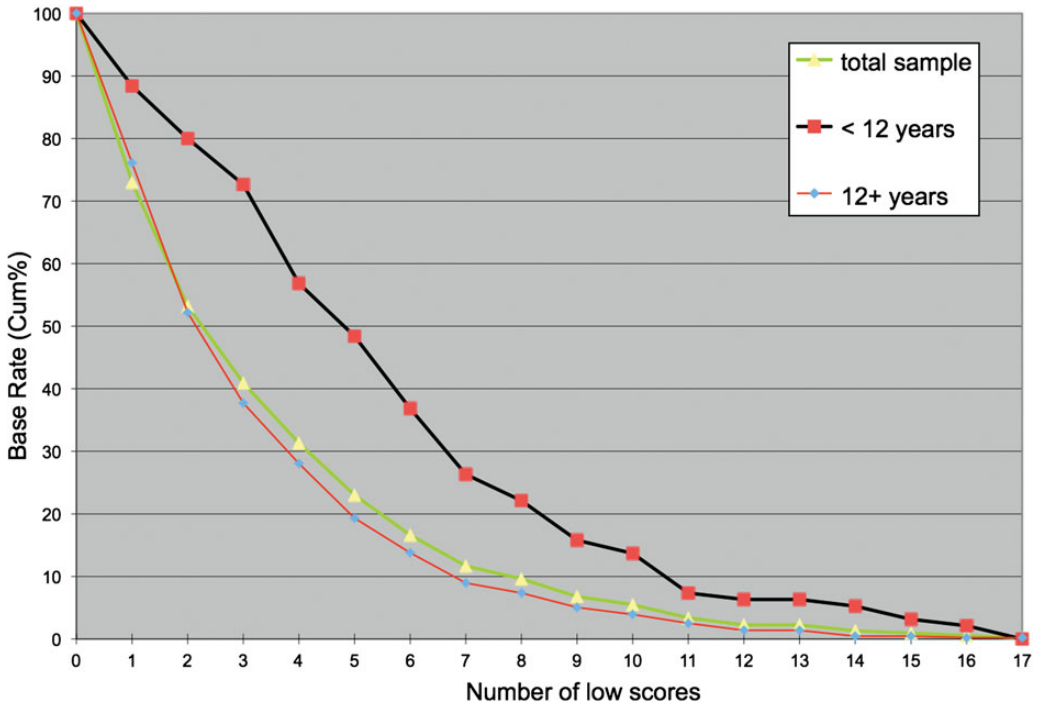

Fig. 1. Base rates of low scores at <16th percentile.

Table 3. Base rates of dispersion (ISD-scores)

ISD-score	Cum%	%
0.400	100.0	0.3
0.425	99.7	0.4
0.450	99.3	1.1
0.475	98.2	1.6
0.500	96.6	1.6
0.525	95.0	2.0
0.550	93.0	2.6
0.575	90.3	3.4
0.600	86.9	4.5
0.625	82.4	5.0
0.650	77.4	6.8
0.675	70.7	7.0
0.700	63.6	5.0
0.725	58.6	6.2
0.750	52.5	7.0
0.775	45.4	6.1
0.800	39.4	6.5
0.825	32.9	6.2
0.850	26.7	4.7
0.875	22.0	4.1
0.900	17.8	3.2
0.925	14.7	2.8
0.950	11.9	2.2
0.975	9.7	1.6
1.000	8.1	1.8
1.025	6.2	1.6
1.050	4.7	0.9
1.075	3.8	0.9
1.100	2.9	0.9
1.125	2.0	0.5
1.150	1.5	0.4
1.175	1.1	0.1
1.200	1.1	0.3
1.225	0.8	0.4
1.250	0.4	0.3
1.275	0.1	0.1
1.300	0.0	0.0

Notes: There are slight variations due to rounding. Cum% = cumulative percentage (percentile rank).
The first line marks the cutoff for abnormal dispersion. 17.8% of the total sample exhibit a dispersion of 0.9 or more, 3.2% showed a score of exactly 0.9. The second line indicates the percentage of individuals with dispersion >1 *SD* of their own mean.

Discussion

The purpose of this paper was to examine the prevalence of low scores and dispersion in the cognitive performance profiles of a large normative sample for a comprehensive neuropsychological test battery. Being aware of this information and to use it as additional interpretive measure can help to reduce the likelihood of misdiagnosing cognitive deficits. The base rate analyses give the clinician more confidence in the interpretation that a patient has acquired neuropsychological deficits, even though he might have a low premorbid intellectual level or an unclear cerebral etiology.

Our analyses showed that low scores are common in the healthy population. This finding is consistent with existing studies investigating this aspect of performance variability (Binder et al., 2009; Brooks et al., 2012; Iverson & Brooks, 2011; Palmer et al., 1998). According to Binder and colleagues (2009), a prevalence of low scores that falls <20% is deemed uncommon, whereas a prevalence falling <10% is unusual. Almost 75% of the normative sample obtained at least one low score when using 1 *SD* below the mean as a cutoff. At more conservative cutoffs, low scores were still relatively common.

The result that the prevalence of low scores varied by level of education reflects the emphasis of Brooks and colleagues (2012) on "the importance of considering the psychometric principle that the number of low scores varies by the demographic characteristics of the examinee and that low scores will increase in those with fewer years of education" (p. 68). Even though

regression-based normative data were used in MNND, adjusting for age, gender, education, and test version, the education correction proved to be insufficient for the analysis of low scores in persons with <12 years of education in the present analysis. Therefore, clinicians are advised to use education-stratified base rate tables to avoid overestimation of cognitive deficits in less educated individuals.

Marked IIV is often associated with acquired neurocognitive deficits. In this case, it is assumed that an abnormal brain condition interferes with a person's ability to perform at a characteristic level of neuropsychological functioning (Hill et al., 2013; Lezak, Howieson, & Loring, 2004). However, previous research showed that IIV is not necessarily a marker of neurocognitive disorder (Binder et al., 2009; Rabinowitz & Arnett, 2013). With our data, we provide a cutoff to identify abnormally high (i.e., possibly pathological) dispersion. An unusually high level of dispersion gives us reason to perform a profile analysis to investigate relative weaknesses (i.e., deficits) and possible relative strengths (i.e., preserved functions). Such information enables us to be more confident about our interpretation that some pathological brain condition must be present. According to our analyses, a dispersion score >0.91 might reflect some abnormal condition. Our analyses revealed that only a small amount of the healthy population shows a clinically relevant dispersion. This result is contradictory to the above-mentioned notion that IIV is common in healthy individuals. A higher amount of dispersion might be interpreted as an indicator for pathology. Dispersion can be seen as an early marker of loss of neural and cognitive integrity that is a harbinger of future decline. It is a clinically meaningful measure of performance variability. Therefore, we recommend that test developers should not only provide information about the prevalence of low scores but also information about the amount of dispersion in their normative data.

In our data, no linear relationship between dispersion and the level of performance was found. This result seems contradictory to the finding of Rabinowitz and Arnett (2013) or Hill and colleagues (2013) that overall performance is negatively correlated with performance variability. The reason why better performance was not associated with less variability or vice versa can only be assumed in the fact that our population is a healthy sample. Nonetheless, dispersion should always be interpreted in consideration of the general level of performance in a clinical sample, because a high dispersion level alone does not detect a pathological profile. A high dispersion level in a high-performance level profile is interpreted differently than a high dispersion level in a low-performance level profile (Hilborn et al., 2009). The presence of an overall low test performance related to a high level of dispersion needs further analysis of the cognitive profile.

Furthermore, a highly significant age difference for the level of dispersion was found in the current investigation. This result is consistent with previous research relating increased dispersion to advancing age (i.e., typically 65 years of age and more; Christensen et al., 1999; Hilborn et al., 2009; Hultsch et al., 2002). However, our population was considerably younger than in past studies. The increase of dispersion seems to already be present in younger individuals. This result corresponds with the statement of Bielak et al. (2014) that increases in IIV is a fundamental behavioral characteristic associated with growing older, even among healthy adults. However, their study was done using a reaction time task as a measure for IIV (i.e., inconsistency). For a differentiated analysis of dispersion across the lifespan, further studies will be needed.

Men showed a minimally higher amount of dispersion than women. This gender effect was almost certainly due to the large sample size and is therefore probably negligible. Bielak et al. (2014) and Dykiert, Der, Starr, and Deary (2012) both found evidence of sex differences in IIV. Again, these studies used 'inconsistency' as a measure of IIV instead of 'dispersion'. Therefore, the results are not completely comparable. So far, alternate explanations for the found gender effect in our study remain unclear.

There are a few methodological issues and limitations to consider. First, in this paper the actual base rates of different profile characteristics are presented. The base rates were analyzed empirically using frequency tables. The disadvantage of such tables is that they are fixed and cannot be adapted. The tables can only be used when a clinician administers all tests that were included in the analyses for the respective base rate table. Another method is the mathematical approach to estimate base rates. For this purpose, Crawford and colleagues (2007) developed a simple and free available computer program involving a Monte Carlo simulation based on the test intercorrelations. The advantage of such a computer program is that the base rates for any combination of co-normed tests can be calculated rapidly. The disadvantage is that they only *estimate* the base rates statistically. According to Brooks and Iverson (2010), these calculations lose accuracy when estimating base rates for individuals with very low or very high education. Nevertheless, the authors suggested that the Monte Carlo simulation program is a good option for determining the prevalence of low scores. However, base rates of low scores should only be estimated mathematically for people who are demographically closer to the mean of the normative sample. In other words, for individuals with rather low or on the other hand very high education, the base rates should be analyzed empirically, with education taken into account (Brooks & Iverson, 2010). Hence, we adopted the latter method for this paper.

Comparing a patient's number of low scores with the base rate of the normative sample makes it impossible to differentiate whether the low scores are distributed randomly or if they accumulate among a specific cognitive domain. This disadvantage could have been resolved by providing separate base rate tables for each cognitive domain. Due to space issues, we refrained from doing so. In the future, the base rate information of different measures of variability shall be integrated in a computer program for MNND (similar to Crawford, Garthwaite, Longman, & Batty, 2011).

Another limitation is that effort was not controlled at the time of data collection in the present study. Based on a comment of a reviewer, we searched for a measure to minimally screen for effort *post hoc*. 'Recognition' and 'true recognition' (recognition minus false positives) of our word list was examined as a validity indicator. According to Boone, Lu, and Wen (2005), a cutoff of ≤ 9 on the recognition score and a cutoff of ≤ 7 for true recognition serve as embedded effort indices. In our study, only one person of the total sample ($n = 569$) fell under the cutoff for recognition ($<0.2\%$) and $<2\%$ of the total sample showed a cutoff of ≤ 7 for true recognition. In our opinion, this is a very low failure rate.

Despite the above limitations, the present paper has some clear clinical implications by providing important additional interpretive measures. Having access to the prevalence of different measures of variability is important to improve accuracy when interpreting a neuropsychological profile (Brooks, 2010). Also, Brooks and colleagues (2009) point out that clinicians, who do not acknowledge the specific characteristics of a normative dataset, are at risk of over- or underestimating cognitive deficits. However, like Iverson and colleagues (2011) emphasized, the base rates of different measures of variability serve as additional information and are not meant to replace clinical judgment.

Funding

This work was financially unfunded but supported by the Department of Science of the Rehabilitation Centre Reha Rheinfelden in Switzerland by providing valuable time resources.

Conflict of Interest

None declared.

Acknowledgements

We are grateful to Markus Stöcklin, PhD, for his assistance with statistical questions. We thank Brigitte Weiermann, PhD, for her helpful commentaries on earlier drafts of the manuscript and Michael McCaskey for his helpful stylistic and grammatical suggestions. Also we thank the anonymous reviewers for their contributions. Last but not least, we thank the Reha Rheinfelden for their provision of valuable resources.

References

Axelrod, B. N., & Wall, J. R. (2007). Expectancy of impaired neuropsychological test scores in a non-clinical sample. *International Journal of Neuroscience*, *117 (11)*, 1591–1602.

Balzer, C., Berger, J.-M., Caprez, G., Gonser, A., Gutbrod, K., & Keller, M. (2011). *Materialien und normwerte fuer die neuropsychologische diagnostik (MNND)*. Rheinfelden: Verlag Normdaten.

Balzer, C., Moeller, K., Willmes, K., Gutbrod, K., & Eggen, C. (2011). Interpretation. In C. Balzer, J.-M. Berger, G. Caprez, A. Gonser, K. Gutbrod, & M. Keller (Eds.), *Materialien und normwerte fuer die neuropsychologische diagnostik* (pp. 58–76). Rheinfelden: Verlag Normdaten.

Benton, A. L., Hamsher, K. d. S., & Sivan, A. B. (1994). *Multilingual aphasia examination* (3rd ed.). San Antonio, TX: Psychological Corporation.

Bielak, A. A., Hultsch, D. F., Strauss, E., MacDonald, S. W., & Hunter, M. A. (2010). Intraindividual variability is related to cognitive change in older adults: Evidence for within-person coupling. *Psychology and Aging*, *25 (3)*, 575–586.

Bielak, A. A. M., Cherbuin, N., Bunce, D., & Anstey, K. J. (2014). Intraindividual variability is a fundamental phenomenon of aging: Evidence from an 8-year longitudinal study across young, middle, and older adulthood. *Developmental Psychology*, *50 (1)*, 143–151.

Binder, L. M., Iverson, G. L., & Brooks, B. L. (2009). To err is human: "abnormal" neuropsychological scores and variability are common in healthy adults. *Archives of Clinical Neuropsychology*, *24 (1)*, 31–46.

Boone, K. B., Lu, P., & Wen, J. (2005). Comparison of various RAVLT scores in the detection of noncredible memory performance. *Archives of Clinical Neuropsychology*, *20 (3)*, 301–319.

Brooks, B. L. (2010). Seeing the forest for the trees: Prevalence of low scores on the Wechsler intelligence scale for children, fourth edition (WISC-IV). *Psychological Assessment*, *22 (3)*, 650–656.

Brooks, B. L., & Iverson, G. L. (2010). Comparing actual to estimated base rates of "abnormal" scores on neuropsychological test batteries: Implications for interpretation. *Archives of Clinical Neuropsychology*, *25 (1)*, 14–21.

Brooks, B. L., Iverson, G. L., Lanting, S. C., Horton, A. M., & Reynolds, C. R. (2012). Improving test interpretation for detecting executive dysfunction in adults and older adults: Prevalence of low scores on the test of verbal conceptualization and fluency. *Applied Neuropsychology*, *19 (1)*, 61–70.

Brooks, B. L., Sherman, E. M. S., Iverson, G. L., Slick, D. J., & Strauss, E. (2011). Psychometric foundations for the interpretation of neuropsychological test results. In M. R. Schoenberg, & J. G. Scott (Eds.), *The little black book of neuropsychology: A syndrome-Based approach* (pp. 893–922). New York: Springer Science+Business Media.

Brooks, B. L., Strauss, E., Sherman, E. M. S., Iverson, G. L., & Slick, D. J. (2009). Developments in neuropsychological assessment: Refining psychometric and clinical interpretive methods. *Canadian Psychology*, *50 (3)*, 196–209.

Christensen, H., Mackinnon, A. J., Korten, A. E., Jorm, A. F., Henderson, A. S., & Jacomb, P. (1999). Dispersion in cognitive ability as a function of age: A longitudinal study of an elderly community sample. *Aging, Neuropsychology, and Cognition*, *6 (3)*, 214–228.

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 (4)*, 419–430.

Crawford, J. R., Garthwaite, P. H., Longman, R. S., & Batty, A. M. (2011). Some supplementary methods for the analysis of WAIS-IV index scores in neuropsychological assessment. *Journal of Neuropsychology*, *6 (2)*, 192–211.

Dykiert, D., Der, G., Starr, J. M., & Deary, I. J. (2012). Sex differences in reaction time mean and intraindividual variability across the life span. *Developmental Psychology*, *48 (5)*, 1262–1276.

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 extended Halstead-Reitan battery: Demographically adjusted neuropsychological norms for African American and Caucasian adults professional manual*. Lutz, FL: Psychological Assessment Resources.

Hilborn, J. V., Strauss, E., Hultsch, D. F., & Hunter, M. A. (2009). Intraindividual variability across cognitive domains: Investigation of dispersion levels and performance profiles in older adults. *Journal of Clinical and Experimental Neuropsychology*, *31 (4)*, 412–424.

Hill, B. D., Rohling, M. L., Boettcher, A. C., & Meyers, J. E. (2013). Cognitive intra-individual variability has a positive association with traumatic brain injury severity and suboptimal effort. *Archives of Clinical Neuropsychology*, *28 (7)*, 640–648.

Holtzer, R., Verghese, J., Wang, C., Hall, C. B., & Lipton, R. B. (2008). Within-person across-neuropsychological test variability and incident dementia. *JAMA*, *300 (7)*, 823–830.

Hultsch, D. F., MacDonald, S. W., & Dixon, R. A. (2002). Variability in reaction time performance of younger and older adults. *Journal of Gerontology*, *57B (2)*, 101–115.

Hultsch, D. F., & MacDonald, S. W. S. (2004). Intraindividual variability in performance as a theoretical window onto cognitive aging. In R. A. Dixon, L. Bckman, & L.-G. Nilsson (Eds.), *New frontiers in cognitive aging* (pp. 65–88). Oxford: Oxford University Press.

Hultsch, D. F., Strauss, E., Hunter, M. A., & MacDonald, S. W. S. (2008). Intraindividual variability, cognition, and aging. In F. I. M. Craik, & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (3rd ed., pp. 491–556). New York: Psychology Press.

Ingraham, L. J., & Aiken, C. B. (1996). An empirical approach to determining criteria for abnormality in test batteries with multiple measures. *Neuropsychology*, *10 (1)*, 120–124.

Iverson, G. L., & Brooks, B. L. (2011). Improving accuracy for identifying cognitive impairment. In M. R. Schoenberg, & J. G. Scott (Eds.), *The little black book of neuropsychology: A syndrome-bases approach* (pp. 923–950). New York: Springer Science+Business Media.

Iverson, G. L., Holdnack, J. A., Brooks, B. L., & Lange, R. T. (2011). *Evidence-based psychometric criteria for memory impairment*. *Paper presented at the 39th annual meeting of the International Neuropsychological Society*, Boston.

Kliegel, M., & Sliwinski, M. (2004). MMSE cross-domain variability predicts cognitive decline in centenarians. *Gerontology*, *50 (1)*, 39–43.

Kramer, J. (1972). *Kramer intelligenztest*. Solothurn: St. Antonius-Verlag.

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

MacDonald, S. W. S., Li, S. C., & Bäckman, L. (2009). Neural underpinnings of within-person variability in cognitive functioning. *Psychology and Aging*, *24 (4)*, 792–808.

Milner, B. (1971). Interhemispheric differences in the localization of psychological processes in man. *British Medical Bulletin*, *3*, 272–277.

Morgan, E. E., Woods, S. P., Delano-Wood, L., Bondi, M. W., & Grant, I. (2011). Intraindividual variability in HIV infection: Evidence for greater neurocognitive dispersion in older HIV seropositive adults. *Neuropsychology*, *25 (5)*, 645–654.

Morgan, E. E., Woods, S. P., & Grant, I., & HNRP (2012). Intra-individual neurocognitive variability confers risk of dependence in activities of daily living among HIV-seropositive individuals without HIV-associated neurocognitive disorders. *Archives of Clinical Neuropsychology*, *27 (3)*, 293–303.

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 (6)*, 503–511.

Rabinowitz, A. R., & Arnett, P. A. (2013). Intraindividual cognitive variability before and after sports-related concussion. *Neuropsychology*, *27 (4)*, 481–490.

Rapp, M. A., Schnaider-Beeri, M., Sano, M., Silverman, J. M., & Haroutunian, V. (2005). Cross-domain variability of cognitive performance in very old nursing home residents and community dwellers: Relationship to functional status. *Gerontology*, *51 (3)*, 206–212.

Regard, M. (1981). Cognitive rigidity and flexibility: A neuropsychological study. Unpublished PhD dissertation. University of Victoria.

Regard, M., Strauss, E., & Knapp, P. (1982). Children's production on verbal and nonverbal fluency tasks. *Perceptual and Motor Skills*, *55*, 839–844.

Rey, A. (1941). L'examen psychologique dans les cas d'encéphalopathie traumatique. *Archives de Psychologie*, *28*, 286–340.

Rey, A. (1958). Mémorisation d'une série de 15 mots en 5 répétitions. In A. Rey (Eds.), *L'examen clinique en psychologie* (pp. 139–193). Paris: Presses Universitaires de France.

Rey, A. (1964). *L'examen clinique en psychologie*. Paris: Presses Universitaire de France.

Ruff, R. M., & Allen, C. C. (1999). *Ruff-light trail learning test*. Odessa, FL: Psychological Assessment Resources.

Schellig, D. (1997). *Block tapping test*. Frankfurt: Harcourt Test Services.

Schellig, D., Drechsler, R., Heinemann, D., & Sturm, W. (2009). *Handbuch neuropsychologischer testverfahren: Aufmerksamkeit, gedächtnis und exekutive funktionen*. Göttingen: Hogrefe.

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 (4)*, 364–367.

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 (6)*, 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 (3)*, 436–445.

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

Spreen, O., & Strauss, E. (1991). A compendium of neuropsychological tests. Administration, norms, and commentary. New York: Oxford University Press.

Taylor, L. B. (1969). Localization of cerebral lesions by psychological testing. *Clinical Neurosurgery*, *16*, 269–287.
Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.
Thurstone, L. L., & Thurstone, T. G. (1941). *Chicago tests of primary mental abilities*. Chicago: Science Research Association.
Tractenberg, R. E., & Pietrzak, R. H. (2011). Intra-individual variability in Alzheimer's disease and cognitive aging: Definitions, context, and effect sizes. *PLoS ONE*, *6 (4)*, e16973.
von Aster, M., Neubauer, A., & Horn, R. (2006). *Wechsler Intelligenztest für Erwachsene (WIE): Deutschsprachige Bearbeitung und adaptation des WAIS-III von David Wechsler*. Frankfurt: Harcourt Test Services.
Wechsler, D. (1987). *Wechsler memory scale-revised*. San Antonio, TX: The Psychological Corporation.
Wechsler, D. (1997a). *Wechsler adult intelligence scale-revised: Adminstration and scoring manual*. San Antonio, TX: The Psychological Corporation.
Wechsler, D. (1997b). *Wechsler memory scale* (3rd ed.). San Antonio, TX: The Psychological Corporation.
Wechsler, D. (1997c). *WMS-III. Administration and scoring manual*. San Antonio, TX: The Psychological Corporation.
Zazzo, R., & Stambak, M. (1964). *Le test des deux barrages: Une épreuve de pointillage*. Paris: Neuchâtel, Delachaux et Niestlé.