Scandinavian Journal of Psychology, 1992, 33, 230–237

Linear versus normalized T scores as standardized neuropsychological test scores

ERIK LYKKE MORTENSEN[1] and ANDERS GADE[2]

[1]*Institute of Preventive Medicine, Kommunehospitalet, Copenhagen, Denmark*
[2]*Psychological Laboratory, University of Copenhagen and Department of Neurology, Rigshospitalet, Copenhagen, Denmark*

Mortensen, E. L. & Gade, A. (1992). Linear versus normalized T scores as standardized neuropsychological test scores. *Scandinavian Journal of Psychology, 33,* 230–237.

In this paper we present and discuss standardized *T* score systems for neuropsychological test data. Both linear and normalized *T* scores were calculated for 141 normal subjects and a group of 141 patients with diffuse or focal brain damage. Many standard neuropsychological tests have skewed raw score and linear *T* score distributions, and we argue that normalized *T* scores have practical advantages because they permit simple descriptions of both patient groups and individual test score distributions. We also argue that skewness can be partially explained by ceiling effects and other test construction artefacts and that skewed raw score distributions do not necessarily reflect skewed distributions of the underlying mental abilities. Consequently, use of normalized *T* scores seems appropriate in many research and clinical contexts.

Key words: Neuropsychological testing, standardized scores, *T* scores, normalized scores.

Erik Lykke Mortensen, Institut for Sygdomsforebyggelse, Kommunehospitalet, 1399 København K, Denmark

One of the purposes of the present paper is to present standardized scores for a standard Danish neuropsychological test battery, called the Basic Battery. Raw score distributions for some of the tests in this battery have previously been described for various samples of non-brain damaged subjects (Andersen, 1976; Mikkelsen *et al.*, 1988; Nielsen *et al.*, 1989). As most Danish neuropsychologists use the complete Basic Battery or some of these tests, there is an obvious need to describe standardized scores, and they will be described for a sample of 141 normal subjects. Raw score distributions for this sample and for various age and educational subgroups will be described separately (Gade & Mortensen, 1992).

The standardized scores will be presented both as linear and normalized *T* scores. Traditionally, the question of linear versus normalized standardized test scores has often been discussed in relation to sampling and test construction problems. Thus, Crocker & Algina (1986, p. 444) discussed normalization in small scale norming studies, and Nunnally (1967, pp. 144–147) showed that the shape of the raw score distribution to a large extent is determined by item difficulties, covariances among items, and the number of items in a test.

In neuropsychology, samples are often very small, and they can rarely be considered random samples of relevant patient populations. Similarly, the typical normal control groups and standardization samples are often much smaller than in non-clinical areas of psychological research. Test construction factors no doubt contribute to skewed raw score distributions of neuropsychological tests which often were constructed without much concern about possible psychometric problems.

In our neuropsychological database we have data for more than one thousand neurological patients. For most diagnostic categories some tests show very skewed raw score

distributions, and some patients show extremely low scores. As can be expected from the above considerations skewed distributions are in fact common for neuropsychological data. This makes the choice between the linear and normalized standardized scores more crucial in neuropsychology than in areas where test score distributions typically show only minor deviations from normality.

Most test batteries consist of different tests, presumably measuring different abilities with different distributions in normal subjects and neurological patients. For some of the tests the distribution of raw scores may be similar to the distribution of measured abilities, but for other tests the raw score distributions will deviate from the ability distributions in different ways. Ideally, the distribution of test scores should approximate the distribution of abilities, and therefore no single raw score transformation can be expected to be ideal for all tests in a battery. Using either linear or normalized T scores for all tests in a battery can only be considered an imperfect and practical solution to complicated measurement and test construction problems.

In view of these considerations, a second purpose of this paper is to present comparisons of linear and normalized T scores, and to examine the consequences in practical terms of choosing one transformation rather than the other.

METHOD

Normal subjects

The selection of the 141 brain healthy subjects was described by Gade *et al.* (1988) and Gade & Mortensen (1992). The age range of the sample was from 20 to 83 years. Since the sample covered several generations, it is difficult to evaluate whether the group can be considered representative of the Danish population with respect to education. However, all possible educational levels were represented in all age groups. The subjects were tested by four psychologists at Rigshospitalet, Copenhagen.

Patient group

By individual matching a group of 141 neurological patients was selected and compared with the normal group. The patient group consisted of 87 patients with diffuse cerebral atrophy and 54 patients with focal brain damage. Eight psychologists at Rigshospitalet participated in data collection for this group, but two of the psychologists together tested 80 percent of the patients.

The composition of the two samples is shown in Table 1. The matching variables were age and educational level (which is the sum of number of school years and occupational training). The matching was successful with respect to educational level, but less so with respect to age. The larger age variance in the normal sample was created deliberately by including a number of relatively young and relatively old subjects (range 20–83 years), while fewer very young or very old neurological patients were available for matching.

It is important to observe that our goal in this study was a comparison of the two types of T scores rather than a direct comparison of test performance of the two groups, and therefore the exact composition of the patient group with respect to background variables as well as type of brain damage or disease is not really important. What is important for our purpose is that the patient group provides neuropsychological test scores that are assumed to be typical of patients with both diffuse and focal brain damage normally tested with this battery.

Test battery

The test battery has been described by Gade *et al.* (1988) and Gade & Mortensen (1992). The Basic Battery is mainly used to evaluate suspected general intellectual impairment. The battery consists of Proverb Interpretation, Classification, Paired Associates, List Learning (Buschke), Digit Span, Sentence Repetition, Symbol Digit Modalities Test (SDMT), Trail Making, Block Design, and Visual Gestalts. Two scores are generated from the Digit Span and the Trail Making tests, and thus the Total Mean of the Basic Battery is calculated from 12 test scores (in Visual Gestalts learning and retention errors were added to yield one score).

Table 1. *Mean age and educational background in 141 normal subjects and 141 patients. Standard deviations are also presented*

	Normal subjects		Patient Group	
	Mean	SD	Mean	SD
Age	49.13	15.72	46.56	11.67
Educational level	11.48	2.58	11.41	2.55
School years	8.80	1.71	8.68	1.77
Occupational training	2.67	1.14	2.73	1.11

Note: The normal group consisted of 63 females and 78 males while the patient group consisted of 39 females and 102 males. This difference in sex ratio is significant, but none of the differences in the educational means are significant in paired t tests. There is a significant difference in mean age and the variance of the age variable is significantly larger in the normal group.

A factor analysis of the test battery suggested four factors, and consequently the means of the corresponding four groups of tests are presented: Visuo-motor, Verbal Learning, Memory Span, and Abstraction test means. The Total Mean of all 12 test scores is also presented because this mean often is used as an index of the general level of cognitive functioning.

Data analysis

In this paper the term "standardizing" is used in the sense of a linear transformation of a test score distribution to obtain a scale with a specified mean and standard deviation in a specified group. The term "normalizing" is used in the sense of a non-linear transformation of a test score distribution to a frequency distribution that approximates a theoretical Gaussian normal curve. Finally the term "T scores" is used to denote test scores that—within rounding errors—have a mean of 50 and a standard deviation of 10 in the normal group.

Computational details and discussion of scoring systems can be found in Crocker & Algina, 1986, chapter 19. For our purpose it is important to note that linear transformations do not change distributional characteristics like skewness and kurtosis or the correlations of a variable with other variables. How much these statistics are changed by normalizing depends on the deviation of the raw score distribution from a theoretical normal distribution. It should also be observed that ties and the discrete distribution of obtained test scores will limit the actual fit of the transformed test score distribution to a Gaussian distribution.

Using the above terminology both linear and normalized T scores were developed for the neuropsychological test battery:

Linear T scores

Based on the results of the normal sample the raw scores of the individual tests were linearly standardized to T scores. Means of test scores were computed for four subgroups of tests and for the sum of all the tests in the Basic Battery. The standard deviation of these composite means will not be 10, and therefore they were restandardized to a mean of 50 and a standard deviation of 10 in the normal sample. The patient group was not used in the calculation of linear T scores, but by using the means and standard deviations of the normal group the test results of the patients were transformed to the same T score system.

Normalized T scores

Using only the normal group to normalize test scores would have the disadvantage that it would not be possible to differentiate between all patient scores worse than the lowest score in the normal group. The combined distributions of single test scores from the normal subjects and the patients were therefore normalized, and these normalized scores were then standardized to a mean of 50 and a standard deviation of 10 in the normal group. These T scores were also used to calculate composite means of test

Table 2. *Mean and SDs for the linear standardization and the normalized T scores in the patient group*

	Linear *T* scores			Normalized *T* scores		
Test	Mean	SD	*F*	Mean	SD	*F*
Total mean	39.63	12.49	74.81	40.26	9.80	94.31
Visuo-motor mean	39.18	14.38	67.69	40.48	10.23	101.89
SDMT	38.51	10.05	117.74	38.67	10.30	117.77
Trial making *A*	37.99	21.00	37.41	41.59	10.37	63.64
Trial making *B*	40.53	14.22	44.07	40.86	9.39	82.62
Block design	42.83	13.34	33.84	44.46	10.76	29.14
Visual gestalts, learning and retention	43.61	12.75	27.39	44.35	9.73	31.49
Verbal learning mean	40.08	14.86	42.34	42.05	12.48	39.81
Paired associates	41.88	12.03	40.17	43.00	11.26	37.41
List learning	39.17	18.42	33.38	42.83	13.25	26.72
Memory span mean	44.63	9.25	20.47	44.87	9.82	21.70
Digit span forward	46.43	8.84	7.27	46.83	9.50	7.07
Digit span backward	44.64	9.69	19.81	43.89	11.41	24.98
Sentence repetition	45.57	9.84	13.40	46.05	9.61	14.28
Abstraction mean	44.28	10.91	22.23	44.78	10.74	21.11
Proverb interpretation	43.86	11.78	27.95	44.69	11.50	25.58
Classification test	46.31	10.92	7.16	46.97	10.50	7.00

Note: The *F* values correspond to a paired *t* test and because of a few missing data they are not all based on exactly 141 pairs. The Classification test score is significant at the 5 percent level while all other *F* values are significant at or below the 1 percent level.

scores and the calculated means were then restandardized to a mean of 50 and a standard deviation of 10 in the normal group. Note that the composite means of test scores were not normalized, and in this aspect we treat our means of test scores like Wechsler IQs: The IQs were only standardized while the scaled scores were normalized and standardized (Wechsler, 1955).

RESULTS

Table 2 presents the means and standard deviations of the composite means of test scores and the corresponding individual tests in the patient group (it will be remembered that the means and standard deviations of the normal sample are 50 and 10). The results are shown for both linear and normalized *T* scores, and as a measure of test sensitivity *F* values corresponding to paired *t* tests between the patient and the normal groups are presented. The pattern of these *F* values are the same for both types of *T* scores. While some *F* values are higher for the linear *T* scores, this tendency is reversed for a number of tests with highly skewed raw scores, and in a few cases the superiority of the normalized *T* scores is considerable. It is, however, important to realize that the *F* values in Table 2 are primarily presented to demonstrate that the brain damage sensitivity of the normalized *T* scores is not substantially less than the linear *T* scores.

It should be observed that the statistical test is valid only if the distribution of the paired differences is approximately normal, and Tables 3 and 4 show that this assumption is much more likely to be violated for some of the highly skewed linear *T* score distributions. Table 2 also demonstrates that the variance of the patient group is much closer to that of the

Table 3. *Kurtosis, skewness, minimal and maximal T score in the normal group*

	Linear *T* scores				Normalized *T* scores			
Test	Kurt.	Skewn.	Min.	Max.	Kurt.	Skewn.	Min.	Max.
Total mean	1.37	−1.02	15	69	−0.15	−0.41	22	69
Visuo-motor mean	4.35	−1.79	8	65	0.31	−0.45	21	72
SDMT	−0.04	−0.05	20	74	0.11	−0.14	20	76
Trial making *A*	3.25	−1.57	11	65	−0.32	0.05	27	76
Trial making *B*	14.98	−3.39	−12	60	0.31	−0.23	19	75
Block design	1.97	−1.28	13	63	0.04	0.07	23	77
Visual gestalts, learning and retention	3.49	−1.67	9	60	−0.33	−0.08	25	70
Verbal learning mean	0.30	−0.87	19	64	−0.20	0.01	25	71
Paired associates	−0.16	−0.88	24	63	0.06	0.04	24	77
List learning	0.98	−1.05	13	64	−0.32	0.21	26	75
Memory span mean	0.70	0.53	27	84	−0.16	0.23	26	79
Digit span forward	0.60	0.45	29	81	−0.25	0.08	29	75
Digit span backward	2.53	1.19	36	87	−0.23	0.14	33	76
Sentence repetition	−0.82	−0.27	24	67	−0.44	−0.06	24	71
Abstraction mean	−0.32	−0.72	20	65	−0.45	−0.37	23	74
Proverb interpretation	−0.40	−0.78	26	63	−0.45	−0.19	27	69
Classification Test	−0.60	−0.54	22	65	−0.64	−0.11	26	74

Note: The measure of kurtosis was adjusted so that both kurtosis and skewness would be zero in a normal distribution. For 141 subjects kurtosis about 0.74 and skewness about 0.40 will give unit normal deviates about 2 according to the formulas in Bock (1975, chapter 3). The presented skewness values are the square root of Bock's b1 statistic.

Table 4. *Kurtosis, skewness, minimal and maximal T score in the patient group*

	Linear *T* scores				Normalized *T* scores			
Test	Kurt.	Skewn.	Min.	Max.	Kurt.	Skewn.	Min.	Max.
Total mean	0.15	−0.58	2	63	−0.13	−0.07	15	66
Visuo-motor mean	2.70	−1.39	−19	63	−0.25	−0.02	15	67
SDMT	0.35	0.28	14	70	0.33	0.22	13	72
Trail making *A*	12.88	−2.92	−97	62	−0.13	0.06	16	70
Trail making *B*	3.18	−1.69	−24	57	−0.18	0.09	16	65
Block design	−0.36	−0.68	3	62	−0.52	−0.02	18	68
Visual gestalts, learning and retention	0.99	−1.21	3	60	−0.23	−0.02	19	70
Verbal learning mean	−0.49	−0.48	0	64	−0.02	0.10	13	74
Paired associates	−1.30	0.10	22	62	−0.35	0.20	19	71
List learning	0.28	−0.88	−18	65	−0.31	0.24	13	80
Memory span mean	0.01	0.12	20	69	−0.01	−0.14	17	66
Digit span forward	−0.02	−0.00	24	70	−0.22	−0.11	22	70
Digit span backward	1.36	1.08	31	82	−0.21	0.20	23	72
Sentence repetition	−0.55	−0.03	20	67	−0.08	0.08	22	71
Abstraction mean	−0.77	−0.25	20	65	−0.44	0.11	23	74
Proverb interpretation	−1.21	−0.09	26	63	−0.45	0.31	27	69
Classification test	−0.13	−0.55	13	65	−0.15	−0.04	20	74

normal sample for the normalized T scores. This is an advantage because it simplifies description and statistical testing of group differences.

When test scores approximate a normal distribution group performance can be characterized by the mean and the standard deviation, and with equal variances, differences in group means can be considered a sufficient description of between group differences in test performance. The standard deviations in Table 2 and the measures of kurtosis and skewness in Tables 3 and 4 clearly show that this is not the case for many of the linear T scores while the normalized T scores approximate a normal distribution in both the normal group and the patient group (because the scores were normalized for the combined distribution the scores in either group may deviate from a normal distribution).

DISCUSSION

One may argue that other parameters—such as the median—can be used as measures of central tendency in highly skewed distributions and that nonparametric statistical tests can be used in group comparisons. However, the interpretation of any measure of central tendency is complicated in highly skewed distributions and the interpretation of nonparametric tests is complicated if the distributions of the compared groups are very dissimilar. Therefore, normalized T scores often will be more appropriate for both simple descriptive purposes and inferential statistics.

The clinician will also be interested in what kind of T scores best describes the results of the individual patient. The minimum and maximum scores in Table 3 and 4 show that a major difference between the linear and normalized T scores is in the lowest obtained scores: In the normal group the lowest obtained scores are -12 and 19 for the two types of scores, and in the patient group the lowest obtained scores are -97 and 13, respectively. The very low linear T scores may be an advantage in the sense that a very low score may truly reflect a patient's performance in a test, but it may also be a disadvantage when means of test scores are used to characterize overall performance. A single very low score may affect the overall mean to an unreasonable degree, and it can in fact be argued that the problems with very skewed individual distributions are very similar to the problems with very skewed group distributions (it should be remembered that specific disabilities such as a slight paresis or aphasia may affect some neuropsychological tests).

In a clinical context the fact that normalized T scores do not permit very low scores should rarely be a problem because patients with such low overall performance are usually referred to neuropsychologists for other than diagnostic purposes. It is therefore concluded that normalized T scores also seem to be most appropriate in the evaluation of individual patients.

Although normalized T scores have desirable properties for the description of both groups and individual patients, there are no *a priori* reasons to expect normal distributions of mental abilities in our groups, including our normal sample. We therefore have to face the question of whether the distributions of the linear or the normalized T scores are closer to the distributions of the mental abilities that the tests are supposed to measure. This question may be clarified by using the terminology of latent trait theory although most neuropsychological tests probably cannot be considered unidimensional, and although latent trait theory has played no role in the development of these tests. The "test-characteristic curve" represents the regression function relating the observed scores to the underlying latent trait or mental ability (Lord, 1980), and the important point is that this relation is typically not linear. This is illustrated in the examples of Lord & Novick (1986, p. 387–392) that also include an example of how test scores may be highly skewed although the latent trait has a symmetrical

distribution. The important point from our perspective is that distributions of mental abilities cannot easily be predicted from distributions of observed test scores, and it can indeed be argued that for most neuropsychological tests no definitive answer can be given to the question of the relation between *T* scores and the underlying mental abilities.

However, the skewness of individual tests may be artefacts of ceiling- and floor effects or a consequence of the chosen response measures. Inspection of the distributions of the normal group has suggested to us that ceiling effects may partially explain the skewness in tests like Proverb Interpretation, Paired Associates, List Learning, Sentence Repetition, and Visual Gestalts (note that all these tests are less skewed in the patient group). Typical tests in this respect are the verbal learning tests and the Visual Gestalts where the raw score is the number of errors. In these tests a number of subjects commit no or very few errors, and clearly the distribution is truncated in the upper end. Therefore, the skewed raw score distributions do not necessarily indicate that the underlying trait has a skewed distribution, and it is indeed not unlikely that the distribution of the normalized *T* scores are closer to the distribution of "verbal learning ability". Obviously, ceiling effects must of course be considered short-comings in test construction. The best solution to this problem is of course to construct tests without ceiling effects, and this should indeed not be difficult for verbal learning tests.

Another group of tests with highly skewed distributions are Trial Making *A* and *B* and the Block Design test. In these tests completion time in seconds is the raw score, and it is well known that time measures often have skewed distributions. It could therefore be argued that some underlying mental ability ("psychomotor speed") inherently has a very skewed distribution. However, comparison with the SDMT does not support this idea. SDMT presumably measures the same mental ability. It is unaffected by ceiling effects, and although the response measure is affected by a time constraint, the raw score distribution seems close to normality in the normal group. This again demonstrates that skewed raw distributions do not necessarily indicate skewed distributions of the underlying mental abilities.

In conclusion, both practical and theoretical concerns lead us to conclude that it will be most appropriate to use normalized *T* scores in many clinical and research contexts. The same concerns, however, lead us to conclude that psychometric considerations should play a much more central role in the development of new neuropsychological tests. With better tests and more relevant response measures the choice between linear and normalized *T* scores may not be so important because raw score distributions will probably deviate much less from normality.

This study was supported by grants from the Danish Medical Research Council to Anders Gade.

REFERENCES

Andersen, R. (1976). Verbal and visuo-spatial memory. Two clinical tests administered to a group of normal subjects. *Scandinavian Journal of Psychology, 17,* 198–204.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research.* New York: McGraw-Hill.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston.

Gade, A. & Mortensen, E. L. (1992). Standardization of a neuropsychological test battery: Methods and basic results (in preparation).

Gade, A., Mortensen, E. L. & Bruhn, P. (1988). "Chronic painter's syndrome". A reanalysis of psychological test data in a group of diagnosed cases, based on comparisons with matched controls. *Acta Neurologica Scandinavica, 77,* 293–306.

Lord, F. M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, New Jersey: Erlbaum.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Massachusetts: Addison-Wesley.

Mikkelsen, S., Jørgensen, M., Browne, E. & Gyldensted, C. (1988). Mixed solvent exposure and organic brain damage. *Acta Neurologica Scandinavica, 78,* Supplementum 118.

Nielsen, H., Knudsen, L. & Daugbjerg, O. (1989). Normative data for eight neuropsychological tests based on a Danish sample. *Scandinavian Journal of Psychology, 30,* 37–45.

Nunnally, J. (1967). *Psychometric theory*. New York: McGraw-Hill.

Wechsler, D. (1955). *Wechsler adult intelligence scale*. Manual. New York: Psychological Corporation.

Received 18 December 1990