Review Article

Open Access

# Better assessment of neuropsychological tests

## Abstract

**Background:** Neuropsychological tests (NTs) with different number of items (test length), number of response-categories/levels (width), dimensions covered, different values of reliability, validity, responsiveness, discriminating value, etc. are not comparable.

**Aim:** To address methodological issues of neuropsychological testing and suggest remedial measures by transforming discrete item scores to normally distributed continuous scores for meaningful evaluation of measurement properties and better utilization of such tests.

**Methods:** Using data driven weights to different response-categories of different items, ordinal item scores are converted to equidistant score (*E*-scores) in ratio scale with fixed zero point. Proposed scores (*P*-scores) are obtained via standardization of *E*-scores followed by further transformation to follow normal distribution. Sub-class/dimension scores and test scores are obtained as sum of item-wise *P*-scores.

**Results:** Normally distributed *P*-scores facilitate arithmetic aggregation with cardinal measures like number of errors, time taken, etc. and offer platform for parametric analysis including statistical testing. The proposed *P*-scores also help to find theoretically defined reliability, factorial validity avoiding criterion scale, discriminating value, assessment of progress/deterioration, efficiency of classification, equivalent scores of neuropsychological tests.

**Conclusion**: Proposed scores following normal distribution and satisfying desired properties of measurement is recommended. Practicing psychiatrists and researchers can derive benefits of the proposed score for more meaningful comparisons, classifications, equating boundary scores for classification, and assessment of progress or deterioration.

**Keywords:** neuropsychological test battery, normal distribution, ratio scale, equivalent scores, theoretical reliability, factorial validity

Volume 15 Issue 5 - 2025

Satyendra Nath Chakrabartty
Indian Statistical Institute, Indian Ports Association, India

**Correspondence:** Satyendra Nath Chakrabartty, Indian Statistical Institute, Indian Ports Association, Flat 4B, Cleopatra, DC 258, Street No. 350, Action Area 1, New Town, Kolkata 700156, India, Tel 919831597909

**Received:** August 05, 2025 | **Published:** October 16, 2025

## Introduction

Measurement issues and associated statistics and psychometrics are foundational elements in neuropsychological practice. Neuropsychological tests (NTs) are used for diagnostic purpose and treatment insights for mental health disorders. NTs involve among others, *K*-point items in Numeric Rating scales (NRS) marked as 1, 2, 3, …., *K* pertaining to sub-classes or dimensions of Motor functions, Perceptions, Ability of problem-solving and decision-making, Verbal ability, Memory, Intelligence, Executive Functions, Language, Visuo-spatial, Multiple Functions, etc. For example, in the Category test of Halstead-Reitan Battery (HRB), participants decide best correspondence between geometric figures and numbers 1, 2, 3, or 4.

Aggregating ordinal scores with cardinal measures like number of errors (Seashore Rhythm Test of HRB) or Time taken to complete (Tactual Performance Test of HRB) etc. are problematic. Depending on patient symptoms, neuropsychologists decide the tests for assessment of the patient's cognitive abilities/disorders, better understanding of current health picture and medical needs and also to know whether the problems have causal relationships with the following:

i. Alzheimer's disease, Dementia, Parkinson's disease and other disorders.

ii. Ischemic attack and stroke.

iii. Traumatic brain injury.

iv. Neurological issue like epilepsy

v. Disease of the central nervous system (CNS) like Multiple sclerosis (MS).

vi. Tumors or infections in brain and spinal cord.

vii. Injury of brain

viii. Depression, anxiety and other emotional disorders

ix. Effect of age on brain changes, etc.

Large numbers of NTs are there to evaluate cognitive functions or cognitive disorders among adolescents, elderly persons and children. Choices of the tests are usually made based on dimensions covered and quality measures like reliability and validity. However, no test uses definition of test reliability as

$$\frac{True\, score\, variance\left(S_T^2\right)}{Observed\ score\, variance\ (S_X^2)}.$$

Test-retest reliability, Cronbach alpha have inherent problems. Test validity computed by correlation between test scores and scores of chosen criterion variable may also be interpreted as validity of the criterion variable. Other quality measures like Discriminating value of test, responsiveness (ability to assess improvement or deterioration), etc. could also be critical in selection of tests. Reliability, validity, for sub-scale and test are influenced by different values of *k*- point scales, *k*= 2, 3, 4, 5, ……. etc.

The paper highlights methodological issues of neuropsychological testing and suggests remedial measures by transforming discrete and ordinal item scores to monotonically increasing, normally distributed continuous scores for meaningful evaluation of properties and better utilization of such tests.

## Problem areas

NTs with different number of items (test length), number of response-categories/levels (width), dimensions covered, different values of psychometric qualities are not comparable. Reporting of results by mean $\pm$ standard deviation (SD) assuming admissibility of addition of ordinal data and equidistant response-categories, testing equality of means by *t*-test, analysis of variance (ANOVA), requiring normally distributed scores, etc. are common.

NTs have been criticized from insides[1] and also outside the discipline.[2] Neuropsychological Test Battery (HRNB) lacks basic psychometric documentations to facilitate interpretations of results.[3] Methodological errors in construction of Luria–Nebraska Neuropsychological Battery (LNNB) are significant.[4] Similar problems exist for normative data of Continuous Performance Test (CPT), National Adult Reading Test, Benton Facial Recognition Test (BFRT), Purdue Pegboard Test, Rey Complex Figure Test (ROCF), Stroop Color and Word Test (SCWT), etc. and are far below contemporary standards.[5]

Major methodological problem areas are:

## Nature of data

Data generated by NRS/Likert items consist of frequency (count of responses) of each response-category (level) which are assigned successive numbers in integers but the levels are not equidistant in terms of the trait differences. If $d_{j(j+1)}$ denotes the distance between *j*-th and *(j+1)*-th levels of an item then equidistant property requires $d_{j(j+1)}$ to have constant value for all *j* =1, 2, 3, 4 for a 5-point item.

Rating data are often skewed (asymmetric) and require normality checks for analysis and inferences.[6] Distribution of ordinal data emerging from rating scales are unknown and may not follow normal distribution, which violates basic assumptions of a number of statistical procedures.[7] Ordinality, discreteness, nonlinearity, skew of NRS data with ceiling and floor effects suffer from methodological limitations for parametric statistical analysis.[8] For measurement, a scale needs to have important features like well-defined zero point, metric characteristics and defined operational procedure.[9]

## Scoring

Failure to satisfy equidistant property makes addition of ordinal item scores non- meaningful.[10] Distance between ordered response-categories is unknown and not uniform.[11] Test score as sum of independent dimensions amounts to adding apples with oranges. Assigning equal importance to the items and dimensions contradict different contributions of items/dimensions to total score, different values of item-total correlations, factor loadings, etc.[12] Non-meaningful addition implies computation of mean, variance, correlation, Cronbach alpha and analysis like regression, ANOVA, Factor Analysis (FA), Principal Component Analysis (PCA) and statistical inferences covering estimation, testing, etc. are not meaningful.[13] Parametric statistical analysis of ordinal data with non-meaningful addition and non-satisfaction of normality assumption is taken as one of the seven deadly sins.[14]

Interval scales like BMI scores are equidistant but the zero point is not fixed. Thus, difference between two measurements has meaning, but their ratio does not.[15] Ratio scale measurements like height, weight, pulse rate, respiratory rate, etc. have fixed zero point indicating absence of the quantity being measured and allows divisions. Different patterns of responses to different levels of items often results in tied scores i.e. the same aggregate test score for a sub-group of subjects taking the test. Thus, commonly used summative scores cannot distinguish the subjects with tied score which in turn lowers the discriminating value of the test.

## Distribution of score

Probability distribution of test scores are unknown for NTs. Item scores depend on endorsed response-categories and follow dissimilar distributions. Interpretation of sum of scores of two items $X + Y = Z$ is most meaningful when we know probability density function (pdf) of *X* and *Y* and find pdf of *Z* say by convolution, enabling us to find $P(Z = z) = P(X = x,\ Y = z - x)$ for discrete case and $P(Z \le z) = P(X + Y \le z) = \int_{-\infty}^{\infty} (\int_{-\infty}^{z} f_{X,Y}(x, t-x)dt)dx$ for continuous case. However, if each of *X* and *Y* follows log-normal, then distribution of Z as $(X + Y)$ cannot be obtained as such and requires use of complex splitting method of Lie-Trotter operator.[16] Problems of parametric statistical analysis with ordinal, discrete, skewed NRS data were addressed by Šimkovic & Träuble.[8]

## Non-satisfaction of assumptions

Scores emerging from NTs may not satisfy assumptions of statistical techniques used in analysis of data. Statistical techniques like PCA, FA, *t*-test, paired *t*-test, *F*-test, etc. are based on normal distribution of the study variables and may give distorted results in case of violation of the respective assumption. For example, high value of $r_{XY}$ is taken as linearity of *X* and *Y* and ordinary least-squares (OLS) regression of the $(Y = \alpha + \beta X + \in)$ is fitted. However, $r_{XY}$ may be high even if *Y* is non-linearly related with *X*. For example, if *X* takes values 1, 2, 3…30, $r_{X,X^2} = 0.97; r_{X,X^3} = 0.92;$ $r_{X,\log_{10}^{x}} = 0.92$ despite each of $X^2, X^3, log_{10}^{X}$ is non-linear function of *X*. Clearly, linearity implies high correlation but not the converse. Possible action could be to fit $Y = \alpha + \beta X + \grave{o}$ and check whether error score $E = (Y - \hat{Y})$ follows normal distribution by say Anderson–Darling test or testing the hypothesis $H_0 : S_E^2 = 0$ where $S_E^2$ denotes variance of error scores and is computed by $S_E^2 = \frac{1}{n}\sum(Y_i - \hat{Y}_i)^2$ for a sample size *n*. Error score of $(Y = \alpha + \beta X + \in)$ for Y=X$^2$ or X$^3$ or $log_{10}^{X}$, did not follow normal and violated the OLS assumptions.

It will be highly desirable to transform ordinal item scores to continuous equidistant scores (*E*-scores) followed by transforming *E*-scores to normally distributed proposed scores (*P*-scores) such that $1 \le P_i \le 100$

## Cut-off point and classification

Consider another example where $Y = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}X^2}$ for $0 \le X \le 3.9$. Here, $r_{XY} = 0.93$. But, if $-3.9 \le X \le 3.9$, $r_{XY} = 0.00036$. Thus, truncated values of either *X* or *Y* or both can affect the correlation significantly. The point is relevant for classification of a group of subjects under "Healthy" (not suffering from the disease) and "With disease" assuming higher the score, higher are the dysfunctions i.e. to decide cut-off point $(X_0)$ ensuring that subjects with scores $\le X_0$ are healthy and those with scores exceeding $X_0$ have the disease. Samples with high proportion of healthy persons or patients suffering from the disease will tend to truncate distribution of test score at right or left. Thus, $X_0$ obtained from study-specific populations may affect clinical heterogeneity in different investigations and not allow comparisons between studies. Selection of the sample should also consider demographic variables like age, education level, intellectual ability, gender, ethnicity, etc. and anthropometric factors like handedness, non-demographically based factors like smoking habits, lateralized tremor in Parkinson's disease (PD) patients, etc. since they can affect test results.[17]

Diagnosis of neurological disorder is complex since different combinations of different disorders may give rise to similar symptoms. Moreover, causes of many disorders are not known and biomarkers, NTs for such disorders are not specific. Biomarkers have limitations too. For example, a case study by Bouter et al.[18] indicated missing of core clinical features of Dementia with Lewy Body (DLB) and negative indicative biomarkers, but NTs and Positron emission tomography (PET) imaging provided evidence of early stage of DLB. After follow-up and treatment, core symptoms and indicative biomarkers appeared in later stages of the disease. Testing of cerebrospinal fluid (CSF) by standard two-tier testing algorithm (STTA), enzyme immunoassays (EIAs), and/or immunoblots for diagnosis of central nervous system Lyme disease; culture of CSF for Lyme Borreliais are not recommended.[19]

Diagnostic and Statistical Manual of Mental Disorders (DSM)-IV is a categorical classification system which is prototypes. A patient belonging to the prototype is said to have that disorder. In addition, classifications in several classes are made like mild, moderate, severe forms of a disorder. However, diagnosis based on a function of data emerging from a NT is inefficient. For example, in case of dementia, multiple additional criteria are required to meet for which relevant information are obtained from other sources.

## Reliability

Commonly used measures of reliability of NTs are Cronbach alpha, Test-retest reliability. Test-retest reliability reflecting stability of scores is calculated as correlation between two administrations of the test at two different time points on the same sample with same testing conditions. But there is no agreed choice of time between the two administrations.[20] It does not consider rank-order stability of individuals in two administrations and may fail to show degree of agreement between two administrations. Difference between correlation and agreement was demonstrated empirically by Streiner, Norman & Cairney.[21] Test-retest reliability may be high despite rejection of hypothesis $H_0 : \mu_{X\ Test} = \mu_{X\ \text{Retest}}$ by paired *t*-test.[22] In addition, correlation is influenced by heterogeneity of samples which implies that test-retest reliability is relative to the sample.

Reliability by Cronbach's alpha works best for unidimensional test i.e. all items measure the same construct. Violation of the assumption may make the coefficient α biased[23] by distorting variance-covariance matrix for non-symmetric distribution of observed responses.[24] However, there are instances of reporting Cronbach alpha despite several factors emerged from PCA or FA. Against suggestion of two-factor solution (memory factor and visuospatial factor) for Repeatable Battery for Assessment of Neuropsychological Status (RBANS) with 12 sub-tests, five index scores and a total scale score, Cronbach alpha = 0.92 of RBANS was found.[25] Eight independent factors of expanded Halstead-Reitan Battery (eHRB) out of over forty test measures was found by Patt et al.,[26] despite caution about inferring complex mental abilities by FA or structural equation modeling).[27] Test reliability is different from sum or average of sub-test reliabilities. Internal reliability coefficients of the Wechsler Adult Intelligence scale – Fourth Edition (WAIS-IV) was 0.98 against reliability of Verbal comprehension (0.96), Working memory (0.94), Perceptual reasoning (0.95) and Processing speed (0.90).[28]

Avoiding unidimensionality assumption of Cornbrash alpha, Chakrabartty[29] proposed finding reliability as per theoretical definition $(r_{tt-theoretical})$ and Error variance $S_E^2$ in terms of two parallel subtests (say *g*-th and *h*-th) obtained by dichotomization of the test.

Here, $$S_E^2 = \frac{1}{N}[\left\|X_g\right\|^2 + \left\|X_h\right\|^2 - 2\left\|X_g\right\|\left\|X_h\right\|\cos\theta_{gh} \qquad (1)$$

where *N* denotes sample size; length of the *g*-th and *h*-th vector are $\left\|X_g\right\| = \sqrt{\sum_{i=1}^{N} X_{ig}^2}$ and $\left\|X_h\right\| = \sqrt{\sum_{i=1}^{N} X_{ih}^2}$ respectively and $\theta_{gh}$ is the angle between the vectors(*g*-th and *h*-th). Thus,

$$r_{tt-theoretical} = \frac{S_T^2}{S_X^2} = 1 - \frac{S_E^2}{S_X^2} = 1 - \frac{\frac{1}{N}\left[\left\|X_g\right\|^2 + \left\|X_h\right\|^2 - 2\left\|X_g\right\|\left\|X_h\right\| Cos\theta_{gh}\right]}{NS_X^2}$$

(2)

Reliability of a battery of test with *K*-number of subtests could be found in terms of sub-test reliabilities (without weights) by

$$r_{tt(Battery)} = \frac{\sum_{i=1}^{K} r_{tti} S_{Xi}^2 + \sum_{i=1 i\neq j}^{K} \sum_{j-1}^{K} 2Cov(X_i, X_{j)}}{\sum_{i=1}^{K} S_{Xi}^2 + \sum_{i=1 i\neq j}^{K} \sum_{j-1}^{K} 2Cov(X_i, X_{j)}} \qquad (3)$$

## Validity

For two different tests *X* and *Y*, criterion validity of *X* is given by $r_{XY}$. However, $r_{XY}$ could also be taken as validity of *Y*. High value of $r_{XY}$ may mean that instrument X is not required. Moreover, such validity assumes similarity of latent variables including similar factor structure of tests *X* and *Y*, both administered to the same sample. For high positive skew of the test score(*X*) and/or criterion score(*Y*), validity may be lower if the data contains high proportion of high performers.[30] To avoid such problems, factorial validity of normally distributed transformed scores was preferred.[31]

# Proposed method

## Pre-adjustment of data

(1) Convert levels of a *k*-point item to 1, 2, 3, …., *k* avoiding zero.

(2) Ensure each item is positively related to the test score i.e. higher item score indicates higher level of disorder.

## Transformations of item scores

Using different positive weights $W_1, W_2, W_3, ..........., W_k$ to response-categories of different items transform each item score to equidistant score (*E*) such that $W_1, 2W_2, 3W_3, ..........., \text{K}W_k$ are an arithmetic progression which ensures satisfaction of equidistant property by method suggested by Chakrabartty[31] which is briefly described below for *n*-number of respondents and $k = 5$

I: Find maximum frequency $(f_{max})$ and minimum frequency $(f_{min})$ of the levels of *i*-th item.

Let $\frac{f_{ij}}{n}$ be the initial weight of the *j*-th response-category of *i*-th item $\left(\omega_{ij}\right)$.

Arrange $\omega_{ij}'s$ in increasing order so that where $\omega_{i1} = \frac{f_{min}}{n}$ and $\omega_{i5} = \frac{f_{max}}{n}$.

Choose intermediate weight $W_{i1} = \omega_{i1}$ Compute common difference $\alpha$ so that

$$W_{i1} + 4\alpha = 5\text{W}_{\text{i}5} \Rightarrow \alpha = \frac{5f_{max} - f_{min}}{4n}$$

Find intermediate weights of other response-categories as

$$W_{i2} = \frac{\omega_{i1} + \alpha}{2},\ W_{i3} = \frac{\omega_{i1} + 2\alpha}{3};$$

$W_{i4} = \dfrac{\omega_{i1}+3\alpha}{4}$ ; and $W_{i5} = \dfrac{\omega_{i1}+4\alpha}{5}$ .

Get final weights $W_{ij(Final)} = \dfrac{W_{ij}}{\Sigma_{j=1}^{5} W_j}$ . Here, $\Sigma W_{ij(Final)} = 1$ and

$j.\mathrm{W}_{j(Final)} - (j-1).W_{(j-1)\,(Final)} = constant$ .

Value of the constant will be different for different items.

II: Now, $E$-scores are to be standardized to $Z$-scores using $Z = \dfrac{E-\bar{E}}{SD(E)} \sim N(0,1)$

III: Convert $Z$-scores to proposed scores ($P$-scores) by

$$P = (99)\left[\frac{Z_{ij} - Min_{Z_{ij}}}{Max_{Z_{ij}} - Min_{Z_{ij}}}\right] + 1 \text{ so that } 1 \le P \le 100 \text{ and } P \text{ follows}$$

normal. Dimension score of an individual is the sum of normally distributed $P$-scores of the relevant items and test score is the sum of all dimension scores.

Empirical illustration of calculation of weights is given below in Table-1 for a hypothetical 5-point scale with 7 items, responded by 101 individuals.

**Table 1** Calculation of weights.

| Item | Description | Level -1 | Level -2 | Level -3 | Level -4 | Level -5 | Total |
|---|---|---|---|---|---|---|---|
| 1 | Frequency | 27 | 32 | 10 | 18 | 14 | 101 |
| | Proportions ($\omega_1 j$) | 0.26733 | 0.31683 | 0.09901 | 0.17822 | 0.13861 | 1.00 |
| | Intermediate weights($W_{1i}$) ($\alpha$ =0.37129) | 0.26733 | 0.31931 | 0.33663 | 0.34530 | 0.35049 | 1.6196 |
| | Final weights ($W1j(Final)$ | 0.16511 | 0.19722 | 0.20792 | 0.21327 | 0.21648 | 1.00 |
| 2 | Frequency | 5 | 12 | 11 | 31 | 42 | 101 |
| | Proportions ($\omega_2 j$) | 0.04950 | 0.11881 | 0.10891 | 0.30693 | 0.41584 | |
| | Intermediate weights($W_2 j$) ($\alpha$ =0.50743) | 0.04950 | 0.27846 | 0.35478 | 0.39295 | 0.41584 | 1.4913 |
| | Final weights ($W2j(Final)$ | 0.03319 | 0.18670 | 0.23786 | 0.26345 | 0.2788 | 1.00 |
| 3 | Frequency | 6 | 13 | 7 | 33 | 42 | 101 |
| | Proportions ($\omega_3 j$) | 0.05941 | 0.12871 | 0.06931 | 0.32673 | 0.41584 | 1.00 |
| | Intermediate weights($W_3 j$) ($\alpha$ =0.50495) | 0.05941 | 0.28218 | 0.35644 | 0.39356 | 0.41584 | 1.5076 |
| | Final weights ($W5j(Final)$ | 0.03941 | 0.18719 | 0.23645 | 0.26108 | 0.27586 | 1.00 |
| 4 | Frequency | 29 | 14 | 12 | 25 | 21 | 101 |
| | Proportions($\omega 4j$) | 0.28713 | 0.13861 | 0.11881 | 0.24752 | 0.20792 | 1.00 |
| | Intermediate weights ($W_4 j$) ($\alpha$=0.329208) | 0.28713 | 0.30817 | 0.31518 | 0.31869 | 0.32079 | 1.549 |
| | Final weights ($W4j(Final)$ | 0.18525 | 0.19882 | 0.20335 | 0.20561 | 0.20697 | 1.00 |
| 5 | Frequency | 9 | 14 | 6 | 32 | 40 | 101 |
| | Proportions ($\omega_5 j$) | 0.08911 | 0.13861 | 0.05941 | 0.31683 | 0.39604 | 1.00 |
| | Intermediate weights($W_5 j$) ($\alpha$ =0.480198) | 0.08911 | 0.28465 | 0.34983 | 0.38243 | 0.40198 | 1.5083 |
| | Final weights ($W5j(Final)$ | 0.05909 | 0.18876 | 0.23199 | 0.25360 | 0.26657 | |
| 6 | Frequency | 4 | 9 | 3 | 36 | 49 | 101 |
| | Proportions ($\omega_6 j$) | 0.03960 | 0.0891 | 0.02970 | 0.35644 | 0.48515 | 1.00 |
| | Intermediate weights($W_6 j$) ($\alpha$ = 0.599) | 0.03960 | 0.31930 | 0.41253 | 0.45915 | 0.48712 | 1.717 |
| | Final weights ($W6j(Final)$) | 0.02306 | 0.18589 | 0.24016 | 0.26730 | 0.28359 | 1.00 |
| 7 | Frequency | 49 | 25 | 6 | 13 | 8 | 101 |
| | Proportions ($\omega 7j$) | 0.48515 | 0.24753 | 0.05941 | 0.12871 | 0.07921 | 1.00 |
| | Intermediate weights($W7j$) $\alpha$ =0.59158) | 0.48515 | 0.53837 | 0.55611 | 0.56497 | 0.57030 | 2.714S4 |
| | Final weights ($W_7 j$ ($Final$)) | 0.17870 | 0.19830 | 0.20483 | 0.20810 | 0.21006 | 1.00 |

*E*-scores of item 1 = 1(0.16511) +2(0.19722) +3(0.20792) +4(0.21327) +5(0.21648) = 3.11879

For each of other items, *E*-scores can be computed similarly. Item-wise *E*-scores $\left(E_i's\right)$ can be standardized to *Z*-scores

$Z_i = \frac{E_i - \overline{E}}{SD(E)} \sim N(0,1)$. However, *E*-scores of the indicators in ratio scale are not required. Such indicators can be standardized straight to *Z*-scores. *Z*-scores of each indicator (ordinal or in ratio/interval scale) can be further transformed to proposed scores (*P*-scores) by

$P = (99)\left[\frac{Z_{ij} - Min_{Z_{ij}}}{Max_{Z_{ij}} - Min_{Z_{ij}}}\right] + 1$ so that $1 \le P \le 100$ and (*P*-scores) follows normal distribution.

## Properties

i) Equidistant scores (*E*-scores) can be taken in ratio scale where fixed zero point is obtained when $f_{ij} = 0$ for the *i*-th item.

ii) The method is applicable for items with different width i.e. different values of *k*.

iii) *P*-scores of the *i*-th indicator ($P_i$) follows $N\left(\mu_i, \sigma_i^2\right)$ where $\mu_i$ and $\sigma_i^2$ can be estimated from the data.

iv) Scores of a dimension $\left(D_i\right)$ is $\sum \text{Re}\,levant\,P_i$ and the scale score $\left(S_i\right) = \sum D_i = \sum All\,P_i$ and unweighted battery score is equal to sum of all $S_i s$. Similar probability distribution of $P_i s$ give meaningful arithmetic aggregation. Here, $S_i's$ are monotonically increasing normally distributed continuous scores and facilitate parametric analysis including testing of statistical hypothesis like $H_0 : \mu_1 = \mu_2$ or $H_0 : \sigma_1^2 = \sigma_2^2$ etc. for longitudinal or cross sectional data.

v) Progress or deterioration of the *i*-th person in two successive time-periods is indicated by $S_{i(t)} > S_{i(t-1)}$ or $S_{i(t)} < S_{i(t-1)}$ respectively. Quantification of progress is given by $\frac{S_{i(t)} - S_{i(t-1)}}{S_{i(t-1)}} X 100$ reflecting effectiveness of treatment plan. Similarly, progress for a sample of patients is reflected if $\overline{S_{i(t)}} > \overline{S_{i(t-1)}}$. Deterioration reflected by $\overline{S_{i(t)}} < \overline{S_{i(t-1)}}$ may be probed to find extent of deterioration in dimension scores for necessary actions including change of treatment plan.

vi) Normality of $S_i$ helps to test $H_0 : \mu_{st} = \mu_{s(t-1)}$ or $H_0 : \text{Progress}_{(t+1)\,\text{over}\,t} = 0$. This may avoid finding minimal important difference (MID) of a scale or testing $H_0 : \mu_{S_{pre-group}} = \mu_{S_{post-group}}$ using paired *t*-test.

vii) Plotting of progress/deterioration of one patient or a group of patients across time gives progress pattern i.e. response to treatments from the start. Such progress patterns can be meaningfully compared.

## Benefits

Normally distributed *S*-scores also help to find reliability, validity of the scale in better fashion.

## Factorial validity

Factorial validity (FV) of a scale is defined as ratio of the first eigenvalue to the sum of all eigenvalues i.e. Factorial Validity = $\frac{\lambda_1}{\sum \lambda_i}$, where $\lambda_1$ denotes the highest eigenvalue associated with the first principal component reflecting the main factor for which the test was developed.[32] FV accounts for $\frac{\lambda_1}{\sum \lambda_i} \times 100$ percent of overall variability. FV avoids the problems of selection of criterion scale with similar factor structure and can better be used to find validity of multidimensional tests.

## Reliability

Normality helps to have population estimates of variance of item and test variance and can be used to find population estimate of Cronbach alpha for a domain/sub-class with *n*-items as

$$\hat{\alpha} = \left(\frac{n}{n-1}\right)\left(1 - \frac{\Sigma \sigma_i^2}{\sigma_T^2}\right) \qquad (4)$$

where $\sigma_T^2$ is the estimated variance of the test and $\sigma_i^2$ is the estimated variance of the i-th item

Theoretically defined Reliability of a sub-scale can be obtained by dichotomizing the sub-scale in two parallel halves and find test reliability $\left(r_{tt-theoretical}\right)$ by equation (2) and battery reliability $\left(r_{tt-Battery}\right)$ by equation (3).

## Discriminating value

Average score for a group of healthy adults < average score of the group of patients suffering from the disease may not suffice to conclude that discriminating value of the test is good. Discriminating value of a test reflecting ability of the test to distinguish among the individuals that have different degrees of the underlying construct (e.g. more or less severe disease), needs to be quantified. Discriminating value of a NRS item $Disc_i$ and test $Disc_{Test}$ can be computed by Coefficient of variation (CV) as:

$$Disc_i = \frac{SD_i}{Mean_i} \text{ and } Disc_{Test} = \frac{SD_{Test}}{Mean_{Test}}.$$

Clearly, $Disc_i^2 = \frac{S_{X_i}^2}{\overline{X}_i^2} \Rightarrow S_{X_i}^2 = \overline{X}_i^2 . Disc_i^2$

$\Rightarrow \sum_{i=1}^{m} S_{X_i}^2 = \sum_{i=1}^{m} \overline{X}_i^2 . Disc_i^2$ for a test with *m*-items.

and Test variance $S_X^2 = \overline{X}^2 .\ Disc_T^2$

$$\text{Thus, } \alpha = \left(\frac{m}{m-1}\right)\left(1 - \frac{\sum_{i=1}^{m} \overline{X}_i^2 . Disc_i^2}{\overline{X}^2 . Disc_T^2}\right) \qquad (5)$$

and $$\left(Disc_{Test}\right)^2 = \frac{CV_{Truescores}{}^2}{r_{tt}} \quad \text{where } r_{tt} = \frac{S_T^2}{S_X^2} \qquad (6)$$

Clearly, theoretical test reliability and $Disc_{Test}$ have inverse non-linear relationship.

## Classification

Classifications of individuals to a number of mutually exclusive classes is an exercise to find boundary points where members within a class are similar (i.e. small within group variance) and members between classes are dissimilar (i.e. high between group variance). Quartile clustering helps in classification of a group of individuals in four mutually exclusive classes viz. the quartiles $Q_1, Q_2, Q_3, Q_4$ ).[33] Quartile clustering of normally distributed *S*-scores is simple and appealing with well-defined cut-off scores and equal probability to each class i.e.

$$\int_1^{Q_1} f(x)dx = \int_{Q_1}^{Q_2} f(x)dx = \int_{Q_2}^{Q_3} f(x)dx = \int_{Q_3}^{Q_4} f(x)dx \qquad (7)$$

If needed, decile clustering may be used to classify individuals in 10 classes instead of quartile clustering. Efficiency of classification may be assessed by Davies-Bouldin Index (DBI)[34] defined as

$$DBI_K = \frac{1}{K}\sum_{i=1}^{K}\sum_{j=1(i\neq j)}^{K} Max\left[\frac{DiamC_i - DiamC_j}{C_i - Cj}\right] \text{ where}$$

diameter of *i*-th class $DiamC_i = \sqrt{\frac{\sum_{x_i \in c_i} \| x_i - c_i \|^2}{n_i}}$

$C_i$ : Centroid or mean of the *i*-th class;

*K*: Number of classes;

$n_i$ : Number of members in the *i*-th class.

Upper limit of DBI is 1 and lower DBI value implies higher efficiency of classification.

Classification thresholds ensuring efficiency of classification indicate a specific neurological condition in homogeneous fashion. These thresholds can be used for objectively classify patients with respect to their neurological status into relevant categories. However, each classification needs to be evaluated in terms of clinical meaningfulness.

## Equivalent scores

Different tests have different cut-off points. Hence, for two tests *A* and *B*, one needs to ensure that cut-off points $X_{0A}$ and $X_{0B}$ are equivalent. Chakrabartty,[35] suggested that if scores of *A* and *B* are transformed to follow normal distributions, then transformed score $T_{0A}$ corresponding to $X_{0A}$ and $T_{0B}$ corresponding to $X_{0B}$ are equivalent if

$$\int_{-\infty}^{T_{0A}} f(x)dx = \int_{-\infty}^{T_{0B}} g(y)dy \qquad (8)$$

Where $f(x)$ and $g(y)$ denote pdf of transformed scores of test *A* and test *B* respectively.  Equation (8) can be solved using Standard Normal probability table and can also be used to integrate different neuropsychological tests i.e. to find score combinations $\{X_{01}, Y_{Y_{02}}\}$ such that for a given score of $X_{01}$ in Test-1, equivalent score of $Y_{02}$ in Test-2 is

$$\int_{-\infty}^{X_{01}} f(x)dx = \int_{-\infty}^{Y_{02}} g(y)dy \qquad (9)$$

## Limitations

The study considered availability of complete responses from each respondent to neuropsychological tests. In practice, few respondents may give incomplete or non-responses to few items and elimination of entire data of those respondents reduce the sample size and statistical power. Understanding structure and characteristics of missing data in clinical datasets is important for selecting appropriate imputation methods. For neuropsychological data, multiple imputations are often preferred to handle data missing at random (MAR) compared to simple deletion or single imputation methods. Afkanpour et al.[36] provided a guideline of selection of the appropriate imputation methods in data pre-processing stages for clinical datasets.

## Conclusion

Proposed equidistant scores (*E*-scores) using data driven weights to response-categories of different items are in ratio scale with fixed zero point. *P*-scores obtained from *E*-scores via standardization and further linear transformation are continuous and normally distributed. Sub-class/dimension scores and test scores are obtained by adding item-wise *P*-scores facilitating aggregation with cardinal measures like number of errors, time taken, etc. and offer platform for parametric analysis including statistical testing.

In addition, the proposed method helps to find theoretically defined reliability, factorial validity avoiding criterion test, discriminating value of test, assessment of progress/deterioration of one or a sample of patients, efficiency of classification, equivalent scores of two neuropsychological tests, etc. Practicing psychiatrists and researchers can derive benefits of the proposed score in ratio scale for better comparisons and prognosis. Future investigations to evaluate merits and robustness of the proposed approach by simulation with wide range of datasets are suggested.

## Acknowledgements

None

## Funding

No funds, grants, or other support was received

## Conflicts of interest

The author declares that there are no conflicts of interest.

## References

1. Prigatano GP. Challenging dogma in neuropsychology and related disciplines. *Arch Clin Neuropsychol.* 2003;18(8):811–825.
2. Reschly D, Gresham FM. *Current neuropsychological diagnosis of learning problems: A leap of faith.* In: Reynolds CR, Fletcher-Janzen E, editors. Handbook of Clinical Child Neuropsychology. New York, NY: Plenum Press; 1989:503520.
3. Dean RS. Review of the Halstead- Reitan neuropsychological test battery. In: Mitchell JV, editor. The Ninth Mental Measurements Yearbook. Lincoln, NE: University of Nebraska Press; 1985.
4. Adams RL. Review of the Luria- Nebraska neuropsychological battery. In: Mitchell JV, editor. The Ninth Mental Measurements Yearbook. Lincoln, NE: University of Nebraska Press; 1985.
5. Reynolds C, Mason B. Measurement and statistical problems in neuropsychological assessment of children. In: [Book/Source Title not provided]. 2009.
6. Clason DL, Dormody TJ. Analyzing data measured by individual Likert-type items. *J Agric Educ*. 1994;35(4):31–35.

7. Harwell MR, Gatti GG. Rescaling ordinal data to interval data in educational research. *Rev Educ Res.* 2001;71:105–131.

8. Šimkovic M, Träuble B. Robustness of statistical methods when measure is affected by ceiling and/or floor effect. *PLoS One*. 2019;14(8).

9. Yusoff R, Janor RM. Generation of an interval metric scale to measure attitude. *SAGE Open.* 2014;1–16.

10. Lewin RJP, Thompson DR, Martin CR, et al. Validation of the cardiovascular limitations and symptoms profile (CLASP) in chronic stable angina. *J Cardiopulm Rehabil.* 2002;22(3):184–191.

11. Munshi J. A method for constructing Likert scales. *SSRN Electron J.* 2014.

12. Parkin D, Rice N, Devlin N. Statistical analysis of EQ-5D profiles: does the use value sets bias inferences? *Med Decis Making*. 2010;30(5):556–565.

13. Marcus-Roberts HM, Roberts FS. Meaningless statistics. *J Educ Stat*. 1987;12(4):383–394.

14. Kuzon WM, Urbanchek MG, McCabe S. The seven deadly sins of statistical analysis. *Ann Plast Surg.* 1996;37(3):265–272.

15. Campbell MJ, Machin D, Walters SJ. Medical Statistics: A Textbook for the Health Sciences. 4th ed. Chichester, England: John Wiley and Sons; 2010.

16. Lo CF. The sum and difference of two lognormal random variables. *J Appl Math*. 2012.

17. Alamri Y. Scoring neuropsychological tests: What corrections need to be considered? *Eur Neurol.* 2017;78(12-):84–85.

18. Bouter C, Hansen N, Timäus C, et al. Case report: The role of neuropsychological assessment and imaging biomarkers in the early diagnosis of Lewy body dementia in a patient with major depression and prolonged alcohol and benzodiazepine dependence. *Front Psychiatry.* 2020;11:684.

19. Theel ES, Aguero-Rosenfeld ME, Pritt B, et al. Limitations and confusing aspects of diagnostic testing for neurologic Lyme disease in the United States*. J Clin Microbiol*. 2019;57:e01406–e01418.

20. Chmielewski M, Watson D. What is being assessed and why it matters: The impact of transient error on trait research. *J Pers Soc Psychol*. 2009;97(1):186–202.

21. Streiner DL, Norman GR, Cairney J. Health Measurement Scales: A Practical Guide to Their Development and Use. Oxford, United Kingdom: Oxford University Press; 2014.

22. Altman DG, Bland JM. Statistics notes: Absence of evidence is not evidence of absence. *BMJ.* 1995;311(7003):485.

23. Sheng Y, Sheng Z. Is coefficient alpha robust to non-normal data? *Front Psychol.* 2012;3:34.

24. Flora DB, Curran PJ. An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychol Methods.* 2004;9(4):466–491.

25. De la Torre GG, Suárez-Llorens A, Caballero FJ, et al. Norms and reliability for the Spanish version of the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) Form A. *J Clin Exp Neuropsychol*. 2014;36(10):1023–1030.

26. Patt VM, Brown GG, Thomas ML, et al. Factor analysis of an expanded Halstead-Reitan battery and the structure of neurocognition. *Arch Clin Neuropsychol.* 2018;33(1):79–101.

27. McFarland D. Evaluation of multidimensional models of WAIS-IV subtest performance. *Clin Neuropsychol.* 2017;31(7):1127–1140.

28. Sherman E, Brooks BL, Iverson GL, et al. Reliability and validity in neuropsychology. In: [Book/Source Title not provided]. 2011.

29. Chakrabartty SN. Reliability of test battery. *Methodol Innov.* 2020;1-8.

30. Vaughan ED. *Statistics: tools for understanding data in the behavioral sciences*. Upper Saddle River, NJ: Prentice-Hall; 1998.

31. Chakrabartty SN. Improved quality of pain measurement. *Health Sci.* 2020;1:1–6.

32. Parkerson HA, Noel M, Gabrielle MP, et al. Factorial validity of the English-language version of the pain catastrophizing scale–child version. *J Pain.* 2013;14(11):1383–1389.

33. Goswami S, Chakrabarti A. Quartile clustering: a quartile based technique for generating meaningful clusters. *J Comput.* 2012;4(2):48–55.

34. Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell*. 1979;1(2):224–227.

35. Chakrabartty SN. Integration of various scales for measurement of insomnia. *Res Methods Med Health Sci*. 2021;2(3):102–111.

36. Afkanpour M, Hosseinzadeh E, Tabesh H. Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review. *BMC Medical Research, Methodology*. 2024:188.