WINSOME PUBLISHING LLC

New Healthcare Advancements and Explorations

OPEN ACCESS **RESEARCH ARTICLE**

Sound Measurement of Patients Reported Outcomes

Satyendra Nath Chakrabartty

Indian Ports Association, Indian Statistical Institute, India.

***Corresponding Author:** Satyendra Nath Chakrabartty, Indian Ports Association, Indian Statistical Institute, India.

Received date: **March 26, 2024:** **Accepted date:** **April 26, 2024:** **Published date:** **May 01, 2024**

Citation: Satyendra Nath Chakrabartty , Sound Measurement of Patients Reported Outcomes , 1(1). **New Healthcare Advancements and Explorations (NHAE) DOI: 10.1875 nhae.2024/001**

Abstract:

Patient-reported outcome scales (PROs) with different number of items containing K-point items, K= 2, 3, 4, 5, and different scoring systems suffer from methodological limitations.

The paper gives method to convert item scores to continuous, monotonic, equidistant scores following normal distribution satisfying desired properties and facilitating parametric analysis

Ordinal item-score are transformed to equidistant scores (E_i-scores) by assigning different weights to the levels of different items followed by standardization and further transformation to proposed item-scores (P_i-scores) irrespective of length of scale and width of items. Score of i-th dimension (D_i) is the sum of P_i-scores of items belonging to the dimension and scale score (P) is the sum of all D_i s. Normally distributed D_i-scores and P-scores facilitate meaningful comparison of patients and group of patients including assessment of progress or effectiveness of treatment plans, drawing of path of progress across time for better prognostication, testing statistical hypothesis of equality of mean using t-statistic for independent samples or using paired t-statistic for dependent samples e.g. pre-treatment and post-treatment to a group, finding equivalent scores of PRO-1 and PRO-2, so that area under normal curve up to P_(PRO-1)^0 = area under normal curve up to P_(PRO-2)^0. Methodological novelties include among others use of highest eigenvalue λ_1 to find factorial validity (FV) reflecting the main factor being measured by the questionnaire; maximum value of test reliability α_PCA; finding relationship between 〖FV〗_(Z-scores) and α_PCA and also relationship between r_(tt(theoretical)) and FV.

Key Words: Patient-reported Outcome Scale; Normal distribution; Progress path; Equivalent scores, Factorial validity, Reliability

Introduction

Recent advances in treatment, management and diagnosis of diseases have considerably improved health care by improving patient outcomes and quality of life (QoL) (Jordan & Tchantchaleishvili, 2021; Wang & Jang, 2022; Wang et al. 2022). Healthcare outcomes improvements require methodologically sound measures of outcomes. Need of rigorous assessment of outcomes for patient safety and treatment quality has been highlighted (MacGillivray, 2020). However, methodological issues in measurement of health outcomes have not been addressed adequately (Streiner and Norman, 2008). Major issues to be resolved include among others scoring system facilitating meaningful aggregation of items and dimensions, finding distribution of tests scores, parametric statistical analysis, etc. for better evaluation and differentiations among subjects and existing tools (Panagiotakos, 2009).

A number of tools are being used to assess pertinent outcome measures for diagnosis, therapeutic and rehabilitation approach (Okkersen et al. 2018). Outcome measures used in clinical set up could be (i) Patient-reported, using disease specific or generic questionnaires where score of an individual is taken as sum of item scores – in ordinal scale (Kyte et al. 2015), (ii) Performance-based, primarily for physiologic factors where patients perform a set of movements/tasks and scores are assigned either based on an objective measurement (like time to complete a task- in ratio scale) or a qualitative assessment (like normal or abnormal for a given task),(iii) Observer-reported, completed by parents, caregivers who regularly observes the patient on a daily basis and (iv) Clinician-reported, completed by a health care professional using clinical judgments and signs. For the same disease, different types of outcome measures may be used. For example, outcomes measures of non-curable Myotonic Dystrophy type 1 (DM1) include:

Operator dependent 5-point ordinal muscle impairment rating scale (MIRS) involving manual muscle testing (MMT) of 11 muscle groups for identification of stages and progression of DM1 and covers five different stages: MIRS-1 (no muscular impairment); MIRS-2 (myotonia, jaw and temporal wasting, facial weakness, neck flexors weakness, ptosis, nasal speech, no distal weakness except isolated digit flexor weakness); MIRS-3 (distal weakness, no proximal weakness except isolated elbow extensor weakness); MIRS- 4 (mild to moderate proximal weakness); MIRS-5 (severe proximal weakness) (Mathieu et al. 2001).

Performance based outcome measures in DM1 are: The Six-Minute Walk Test (6-MWT) (walking capacity over longer distances); The 10-meterWalk Test (10-mWT) (walking speed over a short distance); The 30-second chair-stand test (30-sCST) (lower limb strength and dynamic balance); The Nine-Hole Peg Test (9-HPT) (upper extremity function, specifically fine dexterity and coordination), etc. (Gagnon et al. 2015).

To evaluate characteristics of gait alterations in ambulant patients, Saggio et al. (2021) suggested two types of severity index viz. *SI-1* and *SI-2* based on plot of elapsed times in both plantar-flexion (PI) (negative- angles) and dorsi-flexion (DI) (positive – angles) in Y-axis within a "narrow" time interval in X-axis.

Results of such outcome measures vary. Thus, selection of appropriate outcome measures is critical for better understanding of current status and progress/decline, relapse or development of adverse reaction or a new disease entity (like infection) of patients over time (Hefford et al. 2011).

From the angle of measurement, outcome measure in ratio scale with fixed zero point facilitates systematic addition, subtraction, multiplication, division and undertaking parametric statistical analysis. However, outcome measures in ordinal scales containing K-point items (K= 2, 3, 4,5 ……) suffer from methodological limitations. For example, Likert scales assume that distance between two successive levels of an item is equal i.e. for a 7-point item, it assumes constant value of distance between j-th and (j+1)-th levels ∀ j =1, 2, 3, 4, 5, 6. Equal psychological distance between levels will provide exact measurements of the psychological trait being assessed (Wakita, et al. 2012). Arithmetic averages requiring equidistant scores are not meaningful for ordinal item scores (Jamieson, 2004) and (X)‾ > or <Y‾is meaningless (Hand, 1996). Non-meaningful addition imply computation of statistics like standard deviation (SD), coefficient of variation (CV), correlation, Cronbach α , etc. are not meaningful and analysis like regression, Principal component analysis (PCA), Factor analysis (FA), etc. with ordinal item scores may result in distorted results. Summative scores for dimensions and test giving equal importance to items and dimensions are un-justified due to different contributions of items or dimensions to total score, different values of inter-item correlations, item-total correlations and factor loadings of the items and dimensions (Parkin et al.2010). Addition of scores of independent dimensions may not be meaningful. Manual of the 36-Item Short Form Health Survey questionnaire (SF-36) does not support calculation of overall score (http://www.webcitation.org/6cfeefPkf) since several independent dimensions are being measured. Mean and SD tend to increase with increase in number of levels and may influence mean more than the underlying variable (Lim, 2008).

Distribution of scores of items, dimensions and test are different and skewed. For two variables X and Y, $X \pm Y = Z$ is meaningful if X and Y follow similar probability distribution and distribution of Z is known for further operations. Thus, it is necessary to know probability density function (pdf) of X and Y and their convolution.

PCA, FA, *t*-test, paired *t*-test, *F*-test, etc., assume normal distribution of the variables under study. Results may go wrong if assumptions of the techniques are violated. Outcome scores emerging from questionnaires do not satisfy the normality assumption.

High r_{xy} may not imply linearity between X and Y. Chakrabartty (2020) gave an example of $r_{X,X^2} > 0.9$ and $r_{X,X^3} > 0.9$ despite each of X^2, X^3 is non-linear function of X, due to non-satisfaction of assumptions of linear regression of Y on X where the error score $E = \left(Y - \hat{Y}\right)$ did not follow normal distribution. One possible solution to the above said problem areas is to transform item scores to follow similar distribution say Normal distribution.

The paper gives a multi-stage method to convert item scores to continuous, monotonic, equidistant scores followed by standardization and further linear transformation to ensure fixed score range from 1 to 100 and normality and scale score is taken as sum of all normally distributed item scores.

2. Literature survey:

Attempts have been made to transform K-point scales to L-point scales where K < or <L. For example, Grassi et al. (2007) transferred scores of Likert items of SF-36 to binary formats using Multiple Correspondence Analysis (MCA). However, MCA does not give a unique way to transform. Converting K-point scales to binary scores involves loss of information due to the reduction of response possibilities.

Scoring of PROs involve different methods to obtain dimension/scale scores. While dimension score of MacNew Heart Disease Health–Related Quality of Life Questionnaire (MacNew) is taken as arithmetic average of the responses in that dimension, Cardiovascular Limitations and Symptoms Profile (CLASP) scores are weighted to provide a total for each subscale. Each dimension of Myocardial Infarction Dimensional Assessment Scale (MIDAS) is scored separately. Such dimension scores create difficulties in meaningful computation of mean, SD, distribution of scale scores for meaningful comparisons, ranking, classifying individuals, and statistical inferences.

Cronbach alpha for test reliability assumes that each item measures the single latent trait on the same scale. PROs involving multiple factors violate the assumption and thus, Cronbach alpha may underestimate reliability of a PRO (Daniel, 1990). The coefficient alpha is influenced by variance sources, unknown-direction of sampling errors (Terry & Kelley, 2012), sample size (Charter, 1999).and even number of items (Luh, 2024). Moreover, Friedman's nonparametric tests cannot quantify interaction effects (Luepsen, 2017). Aligned Rank Transform (ART), a non-parametric factorial ANOVA analyzes the interaction and also the main effects, by aligning the data for each effect (main or interaction), followed by assignments of ranks. Alignment works best for completely randomized designs (King et. al. 2003).

List of PRO scales is too long. The Australian Commission on Safety and Quality in Health Care (2016) reviewed Patient-Reported Outcome measures (www.safetyandquality.gov.au).

PROs vary in terms of number of items (length) and number of levels (width) as can be seen from the illustrative scales for insomnia given below:

- Insomnia Severity Index (*ISI*): Consists of 7- number of 5-point items marked as 0 to 4 Individuals with score ≤14 are taken as Normal and those scoring > 14 are considered as having insomnia (Chahoud et al. 2017).

- Pittsburgh Sleep Quality Index (*PSQI*): Total 19-items, where first four items are open and each of the rest items is in 4-point scale from 0 to 3(Buysse et al. 1989). A score > 5 implies poor sleep quality and higher score implies worse sleep quality.

- Insomnia Symptom Questionnaire (*ISQ*): 13- Items. Items 1 – 5 are 6-point from 0 to 5 and Item 6 -13 are in 5-point scale from 0 to 4 (Okun et al. 2009).

Following major problem areas may please be noted:

- Each of ISI, PSQI, and ISQ generates ordinal scores and their distributions are unknown. Lack of meaningful addition of item scores to get dimension scores and scale scores fails to satisfy many desired properties.
- Different length and width of ISI, PSQI, and ISQ result in different contributions of dimensions covered by the scales. Mean, variance of PSQI with 19 items exceed the same of ISI and ISQ.
- Psychometric properties of multidimensional ISI, PSQI, and ISQ are different. Assumptions of Cronbach alpha are violated by scales measuring more than factor. Validity as correlation between a multidimensional scale score and criterion scores is the validity of which dimension /factor? Can we have validity of a scale for the main factor for which the scale was developed? Is it possible to have relationship between test reliability and test validity?
- Use of zero as an anchor value does not help to define expected values (value of the variable × probability of that value) of level-wise scores, unnecessarily reduces mean and variance of the scale, item-total correlations, regression or logistic regression may be inappropriate due to presence of many zeroes. If each respondent of a sub-group selects the level marked as "0" to an item then computation of between group variance will be difficult since mean = variance = 0 for the sub-group and correlation with that item is undefined. Stucki et al. (1995) found more than 40% of the patients scored zero in 10 subscales of Sickness Impact Profile (SIP) and in one subclass of SF-36. Better is to mark the anchor values as 1, 2, 3,….. and so on, keeping the convention of higher score ⇔ higher value of the variable being measured.
- Higher score in each of Nottingham Health Profile (NHP), Minnesota Living with Heart Failure (MLHF) indicate higher health problems, unlike Sickness impact Profile (SIP). Thus, directions of scores are different for different scales.
- Different PROs suggest different cut-off scores. For example, cut-off score of Stroke-Adapted Sickness Impact Profile (SA-SIP30) with 30 items covering 8 subscales is >33 and the same for Sickness Impact Profile (SIP136) with 136 "Yes–No" type items distributed over 12 domains is > 22. Question arises whether, a score of 33 in SA-SIP30 is equivalent to the score of 22 in SIP136? Similarly, score of 14 in ISI indicating "no insomnia" is equivalent to which score in PSQI or ISQ? Such questions highlight need of comparing the PROs with special emphasis on finding equivalent scores of two scales for the purposes of diagnosis and classification of individuals. Silva et al. (2014) observed that comparison of cut-off points of PROs or QoL questionnaires is not possible and suggested further investigations on different cut-off points for better comparisons. Based on treatment status for Cancer Core Questionnaire (EORTC QLQ-C30), four different cut-off scores were found (Lidington et al. 2022)

3. Suggested Remedial action:

3.1: Pre – adjustment of data:

i) Ensure that each item is positively related to intensity of the trait in question i.e. higher the item score, higher is the intensity of the disease or the dimension. For the variables like Platelet count, WBC count, % Myeloid cells in peripheral blood, etc. where lower value indicates higher risk to cancer, reciprocal of such variables are taken. For variable like Basophils, a type of white blood cell, a single value is given in the reference range instead of a range; an agreed particular value may be taken as the standard.

ii) Assign 1, 2, 3, 4, 5…to the levels or response-categories of items avoiding zero.

3.2 Converting ordinal score:

Let X_{ij} be the raw score of the i-th patient in the j-th item, for $i = 1,2, \dots .,n$ and $j = 1,2, \dots \dots ,m$. X_{ij} takes discrete value 1, 2, 3, 4 and 5 for a 5-point item. Let f_{ij} be the frequency of X_{ij}. Ordinal $X_{ij}s$ can be converted to normally distributed continuous, monotonic, equidistant scores by following stages:

Stage I. Equidistant scores:

For a 5-point item, find weights $W'_{ij}s$ for different values of i and j so that $W_{ij} > 0$; $\sum_{\mathrm{j}=1}^{5} W_{ij} = 1$. Equidistant property and monotonic condition will be satisfied if $W_1, 2W_2, 3W_3, 4W_4, 5W_5$ forms an arithmetic progression with a positive value of the common difference. Two ways to find such weights are as follows:

Method 1: Procedure for obtaining $W_j's$ of an item considering area under $N(0,1)$ is illustrated in Table 1

Table – 1: Calculation of weights based on area under N (0, 1)

Response Category	Proportion (p_i)	Cumulative Proportions (C_i)	Area under the standard Normal curve	Initial Weights
1	$p_1 = \frac{f_1}{mn}$	p_1	$A_1 = Upto\ p_1$	$\omega_1 = \frac{A_1}{\sum A_i}$
2	$p_2 = \frac{f_2}{mn}$	$p_1 + p_2$	$A_2 = Up\ to\ p_1 + p_2$	$\omega_2 = \frac{A_2}{\sum A_i}$
3	$p_3 = \frac{f_3}{mn}$	$p_1 + p_2 + p_3$	$A_3 = Upto\ p_1 + p_2 + p_3$	$\omega_3 = \frac{A_3}{\sum A_i}$
4	$p_4 = \frac{f_4}{mn}$	$p_1 + p_2 + p_3 + p_4$	$A_4 = Upto\ p_1 + p_2 + p_3 + p_4$	$\omega_4 = \frac{A_4}{\sum A_i}$
5	$p_5 = \frac{f_5}{mn}$	$p_1 + p_2 + p_3 + p_4 + p_5$=1.00	$A_5 = Upto\ p_1 + p_2 + p_3 + p_4 + p_5$	$\omega_5 = \frac{A_5}{\sum A_i}$
Total	1.00		$\sum_{i=1}^{5} A_i > 1$	1.00

Here, $\omega_j > \omega_{j-1}$ for j= 2,3,4,5 implying satisfaction of monotonic condition. To make the transformed scores equidistant for a 5-point scale, divide the difference between Maximum area and the Minimum area by 3 and call it the correction factor α. Determine the modified areas $\Delta_1, \Delta_2, \Delta_3, \Delta_4$ and Δ_5 as follows:

$\Delta_1 = A_1$(unchanged), $\Delta_2 = \Delta_1 + \alpha$; $\Delta_3 = \Delta_2 + \alpha$; $\Delta_4 = \Delta_3 + \alpha$; $\Delta_5 = \Delta_4 + \alpha$

Define corrected weights$W_j = \frac{\Delta_j}{\sum_{j=1}^{5} \Delta_j}$. Equidistant scores based on corrected weights are $E_i = \sum_{j=1}^{m} X_{ij} W_{ij}$ satisfying the monotonic condition.

Method 2: Find maximum $(f_{i\,max})$ and minimum frequency $(f_{i\,min})$ of the response-categories. Let initial weights $\omega_{ij} = \frac{f_{ij}}{n}$ and arrange the $\omega'_{ij}s$ so that

$\omega_{i1} = \frac{f_{i\,min}}{n} < \omega_{i2} < \omega_{i3} < \omega_{i4} < \omega_{i5} = \frac{f_{i\,max}}{n}$

Let intermediate weight $W_{i1} = \omega_{i1}$

Take common difference α as $\alpha = \frac{5f_{imax} - f_{i\,min}}{4n}$ since $W_{i1} + 4\alpha = 5W_{i5}$

Define other intermediate weights as:

$W_{i2} = \frac{1}{2}(\omega_{i1} + \alpha)$; $W_{i3} = \frac{1}{3}(\omega_{i1} + 2\alpha)$; $W_{i4} = \frac{1}{4}(\omega_{i1} + 3\alpha)$ and $W_{i5} = \frac{1}{5}(\omega_{i1} + 4\alpha)$.

Consider the final weights $W_{ij(Final)} = \frac{W_{ij}}{\sum_{j=1}^{5} W_j}$ enabling $\sum W_{ij(Final)} = 1$ and $kW_{ik(Final)} - (k-1)W_{i(k-1)(Final)}$= Constant, value of which may be different for different items.

Stage II. Standardization of E-scores: Standardize E-scores by $Z_{ij} = \frac{X_{ij} - \overline{X_i}}{SD(X_i)} \sim N(0,1)$.

Stage III. Transformation of Z-scores: Take further linear transformation of Z-scores to normally distributed proposed scores (P-scores) by:

$$P = \left[\frac{99(Z_{ij} - Min(Z_{ij})}{Max\,(Z_{ij}) - Min(Z_{ij})}\right] + 1 \qquad (1)$$

Parameters of the distribution of the i-th item, $P_i \sim N\,(\mu_i, \sigma_i^2)$ and $1 \le P_i \le 100$ can be estimated from the data. Item-wise P-scores as per (1) are applicable irrespective of length of scale and width of items. Thus, all items have same score range.

Dimension score is taken as sum of normally distributed P-score of relevant items contained in the dimension following normal with mean $\sum_i \mu_i$ and SD $= \sqrt{\sum \sigma_i^2 + 2\sum_{i \ne j} Cov(P_i, P_j})$ and the Scale score is sum of dimension scores (or item scores) each following normal.

Properties:

Continuous, equidistant and monotonic E-scores obtained by assigning different weights to the levels of different items by method-1 and method-2 are highly correlated. However, method-2 avoiding Standard Normal Table appears to be straightforward.

$f_{ij} = 0$ can be taken as zero value for scoring K-point items as weighted sum.

Equal importance to items and dimensions are avoided by item-wise *E*-scores and scale scores (*P*-scores). Normality ensures meaningful admissibility of arithmetic aggregation.

P-scores offer practically zero tied scores and thus, can better discriminate the respondents with tied raw scores and assign unique ranks to individuals and facilitate parametric analysis.

For items in ratio scales, transformation to *E*-scores are not required and can be standardized and transformed to follow normal distribution in the score range [1, 100].

3.3 Benefits of P-scores:

Benefits of proposed scores:

Dimension score (D_i) and proposed scale scores(P) are continuous, monotonic, normal and enable undertaking parametric analysis including estimation of population mean (μ),

population variance (σ^2), confidence interval of μ, testing statistical hypothesis like $H_0: \mu_1=\mu_2$ or $H_0: \sigma_1^2=\sigma_2^2$ etc. for snap-shot data and also for longitudinal data.

Evaluate progress of i-th patient in time-period (t) over the previous period by $\frac{P_{i(t)}-P_{i(t-1)}}{P_{i(t-1)}} \times 100$. Decline is indicated if $P_{i(t)}-P_{i(t-1)}<0$. For a group of patients, $\overline{P_{i(t)}} > \overline{P_{i(t-1)}}$ indicates progress. Normally distributed P_i satisfying assumptions of t-test, paired t-test helps to test $H_0: \mu_{P_t} = \mu_{P_{(t-1)}}$ and also $H_0: \text{Progress}_{(t+1)\text{over } t} = 0$, reflecting effectiveness of the treatment plans. Decline if any, may be probed to find the critical dimension(s) where $D_{i(t)}-D_{i(t-1)} <0$ and initiate appropriate corrective actions.

Graph depicting progress/decline of a patient or a sample of patients at various time points is analogous to hazard function and can be used to compare response to treatments from the start. Such trajectories can help to identify high-risk groups.

Effect of small change in D_i to scale score (P) can be expressed by percentage change of P due to small change in D_i i.e. elasticity indicating relative importance of the dimensions. The dimensions can be ranked in terms of elasticity.

For two scales X with normal pdf f(x) and Y with normal pdf g(y), one can find regression equation of the form $Y=\alpha_1+\beta_1 X$ to predict Y from X or $X=\alpha_2+\beta_2 Y$ to predict X from Y. However, the two regression lines differ and thus, empirical relationship between X and Y will not be unique. For a given value say x_0, better is to find equivalent score combinations (x_0, y_0) of two scales by solving the equation

$$\int_{-\infty}^{x_0} f(x)dx=\int_{-\infty}^{y_0} g(y)dy \qquad (2)$$

This avoids the problems of linear equating or percentile equating. The equation (2) can be solved using standard normal table (Chakrabartty, 2021). The method of finding equivalent score-combinations is possible even if the scales have different length, width and dimensions.

Normally distributed scores satisfy the assumptions of PCA, FA and enable finding Factorial Validity (FV) $= \lambda_1/\sum\lambda_i = \lambda_1/\sum S_{X_i}^2$ where λ_1 is the highest eigenvalue. FV reflects the main factor being measured by the questionnaire (Parkerson et al. 2013). Item validity can be computed as the correlation of the item with the principal component or item validity. Here, sum of item validities ≠ Scale validity. Eigenvalue ≈0 indicates existence of multicolinearity among the items. Test of significance of the largest eigenvalue can be done by Tracy–Widom (TW) test statistic $U = \lambda_1/\sum\lambda_i$ which follows a TW-distribution i.e. distribution of the normalized $\lambda_1$of a Hermitian matrix (Nadler, 2011). Such FV avoids the shortcomings of construct validity and selection of criterion scale with matching constructs and administration of the scale and also the criterion scale.

For standardized item scores, $FV_{(Z\text{-scores})}$ of a test is λ_1/m and the test variance S_X^2 can be written as

$$S_X^2=\sum\lambda_i + 2\sum_{i\neq j=1}^{m} Cov(X_i,X_j) = \lambda_1/FV+2\sum_{i\neq j=1}^{m} Cov(X_i,X_j) \qquad (3)$$

Thus, theoretical reliability

$$r_{tt(theoretical)} = \frac{S_T^2}{S_X^2} = \frac{S_T^2}{\lambda_1/FV+2\sum_{i\neq j=1}^{m} Cov(X_i,X_j)} \qquad (4)$$

Equation (4) gives non-linear relationship between $r_{tt(theoretical)}$ and factorial validity.

Maximum value of test reliability (α_{PCA})as a function of λ_1 derived from the correlation matrix of m-number of items was given by Ten Berge and Hofstee (1999) as

$$\alpha_{PCA}= \frac{m}{m-1}\left(1-\frac{1}{\lambda_1}\right) \qquad (5)$$

Relationship between FV and α_{PCA} as given in equation (5) is:

$$\alpha_{PCA}= \frac{m}{m-1}\left(1-\frac{1}{\lambda_1}\right) = \frac{m}{m-1}\left(1-\frac{1}{FV.\sum\lambda_i}\right) = \frac{m}{m-1}\left(1-\frac{1}{m.\,FV_{(Z\text{-scores})}}\right) \qquad (6)$$

As per (6), higher value of $FV_{(Z\text{-scores})}$ increases α_{PCA}

Normality helps to estimate variance of each item, dimension and questionnaire, enabling estimation of Cronbach alpha for a dimension at population level as

$$\hat{\alpha}=\frac{n}{n-1}\left(1-\frac{\text{Sum of estimates of variance of items in the dimension}}{\text{Estimate of variance of the dimension}}\right) \qquad (7)$$

Cronbach alpha of a battery consisting of K-dimensions can be obtained as a function of dimension reliabilities by

$$\hat{\alpha}_{Battery} = \frac{\sum_{i=1}^{K} r_{tt(i)} S_{Xi} + \sum_{i=1,i\neq j}^{K}\sum_{j=1}^{K} 2COV(X_i,X_j)}{\sum_{i=1}^{K} S_{Xi} + \sum_{i=1,i\neq j}^{K}\sum_{j=1}^{K} 2COV(X_i,X_j)} \qquad (8)$$

where $r_{tt(i)}$ and S_{xi} denote respectively reliability and SD of the i-th dimension.

Population estimates of dimension and battery by (7) and(8) respectively are simple and avoid

complex methods of Heo et al (2015) assuming parallel measures and involving estimation of unbiased sample covariance matrix; variance-covariance matrix of the population.

4. Discussion:

The proposed method of transforming ordinal item score to follow normal distribution ensures admissibility of the operation "addition". Sum of normally distributed scores of all items belonging to the i-th dimension is taken as the dimension score

(D_i) and scale score (P) is the sum of scores of all the dimensions (or equivalently the sum of scores of all the items). Each D_i and P follows normal even if the items differ in length and width. Normally distributed P-scores with data driven estimates of the parameters facilitate meaningful comparison of patients and group of patients including assessment of progress or effectiveness of treatment plans, drawing of path of progress across time for useful conclusions and better prognostication, testing statistical hypothesis H_0 : μ_1=μ_2 against H_1 : μ_1≠μ_2 using t-statistic for independent samples or using paired t-statistic for dependent samples e.g. pre-treatment and post-treatment to a group, finding equivalent scores of two PROs, finding equivalent score combinations (P_(PRO-1)^0,P_(PRO-2)^0) of two PRO scales (and or equivalent class-boundaries in case of classification of individuals by each of the two scales) can be found by ∫_(-∞)^(P_(PRO-1)^0)▒〖f(x)dx=∫_(-∞)^(P_(PRO-2)^0)▒g(y)dy〗 i.e.

area under normal curve corresponding to f(x) up to P_(PRO-1)^0 = area under normal curve corresponding to g(y) up toP_(PRO-2)^0. Such equivalent cut-off scores also satisfy

〖Var.of group〗_(Score ≥ P_(PRO-1)^0)/(Variance of PRO-1)=
〖Var.of group 〗_(Score ≥P_(PRO-2)^0)/(Variance ofPRO-2)
and can be used to evaluate efficiency of classification, say in terms of within group variance and between group variance.

Methodological novelties include among others use of highest eigenvalue λ_1 to find factorial validity (FV) reflecting the main factor being measured by the questionnaire; maximum value of test reliability α_PCA as a function of λ_1; finding relationship between 〖FV〗_(Z-scores) and α_PCA and also non-linear relationship between r_(tt(theoretical)) and FV. In addition, normally distributed D_i scores with estimated parameters help to find population estimate of Cronbach alpha for a dimension and Cronbach alpha of a battery consisting of K-dimensions.

Conclusions:

Methodologically sound approach given in the paper with wide application areas help significantly in evaluation of better assessment of outcomes and comparison of subjects and PROs along with measures of psychometric properties like reliability, validity, of tests and their relationships including derived relationship between reliability and validity, each as a function of largest eigenvalue. Such relationships may be used to find optimal value of one psychometric parameter to maximize another parameter. Future studies may explore such potentials with empirical investigations, extension of factorial validity to battery of tests and construction of psychometric quality index of test and battery, in addition to empirical verification of the properties of proposed methods using real life data.

Declarations:

Acknowledgement: Nil

Conflict of interests: Nil

Funding: Did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

Informed Consent: Not applicable. The paper did not collect data from human participants.

Data availability statement: The paper did not analyse or generate any datasets, because the work proceeds within a theoretical and mathematical approach

CRediT statement: Conceptualization; Methodology; Analysis; Writing and editing the paper by the Sole Author

References:

1. Buysse DJ, Reynolds CF, Monk TH, et al. (1989): The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. *Psychiatry Res*; 28: 193–213.

2. Chahoud M, Chahine R, Salameh P, Sauleau EA.(2017): Reliability, factor analysis and internal consistency calculation of the Insomnia Severity Index (ISI) in French and in English among Lebanese adolescents. *e-Neurological Sci*.18;7:9-14. doi: 10.1016/j.ensci.2017.03.003

3. Chakrabartty, Satyendra Nath (2021). Integration of various scales for Measurement of Insomnia. *Research Methods in Medicine & Health Sciences*, 2(3), 102-111, DOI: 10.1177/26320843211010044

4. Chakrabartty, Satyendra Nath (2020). Improved Quality of Pain Measurement. *Health Sciences*, Vol. 1, 1 -6, DOI: 10.15342/hs.2020.259

5. Charter, RA.(1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology*, 21(4), 559–566.https://doi.org/10.1076/jcen.21.4.559.889

6. *Daniel, Wayne W. (1990): Friedman two-way analysis of variance by ranks. Applied Nonparametric Statistics (2nd ed.). Boston: PWS-Kent. 262–274.* ISBN 978-0-534-91976-4.

7. Gagnon C, Heatwole C, Hébert LJ, Hogrel JY, Laberge L. Leone M, Meola G, Richer L, Sansone V and Kierkegaard M.(2018). Report of the Third Outcome Measures in Myotonic Dystrophy Type 1 (OMMYD-3). *J. Neuromuscul. Dis. 5*, 523–537.

8. Grassi, M., Nucera,A., Zanolin, e. et al.(2007): Performance Comparison of Likert and Binary Formats of SF-36 Version 1.6 Across ECRHS II Adults Populations, *Value in Health,* 10 (6), 478 – 488; https://doi.org/10.1111/j.1524-4733.2007.00203.x

9. Hand, D. J. (1996).Statistics and the Theory of Measurement, *J. R. Statist. Soc. A*; 159, Part 3, 445-492

10. Hefford C, Abbott JH, Baxter GD and Arnold R.(2011). Outcome measurement in clinical practice: practical and theoretical issues for health related quality of life (HRQOL) questionnaires. *Physical Therapy Reviews*;16(3):155-167.

11. Heo, M., Kim, N., & Faith, M. S. (2015). Statistical power as a function of Cronbach's alpha of instrument questionnaire items. *BMC Medical Research Methodology*, 15, Article 86. https://doi.org/10.1186/s12874-015-0070-6

12. Jamieson, S. (2004): Likert scales: How to (ab) use them. *Medical Education,* 38, 1212 -1218

13. Jordan, A., & Tchantchaleishvili, V. (2021). Recent progress in the field of Artificial Organs. *ARTIFICIAL ORGANS*, *45*(3), E38-E38.

14. King, G, Murray, C. J. L., Salomon, J. A., & Tandon, A. (2003): Enhancing the validity of Cross-cultural comparibility of measurement in survey research, *American Political Science Review,* 97, 567 – 583

15. Kyte DG, Calvert M, Vander Wees PJ, Ten Hove R, Tolan S, Hill JC. (2015). An introduction to patient-reported ontcome measures (PROMs) in physiotherapy. *Physiotherapy*; 101 (2); 119-125

16. Lidington E, Giesinger JM, Janssen SHM, Tang S, Beardsworth S, Darlington AS *et al.* (2022). Identifying health-related quality of life cut-off scores that indicate the need for supportive care in young adults with cancer. *Qual Life Res.* 31, 2717–27. DOI.10.1007/s11136-022-03139-6

17. Lim, H. E. (2008). The use of different happiness rating scales: bias and comparison problem? *Social Indicators Research,* 87, 259–267. https://doi.org/10.1007/s11205-007-9171-x.

18. Luepsen, Haiko (2017) The aligned rank transform and discrete variables: A warning, *Communications in Statistics - Simulation and Computation*, 46:9, 6923-6936, DOI: 10.1080/03610918.2016.1217014

19. Luh, Wei-Ming (2024). A General Framework for Planning the Number of Items/Subjects for Evaluating Cronbach's Alpha: Integration of Hypothesis Testing and Confidence Intervals. *Methodology*; Vol. 20(1), 1–21, https://doi.org/10.5964/meth.10449

20. MacGillivray TE. (2020). Advancing the Culture of Patient Safety and Quality Improvement. *Methodist Debakey Cardiovasc J.*;16(3):192-198. doi: 10.14797/mdcj-16-3-192.

21. Mathieu, J.; Boivin, H.; Meunier, D.; Gaudreault, M.; Bégin, P.(2001). Assessment of a Disease-Specific Muscular Impairment Rating Scale in Myotonic Dystrophy. *Neurology*, *56*, 336–340. Doi: 10.1212/WNL.56.3.336

22. Nadler, Boaz (2011): On the distribution of the ratio of the largest eigenvalue to the trace of a Wishart matrix. *Journal of Multivariate Analysis*, 102; 363-371

23. Okkersen, K.; Jimenez-Moreno, C.; Wenninger, S.; Daidj, F.; Glennon, J.; Cumming, S.; Littleford, R.; Monckton, D.G.; Lochmüller, H.; Catt, M.; et al. (2018). Cognitive Behavioural Therapy with Optional Graded Exercise Therapy in Patients with Severe Fatigue with Myotonic Dystrophy Type 1: A Multicentre, Single-Blind, Randomised Trial. *Lancet Neurol. 17*, 671–680.

24. Okun, M.L., Kravitz, H.M., Sowers, M.F., Moul, D.E., Buysse, D.J., & Hall, M.(2009). Psychometric evaluation of the Insomnia Symptom Questionnaire: A self-report measure to identify chronic insomnia. *Journal of Clinical Sleep Medicine*, 5(1), 41-51

25. Parkerson, H. A., Noel, M., Gabrielle M. P., Fuss, S., Katz, J., Gordon J. G. Asmundson (2013): Factorial Validity of the English-Language Version of the Pain Catastrophizing Scale–Child Version, *The Journal of Pain,*14(11),1383-1389. https://doi.org/10.1016/j.jpain.2013.06.004

26. Panagiotakos D.(2009). Health measurement scales: methodological issues. *Open Cardiovasc Med J.*;3:160-5. doi: 10.2174/1874192400903010160.

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

28. Saggio, G; Manoni, A; Errico, V; Frezza, E; Mazzetta, I; Rota, R; Massa, R; Irrera, F. (2021). Objective Assessment of Walking Impairments in Myotonic Dystrophy by Means of a Wearable Technology and a Novel Severity Index. *Electronics*, 10,708. https://doi.org/10.3390/electronics10060708

29. Streiner, D.L. and Norman, G.R. (2008). *Health measurement scales: A practical guide to their development and use*. 4th Edition, Oxford University Press, Oxford.

30. Stucki G, Liang MH, Phillips C, Katz JN. (1995): The Short Form-36 is preferable to the SIP as a generic health status measure in patients undergoing elective total hip arthroplasty. *Arthritis Care Res*.; 8(3):174-181.10.1002/art.1790080310

31. Ten Berge, J. M. F. & Hofstee, W. K. (1999): Coefficients alpha and reliabilities of unrotated and rotated components. *Psychometrika*, 64, 83–90. doi: 10.1007/BF02294321

32. Terry, L., & Kelley, K. (2012). Sample size planning for composite reliability coefficients: Accuracy in parameter estimation via narrow confidence intervals. *British Journal of Mathematical & Statistical Psychology*, 65(3), 371–401. https://doi.org/10.1111/j.2044-8317.2011.02030.x

33. Wakita, T., Ueshima, N. and Noguchi, H. (2012). Psychological Distance between Categories in the Likert Scale Comparing Different Numbers of Options. *Educational and Psychological Measurement,* 72, 533-546. https://doi.org/10.1177/0013164411431162

34. Wang Y, Jang YY.(2022). From Cells to Organs: The Present and Future of Regenerative Medicine. *Adv Exp Med Biol.*; 1376:135-149. doi: 10.1007/5584_2021_657.

35. Wang X, Li C, Wang Y, Chen H, Zhang X, Luo C, Zhou W, Li L, Teng L, Yu H, Wang J.(2022). Smart drug delivery systems for precise cancer therapy. *Acta Pharm Sin B;* 12(11):4098-4121. doi: 10.1016/j.apsb.2022.08.013. E