Recebido: 15-10-2024 | Aprovado: 12-01-2025 | DOI: https://doi.org/10.23882/rmd.25261

Normally distributed Health Related Quality of Life Measures

Medidas de qualidade de vida relacionadas à saúde normalmente distribuídas: Pontuações de QVRS normalmente distribuídas

Satyendra Nath Chakrabartty,
Indian Ports Association, Indian Statistical Institute
(chakrabarttysatyendra3139@gmail.com)

Abstract:
Background: Non-normal distributions of Health Related Quality of Life (HRQoL) measures violate basic assumption of parametric statistical analysis. Testing normality of the data is a pre-requisite for selection of statistical tests and techniques. Different methods of testing normality give contrasting results.
Objective: Avoiding the problems of discrepancies of tests of normality, the paper describes methods of transforming ordinal item scores of HRQoL tools to equidistant scores facilitating meaningful addition and further linear transform to proposed scores which can be added to get dimension scores and test scores, each following normal distribution, parameters of which can be estimated from the data.
Results: Distribution of *HRQoL* scores as convolution of normally distributed item scores facilitates meaningful arithmetic aggregation and provides platform to perform parametric analysis with desired properties like plotting of progress/decline of HRQoL across time, statistical test of hypothesis, identification of critical indicators, finding equivalent scores of two or more HRQoL tools, etc.
Conclusions: The proposed methods of *HRQoL* scores with normality and wide application areas with better measures of reliability, validity in terms of largest eigenvalue is recommended.

Keywords: HRQoL; Arithmetic aggregation; Normal distribution; Equivalent scores; Progress path; Factorial validity

Resumo:
Antecedentes: Distribuições não normais de medidas de Qualidade de Vida Relacionada à Saúde (QVRS) violam pressupostos básicos da análise estatística paramétrica. Testar a normalidade dos dados é um pré-requisito para a seleção de testes e técnicas estatísticas. Diferentes métodos de teste de normalidade fornecem resultados contrastantes.

Objetivo: Evitando os problemas de discrepâncias de testes de normalidade, o artigo descreve métodos para transformar pontuações de itens ordinais de ferramentas de QVRS em pontuações equidistantes, facilitando a adição significativa e posterior transformação linear às pontuações propostas que podem ser adicionadas para obter pontuações de dimensão e pontuações de testes, cada uma seguindo distribuição normal, cujos parâmetros podem ser estimados a partir dos dados.
Resultados: A distribuição das pontuações de QVRS como convolução de pontuações de itens normalmente distribuídas facilita a agregação aritmética significativa e fornece plataforma para realizar análises paramétricas com propriedades desejadas, como plotagem de progresso/declínio de QVRS ao longo do tempo, teste estatístico de hipótese, identificação de indicadores críticos, encontrar equivalentes pontuações de duas ou mais ferramentas de QVRS, etc.
Conclusões: Os métodos propostos de escores de QVRS com normalidade e amplas áreas de aplicação com melhores medidas de confiabilidade, recomendam validade em termos de maior autovalor.

Palavras-chave: QVRS; Agregação aritmética; Distribuição normal; Pontuações equivalentes; Caminho de progresso; Validade fatorial

Introduction:

Health Related Quality of Life (HRQoL) measures are increasingly being used in clinical trials as primary endpoints. For designing a study to compare the outcomes of an intervention, an important step is to find the sample sizes to allow a reasonable chance of detecting a pre-determined difference (effect size) in the outcome variable, depending on the outcome measure and probability distribution of aggregated item scores, based on which the test statistic is chosen (Machin et al. 2011). Most desired distribution of HRQoL measure is normal distribution which is the basic requirement of many parametric statistical analyses like regression, *t*-tests, F-test, analysis of variance, and techniques like Principal component analysis (PCA), Factor analysis (FA), Analysis of variance (ANOVA), etc. Computation of regression coefficients of equations of the form $Y = \alpha_1 + \beta_1 X + \epsilon_{YX}$ or $X = \alpha_2 + \beta_2 Y + \epsilon_{XY}$ do not require normal distribution but ϵ_{XY} or ϵ_{YX} must follow normal with mean = 0 and constant variance (homoscedasticity). If the data are non-normally distributed, such techniques cannot be undertaken. Thus, testing normality of the data is a pre-requisite for selection of statistical tests and techniques (Mishra et al. 2019). Different methods of testing normality of data have their advantages and disadvantages.

Commonly used tests of normality include Shapiro-Wilk (SW), Kolmogorov-Smirnov (KS), Lilliefors (LF), Anderson-Darling (AD) tests, Jarque–Bera (JB) test, etc. The tests differ in terms of test statistics and generated power (1- Prob. of type II error). For example, null hypothesis of Shapiro-Wilk test is H_0: The sample came from a population following $N(\mu, \sigma^2)$ and test statistic is $W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{Sample\ Variance}$ where $x_{(i)}$ denotes the i-th ordered sample value and the coefficients a_i are obtained as $(a_1, a_2, \dots\dots, a_n) = \frac{m^T V^{-1}}{\|V^{-1} m\|}$ with $m = (m_1, m_2, \dots\dots, m_n)^T$ is the expected values of the ordered statistics following $N(0,1)$ and V is the covariance matrix of the order statistics. Instead of distribution of W, the cutoff values for W are calculated through Monte Carlo simulations.

The nonparametric Kolmogorov-Smirnov test of normality is used for sample size ≥ 50 to test to test H_0: the set of data comes from a Normal distribution. Test statistic (D) for a one-sample Kolmogorov-Smirnov test is the maximum vertical deviation between the empirical distribution function (EDF) of the sample and the cumulative distribution function (CDF) of the reference distribution.

Lilliefors (LF) test of normality finds the maximum discrepancy between the EDF and CDF of the normal distribution with the estimated values of mean and variance. Here, the "null distribution" of the test statistic is stochastically smaller than the Kolmogorov–Smirnov distribution. Tables of the Lilliefors distribution are computed by Monte Carlo methods.

Anderson-Darling tests of normality are used to test if a sample data came from a population with normal distribution. It assigns more weight to the tails than does the K-S test and requires computation of critical values.

Jarque-Bera tests whether the skewness and kurtosis of the sample data are matching with a normal distribution. The test statistic is a function of estimated value of second, third and fourth central moments from the data. Under H_0 the JB statistic asymptotically has a chi-squared distribution with two degrees of freedom.

Razali et al. (2012) compared four tests of normality viz. SW, KS, LF and AD test based on HRQoL data for small, moderate and large samples through three statistical packages: SPSS, SAS (using Fisher's definition of skewness and kurtosis), and MINITAB (using

sample skewness and kurtosis defined by Pearson) and found SW test worked best in all large, moderate and small sample sizes; AD test are comparable to results of SW test most of the time. However, for small samples, KS test did not perform very well. Using SF-36 data, Walters (2004) considered different methods with respect to estimated sample size and power, to compare effectiveness of two treatment plans and concluded that results are applicable only to the data considered and cannot be generalized for other HRQoL outcomes.

Avoiding the problems of discrepancies of tests of normality, the paper describes methods of transforming ordinal item scores of HRQoL tools to equidistant scores facilitating meaningful addition and further linear transform to proposed scores which can be added to get dimension scores and test scores, each following normal distribution, parameters of which can be estimated from the data. Benefits of normally distributed scores for better evaluation of ranks, responsiveness, and psychometric properties are also addressed.

HRQoL instruments:

Popular HRQoL measuring tools like SF-36, Nottingham Health Profile (NHP), European Organisation for Research and Treatment of Cancer (EORTC), QLQ-C30, Psychological General Well-Being Index (PGWB), etc. are multidimensional covering health dimensions (like mobility, ability to perform certain activities, emotional state, sensory function, and cognition), social functions, pain dimensions, psychological state (anxiety, depression, emotional reaction, sleep, etc.) and are measured through responses given by the patients. Patient reported scales are popular despite empirical evidence of no correlation of endometriosis with subjective complaints (Chmaj-Wierzchowska et al. 2020).

The 36-Item Short Form Health Survey questionnaire (*SF-36*) is a patient-reported generic questionnaire consisting of 28 number of K-point items (K=3, 5, 6), seven binary items and another item regarding reported health transition over the last year. The items are distributed over eight sub-scales as follows:

- Physical functioning: 10 items (3-point).
- Energy/ Fatigue: 3 items (6-point)
- Emotional well-being: 6 items (6-point)

- Social functioning: 2 items (5-point)
- Pain: 1 item (6-point) and another item (5-point)
- General health: 5 items (5-point)
- Role limitations due to physical health: 4 items (Yes – No type)
- Role limitations due to emotional problems: 3 items (Yes – No type)

The *SF-36* scoring manual does not support calculation of SF_{Total} since several independent dimensions are being measured by the scale (http://www.webcitation.org/6cfeefPkf). Mean, Standard deviation (SD), reliability, validity is calculated separately for each sub-scale. Mean, SD are more for higher-point items. Reliability, validity, are different for *K*-point scales for *K*= 3, 5, 6, and so on (Preston and Colman, 2000; Chakrabartty, 2023)

Nottingham Health Profile(NHP) contains 38 items of "Yes – No" type and covers six domains viz. physical abilities, pain, sleep, social isolation, emotional reactions, and energy level (Hunt, *et al*. 1985). The optional Part II contains seven items reflecting how health problems affect occupation, jobs around the house, personal relationships, social life, sex life, hobbies, and holidays. NHP items are scored as 0 to a 'no' response and 1 to a 'yes' response. Scores for each of the sections range between 0 (worst health) and 100 (best health). NHP score is taken as the mean of the domain scores. NHP fails to detect "milder forms of distress" and is difficult to compare the general population and detect change (Hunt et al. 1985). Lack of knowledge of distributions of domain scores makes it difficult to compare the domains or to test significance of changes in pre- and post-intervention studies. Improvements for those with zero score in pre-administration cannot be evaluated, as zero scores may not indicate total absence of distress. While comparing SF-36 and NHP, Wann-Hansson et al. (2004) observed that the two tools showed conflicting results in patients with chronic lower limb ischaemia in a longitudinal perspective. SF-36 with better psychometric properties was more suitable for patients with intermittent claudication and NHP had higher discriminating power among severity of ischaemia and was more responsive in patients with critical ischaemia. Comparison of health-status measures by NHP, SF-36, and two other scales found that no scale performed uniformly as "best" or "worst" (Essink-Bot, 1997).

The specific *EORTC QLG Core Questionnaire (EORTC QLQ-C30*) is a 30-item instrument designed to measure physical, psychological and social functions of cancer patients. It is composed of 5-functional scales (cognitive, emotional, physical, role, and

social functioning), 3-symptom scales (fatigue, nausea/vomiting, and pain), a global health status scale and 5 single items assessing additional symptoms (appetite loss, constipation, diarrhea, dyspnea, and sleep disturbance) and perceived financial impact. Factor analysis resulted in six independent factors explaining 76.85% of the total variance (Mystakidou et al. 2001). The QLQ-C30 summary score is taken as the mean of the combined 13 QLQ-C30 scale and item scores (excluding global QoL and financial impact), where higher score indicates better HRQoL (Kasper, 2020). However, prognostic value of such summary score is yet to be confirmed empirically. Lidington et al. (2022) found four cut-off scores of EORTC QLQ-C30 for different treatment status.

With 22 number of 6-point self-administered items (0 to 5), *Psychological General Well-Being Index (PGWBI)* assess the psychological and general well-being of subjects in six HRQoL domains like anxiety, depression, positive well-being, self-control, general health, and vitality. It enables computation of a single measure of psychological well-being as arithmetic average of unweighted responses to individual items of all domains where reversed scoring are done for items 1, 4, 6, 7, 9, 10, 14, 16, and 21 (http://www.opapc.com/uploads/documents/PGWBI.pdf). Against the envisaged six domains, exploratory factor analysis indicated a two-factor solution with mediocre fit (RMSEA 0.095) (Lundgren-Nilsson et al. 2013).

Carotenuto et al. (2013) used PGWBI to 162 male seafarers on board of 7 tankers and found no significant differences with respect to general index of well-being (GWBI) against significantly higher anxiety levels for engine officers than the deck or engine crew. In other words, comparisons with total scores and component scores were different.

Observations:

HRQoL tools differ in dimensions, number and format of items, scoring methods and score ranges. Different HRQoLs in same sample may give different conclusions and thus, HRQoLs are not comparable.

HRQoL tools are measured on an ordered categorical scale with different number of response-categories (levels). Such ordinal item scores with unknown distributions do not lead to meaningful arithmetic aggregation to obtain dimension scores and scale scores

facilitating knowledge of distribution of scale scores. For two random variables *X* and *Y*, X + *Y* = *Z* demands to find $P(Z = z) = P$ (X= x, Y= z-x) for discrete case and$P(Z \leq z) = P\ (X + Y \leq z) = \int_{-\infty}^{\infty}(\int_{-\infty}^{z} f_{X,Y}(x, t-x)dt)\ dx$ for continuous case. Thus, it is necessary to know probability density function (pdf) of each item scores and convolution of summated dimension scores and scale scores as sum of item scores.

Mean, SD of rating data are not meaningful (Gail and Artino, 2013; Reeves et al. 2020) due to non-satisfaction of equidistant property (Bastien *et al*. 2001). Distance between successive response-categories is not uniform and unknown (Munshi, 2014). The equidistant property demands constant distance between two successive response-categories. Unknown and different distributions of item scores and resultant dimension/test scores make it difficult to interpret $X \pm Y$ and to find joint distribution of $X \pm Y$ of the random variables being added and their convolution.

Discrete ordinal data are not normally distributed; violets assumptions of many statistical procedures (Harwell et al. 2001) and parametric statistical analysis are problematic. A scale needs to have features like: metric, presence of zero point, and clearly defined operational procedure as the basis for measurement (Yusoff and Janor, 2014):

Response-categories like *very often, often, once in a while, almost never* and *never* could be misleading as individuals differ in perception on how frequently an action is to occur to consider it as often. Pertinent question is "How often is *often"*? (Gu, *et al.* 1995)

Subjective responses endorsed by patients may be different from the true situations. Edinger *et al.* (2000) observed that disturbed sleep reported by few subjects showed normal sleep-patterns when monitored objectively. Individuals also differ with respect to their subjective views on physical, emotional and social functions.

Summative scores assigning equal importance to the items and dimensions may not be justified due to different values of correlations of item/dimension scores with total score and different factor loadings.

Use of "Zero" as an anchor value (e.g. PGWBI, NHI) may distort the distribution of scale scores. Frequent zero responses to an item unnecessarily lowers mean, SD, correlation with that item and does not enable computation of expected values of level-wise score.

Better could be to assign numbers 1, 2. 3. 4. 5 etc. avoiding zero to the levels. Decomposition of Likert scores using multipoles for reduced heterogeneity of responses of the respondents or raters was proposed (Lipovetsky and Conklin, 2018), where anchor values were changed suitably before calculation of multipoles moments. Nature of generated data remains invariant if the anchor values are replaced by linear transformation of such numbers.

Summative rating scores do not consider patterns of getting a particular score. Different responses to different items can generate tied test score for several persons. Thus, the scale fails to discriminate the respondents with tied score.

Different values of *K* distorts shape of distribution of scores and influence item/test parameters like Reliability, validity, more by number of levels than the underlying variable (Lim, 2008). Wakita *et al*. (2012) administered 4, 5, and 7-point scales of the same items and found that number of options influenced the psychological distance between options, particularly for the 7-point scale. Studies to find optimum number of response-categories considering maximum reliability and/or validity produced contrasting results.

Proposed method:

Pre-processing of data:

(i) Consider levels of each item as 1, 2… K. For example, items with levels 0 to 5 to be taken as 1 to 6.

(ii) Ensure uniformity in direction of each item i.e. higher score of an item implies higher score in the dimension containing the item.

Proposed scores:

Proposed scores following normal distribution can be found by transforming raw item scores to equidistant scores followed by Z-transformation and further linear transformation.

Stage I: Equidistant scores:

Suppose a HRQoL scale has been administered once to a sample of size *n*. Equidistant scores (E-scores) may be obtained by following any of the method given below:

Method-1. Different weights based on frequency of different levels of different items

Find frequency f_{ij} of the *j*-th level of the *i*-th item. For each item, find maximum (f_{max}) and minimum frequency (f_{min}). Find proportions: $\omega_{ij} = \frac{f_{ij}}{n}$ for *j*=1, 2, ……K

For the *i*-th item, put initial weights$W_{i1} = \omega_{i1} = \frac{f_{min}}{n}$.

Find the common difference $\alpha = \frac{Kf_{max} - f_{min}}{(K-1)n}$. For K=5, $\alpha = \frac{5f_{max} - f_{min}}{4n}$.

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

Here, $W_{ij} > 0$ and $\sum_{j=1}^{5} W_j \neq 1$.

Get finally selected weights $W_{ij(Final)} = \frac{W_{ij}}{\sum_{j=1}^{5} W_j}$ so that $\sum W_{ij(Final)} = 1$

Method-2. Based on area under$N(0, 1)$:

For the *i*-th item, find proportions $p_{ij} = \frac{f_{ij}}{n}$ and cumulative frequency C_i

Find area (A_i) under the standard Normal curve for each C_i

Take initial weights as $\omega_{ij} = \frac{A_i}{\sum A_i}$. Here, $\omega_{ij} > 0$ and $\sum_{j=1}^{5} \omega_{ij} = 1$

Find correction factor $\beta = \frac{Max\ area - Min\ area}{3}$ since 3-levels are contributing to the numerator.

Consider modified areas as $\nabla_1 = A_1$(unchanged),

$\nabla_2 = \frac{\nabla_1 + \beta}{2}$, $\nabla_3 = \frac{\nabla_1 + 2\beta}{3}$, $\nabla_4 = \frac{\nabla_1 + 3\beta}{4}$ and $\nabla_5 = \frac{\nabla_1 + 4\beta}{5}$

Get finally selected weights $W_{ij(Final)} = \frac{\nabla_j}{\sum_{j=1}^{5} \nabla_j}$ so that $\sum W_{ij(Final)} = 1$

Each of the two methods considers initial weights as empirical probabilities based on the frequencies of Item – Response categories. Thus, item and individual scores are obtained as expected values and generate cardinal data. In each method, $5W_{i5} - 4W_{i4} = 4W_{i4} - 3W_{i3} = 3W_{i3} - 2W_{i2} = 2W_{i2} - W_{i1} = C > 0$, where value of *C* is different for different items. This ensures satisfaction of equidistant property (to facilitate addition) with zero ties (to distinguish the respondents with same raw score) and fixed zero point (when $f_{ij} = 0$ for a particular level of an item).

Empirical illustrations (hypothetical data):

The hypothetical data consist of responses from 463 respondents to a questionnaire consisting of 30 number of 5-point items. Steps for calculation of weights by each of the two methods are given below.

Table 1 - Calculation of weights to different levels of different Items

Item	Description	Level-1	Level-2	Level-3	Level-4	Level-5	Total
Method-1							
1	Frequency	27	32	10	18	14	101
	Proportions (ω_{1j})	0.26733	0.31683	0.09901	0.17822	0.13861	1.00
	Intermediate weights(W_{1j}) (α =0.37129)	0.26733	0.31931	0.33663	0.34530	0.35049	1.6196
	Final weights ($W_{1j(Final)}$)	0.16511	0.19722	0.20792	0.21327	0.21648	1.00
	Y-scores for Item 1 *(Raw scores)*	0.16511 *(1)*	0.39444 *(2)*	0.62376 *(3)*	0.85308 *(4)*	0.21648 *(5)*	
2	Frequency	5	12	11	31	42	101
	Proportions(ω_{2j})	0.04950	0.11881	0.10891	0.30693	0.41584	
	Intermediate weights(W_{2j}) (α =0.50743)	0.04950	0.27846	0.35478	0.39295	0.41584	1.4913
	Final weights ($W_{2j(Final)}$)	0.03319	0.18670	0.23786	0.26345	0.2788	1.00
	Y-scores for Item 2 *(Raw scores)*	0. 03319 *(1)*	0.3734 *(2)*	0.71358 *(3)*	1.0538 *(4)*	1.394 *(5)*	
Method-2							
1	Cumulative Proportions	0.26733	0.58416	0.68317	0.86139	1.00	

	Area under N(0,1)	0.6064	0.7190	0.7517	0.8051	0.8413	
	Modified area(Δ_1) (β =0.0783)	0.6064	0.34235	0.25433	0.21032	0.18392	1.59733
	Final weights ($W_{1j(Final)}$)	0.37963	0.21433	0.15922	0.13167	0.11514	1.00
	Y-scores for Item 1 *(Raw scores)*	0. 37963 *(1)*	0.42866 *(2)*	0. 47766 *(3)*	0.52668 *(4)*	0. 5757 *(5)*	
2	Cumulative Proportions	0.04950	0.16831	0.27723	0.58416	1.00	
	Area under N(0,1)	0.5199	0.5675	0.6103	0.7190	0.8413	
	Modified area(Δ_2) (β =0.507426)	0.5199	0.31352	0.24472	0.21032	0.18969	1.47815
	Final weights ($W_{2j(Final)}$)	0.35172	0.21210	0.16556	0.14229	0.12833	1.00
	Y-scores for Item 2 *(Raw scores)*	0. 35172 *(1)*	0.4242 *(2)*	0. 49668 *(3)*	0.56916 *(4)*	0. 64165 *(5)*	

Stage II: Normalization:

Normalize equidistant scores by $Z_{ij} = \frac{E_{ij} - Mean\ (E_i)}{SD(E_i)} \sim N(0,1)$

Stage III: Further Transformation:

To avoid negative scores, transform Z_{ij} to Y by linear transformation:

$$Y = (99)[\frac{Z_{ij} - Min(Z_{ij})}{Max\ (Z_{ij}) - Min(Z_{ij})}] + 1 \sim N(\mu, \sigma^2) \quad (1)$$

Here, $1 \leq Y \leq 100$ ensures uniformity in item score–range. The parameters μ and σ^2 can be estimated from the data

Stage IV: Define dimension score as arithmetic aggregation $Dim_i = \sum Y_i$ where summation is taken overall all items belonging to the dimension and HRQoL score is sum of all dimension scores or $HRQol = \sum_{All\ items} Y_i$

Normally distributed item scores can be added to get sub-scale scores and scale scores. Clearly,

For an individual, *Scale score* = $\sum Sub - scale\ scores = \sum Item\ scores$

Correlation of *Y*-scores by Method 1 and Method 2 was found to be 0.986 implying similar clusters of individuals taking the test.

Descriptive statistics:

Table 2 Descriptive statistics for various Methods

Description	Summative scores (Raw scores)	Method – 1 (Based on frequency of different levels of different items)	Method - 2 (Based on area under N(0,1)
Test Mean	90.48	21.46	21.74
Test Variance	63.41	6.62	7.53
Range of Item Variance	Max:2.39 Min: 0.74	Max: 0.25 Min:0.08	Max: 0.29 Min:0.09

Observations: Method 1 and 2 reduced test average and test variance significantly. Range of item variances also got reduced in Method 1 and 2.

Regression equations of *Y*-scores by each method on summative raw score (X_0) are shown below:

$$Y_1 = 0.5660(X_0) - 10.29 \text{ , corresponding } R^2 = 0.9623$$

$$Y_2 = 0.34(X_0) - 9.023 \text{ , corresponding } R^2 = 0.9731$$

High value of R^2 indicate goodness of fit of the data to the linear model.

Properties:

- Distribution of *HRQoL* scores is the convolution of normally distributed item scores and provides platform to perform parametric analysis. *HRQoL* score can be computed even if different items have different number of levels.
- Methods to find *E*-scores start with f_{ij}s and are highly correlated. However, Method-2 involving standard normal probability table is likely to have lower variance. Method-1 appears to be more straightforward.
- Linear transformations of *E*-scores following normal distribution facilitate meaningful arithmetic aggregation to get *HRQoL* scores which are monotonically

increasing continuous variable, facilitating better ranking and classification of individuals, and finding relative importance of j-th dimension by $\frac{\nabla(HRQoL)}{\nabla Dim_j}$

- Improvement of *HRQoL* in successive periods can be reflected by: $(HRQoL_t - HRQoL_{(t-1)}) > 0$ or by $\frac{HRQoL_t}{HRQoL_{(t-1)}} > 1$. Percentage improvement is given by $\frac{HRQoL_t - HRQoL_{(t-1)}}{HRQoL_{(t-1)}} * 100$. Progress/decline path of *HRQoL* can be plotted considering percentage improvement against time. Such progress/decline in pre- and post- administration of HRQoL scale reflects effectiveness of interventions or treatment plans.
- i-th dimension is critical if $\frac{Dim_{i_t}}{Dim_{t_{(t-1)}}} < 1$. Critical dimension(s) merit managerial attention for necessary corrective action.
- Normally distributed *HRQoL* scores satisfying the basic assumption of statistical techniques helps to estimate population parameters and test H_0: $(HRQoL_t - HRQoL_{(t-1)}) = 0$. Similarly, for two HRQoL scales, one can test H_0: $\mu_{HRQoL_1} = \mu_{HRQoL_2}$ or H_0: $\sigma_{HRQoL_1} = \sigma_{HRQoL_2}$ by t-test and F-test respectively
- Normality also helps to find equivalent scores (X_0, Y_0) for two HRQoL-scales such that

$$\int_{-\infty}^{X_0} f(x)dx = \int_{-\infty}^{Y_0} g(y)dy \quad (2)$$

where $f(X)$ and $g(Y)$ denote normal probability density function of $HRQL_1$ and $HRQoL_2$ i.e. area under $f(X)$up to X_0= area under $g(Y)$ up to Y_0. For a given value of X_0 (or Y_0) equation (2) can be solved using Standard Normal Probability Table (Chakrabartty, 2021). It is possible to find all equivalent combinations $\{X_0, Y_0\}$ including cut-off scores of two scales for better comparison of HRQoL scales.

- Floor or ceiling effects are taken in arbitrary fashion. For SF-36, Busija et al. (2008) considered that a sub-scale has floor or ceiling effects if at least 15% of respondents reported the worst (0) or best (100) possible scores, respectively. Normal distribution of scores may help to consider responses lying outside $Mean \pm 3SD$ as effect of floor or ceiling.

- Purpose of Minimal Detectable Change (MDC) is to assess changes which exceed the measurement error by 1.96 *(SEM) presumes normality (Ware et al.2005).
- Proposed *HRQoL* scores may be used in classification of the subjects in four mutually exclusive classes viz. the quartiles Q_1, Q_2, Q_3, Q_4 (Goswami & Chakrabarti, 2012) and assigning equal probability to each quartile/class i.e. $\int_0^{Q_1} f(x)dx = \int_{Q_1}^{Q_2} f(x)dx = \int_{Q_2}^{Q_3} f(x)dx = \int_{Q_3}^{Q_4} f(x)dx$ where $f(x)$ denotes the pdf of normally distributed *HRQoL* scores.
- Reported reliability of HRQoL scales are specific to sample. Normally distributed scores, helps to have population estimate of scale variance and variance for each item to obtain population estimate of Cronbach alpha and also enable to find λ_1 the first principal component with highest eigenvalue reflecting the main factor for which the scale was developed. Based on λ_1 and other eigenvalues (λ_is), one can compute Cronbach alpha in terms of λ_1 (α_{PCA}) (Ten Berge and Hofstee, 1999) and factorial validity as $\frac{\lambda_1}{\sum \lambda_i}$. Such factorial validity avoids the problems of construct validity and selection of criterion scale (Parkerson, et al. 2013).

Applications:

The method can accommodate all dimensions and items of different formats irrespective of their inter-correlations and facilitates computation of *HRQoL* scores for properly defined different sub-groups say rural or urban groups, economically backward groups, educated or uneducated groups, etc. without undertaking tests of normality. Effect of *HRQoL* on health, disability, insomnia, well-being, etc. can be investigated by finding empirical relationship of *HRQoL* scores with the contrast of interest. The proposed method is well applicable to all social science areas using rating scales with better measures of reliability, validity.

Limitations:

Did not consider missing data and stability of HRQoL on deletion of highly or poorly correlated items/dimensions, which are proposed as future studies.

Discussions:

The paper describes measurement of *HRQoL* by arithmetic aggregation of normally distributed item scores where *HRQoL* scores also follow normal distribution. This avoids the problems of testing normality. The proposed method avoiding scaling and satisfying desired properties like meaningful summative scores, plotting of progress/decline on year-to year basis, statistical test of hypothesis, identification of critical indicators, etc. is preferred based on theoretical advantages.

Methodological novelties include among others: Factorial validity of a test and Cronbach's alpha in terms of λ_1 (α_{PCA}).

Conclusions:

The proposed methods of *HRQoL* scores with normality and wide applications help enables parametric statistical analysis along with better measures of reliability, validity, and their relationships as a function of largest eigenvalue. The method avoiding major limitations of tests of normality is recommended.

Future studies with multi-data sets may be undertaken for further investigation of stability of HRQoL scores on deletion of highly or poorly correlated items/dimensions, comparison of progress paths of HRQoL registered by two or more units along with optimal value of reliability to maximize factorial validity.

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.

Data availability: Not applicable as no datasets were generated or analyzed in the study.

Code availability: No application of software package or custom code

Authors' contribution: The single author is involved in Conceptualization, Methodology, Writing- Original draft preparation, Writing- Reviewing and Editing.

References:

Bastien, C. H., Vallieres, A. and Morin, C. M. (2001). Validation of the Insomnia Severity Index as an outcome measure for insomnia research. *Sleep Medicine, 2*(4),297–307

Busija, L., Osborne, R.H., Nilsdotter, A. et al. (2008). Magnitude and meaningfulness of change in SF-36 scores in four types of orthopedic surgery. *Health Qual Life Outcomes,* 6**,** 55 https://doi.org/10.1186/1477-7525-6-55

Carotenuto, A., Fasanaro, A. M., Molino, I., Sibilio, F., Saturnino, A., Traini, E., & Amenta, F. (2013). The Psychological General Well-Being Index (PGWBI) for assessing stress of seafarers on board merchant ships. *International maritime health*, 64(4), 215-220.

Chakrabartty, S. N. (2021). Integration of various scales for Measurement of Insomnia. *Research Methods in Medicine & Health Sciences*, 2(3), 102-111, https://doi.org/10.1177/26320843211010044

Chakrabartty, S. N. (2023). Optimum number of Response categories. *Curr Psychol.* 42, 5590–5598. https://doi.org/10.1007/s12144-021-01866-6

Chmaj-Wierzchowska, K., Rzymski, P., Wojciechowska, M., Parda, I., Wilczak, M. (2020). Health-related quality of life (Nottingham Health Profile) in patients with endometriomas: correlation with clinical variables and self-reported limitations. *Archives of Medical Science*, 16(3), 584-591. https://doi.org/10.5114/aoms.2019.82744

Edinger, J.D., Fins, A. I, Glenn, D. M., Sullivan, R. J., Bastian, L. A, Marsh, G. R., et al.: (2000). Insomnia and the eye of the beholder: Are there clinical markers of objective sleep disturbances among adults with and without insomnia complaints? *J Consult Clin Psychol*. 68: 586-593.

Essink-Bot, M. L., Krabbe, P. F., Bonsel, G. J., Aaronson, N. K. (1997). An empirical comparison of four generic health status measures. The Nottingham Health Profile, the Medical Outcomes Study 36-item Short-Form Health Survey, the COOP/WONCA charts, and the EuroQol instrument. *Med Care;* 35(5):522–537.

Gail, M.S., & Artino AR (Jr.) (2013). Analyzing and interpreting data from likert-type scales. *Journal of graduate medical education*; 5(4):541-542. https://doi.org/10.4300/JGME-5-4-18

Goswami, S. & Chakrabarti, A. (2012). Quartile Clustering: A quartile based technique for Generating Meaningful Clusters, *Jr. of Computing*, 4(2), 48-55.

Gu, Y., Wen, Q. and Wu, D. (1995). How Often Is Often? *English Language Teaching;* 5, 19-35.

Harwell, M. R and Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research*, 71, 105-131.

Hunt, S. M., McEwen.J. and McKenna, S. P. (1985). Measuring health status: a new tool for clinicians and epidemiologists, *Journal of the Royal College of General Practitioners*, 35, 185-188.

Kasper, B. (2020). The EORTC QLQ-C30 Summary Score as a Prognostic Factor for Survival of Patients with Cancer: A Commentary. *Oncologist*. 25(4). e610-e611. https://doi.org/10.1634/theoncologist.2019-0749.

Lidington, E., Giesinger, J. M., Janssen, S. H. M., Tang, S., & Beardsworth, S. (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, 2717–2727. doi.org/10.1007/s11136-022-03139-6

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.

Lipovetsky, S., & Conklin, M. (2018). Decreasing Respondent Heterogeneity by Likert Scales Adjustment via Multipoles. *Stats.* 1.169-175. https://doi.org/10.3390/stats1010012

Lundgren-Nilsson, Å., Jonsdottir, I.H., & Ahlborg, G. (2013). Construct validity of the psychological general wellbeing index (PGWBI) in a sample of patients undergoing treatment for stress-related exhaustion: a Rasch analysis. *Health Qual Life Outcomes*, 11, 2. https://doi.org/10.1186/1477-7525-11-2

Machin, D., Campbell, M. J., Tan, S. B., & Tan, S. H.: *Sample size tables for clinical studies*. John Wiley & Sons.

Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., Keshri, A. (2011). Descriptive statistics and normality tests for statistical data. *Ann Card Anaesth*; 22(1):67-72, (2019). https://doi.org/10.4103/aca.ACA_157_18.

Munshi, J. (2014). A Method for Constructing Likert Scales. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2419366.

Mystakidou, K., Tsilika, E., Parpa, E., Kalaidopoulou, O., Smyrniotis, V., & Vlahos, L. (2001). The EORTC Core Quality of Life Questionnaire (QLQ-C30, Version 3.0) in Terminally Ill Cancer Patients under Palliative Care: Validity and Reliability in a Hellenic Sample. *Int. J. Cancer*, *94*, 135–139.

Parkerson, H. A., Noel, M., Page, M. G., Fuss, S., Katz. J., & Asmundson, G. J.G. (2013). Factorial Validity of the English-language Version of the Pain Catastrophizing Scale-child Version. *J Pain*; 14: 1383-1389.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences, *Acta Psychologica* 104, 1-15.

Razali, N. M., Shamsudin, N. R., Maarof, N. N. N. A., Hadi, A. A., & Ismail, A. (2012). A comparison of normality tests using SPSS, SAS and MINITAB: An application to Health Related Quality of Life data. *International Conference on Statistics in Science, Business and Engineering (ICSSBE)*, https://doi.org/10.1109/ICSSBE.2012.6396570

Reeves, A.J., Baker, R.T., & Casanova, M.P. (2020). Examining the factorial validity of the Quality of Life Scale. *Health Qual Life Outcomes,* 18**,** 32.

Ten Berge, J.M.F. & Hofstee, W.K.B. (1999). Coefficients alpha and reliabilities of unrotated and rotated components. *Psychometrika,* 64(1), 83–90. https://doi.org/10.1007/BF02294321

Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological distance between categories in the Likert scale: Comparing different numbers of options. *Educational and Psychological Measurement, 72*(4), 533-546. https://doi.org/10.1177/0013164411431162

Walters, S. J. (2004). Sample size and power estimation for studies with health related quality of life outcomes: a comparison of four methods using the SF-36. *Health Qual Life Outcomes*; 2:26. https://doi.org/10.1186/1477-7525-2-26.

Wann-Hansson, C., Hallberg, I.R., Risberg, B., & Klevsgård, R. (2004). A comparison of the Nottingham Health Profile and Short Form 36 Health Survey in patients with chronic lower limb ischaemia in a longitudinal perspective. *Health Qual Life Outcomes*.17; 2:9. https://doi.org/10.1186/1477-7525-2-9.

Ware, J. E., Kosinski, M. A. & Gandek, B. (2005). *SF-36 Health Survey: Manual and interpretation guide*. Lincoln: Quality Metric Inc.

Yusoff, R., & Janor, R. M. (2014). Generation of an Interval Metric Scale to Measure Attitude, *SAGE Open*, 1-16. https://doi.org/10.1177/2158244013516768