Online at: www.multiarticlesjournal.com

ISSN No. 2583-7397

International Journal of Contemporary Research in Multidisciplinary; 2023: 2(3):01-09

Research Paper

A Single Measure of SF 36

Author(s): Prof. Satyendra Nath Chakrabartty*[1]

[1*]Indian Statistical Institute, Indian Maritime University, Indian Ports Association, India

Corresponding Author: *****Prof. Satyendra Nath Chakrabartty

Abstract

Manual of SF-36 does not support computation of total score of the scale ($SF36_{Total}$) unlike multidimensional Well-being index, Human development index, etc. where total score are found for an individual or a country/region. The paper gives a method to transform scores of each item to follow normal distribution and find sub-scale scores and $SF36_{Total}$as convolution of normally distributed item scores, parameters of which can be estimated from the data. Addition and arithmetic aggregation of item scores is meaningful due to normality of such scores following same distribution with different parameters. Normally distributed $SF36_{Total}$ scores avoid limitations of summative Likert scores and facilitate parametric statistical analysis including statistical tests of equality of average of $SF36_{Total}$ in two groups for cross-sectional as well as longitudinal data. In addition, $SF36_{Total}$helps to find responsiveness of SF-36 by assessing changes in two time periods either for an individual or a sample of individuals which in turn helps to draw progress-paths. Normality of proposed scores also helps to find psychometric features like factorial validity, discriminating value and reliability. The proposed method is illustrated with hypothetical data.
.

Manuscript Information

Received Date: 17-04-2023
Accepted Date: 20-04-2023
Publication Date: 19-05-2023
Plagiarism Checked: Yes
Manuscript ID: IJCRM:2-3-1
Peer Review Process: Yes

How to Cite this Manuscript

Satyendra Nath Chakrabartty. A single measure of SF 36. International Journal of Contemporary Research in Multidisciplinary. 2023: 2(3):01-09.

Keywords: Socioeconomic SF 36; Equidistant; Normal distribution; Transformation; Assessment of progress

Introduction:

SF-36 is the short form of the Health Survey questionnaire (SF-36). With 36 items distributed over eight sub-scales, the SF-36 is a self-reported, patient-reported questionnaire and is a popular tool to evaluate health-related quality of life (HRQoL). Items in SF 36 have different formats: 28 items in Likert format; seven items are binary, and another item relates to health transitions over the previous year. Raw item scores are rescaled to ranges between 0 and 100, and a high score indicates a more favourable health state. This requires reverse scoring for negatively worded items to have a uniform direction of scores. A zero score on an item is problematic as it tends to reduce the values of mean, variance, covariance, and correlation with that item. Analysis like expected values (score multiplied by the corresponding probability) is not meaningful when zero is attached to one level of an item. The sub-scale wise distribution of items in SF-36 is shown in Table 1.

Table 1: Sub-sales of SF 36, corresponding items and Recorded values

Sub-scales	Total no. of items & scale points	Item numbers & Recorded values of Response choices
Physical functioning	10 (3-points)	3, 4, 5, 6, 7, 8, 9, 10, 11, 12(choice of 1 is recorded as 0 and choice of 3 is recorded as 100)
Role limitations due to physical health	4 (Yes – No type)	13, 14, 15, 16 (choice of 1 is recorded as 0 and choice of 2 is recorded as 100)
Role limitations due to emotional problems	3 (Yes – No type)	17, 18, 19 (choice of 1 is recorded as 0 and choice of2 is recorded as 100)
Energy/ Fatigue	3 (6-point)	23, 27(choice of 1 is recorded as 100 and choice of 6 is recorded as 0); 29 31 (choice of 6 is recorded as 100 and choice of 1 is recorded as 0)
Emotional well-being	6 (6-points)	24, 25, 26, 28, 30 (choice of 6 is recorded as 100 and choice of 1 is recorded as 0) and 26 (choice of 1 is recorded as 100 and choice of 6 is recorded as 0)
Social functioning	2 (5-point)	20 (choice of 1 is recorded as 100 and choice of 5 is recorded as 0) and 32 (choice of 1 is recorded as 0 and choice of 5 is recorded as 100)
Pain	Item 21(6-point) & Item 22 (5-point)	21 (choice of 1 is recorded as 100 and choice of 6 is recorded as 0) and 22 (choice of 1 is recorded as 100 and choice of 5 is recorded as 0)
General health	5 (5-point)	1, 34, 36 choice of 1 is recorded as 100 and choice of 5 is recorded as 0) and 33, 35 choice of 1 is recorded as 0 and choice of 5 is recorded as 100)

Since the total score of an individual by a single number () is not supported by the Manual of SF-36 (http://www.webcitation.org/6cfeefPkf), due to several dimensions being measured by the scale, descriptive statistics and psychometric qualities like reliability, validity, etc. are reported separately for each of the eight sub-scales. Different item formats and different numbers of items in sib-classes result in different values of the mean, standard deviation (SD), distribution of item scores, and reliability, validity, and discriminating power of the SF-36.

Studies to investigate the factor structure of the SF-36 using factor analysis (Guermazi et al. 2012) and structural equation model analysis (Anagnostopoulos et al. 2005) confirmed the multidimensional structure of the SF-36. However, there are measures of variables having a multidimensional nature, like the well-being index, the human development index, etc., where a total score is found for an individual or a country or region.

Physical Component Summary (PCS) and Mental Component Summary (MCS) are the two distinct concepts measured by SF-36. Calculation of PCS and MCS requires the use of special algorithms, which are controlled by a private company (http://www.webcitation.org/6cfdiZOJI).

Lins & Carvalho (2016) found a number of articles suggesting calculation of the total score of SF-36 in different ways, including the arithmetic average of eight sub-classes (Arefnasab et al. 2013; Boccard et al. 2014; Boi et al. 2012) or MCS and PCS (Barnett et al. 2013; Pekmezovic et al. 2015). One major problem with summative item scores is the assumption of equidistance between response categories, despite the fact that additions of ordered categorical data are not strictly appropriate for ordinal data (Svensson, 2001). Thus, the arithmetic average would be meaningful if each item's scores were transformed into equidistant scores following a similar distribution. Available norms based on scores of SF-36 for various populations in several countries, including sub-populations by age and gender, can be compared with the general population (Ware et al., 2005; Australian Bureau of Statistics, 1997). However, this requires a sound scoring system for each sub-scale of the SF-36 questionnaire and a method of aggregation of item scores to sub-scale scores and sub-scale scores to get the questionnaire or scale score for each individual.

The paper aims at transferring raw item scores of different sub-scales of SF 36 to follow normal distribution in a desired score range and finding the total score of SF-36 scores as the sum of normally distributed item scores or, equivalently, sub-class scores obtained as the sum of scores of relevant items of a sub-class. Such transformed scores satisfy desired properties and facilitate the assessment of changes in scale scores over time, either for an individual or a sample of individuals, and the drawing of progress paths.

After a literature survey highlighting the limitations of existing methods, a proposed method is introduced and its associated properties are elaborated. This is followed by empirical illustrations and rounded up by discussing salient outcomes and recommendations.

Literature survey:

Grassi et al. (2007) transferred scores of Likert items on the SF-36 to binary formats using Multiple Correspondence Analysis (MCA). However, MCA does not provide a unique way to transform. Questions can be raised on why to transform Likert scores to binary scores, which amounts to a restriction of data fluctuations due to the reduction of response categories. Why not transform the original scores of each item into a K-point scale where K = 2, 3, 4, 5, or 6? All such questions can be avoided if item scores are transformed to follow a similar distribution, say "normal.

Taft et al. (2001) investigated the relative contribution of the subscales to PCS or MCS using factor scoring coefficients derived from PCA and found an inverse relationship between PCS and MCS, implying that good physical health presupposes poor mental health and vice versa. However, assumptions of PCA like at least interval-level measurements (i.e., equidistant), linearity between observed variables, bivariate normal distribution for each pair of observed variables, etc. are

not satisfied by ordinal item scores emerging from SF-36.

Reporting of the analysis of SF-36 data starts with descriptive statistics showing the mean, SD, etc. for each sub-class. But addition is not meaningful for ordinal data since Likert scores fail to satisfy the equidistant property. Thus, X>Y oXY isis meaningless since the arithmetic mean is not defined for ordinal scales (Hand, 1996). Non-admissibility of meaningful addition implies SD, coefficient of variation (CV), correlation, Cronbach, may not be meaningful, and analysis like regression, PCA, FA, SEM, etc. may give distorted results. In addition, if the assumptions of the techniques used to analyse the data are not satisfied, the results may go wrong. For example, a high correlation (between two variables X and Y) is taken as a linear relationship between X and Y. However, linearity implies high correlation, but the converse is not true. For example,if X takes integer values from 1 to 30, it is 0.97; tatistics showing the mean, SD, etc. for each sub-class. But addition is not meaningful for ordinal data since Likert scores fail to satisfy the equidistant property. Thus, X>Y or XY is meaningless since the arithmetic mean is not defined for ordinal scales (Hand, 1996). Non-admissibility of meaningful addition implies SD, coefficient of variation (CV), correlation, Cronbach, may not be meaningful, and analysis like regression, PCA, FA, SEM, etc. may give distorted results. In addition, if the assumptions of the techniques used to analyse the data are not satisfied, the results may go wrong. For example, a high correlation (between two variables X and Y) is taken as a linear relationship between X and Y. However, linearity implies high correlation, but the converse is not true. For example, if X takes integer values from 1 to 30, it is 0.97; 0.92, despite each being non-linearly related to X. To fit the regression line of Y on X (or X on Y), it is necessary to test the hypothesis where n denotes the number of observations and the test of homoscedasticity, reflecting that the residuals are equally distributed. Hawkin's test is a test of homoscedasticity as well as multivariate normal. Here, the error score for OR did not follow the normal distribution, indicating a violation of the assumption of OLS. This is an example to show how violations of assumptions in statistical analysis may mislead the results.

Major shortcomings of summative Likert scores are:

Addition is not meaningful with ordinal Likert scores (Jamieson, 2004). Do not satisfy equidistant property (Bastien and Morin, 2001). Item levels may mean different things to different subjects responding to the scale (Kampen and Swyngedouw, 2000). Items with varying contributions to total scores, different reliabilities as item-total correlations, different factor loadings, etc. do not justify assigning equal importance to them (Parkin et al. 2010). This often results in tied scores as different individuals may get the same scale score based on different patterns of responses to the items. Thus, a sub-scale cannot discriminate between individuals with the same score.

Unknown and different distributions of it on the on score. A score of 50 in sc with aiumber 5-average but the same score in scale Y with 10 items, each in 5-fformat is the maximum possible score. Interpretation and further operations of sum of X and Y are problematic when X and Y follow different unknown distributions.

A questionnaire may have several scales (a battery of Likert scales) where scale length (number of items) and item formats (number of levels) vary. Here, joint distribution of scale scores is problematic without knowledge of scale distributions.

The distributions of scores emerging from Likert scales are skewed and do not follow a normal distribution. Normality is the common assumption of statistical techniques such as PCA, AVOVA, goodness of fit of regression equations, estimation, testing, etc. (Montgomery and Runger 2006).

The need to consider response a category along with the format of the questionnaire was suggested (Khadka et al. 2012). An increase in the number of response categories usually increases Cronbach's alpha and factorial validity (Lozano et al. 2008). Su et al. (2014) found that WHOQOL-BREF scores were more reliable than the SF-36 scores for assessing people with schizophrenia. The reliability of scales is sample-specific. For normally distributed scores, it is possible to have a population estimate of scale variance and also variance for each item and obtain a population estimate of Cronbach alpha. The generic scale SF-36 has been applied to various disease groups. Empirically, studies showed adequate internal consistency and reliability in terms of Cronbach's for most of the sub-scales except social functioning and general health. For example, alpha for social functioning in patients with brain tumours was 0.53 (Bunevicius, 2017) and 0.45 in patients with coronary artery disease (Alonso, 1995). Cronbach's alpha works best if the scale has one dimension and satisfies the assumptions of uncorrelated errors, tau-equivalence, and normality (Sijtsma and van der Ark, 2015). No attempt was found to report the reliability of the SF-36 scale as a battery consisting of eight sub-scales.

For longitudinal data, Busija et al. (2008) considered that a sub-scale of the SF-36 may have floor or ceiling effects if respondents reported the worst (0) or best (100) possible scores of at least 15%. Here, 15% is arbitrary. It could be better to find the distribution of SF-36 scores, convert it to a symmetrical distribution like the normal distribution, and consider respondents lying outside

Changes in scores at different time periods could be either due to a real change in health status, the effects of random error, or both. Sensitivity of sub-scales using Minimal Detectable Change (MDC) by de Vet et al. (2001) requires partitioning of the within-person variations as between-assessment variance plus the residual variance and computation of the standard error

of measurement (SEM) as the square root of this residual within-person variance (Masse et al. 1997).
MDC for a group obtained as standard errors of the sample means is influenced also by sample size. An individual is categorized under deteriorated if amount of decrease of his/her score exceeds the MDC at individual level. Purpose of MDC is to assess changes which exceed the measurement error by 1.96* SEM presumes normality (Ware et al.2005).Busija et al. (2008) concluded against use of SF–36 due to low sensitivity of SF-36 subscales (except General Health across the intervention groups) for patients undergoing orthopedic surgery. Ware et al, (2005) attempted to compare the published norms for SF-36 over age and gender. However, this may be problematic, in case of small value of the standard errors of published norm scores.

Proposed method:
Pre-adjustment: Ensure that response categories for each item are ordered from low to high, i.e., the lowest level is marked as 1, the second lowest level is marked as 2, and so on. This requires reverse scoring of each "negatively phrased" item.
As per the method given by Chakrabartty (2021), raw discrete item scores (X) are transformed to continuous, equidistant sores (E), which are standardised to follow and further transformed by linear transformation to follow Normal in the desired score range [0, 100].
Equidistant scores:
Let's denote the raw score of a respondent in the i-th item if he or she chooses the j-th response category. If the item is 5-point, weighted score (WS) = where are different for different levels of the i-th item satisfying Scores of the i-th item will be equidistant and monotonic if,, and form an arithmetic progression (AP) with a common difference (CD)> 0.
Forthe i-th item, find maximum ($f_{i\,max}$) and minimum frequency($f_{i\,min}$) of the levels. Find initial weights$\omega_{ij} = \frac{f_{ij}}{n}$. Arrange the $\omega_{ij}^{'}s$ so that ω_{i1}<ω_{i2}<ω_{i3} < ω_{i4}< $\omega_{i5}$where ω_{i1}= $\frac{f_{i\,min}}{n}$and $\omega_{i5} = \frac{f_{i\,max}}{n}$. Let intermediate weight $W_{i1} = \omega_{i1}$
The common difference αcan be found as $\alpha = \frac{5f_{imax}-f_{i\,min}}{4n}$ since $W_{i1} + 4\alpha = 5W_{i5}$
Define other intermediate weights as $W_{i2} = \frac{\omega_{i1}+\alpha}{2}$, $W_{i3} = \frac{\omega_{i1}+2\alpha}{3}$;$W_{i4} = \frac{\omega_{i1}+3\alpha}{4}$; and
$W_{i5} = \frac{\omega_{i1}+4\alpha}{5}$. Get final weights $W_{ij(Final)} = \frac{W_{ij}}{\sum_{j=1}^{5} W_j}$ enabling $\sum W_{ij(Final)} = 1$and
$j.W_{j(Final)} - (j-1).W_{(j-1)(Final)}$= constant
However, value of constant will be different for different items, when the process is repeated for each item
Observations:
i) $W_{j(Final)}$are based on empirical probabilities.
ii) $f_{ij} = 0$ is the zero value of the transformed scores.
iii) Generated scores (E) as weighted sum are equidistant and continuous.
iv) The method can be used for items with different number of response-categories including binary items.
Transform E-scores of the i-th item by$Z_{ij} = \frac{X_{ij}-\overline{X_i}}{SD(X_i)} \sim N(0,1)$.
Take linear transformation of Z-scores to P-scores by:

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

For the i-th item, $P_i \sim N(\mu_i, \sigma_i^2)$ and $1 \le P_i \le 100$where estimates of $\mu_i$andσ_i^2 are obtained from the data. P-score of an item as per equation (1) can be obtained irrespective of length of scale and width of items.
Sub-class score of an individual is taken as sum of normally distributed P-score of relevant items which will follow normal with mean $\sum_i \mu_i$ and SD $=\sqrt{\sum \sigma_i^2 + 2\sum_{i\neq j} Cov(P_i, P_j)}$. Scale/battery score or total SF-36 score ($SF36_{Total}$) is similarly taken as sum of sub-class scores, which also follows normal.
Properties
Sub-class scores (D_i) and scalescores (S_i) of the i-th individual follow normal distribution.
Normality ensures meaningful computation of arithmetic average, SD, correlation, etc. and facilitates statistical analysis under parametric set up including unbiased estimates of population mean (μ), population variance (σ^2), confidence interval of μ, and testing of null hypothesis like $H_0: \mu_1 = \mu_2$ or $H_0: \sigma_1^2 = \sigma_2^2$ etc. across time and space.
2. Progress registered by the i-th person in two successive time-periods can be quantified in percentage by by $\frac{SF36\,Total_{i(t)}-SF36\,Total_{i(t-1)}}{SF36\,Total_{i(t-1)}} \times 100$ which also quantifies responsiveness of the scale and effectiveness of a treatment plan. Deterioration is indicated when
$SF36\,Total_{i(t)} - SF36\,Total_{i(t-1)} > 0$implying progress in t-th period over (t-1)-th period. The reverse is true for $SF36\,Total_{i(t)} < SF36\,Total_{i(t-1)}$. Deterioration in terms of $SF36_{Total}$scores may be probed to find extent of deterioration in sub-class scores for possible corrective actions. Similarly, progress for a group of persons is reflected if $\overline{SF36\,Total_{i(t)}} > \overline{SF36\,Total_{i(t)}}$. Denoting $SF36_{Total}$scores as S, one can test $H_0: \mu_{S_t} = \mu_{S_{(t-1)}}$
3. The graph of progress and/or deterioration of a patient or sample of patients at various time points can be used to compare pattern of progress or HRQoL of patient(s) from the starting year

Benefits:
In addition, the normality of the proposed method can also help to find psychometric properties of
SF-36 in better fashion.
Normally distributed scores satisfy the assumptions of PCA and enable factorial validity.
in terms of the ratio of the first eigenvalue to the sum of all eigenvalues, i.e., validity =, where is the first principal component with the highest eigenvalue

reflecting the main factor for which the scale was developed. Note that validity accounts for a percentage of overall variability. Such factorial validity avoids the problems of construct validity and the selection of criterion scales (Parkerson et al. 2013).

The generic scale SF-36 resulted in various values of Cronbach's alpha for different groups. Population estimates of the variance of each item and scale are possible for normally distributed sub-class scores. Such estimates can be used to find the population estimate of Cronbach's alpha for a subclass as

$$\hat{\alpha} = \left(\frac{n}{n-1}\right)\left(1-\frac{Sum\ of\ estimates\ of\ variance\ of\ items\ in\ the\ sub-class)}{Estimate\ of\ variance\ of\ the\ sub-class}\right) \quad (2)$$

Reliability of the SF-36 as a battery consisting of eight sub-scales as a function of reliabilities of the subscales can be obtained as follows:

$$r_{tt} = \frac{\sum_{i=1}^{8} r_{tt(i)} S_{Xi} + \sum_{i=1, i\neq jk}^{8} \sum_{j=1}^{8} 2COV(X_i, X_j)}{\sum_{i=1}^{8} S_{Xi} + \sum_{i=1, i\neq jk}^{8} \sum_{j=1}^{8} 2COV(X_i, X_j)} \quad (3)$$

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

Discriminating value of a QOL scale is poorly defined or not defined. Mere observation that average QOL score for a group of healthy adults was higher than the group of patients suffering from chronic illnesses, like cancer, etc. may not be sufficient to conclude that the scale has good discriminating value. Such value needs to be quantified. Discriminating value reflects ability of the scale to distinguish between individuals that have different degrees of the underlying construct (e.g. more or less severe disease). Discriminating value of Likert item ($Disc_i$) and test ($Disc_{Test}$)can be computed by Coefficient of variation (CV) where $Disc_i = \frac{SD_i}{mean_i}$ and $Disc_{Test} = \frac{SD_{Test}}{Mean_{Test}}$. Relationship between Cronbach α and $Disc_{Test}$(with m-items) was derived as

$$\alpha = \left(\frac{m}{m-1}\right)\left(1 - \frac{\sum_{i=1}^{m} \overline{X_i}^2 . Disc_i^2}{\bar{X}^2 . {Disc_T}^2}\right) \quad (4)$$

Since, variance of the i-th item $S_{X_i}^2 = \overline{X_i}^2 . Disc_i^2 \forall$i= 1, 2, ….,m $\Longrightarrow \sum_{i=1}^{m} S_{X_i}^2 = \sum_{i=1}^{m} \overline{X_i}^2 . Disc_i^2$ and Test variance $S_X^2 = \bar{X}^2 . {Disc_T}^2$

It can be proved that $(Disc_{Test})^2 = \frac{{CV\ True\ scores}^2}{r_{tt}}$ where $r_{tt} = \frac{S_T^2}{S_X^2}$ (5)

Thus, test reliability and $Disc_{Test}$ are related by a negative non-linear relationship.

A classification of individuals to a finite number of mutually exclusive categories needs to decide boundary points ensuring that members within a class/cluster are similar (small within group variance) and members between classes/clusters are dissimilar (high between group variance). Efficiency of classification needs to be evaluated. 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 (Goswami and Charabarti, 2012). Quartile clustering of proposed scale scores of SF 36 following normal distribution may be adopted because it is simple, appealing, adds clear meaning to the clusters, and gives equal probability to each quartile i.e. $\int_0^{Q_1} f(x)dx = \int_{Q_1}^{Q_2} f(x)dx = \int_{Q_2}^{Q_3} f(x)dx = \int_{Q_3}^{Q_4} f(x)dx$

Empirical illustration:

Hypothetical data involving 100 individuals in each sub-class of the SF-36 was considered for illustration of the proposed method.

Equidistant scores:

Different weights were assigned to different response categories of different items to get an equidistant score. An example of the computation of weights to get equidistant scores for general health with five items, each in 5-point form (sub-scale 8), is given in Table 2.

Table 2: Different Weights to response-categories of five Items of sub-scale 8

Item	Weights to different response categories					Common difference (CD)
	RC 1	RC 2	RC 3	RC 4	RC 5	
1	0.025121	0.186053	0.239696	0.266518	0.282612	0.346984
2	0.050361	0.188066	0.233967	0.256918	0.270689	0.32577
3	0.082432	0.190623	0.226687	0.244719	0.255538	0.298815
4	0.03319	0.186696	0.237865	0.263449	0.2788	0.340202
5	0.07863	0.19032	0.22755	0.246165	0.257334	0.30201

Legend: RC- j denotes the j-th Response-category∀ j=1, 2,3,4,5

CD = $jW_j - (j-1)W_{j-1}$ $\forall j$ =2,3,4,5

Observations:

E-score of an item is obtained as a weighted sum, i.e.$\sum_{i=1}^{5} i.W_{i(Final)}$ is continuous, monotonic and equidistant since $5W_{5(Final} - 4W_{4(Final)} = 4W_{4(Final)} - 3W_{3(Final)} = 3W_{3(Final)} - 2W_{2(Final)} = 2W_{2(Final} - W_{1(Final)} > 0$. However, values of CD were different for different items. E-score of

A sub-class was the sum of item-wise E-scores.

Observed Score range of sub-classes:

E-score of sub-classes reduced the range of scores. Score range of P-score in [1,100] increased the range of scores. Score rangeof sub-classes under X, E, and Pare shown in Table 3

. **Table 3: Observed range of scores for each sub-scale**

Sub-class and items	Raw score (X)		Equidistant score (E)		P-score in [1,100] following normal	
	Max	Min	Max	Min	Max	Min
Physical functioning (10 items, 3-points)	28	11	10.14867	3.099054	859.5508	63.82017
Physical Role functions (4 binary items)	8	4	4.22	1.86	400.00	4.000008
Emotional Role functions (3 binary items)	6	3	3.167311	1.584366	235.3833	3.000163
Energy/ Fatigue (4 items, 6-point)	22	8	4.716457	1.149574	360.40	79.19447
Emotional wellbeing (5 items, 6-points)	27	7	5.382082	0.87412	440.3357	44.60047
Social Functioning (2items, 5point)	9	2	2.7665	0.075482	200.00	1.999831
Pain(Item 21, 6-point & Item 22, 5-point)	11	2	2.506782	0.154319	200.00	2.000081
General Health (5 items, 5-point)	22	11	5.771311	2.116792	425.75	151.249

Tied scores:
Raw scores resulted in number of tied scores unlike E-scores and Z-scores. For example, seven persons were tied at a raw score of 13 in the 8th sub-scale. The tie was broken by E-score and Z-score. Details areshown in Table 4.

Table 4: Illustrative tied raw score and corresponding E-score and Z-score

Sl. No.	Raw score of items					Total score under		
	Item 1	Item 2	Item 3	Item 4	Item 5	X	E	Z
1	2	5	1	4	1	13	2.940406	-2.74749
2	5	1	1	4	2	13	2.980287	-2.50717
3	3	1	2	2	5	13	2.810761	-1.58281
4	4	1	2	1	5	13	2.817543	-2.94016
5	4	1	2	2	4	13	3.202719	-1.99881
6	3	1	1	3	5	13	2.852148	-2.83989
7	4	1	4	3	1	13	2.887535	-2.7765

Observations:

- If i and j aretwo different persons, $E_i \neq E_j$ and $Z_i \neq Z_j$even if $X_i = X_j$=13 in a sub-class
- E-scores and Z-scores consider the pattern of responses to get a particular score and each helps to assign unique ranks to the subjects responding to SF-36.
- Variance in terms ofX-score for the set of subjects with tied score as 13 was zero and thus, the scale failed to discriminate them. -$\bar{E} \neq \bar{Z} \neq \bar{X}$; SD (E) =0.136438and SD (Z) = 0.505262 for the 7-persons with X= 13 .

Descriptive statistics:
Descriptive statistics for sub-classes and $SF36_{Total}$ are shown in Table 5.

Table 5: Mean, SD, CV and correlation (r_{XP}) of sub-class scores and total SF 36 scores

Sub-class	Raw scores (X)			Normally distributed P-scores			r_{XP}
	Mean	SD	CV	Mean	SD	CV	
Sub-class 1	20.29	3.677106	0.181228	488.3476	171.037	0.350238	0.99609
Sub-class 2	6.11	0.815197	0.13342	312.4843	79.40792	0.254118	0.52294
Sub-class 3	4.32	0.723069	0.121025	162.1078	64.40614	0.397304	0.89897
Sub-class 4	16.66	2.999057	0.180015	254.502	60.05119	0.235956	0.99910
Sub-class 5	18.41	3.621192	0.196697	269.2216	71.84771	0.266872	0.99864
Sub-class 6	6.18	1.629154	0.263617	106.9399	41.04252	0.38379	0.93273
Sub-class 7	7.29	2.066056	0.28341	93.52947	38.24569	0.408916	0.57372
Sub-class 8	16.39	2.870593	0.175143	291.7759	70.86596	0.242878	0.98828
$SF36_{Total}$	95.57	8.932129	0.093462	1666.325	267.0231	0.160247	0.89856

Observations:
Lower value of CV reflects more score consistency. Here,lowest CVwas registered by the sub-class 4 for P-score and sub-class 8 for X-scores,

Theoretically defined test reliability $r_{tt} = \frac{{S_T}^2}{{S_X}^2} = \frac{{S_T}^2/\bar{T}^2}{{S_X}^2/\bar{X}^2} = \frac{CV_T^2}{CV_X^2} \Rightarrow CV_X^2 = \frac{CV_T^2}{r_{tt}}$ where CV_X denotes CV for observed scores and similarly,CV_T stands for CV for true scores. Thus, there is a negative relationship between CV_X^2 and r_{tt} (as per theoretical definition).

High correlation between X and P for each sub-class except sub-class 2 and 7 (moderate correlations) indicate not much disturbance of data structure due to the transformations. Correlation between X and P of $SF36_{Total}$ scores was 0.898. However, poor admissibility of X scores might have distorted the results and extent of distortions is not known.

Correlations:
Inter-subscale correlations and subscale-battery correlations for P-scores are shown in Table 6.

Table 6: Correlation matrix between Sub-classes and Battery

	SC 1	SC 2	SC 3	SC 4	SC 5	SC 6	SC 7	SC 8	Battery
SC 1	1	0.0486	0.0223	0.1201	0.1602	0.0507	-0.0201	-0.0248	0.7234 (0.4944)
SC 2		1	-0.1494	-0.0788	-0.0607	0.0277	0.0210	-0.0769	0.1339 (0.0191)
SC 3			1	-0.0159	-0.0037	0.0345	-0.0347	-0.0002	0.2234 (0.0606)
SC 4				1	0.7005	0.1937	0.0807	0.0319	0.5216 (0.6985)
SC 5					1	0.1540	0.0646	-0.0699	0.5314 (0.7069)
SC 6						1	0.0121	0.5988	0.4453 (0.4624)
SC 7							1	0.2897	0.2402 (0.2072)
SC 8								1	0.3571 (0.4179)

Legend: Figures within brackets represent correlations in terms of X-scores SC-i: i-th sub-class, i=1, 2,…

Observations:
Correlations between sub-classes were mostly poor with the exceptions of 4th and 5th subclasses and also 6th and 8th subclasses. Such poor correlations tend to indicate more than one factor from PCA or FA.
Subclass reliability in terms of correlation with battery scores (in line with item-total correlations) ranged between 0.1339 to 0.7234 for P-scores and between 0.0191 to 0.7069 for X-scores.
Distributions:
LetP-scores of i-th sub-class be denoted by P_i. Parameters of normally distributed P_i and $SF36_{Total}$ are shown in Table 7.

Table 7: Parameters of normally distributed Subscale score and $SF36_{Total}$ score

Score	Normal distribution with parameters	
	Mean	SD
P_1	488.3476	171.0377
P_2	53.47	49.65976
P_3	109.7719	64.40614
P_4	253.2684	60.05119
P_5	269.2216	71.84771
P_6	106.9399	41.04252
P_7	93.52947	38.24569
P_8	291.7759	70.86596
$SF36_{Total}$	1666.325	267.0231

Distribution of each P_i may be further transformed to common mean and SD say 50 and 10.

Discussions:
The proposed method generating normally distributed scores for items, sub-classes and SF-36 scale contributes to improve scoring of the instrument and makes it possible to have meaningful single valued $SF36_{Total}$ scores satisfying desired properties.

Benefits of the proposed methods are:
Better admissibility of arithmetic average, normally distributed scores
Parametric statistical analysis for meaningful comparisons over time and space, classification, testing of statistical hypothesis of equality of average of $SF36_{Total}$ in two groups for cross-sectional as well as longitudinal data
Facilitates estimation of mean (μ), variance (σ^2), confidence interval of μ,Cronbach alpha at population level.
Responsiveness of SF-36 can be quantified reflecting ability of the scale to detect changes for an individual or for clinical samples and drawing of progress-paths. For non-clinical samples, changes in $SF36_{Total}$may be reflect effect of changes in major events of life.
Significance of responsiveness of SF-36 can be tested by χ^2 test since ratio of two normally distributed variables follows χ^2distribution.
A better measure of validity of a multidimensional SF-36 scale is proposed as the ratio of the first eigenvalue to the sum of all eigen values.
Discriminating value of Likert item $(Disc_i)$ and test discriminating value $(Disc_{Test})$ were defined as CV and relationship derived between Cronbach α and $Disc_{Test}$and $Disc_{Test}$and theoretically defined r_{tt}.
Advantages of quartile clustering using normally distributed P-scores discussed. It is simple, appealing, adds clear meaning to the clusters, provides well-defined cut-off scores for the four mutually-exclusive classes and assigns equal probability to each quartile.
Researchers and practitioners in social and behavioral health can take advantages of the proposed method to find $SF36_{Total}$for meaningful comparison, assessment of progress or deterioration registered by patient(s) between successive time periods, avoiding limitations of summative Likert scores. Considering theoretical advantages, the proposed method of transforming raw scores of SF-36 items to normally distributed $SF36_{Total}$ scores is recommended for better inferences.

Declarations:

Acknowledgement: Nil

Funding details: Nil

Conflict of interests: No potential conflict of interest is reported.

Informed Consent: Not applicable

Institutional Review Board Approval: Not applicable for this methodological paper

Ethical Compliance with Human/Animal Study: Not applicable

Data availability: No data set used in the methodological paper

Authors' contributions: Sole author

References:

1. Alonso J, Prieto L, Anto JM. (1995): The Spanish version of the SF-36 Health Survey (the SF-36 health questionnaire): an instrument for measuring clinical results. Med Clin; 104:771–776.[Google Scholar]
2. Anagnostopoulos F, Niakas D, Pappa E.(2005): Construct validation of the Greek SF-36 Health Survey. Qual Life Res.; 14: 1959–1965 [Springer] [Google Scholar]
3. Arefnasab Z, Ghanei M, Noorbala AA, et al. (2013): Effect of mindfulness based stress reduction on quality of life (SF-36) and spirometry parameters, in chemically pulmonary injured veterans. Iran J Public Health; 42: 1026–1033. [Google Scholar]
4. Australian Bureau of Statistics (1995): National Health Survey: SF36 Population Norms, Australia, Cat. no. 4399.0. Canberra: ABS; 1997. [Gogle Scholar]
5. Barnett CT, Vanicek N, Polman RCJ(2013): Temporal adaptations in generic and population-specific quality of life and falls efficacy in men with recent lower-limb amputations. J Rehabil Res Dev; 50: 437–448.[Google Scholar]

6. Bastien CH., Vallieres A, & Morin CM(2001): Validation of the Insomnia Severity Index as an outcome measure for insomnia research. Sleep Medicine; 2: 297-307. [Elsevier]
7. Boccard SG, Pereira EA, Moir L, et al. (2014): Deep brain stimulation of the anterior cingulate cortex: targeting the affective component of chronic pain. Neuroreport; 22: 83–88. [Publisher] [Google Sch.]
8. Boi R, Racca L, Cavallero A, et al. (2012): Hearing loss and depressive symptoms in elderly patients. Geriatr Gerontol Int; 12: 440–445.
9. Bunevicius, A.(2017): Reliability and validity of the SF-36 Health Survey Questionnaire in patients with brain tumors: a cross-sectional study. Health Qual Life Outcomes 15, 92 https://doi.org/10.1186/s12955-017-0665-1
10. Busija, L., Osborne, RH., 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
11. Chakrabartty, Satyendra Nath (2021): Integration of various scales for Measurement of Insomnia, Research Methods in Medicine & Health Sciences; 2(3), 102-111. 10.1177/26320843211010044. [Sage]
12. de Vet HC, Bouter LM, Bezemer PD, Beurskens AJ (2001): Reproducibility and responsiveness of evaluative outcome measures: Theoretical considerations illustrated by an empirical example. Int J Technol Assess Health Care; 17: 479–487. 10.1017/S0266462301106148. [Cambridge Uni. Press]
13. Grassi, M., Nucera,A., Zanolin, 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. [Elsevier]
14. Goswami, S. and Chakrabarti, A.(2012): Quartile Clustering: A quartile based technique for Generating Meaningful Clusters, Jr. of Computing, 4(2), 48-55. [Google Scholar]
15. Guermazi M, Allouch C, Yahia M, et al. (2012): Translation in Arabic, adaptation and validation of the SF-36 Health Survey for use in Tunisia. Ann Phys Rehabil Med; 5: 388–403. [Elsevier]
16. Hand, D. J.(1996): Statistics and the Theory of Measurement, J. R. Statist. Soc. A; 159, Part 3, 445-492. [Google Scholar]
17. Jamieson, S. (2004): Likert scales: How to (ab) use them. Medical Education, 38, 1212 -1218. [Google Scholar]
18. Kampen, J., Swyngedouw, M.(2000): The Ordinal Controversy Revisited. Quality & Quantity 34(1), 87-102. [Springer]
19. Khadka, J., Gothwal, VK., McAlinden, C. et al. (2012):The importance of rating scales in measuring patient-reported outcomes. Health Qual Life Outcomes 10, 80, https://doi.org/10.1186/1477-7525-10-80
20. Lins, L.& Carvalho, FM (2016): SF-36 total score as a single measure of health-related quality of life: Scoping review. SAGE open medicine, 4, https://doi.org/10.1177/2050312116671725
21. Lozano Luis& García-Cueto, Eduardo & Muñiz, José.(2008): Effect of the number of Response Categories on the Reliability and Validity of Ratig Scales, Methodology; 4. 73-79
22. Masse J, Bland JM, Doyle JR, Doyle JM (1997): Measurement error. BMJ; 314: 147.[Google Scholar]
23. Montgomery D and Runger G(2013): Applied Statistics and Probability for Engineers, NY: John Wileyand Sons. [Google Books]
24. Parkin D, Rice N, Devlin N.(2010): Statistical analysis of EQ-5D profiles: does the use value sets bias inferences? Med Decis Making 30(5):556–565. [Sage]
25. Parkerson HA, Noel M, Page MG, Fuss S, Katz J, Asmundson GJG(2013): Factorial Validity of the English-language Version of the Pain Catastrophizing Scale-child Version. J Pain; 14: 1383-1389.[Elsevier]
26. Pekmezovic T, Jecmenica-Lukic M, Petrovic I, et al. (2015): Quality of life in patients with progressive supranuclear palsy: one-year follow-up. J Neurol; 262: 2042–2048. [Springer]
27. Sijtsma, K., and van der Ark, L. A. (2015): Conceptions of reliability revisited and practical recommendations. Nurs. Res. 64 (2), 128–136. doi: 10.1097/NNR.0000000000000077. [Publisher]
28. Su CT, Ng HS. Yang AL & Lin CY (2014): Psychometric evaluation of the Short Form 36 Health Survey (SF-36) and the World Health Organization Quality of Life Scale Brief Version (WHOQOL-BREF) for patients with schizophrenia. Psychological Assessment, 26(3), 980–989. https://doi.org/10.1037/a0036764
29. Svensson E.(2001): Construction of a single global scale for multi-item assessment of the same variable. Stat Med;20:3831–46. [Wiley Online Library]
30. Taft, Charles & Karlsson, Jan & Sullivan, Marianne. (2001). Do SF-36 Summary Component Scores Accurately Summarize Subscale Scores?Quality of life research, 10(5), 395-404. DOI 10.1023/A:1012552211996.[Google Scholar]
31. Ware JE, Kosinski MA, Gandek B: (2005): SF-36 Health Survey: Manual and interpretation guide. Lincoln: [Google Scholar]