International Journal of Integrated Medical Research
Volume 8 Issue 01 January-February-March 2021
https://doi.org/10.57181/ijoimr/vol8i01/27
ISSN (Online): 2393-9869, ISSN (Print): 2350-0360

RESEARCH ARTICLE

IMPROVED SCORING OF PATIENT-REPORTED OUTCOME SCALE: METHODOLOGICAL FRAMEWORK

Satyendra Nath Chakrabartty

Institute Indian Statistical Institute, Indian Maritime University, Indian Ports Association

ABSTRACT

Background: Many Patient-reported outcomes (PROs) use multi-item rating scales consisting of K-point Likert items where K is a positive integer taking values 3, 4, 5,…… and so on. However, measurement properties of PROs vary due to different features of the scales and analysis of ordinal data without checking satisfaction of the assumptions of such analysis.

Objectives: To review limitations of scoring of PROs with different number of items and different number of response categories and to provide an assumption-free method for converting raw item-score to continuous, monotonic, and equidistant score in the score range 1 to 100, following normal distribution.

Method: Converting item-score to equidistant score (E) using different weights to response categories of different items → standardizing E-scores to Z-scores ~N(0,1) → converting Z-scores to proposed scores (P_i) in the range 1 to 100. Scale scores as sum of P_i's follows normal distribution.

Results: P-scores facilitate inferences like estimation and testing of statistical hypothesis on equality of population parameters either for longitudinal or snap-shot data, assessment of progress/deterioration by a patient or a group of patients. Equivalent score combinations (x_(0,) y_0) to integrate two PROs were found. Assumption-free methods described to obtain discriminating values as coefficient of variation and elasticity for better measure of responsiveness.

Discussions: The proposed method resolves ordinal-interval controversy and issues relating to scale length and scale width.

Conclusion: The paper contributes to improve scoring of PRO instruments avoiding limitations of ordinal scores and facilitating analysis under parametric set up for meaningful comparisons.

KEYWORDS: Patient-reported Outcome Scale; Linear transformation; Normal distribution; Ability to detect changes; Elasticity

INTRODUCTION

Large number of patient-reported outcome (PRO) instruments are there in the form of multi-item rating scales consisting of *K*-point Likert items are being used in medical studies where *K* is a positive integer taking values 3, 4, 5,…… and so on. The term PROs is increasingly being used as synonymous with "patient reported outcome measures" (PROMs). PROs are used to measure various concepts like clinical conditions (e.g. disease severity, health-status, quality of life, quality and intensity of pain), psychological behaviors (e.g. depression, anxiety, stress), dietary behaviors, preferences, attitudes, beliefs etc. as the primary end points of clinical trials in health research and guide patient care [1]. Dynamics of changes in patients' symptoms and functional abilities including effect of treatment can be assessed by suitably designed PROs [2]. However, measurement properties of PROs vary due to different features of the scales and analysis of ordinal data without checking satisfaction of the assumptions of such analysis. Scale-length (i.e. number of items) and scale-width (i.e. number of response categories) can influence the results [3]. Consensus is yet to be reached on optimal rating scale format. Increase in number of

response categories resulted in increase of Cronbach alpha and factorial validity [4]. Typology of PROs have been criticized by researchers like [5; 6] etc.

The issue of treating Likert scores as ordinal or interval and the associated statistical analysis has been debated extensively [7; 8]. Without giving empirical support, [7] opined that the top 10 myths about "Likert scales" are wrong. [9] attempted to resolve the controversy of ordinal and interval levels of measurements by Likert scales. Verification and assessment of measurement properties of a PRO-instrument was suggested [10].

Typical analysis of data emerging from PRO using Likert scales starts with descriptive statistics showing mean, standard deviation (SD) etc. But, addition is not meaningful for ordinal data since Likert scores fail to satisfy equidistant property. Non-admissibility of meaningful addition implies mean, SD, coefficient of variation (CV), correlation, regression, Principal Component Analysis (PCA), Factor analysis (FA), Structural Equation Modeling (SEM), estimation of population parameters and testing of statistical hypothesis using *t*-statistics, *F*-statistics, etc. may go wrong [11].

Major shortcomings of summative Likert scores are:

- Addition is not meaningful with ordinal Likert scores [8].
- Do not satisfy equidistant property [11].
- Respondents do not perceive the levels as equidistant [12]. Scale points may mean different things to different subjects responding the scale [13; 14].
- Equal importance to the items for summative score is not justified since items have different contributions to total score, different reliabilities as item-total correlations, different factor loadings, etc. [15].
- Often result in tied scores as different individuals may get same scale-score by different pattern of responses to the items. Thus, the scale cannot discriminate the individuals with same scale score.
- Unknown and different distributions of item scores. Score of 50 in scale X with 20 number 5-point items is average but the same score in scale Y with 10 items, each in 5-point format is the maximum possible score. Interpretation of X ± Y and further operations on X ± Y are problematic when X and Y follow two different distributions, that too unknown.
- A questionnaire may have several scales (battery of Likert scales) where the scales differ in terms of number of items and number of response categories. Here, joint distribution of scale scores is problematic without knowledge of distributions of scale scores.
- Likert scores are often skewed and do not satisfy assumptions of statistical analysis undertaken with such data.

The paper aims at providing an assumption-free method of converting raw score of a Likert item to continuous, monotonic, and equidistant score in the score range 1 to 100, following normal distribution. Such transformed scores will serve as evaluative measures to detect changes, classifying individuals or predicting an outcome, and facilitating the following:

- Analysis under parametric set up including estimation of parameters of the distributions of item scores, scale scores and battery scores and testing statistical hypothesis across samples and time.
- Assessment of progress or deterioration over time for an individual and also for a group of individuals and drawing of progress-paths.
- Identification of critical areas contributing to deterioration in successive time periods.

Literature survey

Differential item functioning (DIF) of a Likert-item depends on several factors including scale length and scale width, frequencies of each level. Need to consider response categories along with format of the questionnaire were suggested [16].Transformation of item-score to consider response pattern of response categories.

There is no consensus on rating scale format. *K*-point scales used in PROs differ in terms of values of *K*. Use of many response categories was suggested by [17] instead of using "Yes -No" type items. [18] found that accuracy of a diagnostic scale increases with increase in number of response categories. [19; 20] suggested use of 11-point items so that summative scores is closure to continuous measurement with reduced skewness and kurtosis. But, [21; 22] suggested respectively 4-point and 5-point items. [23] converted the Likert items of SF – 36 to binary format so

that all 36-items are in same format. There is no optimal number of response categories [24].

Normal distribution is the common assumption of analysis like AVOVA, regression, estimation and testing of population mean etc.[25]. Without considering experiment design behind the data,[26] proposed a method of assigning numerical scores for the response categories and claimed that such scoring is suitable for analysis requiring assumption of normality. However, empirical investigation by [27] showed that distribution of data transformed by Snell's method was not normal.

Satisfaction of assumptions of statistical analysis undertaken with Likert data is often ignored. For example, if two variables X and Y are highly correlated (high r_{XY}) then linearity between X and Y is presumed and linear regression (OLS) of the form $Y = \alpha + \beta X + \epsilon$ is fitted. However, r_{XY} may be high even if $Y=f(X)$ where f is non-linear. For X=1, 2, 3, ….., 30, $r_{X,X^2} = 0.97$; $r_{X,X^3} = 0.92$; $r_{X,log_{10}^X} = 0.92$ despite each of X^2, X^3, log_{10}^X is non-linearly related with X. Clearly, *linearity implies high correlation but not the converse*. Linearity between X and Y can be tested by checking assumptions of OLS viz. normality of error score $E = (Y - \hat{Y})$ by say Anderson – Darling test; or testing the hypothesis H_0: $S_E^2 = 0$ where S_E^2 denotes variance of error scores and is computed by $S_E^2 = \frac{1}{n}\sum(Y_i - \hat{Y}_i)^2$ for n- observations and test of Homoscedasticity reflecting that the residuals are equally distributed. Hawkin's test [28] is a test of homoscedasticity as well as multivariate normality. Error score of $Y = \alpha + \beta X + \epsilon$ for $Y = X^2$ or X^3 or log_{10}^X, did not follow normal distribution indicating violation of assumption of OLS. This is an example to show how violations of assumptions of the techniques used to analyze data may mislead the results. Violation of assumption of interval-level measurement of FA with Likert-item scores may increase the number of independent dimensions [29]. Reduction of score range of X or Y can reduce value of r_{XY}. For example, if $0 \leq X \leq 3.9$ and $Y = \frac{1}{\sqrt{2\pi}}\, e^{\frac{-1}{2}X^2}$ then $r_{XY} = 0.93$. But, if $-3.9 \leq X \leq 3.9$, $r_{XY} = 0.00036$. This shows that truncated values of either X or Y can influence correlation significantly. Thus, the sample should have adequate representation of the response categories and each socio-economic-demographic characteristic.

One assumption of Cronbach alpha is that each item measures the single latent trait on the same scale. Likert scale involving multiple factors violates the assumption and test reliability is underestimated [30]. Moreover, Friedman's nonparametric tests [31] cannot examine interaction effects. Aligned Rank Transform (ART), a non-parametric factorial ANOVA analyzes the interaction as well as 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; it also works for other designs, but effects may not be entirely avoided [32]

Major attempts to transform ordinal data to have properties of interval level measurement using by non-linear transformations are:

i) Alternative least squares optimal scaling (ALSOS), a Model-driven approach assumes that respondents interpret items in a similar way, which may not be true [33]. Violation of the assumption may result in biased scaled variables derived from iterative process of ALSOS.

ii) Item response theory (IRT) involves rigorous assumptions and miss-specification of model may give biased results. [34] used IRT to rescale scores of 5-point items into intervals and found that IRT model was not a good fit to the data.

iii) Anchoring vignettes (AVs), advocated by [33] attempts to minimize the differential item functioning (DIF) using AVs which require respondents to answer questions about hypothetical aspects described in the short vignettes. The method assumes *Vignette Equivalence* (VE) i.e. all respondents understand the vignettes in the same way. However, even when VE holds, different persons may use response categories in different ways i.e. different types of DIF. Considering that assumptions of AV are debatable, [35] opined that VE should not be taken for granted.

iv) Markov Chain Monte Carlo Scaling (MCMC) considers multivariate normal distribution and Bayesian methodology where a prior is specified. MCMC involves a number of assumptions and iterations. It is a simulation technique which is used to find the posterior distribution and to draw samples to obtain the distribution of the parameter of interest. Empirically, MCMC performed better than OLS, ALSOS monotonic, and ALSOS non-monotonic approaches. Using MCMC in conjunction with AVs to improve accuracy of the

MCMC scaling was suggested [36]. However, MCMC approach gives problems of bias due to confounding, information, selection, etc. that are common in health science studies. If the prior is not specified correctly, mean squared error may be high.

The above said model-driven methods for rescaling ordinal data involving set of assumptions are quite complex, difficult to interpret rescaled score and not above criticism. Use of such methods in practical analysis of Likert-type data are rare [37].

List of PRO scales is too long. Validated PRO Measures can be found from EORTC QLQ-C30 http://groups.eortc.be/qol/eortc-qlq; FACIT http://www.facit.org/FACITOrg/Questionnaires;

PROMIS http://www.healthmeasures.net/explore-measurement-systems/promis; PRO-CTCAE http://healthcaredelivery.cancer.gov/pro-ctcae, etc. The Australian Commission on Safety and Quality in Health Care came out with a literature review of Patient-Reported Outcome measures in 2016, prepared by Kathryn Williams, et al. (www.safetyandquality.gov.au)

Illustrative examples for insomnia scales with different number of items and different number of response categories are given below:

- Insomnia Severity Index (ISI): Consists of 7- Likert items, each in 5-point scale marked as 0 to 4 [11]. Persons scoring < 14 are taken as Normal and those scoring more than 14 to be considered as having insomnia [38].
- Pittsburgh Sleep Quality Index (PSQI): Total 19-items. First four items are open. Each of Items 5 – 19 is in 4-point scale from 0 to 3 [39]. Poor sleep quality is reflected by a score > 5 and higher score implies worse sleep quality.
- Insomnia Symptom Questionnaire (ISQ) [40]: 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.

Major points emerging from the brief examples are:

i) *Use of zero as an anchor value*: It reduces scale mean and distorts variance, item-total correlations, etc. Analysis involving expected values (value of the variable × probability of that value) is not meaningful. Use of OLS or logistic regression may be inappropriate due to presence of many zeroes [41]. If each respondent of a sub-group selects the response category with zero value 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.

Suggestion: Assign values 1, 2, 3,….. and so on, keeping the convention of higher score ⇔ higher value of the variable being measured.

ii) *Lack of knowledge of distribution of item scores*: Addition of item- scores may not be justified as joint-distribution of sum of two items is unknown.

Suggestion: Transform item-scores to continuous, equidistant scores following normal distribution. All the *K*-point items with different values of *K* will be transformed to follow normal in a specified score range. Distribution of sum of transformed item scores can be found by convolution property of normal distribution. Sum of transformed scores of the items will give scale score of a person which will also follows normal. Score of a battery can similarly be found by adding transformed scale scores.

iii) *Equivalent scores of two scales*: Finding equivalent threshold values of two different scales is difficult. For example, a score of 14 in *ISI* indicating "no insomnia" is equivalent to which score in *PSQI* or *ISQ*?

Suggestion: Let probability density function (pdf) of transformed *ISI*-scores is $f(X)$ and same for transformed *ISQ* is $g(Y)$. A given score of X_0 in *ISI* will be equivalent to a score of Y_0 in *ISQ* if $\int_{-\infty}^{X_0} f(x)dx = \int_{-\infty}^{Y_0} g(y)dy$

(1)

Proposed method

Assuming that response categories of each item is ordered from low to high where the lowest level is marked as 1, a multi-stage method involving weighted sum followed by linear transformations for converting raw-scores (X) of Likert items to continuous, monotonic, equidistant scores following normal distribution was described [42]. The author also verified computation of such transformed scores and their properties with numerical illustration. For equidistant scores (E) different weights are assigned to response categories of different items satisfying the conditions $0 < W_i < 1$ and $W_1, 2W_2, 3W_3, 4W_4, and\ 5W_5$

forms an Arithmetic progression for a 5-point item. For a sample size *n*, this is attained for the *i*-th item by first finding initial weights (ω_{ij}) followed by intermediate weights (W_{ij}) and final weights ($W_{ij(F)}$) such that $\sum_{j=1}^{5} W_{ij(F)} = 1$ for each item.

Standardized *E*-scores $\sim N(0,1)$ are transformed to proposed scores (P) by a linear transformation $= (100-1)\left[\frac{Z_{ij}-Min(Z_{ij})}{Max(Z_{ij})-Min(Z_{ij})}\right] + 1$ (3)

so that $P \in [1,100]$ and follows normal, since by convolution property, if item scores $X_1, X_2, \ldots\ldots, X_m$ are not independent and each $X_i \sim N(\mu_i, \sigma_i^2)$, then $\sum_{i=1}^{m} X_i \sim Normal$ with mean $= \sum_{i=1}^{m} \mu_i$ and SD $= \sqrt{\sum \sigma_i^2 + 2\sum_{i \neq j} Cov(X_i, X_j)}$. Scale scores are taken as sum of *P*-scores of items. For a battery with say seven scales, where m_3, m_4, m_5, m_6 and m_7 are the number of items with 3, 4, 5, 6 and 7 response categories respectively, items of each sub-scale may be transformed to *P*–scores and battery score is the sum of such sub-scale scores.

Properties and Benefits:

-*E*-scores as weighted sum are continuous, equidistant and monotonic.

- For a particular *j*-th level of an item, $f_{ij} = 0$ can be taken as zero value for scoring Likert items as weighted sum.

-Item-wise *E*-scores and *P*-scores avoid equal importance to items and levels and ensure better admissibility of arithmetic aggregation. *P*-scores of items avoid negative values, follows normal with practically zero tied scores and thus can discriminate respondents with tied *X*-scores and assigns unique ranks to individuals and provides meaningful comparison of scales of different formats (length and width) and facilitate parametric analysis.

-*P*-scores following normal help to estimate population mean μ and population variance σ^2 and 95% confidence limits of μ as $\overline{P_{Scale}} \pm 1.96(\frac{\sigma}{\sqrt{n}})$ for large sample of size *n*.

- Easy computation of contribution of *i*-th sub-scale item $P_{Battery}$

- Progress/decline of *i*-th individual at successive time periods can be assessed by $\frac{P_{i_{t_{j+1}}}}{P_{i_{t_j}}}$ or by $\frac{P_{i_{t_{j+1}}} - P_{i_{t_j}}}{P_{i_{t_j}}} \times 100$ to help monitoring of treatment plan and strategies. Similar assessment can be made by a group of patients including testing of statistical hypothesis of equality of average *P*-scores of two groups or one group at two different time periods.
- Critical areas requiring corrective actions are those dimensions or sub-scales for which $P_{t_{j+1}} < P_{t_j}$
- Equivalent threshold values of two scales (P_i^0 and P_j^0) or equivalent class-boundaries in case of classification of individuals by each of the two scales can be found by $\int_{-\infty}^{P_i^0} f(X)dx = \int_{-\infty}^{P_j^0} g(Y)dy$ using Normal Probability table i.e. area under $f(X)$ up to P_i^0 = area under $g(Y)$ up to P_j^0. Equivalent scores are different from predicted values by regressing [43]

Psychometric properties: In addition to reliability and validity, a measure needs to show good discriminating value and responsiveness i.e. ability to detect change.

Discriminating value: Discriminating value of a test reflects its ability to distinguish between individuals having different degrees of the underlying construct (e.g., more or less severe disease). [44] considered discriminating value of an item $Disc_i$ as co-efficient of variation (CV) of the item i.e. $Disc_i = CV_i = \frac{SD_i}{mean_i} = Disc_i$ and discriminating value of the test as CV of the scale i. e. $Disc_{Test} = CV_{Test} = \frac{SD_{Test}}{Mean_{Test}}$. As par definition, test reliability $r_{tt} = \frac{S_T{}^2}{S_X{}^2} = \frac{S_T{}^2/\bar{T}^2}{S_X{}^2/\bar{X}^2} = \frac{CV_T^2}{CV_X^2}$

$\Rightarrow CV_X^2 = \frac{CV_T^2}{r_{tt}}$ (6)

where CV_X and CV_T denote respectively CV for observed scores and CV for true scores. Equation (6) gives a negative relationship between CV_X^2 and theoretically defined r_{tt} i.e. lower the CV, higher the test reliability. Verification of this requires computation of r_{tt} as per theoretical definition, which is beyond the scope of the paper. Relationship between Cronbach alpha and $Disc_{Test}$ of a test with *m*- number of items is

$$\alpha = (\frac{m}{m-1})(1 - \frac{\sum_{i=1}^{m} \bar{X_i}^2 . Disc_i^2}{\bar{X}^2 . Disc_T{}^2}) \quad (7)$$

since $Disc_i^2 = \frac{S_{X_i}^2}{\bar{X_i}^2}$ and sum of item variances is $\sum_{i=1}^{m} \bar{X_i}^2 . Disc_i^2$. Variance of the test is given by $S_X^2 = \bar{X}^2 . Disc_T{}^2$

Since, scale mean and SD are combined mean and combined SD respectively, $Disc_{Test}$ can be written as a function of $Disc_i's$. CV indicates the extent of variability in relation to the mean. Lower value of CV is desirable. For normally distributed *P*-scores, it is easy to estimate population CV and test statistical hypothesis on equality of CV's.

Ability to detect change

Responsiveness of the scale is quantified by value of progress/decline of one or a group of individuals by $\frac{P_{i_{t_{j+1}}}}{P_{i_{t_j}}}$ or equivalently by $\frac{P_{i_{t_{j+1}}} - P_{i_{t_j}}}{P_{i_{t_j}}} \times 100$. Each can take positive or negative value depending on $P_{i_{t_{j+1}}} > P_{i_{t_j}}$ or $P_{i_{t_{j+1}}} < P_{i_{t_j}}$. Significance of progress/deterioration can be tested statistically since ratio of two normally distributed variable follows χ^2 distribution. In addition, effect of small change in *i*-th dimension (P_i) to scale score (P_{Scale}) can be quantified in terms of elasticity i.e. percentage change of P_{Scale} due to small change inP_i. The dimensions of PRO can be ranked based on such dimension-wise elasticity. Elasticity studies in economics, reliability engineering, often consider model like $logQ_{jt} = \alpha_j + \beta_j logP_{jt}$ where Q_{jt} denotes the quantity demanded of *j*-th industry at time *t* and P_{jt} is industry price relative to the price index of the economy [45]. However, for *P*-scores following normal, logarithmic transformations are not required to fit regression equation of the form $P_{Scale} = \alpha_i + \beta_i P_i + \varepsilon_i$ where $\beta_i = r_{P_{Scale},P_i}[\frac{SD(P_{Scale})}{SD(P_i)}]$. The coefficient β_i reflects the impact of a unit change in the independent variable (*i*-th dimension) on the dependent variable (P_{Scale}). However, these coefficients are not elasticity's. Convention of a meaningful estimate of elasticity is to consider it at the point of means, since all regression lines pass through the point of means. Elasticity of the independent variable P for a regression equation of Q on P, can be written as $\frac{\frac{\Delta Q}{Q}}{\frac{\Delta P}{P}} = \frac{\Delta Q}{\Delta P}\frac{P}{Q} = \beta.\frac{P}{Q}$ where β is the slope of regression line $Q = \alpha + \beta P$. Thus, elasticity is $e = \beta\frac{\bar{p}}{\bar{q}}$ where $\bar{p}\ and\ \bar{q}$ are the mean values of data used to estimateβ. The dimensions can be arranged by increasing order of elasticity (e_i). Policy makers can decide appropriate actions in terms of continuation of efforts towards the dimensions with high values of elasticity and corrective actions for the dimensions with lower elasticity i.e. areas of concern.

DISCUSSIONS

The proposed method coverts score of a scale to normal, even if number of items and number of response categories are different. Thus, it resolves the issues regarding scale length and optimum number of response categories.

Correlation between *P*-scores and *E*-scores will be almost perfect since *P* is obtained from *E* by linear transformations. Weighted sum to get *E*-scores from raw scores (*X*) will make r_{XE} high but less than 1. Thus, r_{XP} will also be high implying data structure of *P*-scores will not be deviated much from the same for *X*-scores. High correlation between *P*-scores and *X*-scores resolves the ordinal–interval controversy of Likert data, in the sense that there may not be much harm of treating Likert scores as interval. However, *P*-scores with theoretical advantages and avoiding most of the criticisms about ordinal scores are recommended.

Progress path obtained from $\frac{P_{i_{t_{j+1}}} - P_{i_{t_j}}}{P_{i_{t_j}}} \times 100$ over time may give zigzag pattern showing improvements and deterioration or occurrences of relapse of cancer.

Testing of significance of progress i.e. testing hypothesis$H_0 : Progress_{(t+1)over\ t} = 0$ may avoid need to find minimal important difference (MID) of a measure that is meaningful for comparing patients. Elasticity shows responsiveness of a dimension for snap-shot data. Arranging the dimensions in increasing order of elasticity (e_i), help Policy makers to decide corrective actions for the dimensions with lower elasticity i.e. areas of concern along with continuation of efforts towards the dimensions with high values of elasticity.

CONCLUSION

The paper contributes to improve scoring of PRO instruments which avoid major limitations of ordinal scores and facilitates analysis under parametric set up for meaningful comparisons. Health care professionals and researchers can take advantages of the proposed method to convert ordinal discrete ordinal scores to normally distributed *P*-scores with many desired properties, including identification of critical dimensions, detection of changes by longitudinal data and dimension-wise elasticity to show changes in snap-shot data. Future studies with multi-data set involving scales with different number of response categories may be undertaken for generalization of findings emerging from the study along with psychometric properties of the proposed transformation and to stimulate approach leading to improved patient care and clinical outcomes.

Declarations:

Acknowledgement: Nil

Funding details: No funds, grants, or other support was received

Conflict of interests: The author has no conflicts of interest to declare that are relevant to the content of this article

Ethical Statement: This is a methodological paper with no data, and thus, ethical approval is not relevant.

Informed Consent: Not relevant for this paper with no data

1. Revicki, DA. (2007): FDA draft guidance and health-outcomes research. *Lancet*, 369: 540–542. 10.1016/S0140-6736(07)60250-5
2. Andrew Bottomley, Jaap C. Reijneveld, Michael Koller, et al. (2019): Current state of quality of life and patient-reported outcomes research, *European Journal of Cancer*, Vol.121, 55-63 https://doi.org/10.1016/j.ejca.2019.08.016.
3. Garratt AM, Helgeland J, Gulbrandsen P. (2011): Five-point scales outperform 10-point scales in a randomized comparison of item scaling for the Patient Experiences Questionnaire. *J Clin Epidemi*, 64: 200–207. 10.1016/j.jclinepi.2010.02.016
4. Lozano, Luis & García-Cueto, Eduardo & Muñiz, José.(2008): Effect of the Number of Response Categories on the Reliability and Validity of Rating Scales. *Methodology*. 4. 73-79. 10.1027/1614-2241.4.2.73.
5. Lawal B. N. J.(2003): Lawrence Erlbaum Associates; 2003. *Categorical Data Analysis with SAS and SPSS Applications*, Mahwah, N.J. : Lawrence Erlbaum Associates.
6. Velleman PF, Wilkinson L.(1993): Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading. *The American Statistician*.47:65–72.
7. Carifio, J. & Perla, R. (2007) Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes. *Journal of Social Sciences, 2*, 106-116
8. Jamieson, S. (2004): Likert scales: How to (ab) use them. *Medical Education*,38, 1212 - 1218
9. Knapp TR. (1990): Treating Ordinal Scales as Interval Scales: An Attempt to Resolve the Controversy. *Nursing Research*. 39:121–3.
10. Mokkink, L.B., Terwee, C.B., Patrick, D.L. et al.(2010*):* The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res* **19,** 539–549. https://doi.org/10.1007/s11136-010-9606-8
11. Bastien, C. H., Vallieres, A., & Morin, C. M.: Validation of the Insomnia Severity Index as an outcome measure for insomnia research. *Sleep Medicine*, 2001; 2(4), 297–307
12. Lee JA and Soutar G. (2010): Is Schwartz's value survey an interval scale and does it really matter? *J Cross-Cultural Psychol*; 41: 76–86.
13. Kampen, J., Swyngedouw, M.(2000): The Ordinal Controversy Revisited. *Quality & Quantity* 34(1), 87-102.
14. Gu, Peter & Wen, Q. & Wu, D. (1995): How Often Is Often? Reference Ambiguities of the Likert-Scale in Language Learning Strategy Research. *Occasional Papers in English Language Teaching*. 5. 19-35.
15. 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

16. Khadka, J., Gothwal, V.K., 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
17. Panagiotakos D.(2009): Health measurement scales: methodological issues. *Open Cardiovasc Med J*. 3:160-165.doi:10.2174/1874192400903010160
18. Kourlaba G, Panagiotakos DB. (2009): Dietary quality indices and human health: a review. *Maturitas*; 62(1):1-8. doi: 10.1016/j.maturitas.2008.11.021
19. Hodge, D. R., & Gillespie, D. F. (2007): Phrase completion scales: A better measurement approach than Likert Scales? *Journal of Social Service Research*, 33(4), 1–12.
20. Leung, S. O. (2011). A comparison of psychometric properties and normality in 4–, 5–, 6–, and 11–point Likert scales, *Journal of Social Service Research*, 37, 412–421.
21. Dillman,D.A., Phelps, G., Tortora, R., et al.(2009): Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet, *Social Science Research,* 38, 1–18
22. Fink, A. (1995): *The Survey Kit.* Thousand Oaks, CA: Sage Publications
23. 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*, Vol.10, Issue 6, 478 – 488. https://doi.org/10.1111/j.1524-4733.2007.00203.x.
24. Chakrabartty, Satyendra Nath and Gupta, Rumki (2016): Test Validity and Number of Response Categories: A Case of Bullying Scale, *Journal of the Indian Academy of Applied Psychology*; 42(2); 344-353
25. Montgomery, D. and G. Runger (2006): *Applied Statistics and Probability for Engineers*, NJ: Wiley, John and Sons. 26. Snell, E. (1964): A Scaling Procedure for Ordered Categorical Data, *Biometrics*, 20(3), 592-607.
26. Wu, Chien-Ho (2007): An Empirical Study on the Transformation of Likert-scale Data to Numerical Scores, *Applied Mathematical Sciences*, 1 (58), 2851 – 2862
27. Hawkins DM. (1981): A new test for multivariate normality and homoscedasticity, *Technometrics*, 23:105–110
28. van der Eijk C. and Rose J. (2015): Risky Business: Factor Analysis of Survey Data – Assessing the Probability of Incorrect Dimensionalisation. *PLoS ONE* 10(3): e0118900. doi:10.1371/journal.pone.0118900
29. Green S, Thompson M. (2005): *Structural equation modeling in clinical psychology research* In: Roberts M, Ilardi S, editors. Handbook of research in clinical psychology. Oxford: Wiley-Blackwell
30. *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.
31. 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
32. 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
33. Harwell, M.R. and Gatti, G.G. (2001): Rescaling Ordinal Data to Interval Data in Educational Research, *Review of Educational Research*, Vol. 71, No. 1,105–131
34. Grol-Prokopczyk H, Verdes-Tennant E, McEniry M, Ispány M. (2015): Promises and Pitfalls of Anchoring Vignettes in Health Survey Research. *Demography,* 52(5):1703-1728. doi:10.1007/s13524-015-0422-1
35. Granberg-Rademacker, J. S. (2010): An Algorithm for Converting Ordinal Scale Measurement Data to Interval/Ratio Scale. *Educational and Psychological Measurement*, 70 (1), 74-90.

36. Lantz, B. (2013). Equidistance of Likert-Type Scales and Validation of Inferential Methods Using Experiments and Simulations. *The Electronic Journal of Business Research Methods*, Vol.11(1), 16 – 28
37. 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.
38. 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.
39. Spielman AJ, Saskin P and Thorpy MJ. (1987): Treatment of chronic insomnia by restriction of time in bed. *Sleep*; 10: 45–56.
40. Goulet J, Buta E, Carroll C, Brandt C. (2015):Statistical Methods for the Analysis of NRS Pain Data. *The Journal of Pain*. 16. 10.1016/j.jpain.2015.01.038.
41. Chakrabartty, Satyendra Nath (2019): Limitations of insomnia severity index and possible remedies. *JSM Neurol Disorders Stroke*; 5: 1–9.
42. Livingston, S. A. (2004). *Equating test scores (without IRT*). Princeton, NJ: ETS.
43. Chakrabartty, Satyendra Nath (2020): Improve Quality of Pain Measurement, *Health Sciences*, Volume 1, ID 259, DOI: 10.15342/hs.2020.259
44. *Sinha, D. K. (1994):Can We Measure Elasticity of Demand From Time-Series Data on Prices and Quantities?, in AP - Asia Pacific Advances in Consumer Research Volume 1, eds. Joseph A. Cote and Siew Meng Leong, Provo, UT : Association for Consumer Research, pp: 213-219*