


Equidistant Likert as weighted sum of Response Categories

Likert equidistante como suma ponderada de categorías de respuesta

<http://dx.doi.org/10.17981/cultedusoc.14.1.2023.04>

Received: November 28, 2021. Accepted: March 4, 2022. Pustuled: November 29, 2022

Satyendra Nath Chakrabartty 
Indian Statistical Institute. New Delhi (India)
chakrabarttysatyendra3139@gmail.com

For cite this article:

Chakrabartty, S. (2022). Equidistant Likert as weighted sum of Response Categories. *Cultura, Educación y Sociedad*, 14(1), 75–92. DOI: <http://dx.doi.org/10.17981/cultedusoc.14.1.2023.04>

Abstract

Introduction: Addition of scores of Likert items may not be meaningful since equidistant property is not satisfied. This implies computation of mean, standard deviation, correlation, regression and Cronbach alpha using sum of item variances and test variance could be problematic. **Objective:** Avoiding limitation of summative Likert scores by transforming raw item scores to continuous monotonic scores satisfying equidistant property and evaluate the methods with respect to desired properties and testing normality of transformed test scores. **Methodology:** The methodological paper gives three methods of transforming discrete, ordinal item scores to continuous scores by weighted sum where weights consider frequencies of different response-categories of different items and generate continuous data satisfying equidistant and monotonic properties. **Results and discussions:** All the proposed methods avoided major limitations of summative Likert scores, generates continuous data satisfying equidistant and monotonic properties. The method based on frequencies of response-categories for different items (Method 3) passed the normality test unlike the Method 1 and Method 2. Normally distributed transformed scores in Method 3 facilitate undertaking analysis under parametric set up. **Conclusions:** Proposed methods having high correlations with summative Likert scores, retained similar factor structure and provides reconciliation to the debate on ordinal vs. interval nature of data generated from a Likert questionnaire. Considering the theoretical advantages, the Method 3 is recommended for scoring Likert items primarily due to Normal distribution of individual scores facilitating meaningfulness of operations and to undertake parametric statistical analysis.

Keywords: Likert items; Weighted sum; Monotonic, Equidistant; Normal distribution

Resumen

Introducción: La suma de puntajes de elementos de Likert puede no ser significativa ya que no se cumple la propiedad de equidistancia. Esto implica que el cálculo de la media, la desviación estándar, la correlación, la regresión y el alfa de Cronbach utilizando la suma de las varianzas de los elementos y la varianza de la prueba podría ser problemático. **Objetivo:** Evitar la limitación de las puntuaciones de Likert sumativas transformando las puntuaciones de los ítems sin procesar en puntuaciones monotónicas continuas que satisfagan la propiedad equidistante y evalúen los métodos con respecto a las propiedades deseadas y prueben la normalidad de las puntuaciones de las pruebas transformadas. **Metodología:** El documento metodológico proporciona tres métodos para transformar puntajes discretos y ordinales de ítems en puntajes continuos por suma ponderada donde los pesos consideran frecuencias de diferentes categorías de respuesta de diferentes ítems y generan datos continuos que satisfacen propiedades equidistantes y monótonas. **Resultados y discusión:** Todos los métodos propuestos evitaron las principales limitaciones de las puntuaciones de Likert sumativas, generando datos continuos que satisfacen las propiedades equidistantes y monótonas. El método basado en frecuencias de categorías de respuesta para diferentes ítems (Método 3) pasó la prueba de normalidad a diferencia del Método 1 y el Método 2. Las puntuaciones transformadas normalmente distribuidas en el Método 3 facilitan la realización de análisis bajo una configuración paramétrica. **Conclusiones:** Los métodos propuestos que tienen altas correlaciones con las puntuaciones de Likert sumativas, conservan una estructura factorial similar y brindan reconciliación al debate sobre la naturaleza ordinal frente a la de intervalo de los datos generados a partir de un cuestionario de Likert. Teniendo en cuenta las ventajas teóricas, se recomienda el Método 3 para puntuar elementos de Likert principalmente debido a la distribución normal de las puntuaciones individuales que facilita la significatividad de las operaciones y para realizar análisis estadísticos paramétricos.

Palabras clave: Ítems tipo Likert; Suma ponderada; Monotónico, Equidistante; Distribución normal

INTRODUCTION

Questionnaire-based survey using K -point Likert items ($K = 3, 4, 5, 6, \dots$) are common in survey research, consumer satisfaction, education, social science, Logistics Performance Index, Human development Index, assessment tool for public's knowledge and awareness of public health, patient reported outcomes, addiction research, market research, Quality of Life, etc. primarily for measuring unobservable individual characteristics or feelings that have no concrete, objective measurements. Major purposes of such tools are to identify cases, screen those at risk of developing mental/cognitive disorder and monitor the progress, classify, compare and rank individuals and also to track impact of interventions/treatments.

However, Likert data suffers from limitations. Numbers assigned to response-categories (levels) of Likert item are not numbers as such, but a way to provide ranking responses. If the numbers 1 to 5 are replaced with the letters A to E, the idea of averaging becomes absurd. Distance between successive response-categories in Likert item is assumed to be equal. But distance between "sometimes" and "occasionally", may not be the same as the distance between "never" and "rarely". Thus, the distance between levels is not uniform and is unknown (Munshi, 2014). Levels of a Likert items are ordered but not equidistant as distance between two values on an ordinal scale is unknown (Arvidsson, 2019). The assumption of equal psychological distance between successive categories of rating scale was not supported and increasing values of K influenced the psychological distance between categories, particularly for the 7-point scale and accuracy of the findings are at risk (Uyumaz & Sırgancı, 2021). If distance between level j and $(j+1)$ is denoted by d_j , then satisfaction of equidistant property requires constant value of $d_j=1,2,3,4$ for a 5-point item. Addition or taking average of scores is not meaningful if equidistant property is not satisfied. If addition is not meaningful, then computation of mean, standard deviation (SD), correlation, regression, ANOVA, estimation, testing, etc. may not be meaningful and Cronbach alpha using sum of item variances and test variance could be problematic. Jamieson (2005) observed requirement of manipulation or transformation of Likert data and satisfaction of assumptions of parametric tests like normally distributed variables. In addition, participants may perceive that distance between successive levels as different and not equidistant (Lee & Soutar, 2010).

Giving equal importance to the items may not be justified as items have different reliabilities, factor loadings, etc. It is well known from the factor analysis (FA) that some items have greater factor loadings (explaining more of the variance) than the other items of the scale and thus reflect lack of justification of equal weights. If factor loadings are taken as weights to the items, it may be noted that data specific factor scores are not unique, score for factors may have different lower and upper limits and factor loadings, a regression coefficient of a factor in predicting an item may be negative. Instead of FA, if Principal Component Analysis (PCA) is conducted, it may be necessary to add the non-negative constraint to ensure each weight is positive.

Huiping and Leung (2017) suggested using 11-point Likert items (0 to 10) which help in closer to normality and interval scales. Simms et al. (2019) found increased value of K tends to increase internal consistency till $K = 6$. Chakrabarty (2021) found no optimum

number of response-categories which maximize validity, reliability or discriminating value of the scale. Scoring of Likert items and scale thus, needs to focus on frequencies of response-categories of the items.

The paper gives three methods of transforming scores of Likert items satisfying equidistant property as weighted sum where weights considering frequencies of response-categories are different for different items. Proposed methods are compared with respect to desired properties and provide a platform to perform analysis under parametric set up. The methods can be applied for general Likert data irrespective of number format i.e. scale length (number of items) and scale width (number of response-categories).

Rest of the paper is structured as follows. After literature survey, the following section deals with methodology for obtaining weights to response-categories of Likert items. Next section describes the proposed methods along with computation and properties of such weighted sums. Empirical verification of the proposed methods are discussed in the next section. The paper is rounded up by recalling the salient outcomes of the work, suggesting the best method and implications of such equidistant scores.

LITERATURE REVIEW

Major limitations of summative Likert scores

Non-admissibility of addition of ordinal data produce strange results involving mean, SD, correlation, etc. (Marcus-Roberts & Roberts, 1987). Assumes equal weight to items despite different factor loadings, item-total correlations, distribution of items, etc. Does not consider patterns of getting a particular score. Different responses to different items can generate the same Likert scores for more than one respondents. Thus, summative scores fail to discriminate the respondents with same Likert score. Mean and variance tend to increase with increase in number of levels. Lim (2008) found that the estimated mean is more influenced by number of response-categories, than the underlying variable.

Consideration of anchor value of zero distorts mean, SD, skew, kurtosis of scales (Dawes, 2007). Too many zero responses to an item artificially lower mean variance of the item and correlation with that item. Distribution of summative Likert scores is often skewed and violates normality assumption. Ordinal, discrete, nonlinear, skews, ceiling and floor effects in Likert data and associated problems for undertaking parametric statistical analysis were addressed by Šimkovic and Träuble (2019).

Seven deadly sins of statistical analysis includes among others, use of parametric statistics on ordinal data with the assumption of normality (Kuzon et al., 1996). Assumption of Normality need to be tested with data generated from Likert Scale. Possible solution is to transform item-wise raw scores suitably so that transformed score follows normal distribution. However, large number of researchers treated Likert responses as an interval scale and applied parametric analysis. Carifio and Perla (2007) even suggested that the Likert-responses approximate ratio data.

Reliability in terms of Cronbach's alpha assumes among others, continuous measurement with uncorrelated errors following normal. Violation of such assumptions may bias the coefficient α (Sheng & Sheng, 2012) and distort the variance-covariance matrix substantively, if distribution of observed responses is not symmetric (Flora & Curran, 2004). Researchers like Yusoff and Janor (2014) or Granberg-Rademacker (2010) proposed different methods for rescaling ordinal data to scales having properties of interval level measurement so that parametric statistics can be used. Besides the complex procedures of such conversion, doubts expressed about quality and accuracy of the rescaled data to represent the actual data. Bürkner and Vuorre (2019) suggested use of ordinal models like Cumulative model, Sequential model and Adjacent category model (for Item Response theory (IRT), each with assumptions about the variables under study (along with finite number of thresholds and predictors to have the same effect on all response-categories and attempts to make a regression model. Yusoff and Janor (2014) suggested that a scale must have the following features: metric, presence of zero point, presence of measurement unit, and clearly defined operational procedure as the basis for measurement. Wu (2007) found that transformation of scale data based on Snell's (1964) scaling procedure may not satisfy the normality test. Harwell and Gatti (2001) used IRT approach to transfer ordinal data to interval by rescaling, but emphasized that IRT models which use transformation to a logit scale are complex requiring satisfaction of rigorous assumptions for the models to be of value. However, from IRT, even large ordinal scales can be radically non-linear.

FA or PCA assign different weights to different items. Weighting Likert items with corresponding Discrimination Index in terms of Spearman's Correlation was suggested (Barua, 2013). Other index of discrimination as weights to items can be attempted where weights are not calculated using sum of scores of Likert items, since addition is not meaningful. However, attempts to have different weights to different item-response category combinations to score Likert items are rather rare.

The assumption of a quantitative structure of psychological attributes for attainment of intervalness of the scale has emerged from Additive Conjoint Measurement (ACM). To put it simply, ACM can be related to a situation where on attribute (say) probability of getting correct answer to an item by a candidate can be expressed as a function of two others (say, A ability of the candidate and B the item difficulty level) such that $P = f(A + B)$ where f is any positive monotonic continuous function (Hinne, 2013). In reality, application of ACM are rare in applied psychological data since the data need to satisfy highly restrictive six axioms of additivity, associativity, commutativity, monotonicity, solvability, positivity and the Archimedean condition—which are difficult to verify (Michell, 1990). Satisfaction of these requirements implies that A and B are additive and are therefore quantitative. However, it is common to quantify cognitive abilities to measure “latent traits” (Markus & Borsboom, 2012). Hinne (2013) opined that it is reasonable to apply more flexible measurement methods.

Sum of factor loadings by PCA could be different from one. Score of a Likert item (and test scores as sum of item scores) depends heavily on distribution of frequencies of response-categories (levels), as can be seen from the hypothetical example in Table 1 and Table 2.

TABLE 1. *Different distributions of level-frequency*

Item	Situation	Frequency of levels for a sample size of 100					Item Score	Item mean	Item variance
		Level 1	Level 2	Level 3	Level 4	Level 5			
1	Uniformly distributed	20	20	20	20	20	300	3	2.0202
2	Bi-polarized	50	-	-	-	50	300	3	4.0404
3	Central tendency	6	11	68	10	5	293	2.93	0.6516
4	Faking good	6	10	8	27	49	403	4.03	1.5243
5	Faking bad	46	23	16	10	5	205	2.05	1.4823
Total for scale with above 5 items		128	64	112	67	129		15.01	7.0605

Source: Authors.

TABLE 2. *Item correlation matrix.*

	Item 1	Item 2	Item 3	Item 4	Item 5	Test
Item 1	1	-0.04243	0.088038	0.051805	-0.1109	0.502814
Item 2		1	-0.43577	0.23606	-0.17335	0.631657
Item 3			1	-0.40328	0.03443	-0.15037
Item 4				1	-0.19588	0.458674
Item 5					1	0.187182

Source: Authors.

Observations

- The five items exhibited different characteristics. Variance of Item 1 where choices to response categories were uniform was not minimum, but greater than item 3, 4 showing respectively central tendency, Faking good and Faking bad.
- Item 2 had 50% responses to level 1 and rest 50% to level 5 i.e. the item is bipolarized. But, its mean Equality of mean value of Item 2 and Item 1 cannot reflect that Item 2 was bipolarized or Item 1 had uniform distribution of frequency across the levels.
- Variance of Item 3 was minimum at the level of 0.65. Natural interpretation is that the sample was homogeneous for the item. Thus, it hides the fact that most of the individuals in the sample avoided extreme response categories so that they are not taken as having extremist views.

- In Item 4, most respondents preferred to choose responses which they believe could evaluate them as having higher strength or lower weakness. Naturally, the Level 5 had maximum frequency which increased the mean score of the item but not the variance.
- Similarly for the Item 5, most respondents preferred to choose Level 1 to show they have weaknesses or high impairments, etc. High frequency of Level 1 lowered the item mean.
- Inter-item correlations were poor and sometimes negative. Maximum positive correlation was 0.236 between the Bi-polarized item and the Faking-good item, possibly due to higher proportion of frequency at Level 5.
- Item-reliability in terms of item-total correlation was negative for Item 3 and poor for Item 5. Proportion of frequency at Level 4 and 5 to an item appears to influence the item-total correlation.

METHODOLOGY

Proposed methodology is transforming Ordinal item scores Continuous equidistant scores. It involves selection of weights based on frequencies of response-categories of an item such that $W_i > 0$ and $\sum W_i = 1$ and the transformed score of the i -th individual for choosing the j -th response category of an item is . However, the transformed scores (for say 5-point item) as weighted sum should satisfy at least the following desirable properties:

- Continuous.
- Monotonic i.e. higher transformed score for response to higher numerical level. For example, if for an item, an individual chooses response category 5, his/her transformed score for the item must exceed the transformed score if he/she had chosen response-category 4 i.e. satisfaction of the following condition for each individual for each item.

$$5W_5 > 4W_4 > 3W_3 > 2W_2 > W_1 \quad (1)$$

- Equidistant i.e.

$$5W_5 - 4W_4 = 4W_4 - 3W_3 = 3W_3 - 2W_2 = 2W_2 - W_1 \quad (2)$$

- Normality*: Standardize Equidistant score to follow $N(0,1)$ which may be further transformed to follow Normal distribution with proposed mean and proposed variance.

Suppose, n -respondents have responded to each of the m -items, each having k -number of levels marked as 1, 2, and 3 ... k of a Likert questionnaire where “1” represents *Least preference or Strongly disagree* and “ k ” represents *Maximum preference or Strongly agree*. Assume that there is no missing data. Without loss of generality, take $k = 5$.

Let X_{ij} be the general element of the basic data matrix of order $n \times m$ where n -individuals are in rows and m -items are in columns. X_{ij} represents score of the i -th individual for the j -th item where $1 \leq X_{ij} \leq 5$ for a 5-point scale.

For the usual summative scoring method, score of the i -th individual is $\sum_{j=1}^m X_{ij}$, score of the j -th item is given by $\sum_{i=1}^n X_{ij}$ and the sum of scores of all the individuals (n) on all the items (m) i.e. total test score is denoted by $\sum_{i=1}^n \sum_{j=1}^m X_{ij}$.

Another matrix $((f_{ij}))$ of order $m \times 5$ showing frequency of i -th item to j -th response category can be obtained where a row total will be equal to the sample size (n) and a column total will indicate total number of times that response category was chosen by all the respondents. Grand total is equal to product of sample size and number of items.

Proposed Methods of obtaining equidistant scores (say 5-point items)

• Method 1

Different weights to the response categories for each item, considering frequency of 1, 2, 3, 4, 5 for all the m -items.

The initial weights are $w_j = \sum_{i=1}^m f_{ij} / \sum \sum f_{ij}$ for $i = 1, 2, \dots, m$.

Thus, initial weight for the j -th response-category is the ratio of total frequency of the category and grand total of the Item-Response Categories frequency matrix $((f_{ij})) = m \times n$ i.e. $w_1 = \text{Freq.of 1 in all the } m\text{-items}/mn$, $w_2 = \text{Freq.of 2 in all the } m\text{-items}/mn$. w_3, w_4 and w_5 are defined accordingly.

After finding the corrected weights, raw scores are transformed as Here, transformed score of the j -th item is given by and transformed score of the i -th individual is Thus, both individual scores and item scores are in terms of expected values and hence each is continuous satisfying conditions of linearity since for constants and, the following are satisfied:

$$E(x + y) = E(x) + E(Y)$$

$$E(ax) = aE(x)$$

$$E(ax + \beta y) = aE(x) + \beta E(y)$$

Observations

1. Initial weight $w_j \geq 0 \forall j = 1, 2, 3, 4, 5$ and equality is attend only if $f_j = 0$ i.e. all respondents do not choose the j -th response category of each item, which may be taken as the zero point of the transformed scores.
2. Sum of initial weights is equal to one.

3. The initial weights do not satisfy the monotonic condition (1). For example, if $f_5 < f_4$ then $5w_5$ may be less than $4w_4$.
4. Transformed scores as weighted sums where weights are w_j 's, may not satisfy the equidistant condition (2).
5. To satisfy the monotonic and equidistant conditions, correction factor is required based on which final weights W_j for $j = 1, 2, 3, \dots, 5$ are to be calculated. The suggested steps are as follows:

Step-1: Arrange the frequency of response-categories (considering all the items) in increasing order. Let the maximum and minimum frequency are f_{\max} and f_{\min} respectively.

Step-2: To satisfy the equidistant property, $W_1, 2W_2, 3W_3, 4W_4, 5W_5$ should form an Arithmetic Progression (AP). Thus, problem is to find common difference β so that:

$$\begin{aligned} W_1 + 4\beta &= 5W_5 \implies \beta = \frac{5W_5 - W_1}{4} \\ &= \frac{5 \cdot \frac{f_{\max}}{mn} - \frac{f_{\min}}{mn}}{4} = \frac{5f_{\max} - f_{\min}}{4mn} \end{aligned} \quad (3)$$

Define:

$$W_1 = f_{\min}/mn, \text{ and } 2W_2 = W_1 + \beta \implies W_2 = W_1 + \beta/2$$

Similarly:

$$W_3 = W_1 + 2\beta/3; W_4 = W_1 + 3\beta/4; \text{ and } W_5 = W_1 + 4\beta/5$$

The transformed scores based on the corrected weights so defined satisfy the monotonic condition and ensures equidistant scores.

6. However, $\sum_{j=1}^5 W_j$ is not always equal to one. For a 5-point scale, $\sum_{j=1}^5 W_j = 1$, if and only if $f_{\min} + 77/48 f_{\max} = mn$.

Proof:

$$\begin{aligned} \sum_{j=1}^5 W_j &= W_1 \left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5}\right) + \beta \left(\frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5}\right) \\ &= W_1 \left(1 + \frac{77}{60}\right) + \beta \left(\frac{77}{60}\right) \\ &= \frac{f_{\min}}{mn} \left(1 + \frac{77}{60}\right) + \frac{77}{60} \left(\frac{5f_{\max}}{4mn}\right) - \frac{77}{60} \left(\frac{f_{\min}}{4mn}\right) \text{ using (1.1)} \\ &= \frac{f_{\min}}{mn} + \frac{77}{60} \left(\frac{5f_{\max}}{4mn}\right) \\ &= \frac{1}{mn} \left[f_{\min} + \frac{77}{48} f_{\max}\right] \end{aligned}$$

Thus, $\sum_{j=1}^5 W_j = 1$ if and only if f_{\min} and f_{\max} are related by (4).

$$f_{\min} + \frac{77}{48} f_{\max} = mn \quad (4)$$

• *Method 2*

W_1, W_2, W_3, W_4, W_5 are based on area under $N(0,1)$ with $W_i > 0$ and $\sum_{j=1}^5 W_j = 1$. Procedure for obtaining W_j 's are illustrated in **Table 3**.

Table 3. Calculation of weights, Method – 2.

Response Category	Proportion (p_i)	Cumulative Proportions (C_i)	Area under the standard Normal curve	Initial Weights
1	$p_1 = f_1/mn$	p_1	$A_1 = \text{Upto } p_1$	$w_1 = A_1/\sum A_i$
2	$p_2 = f_2/mn$	$p_1 + p_2$	$A_2 = \text{Up to } p_1 + p_2$	$w_2 = A_2/\sum A_i$
3	$p_3 = f_3/mn$	$p_1 + p_2 + p_3$	$A_3 = \text{Upto } p_1 + p_2 + p_3$	$w_3 = A_3/\sum A_i$
4	$p_4 = f_4/mn$	$p_1 + p_2 + p_3 + p_4$	$A_4 = \text{Upto } p_1 + p_2 + p_3 + p_4$	$w_4 = A_4/\sum A_i$
5	$p_5 = f_5/mn$	$p_1 + p_2 + p_3 + p_4 + p_5 = 1.00$	$A_5 = \text{Upto } p_1 + p_2 + p_3 + p_4 + p_5$	$w_5 = A_5/\sum A_i$
Total	1.00		$\sum_{i=1}^5 A_i > 1$	1.00

Source: Authors.

Here, for $w_j > w_{j-1}$ for $j = 2, 3, 4, 5$. Thus, the monotonic condition (1) is satisfied. However, to make the transformed scores equidistant for a 5-point scale, take correction factor $\alpha = \text{Area}_{\text{Max}} - \text{Area}_{\text{Min}}/3$. Determine the modified areas $\Delta_1, \Delta_2, \Delta_3, \Delta_4$ and Δ_5 as follows:

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

Define corrected weights $W_j = \Delta_j / \sum_{j=1}^5 \Delta_j$.

The transformed scores based on corrected weights so defined satisfy the monotonic condition (1), ensures equidistant scores (2) and also satisfy $\sum_{j=1}^5 W_j = 1$.

Thus, the Method-2 has clear advantages over Method – 1.

• *Method 3*

Different weights to different response-categories of different items considering frequency of the $(i-j)$ -th cell.

Here, $W_{ij} = f_{ij} / \sum_{j=1}^k j \cdot f_{ij}$ i.e. ratio of number of times the j -th response category of the i -th item was chosen and the score of the i -th item. Clearly, $W_{ij} > 0$ and $\sum_{j=1}^5 W_{ij} = 1$. However, sum of weights for a response-category of all items is different from one. Score of the individual choosing j -th response category of the i -th item will be $X_{ij} \cdot W_{ij}$. Note that for a particular item, weights increase with increase in anchor values associated with response-categories. Thus, monotonic condition as well as the equidistant condition for transformed scores is satisfied for each item. Illustration for 5 points scale is shown below (**Table 4**).

TABLE 4. Calculation of weights, Method – 3.

Response Category/ original score	Weights	Transformed score
1	$W_{i1} = \frac{\text{Frequency of 1's for the } i\text{-th item}}{\text{Score of the } i\text{-th item}} = \frac{f_{i1}}{\sum_{j=1}^5 j \cdot f_{ij}}$	1. W_{i1}
2	$W_{i2} = \frac{\text{Frequency of 2's for the } i\text{-th item}}{\text{Score of the } i\text{-th item}} = \frac{f_{i2}}{\sum_{j=1}^5 j \cdot f_{ij}}$	2. W_{i2}
3	$W_{i3} = \frac{\text{Frequency of 3's for the } i\text{-th item}}{\text{Score of the } i\text{-th item}} = \frac{f_{i3}}{\sum_{j=1}^5 j \cdot f_{ij}}$	3. W_{i3}
4	$W_{i4} = \frac{\text{Frequency of 4's for the } i\text{-th item}}{\text{Score of the } i\text{-th item}} = \frac{f_{i4}}{\sum_{j=1}^5 j \cdot f_{ij}}$	4. W_{i4}
5	$W_{i5} = \frac{\text{Frequency of 5's for the } i\text{-th item}}{\text{Score of the } i\text{-th item}} = \frac{f_{i5}}{\sum_{j=1}^5 j \cdot f_{ij}}$	5. W_{i5}

Source: Authors.

Observations

The monotonic condition and equidistant property are satisfied for each item. If frequency of a particular response-category of an item is zero, the method fails and can be taken as zero value for weighted sum approach.

Summary of observations

Weights to various response-categories remain unchanged across items in Method 1 and Method 2. But, in Method 3, weights to the response-categories are different for different items.

In Method 1, 2 and 3, data driven weights are taken as probabilities, considering the frequencies of Item–Level combinations without involving assumptions of continuous nature or linearity or normality for the observed variables or the underlying variable being measured.

Here, each $W_j > 0$. For method 2, $\sum_{j=1}^5 W_j = 1$. For method 3, sum of weights for each item is equal to one. However, $\sum_{j=1}^5 W_j = 1$ is not always satisfied by the method 1. Equidistant property of transformed scores is satisfied by method 1, 2, and 3 (for each item). Thus, the Method 2 and 3 have clear advantages over Method 1. Between Method 2 and 3, the latter is preferred since it considers variations of item-level combinations by assigning different weights to different response-categories for different items.

Implication

Item-wise equidistant scores (E) by method 3 can be standardized by $Z = (E - \bar{E})/SD(E)$. To avoid negative values, transfer to proposed score (P) by $P_i = (99)[(Z_i - \text{Min}Z_i)/\text{Max}Z_i - \text{Min}Z_i] + 1$ so that $P_i \in [1, 100]$ and P_i follows normal. Item-wise P -scores can be added to get test scores following normal and parameters can be estimated from the data. Normally distributed test score enables meaningful comparisons and undertaking statistical analysis under parametric set up.

Empirical investigation

A Likert questionnaire with 30 items, each with 5 response-categories was administered among a sample of parents to identify relevant factors of the parenting style. Out of these parents, only 463 respondents completed the questionnaire in all respects. Data obtained from these 463 respondents are considered for the empirical investigation.

Calculation of weights

Calculation of weights to item for method 1 and 2 are shown in [Table 5](#).

TABLE 5. Calculation of weights to response categories: Method 1 and Method 2.

Description	Response Category-1 (Score = 1)	Response Category-2 (Score = 2)	Response Category-3 (Score = 3)	Response Category-4 (Score = 4)	Response Category-5 (Score = 5)	Grand Total
Frequency (all the items)	3768	2419	888	3316	3499	13890
Method 1						
Proportions = Initial Weights	$W_1 = 0.27127$	$W_2 = 0.17415$	$W_3 = 0.06393$	$W_4 = 0.23873$	$W_5 = 0.25191$	1.00
Corrected weights $\beta = 0.32311$	$W_1 = 0.06393$	$W_2 = 0.19352$	$W_3 = 0.23672$	$W_4 = 0.25831$	$W_5 = 0.27127$	1.0237
Method 2						
Cumulative Proportions (C_i)	0.27127	0.44543	0.50935	0.74809	1.0000	
Area under $N(0,1)$ up to C_i 's (A_i)	0.60691	0.67199	0.69475	0.77279	0.84135	3.5877
Initial weights	0.16916	0.18730	0.19364	0.21540	0.23450	1.00
Modified area ($\alpha = 0.07815$)	$\Delta_1 = 0.16916$	$\Delta_2 = 0.24731$	$\Delta_3 = 0.32546$	$\Delta_4 = 0.40361$	$\Delta_5 = 0.48176$	1.6273
Corrected weights	$W_1 = 0.10395$	$W_2 = 0.15198$	$W_3 = 0.2000$	$W_4 = 0.24802$	$W_5 = 0.29605$	1.00

Note: Scores obtained by various methods are different but, related by linear relationships, which are clarified in the [Table 6](#). Source: Authors.

TABLE 6. Relationships among scores by various Methods.

Summative score	Score for various response categories				
	1	2	3	4	5
Corresponding score in Method 1	1(0.06393)	2(0.19352)	3(0.23672)	4(0.25831)	5(0.27127)
Corresponding score in Method 2	1(0.10395)	2(0.15198)	3(0.2000)	4(0.24802)	5(0.29605)
Corresponding score in Method 3	$f_{i1}/\sum_{j=1}^5 j \cdot f_{ij}$	$f_{i2}/\sum_{j=1}^5 j \cdot f_{ij}$	$f_{i3}/\sum_{j=1}^5 j \cdot f_{ij}$	$f_{i4}/\sum_{j=1}^5 j \cdot f_{ij}$	$f_{i5}/\sum_{j=1}^5 j \cdot f_{ij}$

Source: Authors.

Observe that Method 1 is times Summative score and Method 3 is K_2 times Summative score where $0.06393 \leq K_1 \leq 0.27127$ and $0.10395 \leq K_2 \leq 0.29605$. However, linear relationship between Method 3 and Summative score involves among others frequency of the j -th response-category of the i -th item and the score of the i -th item. Strong linearity among the methods is likely to result in high correlations between each pair of methods.

Analysis

• Descriptive statistics

Mean, Variance, Skewness, and Kurtosis of the various methods are shown below (Table 7):

TABLE 7. Descriptive statistics for various Methods.

Description	Summative score	Method – 1 (Based on frequencies of Response Categories of all the items)	Method – 2 (Based on area under $N(0,1)$)	Method – 3 (Based on frequency of Response category and score of the item)
Test Mean	90.48	21.46	21.74	0.2345
Range of Item Mean	Max: 4.47 Min: 1.59	Max: 1.18 Min: 0.25	Max: 1.24 Min: 0.26	Max: 0.0100 Min: 0.0049
Test Variance	63.41	6.62	7.53	0.0015
Range of Item Variance	Max: 2.39 Min: 0.74	Max: 0.25 Min: 0.08	Max: 0.29 Min: 0.09	Max: 0.00009 Min: 0.00001
Test Skewness	(-) 0.20	(-) 0.20	(-) 0.32	(-) 0.03
Range of Item Skewness	Max: 2.00 Min: (-)2.11	Max: 2.00 Min: (-)2.11	Max: 2.64 Min: (-)1.46	Max: 2.77 Min: (-)1.37
Test Kurtosis (>3)	0.20	0.20	0.23	(-)0.10
Range of Item Kurtosis (> 3)	Max: 4.83 Min: (-)1.56	Max: 4.83 Min: (-)1.56	Max: 6.39 Min: (-)1.53	Max: 7.08 Min: (-)1.53

Source: Authors.

Observations

1. Method 1, 2, and 3 resulted in significant reduction of test average and test variance, maximum reduction was observed for Method 3.
2. Range of item means and item variances also got reduced in Method 1, 2 and 3.
3. Test skewness for each of the four methods was found to be low negative ranging from (-) 0.32 to (-) 0.03, implying distribution under each method was slightly skewed to the left. Skewness was minimum for Method 3, at the level of (-) 0.03 which is close to perfect symmetry.
4. Test kurtosis (excess of 3) had low positive values for methods other than 3 implying distribution with more outliers than a normal distribution thus fatter tails (*Leptokurtic distributions*). However, Kurtosis of the test was (-) 0.10 for the Method 3 implying similar outlier character as a normal distribution (*Mesokurtic distributions*).

• Correlation and Regression

Correlations of test scores obtained by various methods are shown below (Table 8):

TABLE 8. Correlations between a pair of approaches

	Summative score	Method 1	Method 2	Method 3
Summative score	1.0	0.989	0.986	0.964
Method 1		1.0	0.986	0.964
Method 2			1.0	0.956
Method 3				1.0

Source: Authors.

Observations

1. Very high correlations (> 0.95) between each pair of methods indicate linear relationships among the four methods.
2. Regression equations of each proposed method on summative score (M_0) are shown below:

$$M_1 = 0.5660 (M_0) - 10.29, \text{ corresponding } R^2 = 0.9623.$$

$$M_2 = 0.34 (M_0) - 9.023, \text{ corresponding } R^2 = 0.9731.$$

$$M_3 = 0.0047 (M_0) - 0.191, \text{ corresponding } R^2 = 0.9286.$$
3. High value of indicate goodness of fit of the data to the linear model.
4. Liner relationships among the four methods are likely to give similar clusters of individuals taking the test.

- *Rank Correlation*

Spearman ρ between tests scores of various methods are shown below (Table 9):

TABLE 9. Rank Correlation Matrix (Spearman ρ) between a pair of approaches.

	M_0	M_1	M_2	M_3
M_0	1.0	0.999	0.985	0.959
M_1		1.0	0.981	0.958
M_2			1.0	0.949
M_3				1.0

Source: Authors.

Observations

High value of ρ (was found implying that each of the proposed method retained more or less same ranks with Summative scores. The results are in line with correlation between a pair of methods.

- *Test of Normality*

Anderson–Darling test for Normality to test test score follow Normal distribution. A large p -value ($p > 0.05$) would indicate normality. Values of test statistic and associated p -values for the methods are shown in the Table 10.

TABLE 10. Values of test statistic and associated p -values for test of Normality.

	Value of Test statistics	p – values	Remarks
Summative score	1.709	0.000219	is rejected
Method 1	1.446	0.000970	is rejected
Method 2	1.475	0.000825	is rejected
Method 3	0.396	0.369900	is accepted

Source: Authors.

Observations

Test scores of respondents followed Normal distribution only for Method 3. Thus, the transformed scores as per method 3 offer better platforms for undertaking almost all type of analysis being done for continuous quantitative variable following Normal distribution.

- *Factor structures*

Factor Analysis with orthogonal vari-max rotation was undertaken for each method. Details are given in the [Table 11](#).

TABLE 11. *Results of Factor Analysis.*

Method	Number of Eigen values exceeding one	Cumulative percentage of variance explained
Summative score	13	58.962
Method 1	13	58.962
Method 2	13	59.011
Method 3	13	58.995

Source: Authors.

Observations

Each method gave 13 independent factors explaining cumulative variance of 59% (approx.). The results appear to be in line with high correlation (> 0.95) between each pair of methods implying almost linear relationships among the four methods. However, result of the factor analysis confirms that some of the items had greater factor loadings than the other ones comprising that scale. Thus, equal weight to items may not be justified.

- *Effect on reliability*

Cronbach alpha for the scale did not vary much over the four methods primarily due to linearity. Item reliability in terms of correlation between item score and test score ranged approximately between 0.17 to (–) 0.11 under each of the four methods. Thus, weighted sum proposed in methods 1, 2, and 3 did not have much effect on Cronbach and on Item-Test correlation as can be seen from the [Table 12](#).

TABLE 12. *Test reliability and Range of Item reliability for the four methods.*

	Summative score	Method 1	Method 2	Method 3
Cronbach	0.23832	0.23832	0.22490	0.23614
Range of Item reliability	Max: 0.16958 Min: (–)0.11005	Max: 0.16958 Min: (–)0.11004	Max: 0.16402 Min: (–)0.10737	Max: 0.16177 Min: (–) 0.10629

Source: Authors.

Observations

The test with 13 independent factors failed to satisfy uni-dimensionality assumption of Cronbach alpha. Alpha in each of the four methods was poor. Item reliabilities were poor also.

Changes in Reliability on deletion of an item

Effects of deletion of items on reliability was undertaken for each method. Cronbach got increased upon deletion of 15 items for Summative score and Method 1 and 16 items for Method 2 and 17 items for Method 3. Thus, the proposed methods of scoring Likert test did not have much effect in terms of deletion of an item on the test reliability.

CONCLUSIONS

Avoiding the complex Additive Conjoint Measurement (ACM), three simple methods of scoring Likert scale as weighted sum are proposed where different weights to different response-categories resulted in continuous data satisfying equidistant property. Selection of weights is not model driven. Data driven weights were chosen without making any assumption of nature of underlying or observed variables.

However, sum of weights was not equal to unity in Method 1. Only the Method 3 where weights are taken as ratio of frequency of the j -th response category of the i -th item resulted in different weights to the response-categories for different items, passed normality test. Thus, the Method 3 offers platform for undertaking almost all type of analysis being done for continuous quantitative variable following Normal distribution like the t -test, ANOVA, etc.

Each of the proposed method had strong linear relationship with summative scores. Correlations ranged between 0.964 to 0.989 and emerging from linear regressions on summative score varied between 0.928 to 0.973. Linear relationships among the four methods gave high value of rank correlations implying retention of more or less same ranks, factor structure and reliability with summative scores.

These strong linear relationships provide the bridge between the psychometric issues in the ordinal/interval controversy. Probable reconciliation to the debate on ordinal vs. interval nature of data generated from a Likert questionnaire is the fact that transformed scores as weighted sum are continuous, equidistant, satisfying normality, etc. but correlates very high with summative type Likert score assuming equal weights. However, considering the theoretical advantages including conversion of ordinal Likert data to continuous scale; meaningfulness of operations and comparison; platform to undertake parametric statistical analysis and easiness to compute weights, the Method 3 is recommended for scoring a Likert questionnaire. Simulation studies may be undertaken to evaluate merits of the proposed approach in a wide range of datasets.

DECLARATION

Acknowledgement: Nil.

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

Conflict of interests: The author has no conflicts of interest to declare.

REFERENCES

- Arvidsson, R. (2019). On the use of ordinal scoring scales in social life cycle assessment. *The International Journal of Life Cycle Assessment*, 24(3), 604–606. <https://doi.org/10.1007/s11367-018-1557-2>
- Barua, A. (2013). Methods for Decision–making in Survey Questionnaires Based on Likert Scale. *Journal of Asian Scientific Research*, 3(1), 35–38. <https://archive.aessweb.com/index.php/5003/article/view/3446>
- Bürkner, P.C. & Vuorre, M. (2019). Ordinal Regression Models in Psychology: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 2(1), 77–101. <https://doi.org/10.1177/2515245918823199>
- 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*, 3, 106–116. <http://dx.doi.org/10.3844/jssp.2007.106.116>
- Chakrabartty, S. N. (2021). Optimum number of Response Categories. *Current Psychology*, 104(1), 1–15. <https://doi.org/10.1007/s12144-021-01866-6>
- Dawes, J. (2007). Do data characteristics change according to the number of scale points used? *International Journal of Market Research*, 50(1), 61–77. <https://doi.org/10.1177/147078530805000106>
- Flora, D. B. & Curran, P. J. (2004). An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis with Ordinal Data. *Psychological Methods*, 9(4), 466–491. <https://doi.org/10.1037/1082-989X.9.4.466>
- 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. <https://doi.org/10.1177/0013164409344532>
- Harwell, M. R. & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research*, 71, 105–131. <https://doi.org/10.3102/00346543071001105>
- Hinne, M. (2013). Additive conjoint measurement and the resistance toward falsifiability in psychology. *Frontiers in Psychology*, 4(1), 1–4. <https://doi.org/10.3389/fpsyg.2013.00246>
- Huiping, W. & Leung, S-O. (2017). Can Likert Scales be Treated as Interval Scales?—A Simulation Study. *Journal of Social Service Research*, 43(4), 527–532. <https://doi.org/10.1080/01488376.2017.1329775>
- Jamieson, S. (2005, Aug. 11). Likert scale. *Encyclopedia Britannica*. <https://www.britannica.com/topic/Likert-Scale>
- Kuzon, W. M., Urbanchek, M. G. & McCabe, S. (1996). The seven deadly sins of statistical analysis. *Annals of Plastic Surgery*, 37, 265–272. <https://doi.org/10.1097/00000637-199609000-00006>
- Lee, J. A. & Soutar, G. N. (2010). Is Schwartz’s value survey an interval scale, and does it really matter? *Journal of Cross-Cultural Psychology*, 41(1), 76–86. <https://doi.org/10.1177/0022022109348920>

- 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>
- Marcus-Roberts, H. M. & Roberts, F. S. (1987). Meaningless statistics. *Journal of Educational Statistics*, 12, 383–394. <https://doi.org/10.2307/1165056>
- Markus, K. A. & Borsboom, D. (2012). The cat came back: evaluating arguments against psychological measurement. *Theory & Psychol*, 22(4), 452–466. <https://doi.org/10.1177/0959354310381155>
- Michell, J. (1990). *An Introduction to the Logic of Psychological Measurement*. Erlbaum Associates.
- Munshi, J. (2014). A method for constructing Likert scales. *Social Science Research Network*. <https://doi.org/10.2139/ssrn.2419366>
- Sheng, Y. & Sheng, Z. (2012). Is coefficient alpha robust to non-normal data? *Frontiers in Psychology*, 3(34), 1–13. <https://doi.org/10.3389/fpsyg.2012.00034>
- Šimkovic, M. & Träuble, B. (2019). Robustness of statistical methods when measure is affected by ceiling and/or floor effect. *PloS one*, 14(8), 1–47. <https://doi.org/10.1371/journal.pone.0220889>
- Simms, L. J., Zelazny, K., Williams, T. F. & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, 31(4), 557–566. <https://doi.org/10.1037/pas0000648>
- Snell, E. (1964). A Scaling Procedure for Ordered Categorical Data. *Biometrics*, 20(3), 592–607. <https://doi.org/10.2307/2528498>
- Uyumaz, G. & Sırgancı, G. (2021). Determining the Factors Affecting the Psychological Distance Between Categories in the Rating Scale. *International Journal of Contemporary Educational Research*, 8(3), 178–190. <https://doi.org/10.33200/ijcer.858599>
- Wu, Ch.-H. (2007). An Empirical Study on the Transformation of Likert scale Data to Numerical Scores. *Applied Mathematical Sciences*, 1(58), 2851–2862. <https://doi.org/10.12988/ams>
- Yusoff, R. & Janor, R. M. (2014). Generation of an Interval Metric Scale to Measure Attitude. *SAGE Open*, 4(1), 1–16. <https://doi.org/10.1177/2158244013516768>

Satyendra Nath Chakrabartty. Prof. Satyendra Nath Chakrabartty is an M. Stat. from Indian Statistical Institute. He has taught Post Graduate courses at Indian Statistical Institute, University of Calcutta, Galgotias Business School, etc. He has over 65 publications to his credit. After serving Kolkata Port Trust for 25 years in various managerial positions, he joined Mumbai Port Trust as Director (Planning & Research) and subsequently took over as Director, Indian Institute of Port Management. He retired from the position of Director, Kolkata Campus of the Indian Maritime University. His previous assignment was Consultant, Indian Ports Association, New Delhi. ORCID: <https://orcid.org/0000-0002-7687-5044>