Vol. 02, No. 3 (2020) 311-322, doi: 10.24874/PES02.03.010

Proceedings on Engineering Sciences

www.pesjournal.net

COMBINING LIKERT ITEMS WITH DIFFERENT NUMBER OF RESPONSE CATEGORIES

Satyendra Nath Chakrabartty[1]

Keywords:

Likert items; Weighted sum; Equidistant scores; Z-scores; Coefficient of variation.

A B S T R A C T

Three assumption-free approaches are described to convert and combine scores of Likert items with different response categories to obtain comparable test scores for a single sample. Generated data were continuous, equidistant, normally distributed, avoiding tied scores and thus facilitate ranking the respondents and undertaking analysis under parametric set up.

Empirical verification was undertaken involving 5 items in each of 3, 4, 5 and 7-point scale to 100 subjects. Correlations at the level of 0.99 between a pair of approaches were observed with marginal difference in coefficient of variation (CV). Minimum CV and maximum alpha were found when test scores were obtained as sum of standardized equidistant item scores converted to [1, 5]. Thus, the said Approach appears to have advantages.

Use of such methods of combining scores of Likert items is recommended for clear theoretical advantages and easiness in calculations. Future studies are suggested.

1. INTRODUCTION

Instruments used in survey, behavioral research often contain Likert items with different number of response categories and ordered numerical values attached to these categories. Scores of respondents as discrete summative scores of all such items reduce comparability and even lack meaningfulness. Likert data with different scale format i.e. 4-point, 5-point or 7-point scale differ in mean, standard deviation (SD) and shape. For 3-, 5-, 7- and 9-point rating scales, Finn (1972), reported means as 1.6, 2.2, 4.1 and 4.9 and variances as 0.32, 0.60, 1.32 and 4.0 respectively, implying that mean and variance tended to increase with increase in number of points in rating scales. Test variance and test reliability can be increased by increasing the number of response alternatives (Cook, et al. 2001). Preston and Colman (2000) found that 2-point, 3-point, and 4-point scales performed poorly on reliability, validity, and discriminating power indices in comparison to scales with more levels. Thus, the estimated mean is more influenced by number of response categories, than the underlying variable (Lim, 2008). Dawes, (2007) opined that results of satisfaction surveys – may dependent partially on the choice of scale format. These calls for appropriate methods to transform responses of different scale formats in meaningful fashions attain comparable results.

Collapsing of response categories, say from 7 categories to 5 may arbitrarily treats 1 and 2 on the 7-point item as equivalent and thus ignore the distinction made by the respondents between answers 1 and 2. In addition, consideration of anchor value of zero also distorts mean, SD, skewness, kurtosis of scales (Dawes, 2002; Johnson, Smith, & Tucker, 1982). Too many zero responses to an item artificially lower mean, variance of the item and

[1] Corresponding author: Satyendra Nath Chakrabartty
Email: Snc12@rediffmail.com

covariance and correlation with that item. Techniques like confirmatory factor analysis (CFA), structural equation modeling (SEM), etc. are sensitive to the characteristics of the data (Bentler, 1995). Effect of numerical value attached to various response categories have been investigated (Schwarz et al. 1991; Sangster et al. 2001; Mazaheri and Theuns, 2009). For various types of survey questions and response scales, see Saris and Gallhofer (2007).

Colman, Norris & Preston (1997) compared 5-point and 7-point scales using regression and suggested a pair of equations of the forms $X_7 = \alpha_1 + \beta_1 X_5$ and $X_5 = \alpha_2 + \beta_2 X_7$ for estimating scores in 7-point from 5-point and vice versa. However, equating is different from forecasting (Livingston, 2004). Moreover, two regression equations viz. X_5 on X_7 and X_7 on X_5 violates the axiom that equated scores are interchangeable. Lim, (2008) used estimated transition probabilities (ETP) to compare means of 7-point and 11-point scales. However, ETP based on ordered probit models and a set of assumptions are extremely sensitive to the heterogeneities of sample. Model driven IRT does not work well to detect individual changes for tests with less than 20 items (Jabrayilov, Emons and Sijtsma,2016) The basic idea of probability of a correct response to an item as a function of examinee and item parameters in IRT may not be well applicable for Likert items.

Basic question in this context: “What would have been the pattern of responses if the respondents were asked to response to a 5-ponit item (say) instead of a 3-point or 4-point items and how best to combine those responses to get test score of each respondent”?

Attempts have been made to use linear transformation of levels (ordered numerical values attached to the response categories) where the extreme numbers of an item with lesser number of points say 1 and 5 of a 5-point item are fixed to the extreme numbers of an item with higher number of points say 1 and 7 for a 7-point item and all the intermediate options are given equally distanced numbers in between. Other simple approach to convert a *k*-point scale to (k+Δ)- point scale (where *k*, Δ are positive integers) by multiplying each *k*-point score by $\frac{(K+\Delta)}{k}$ i.e. proportional transformation. Major disadvantages of such methods are:

- If a 3-point item is converted to 5-point, the converted scale has only three values attached to response categories and strictly speaking cannot be taken as a 5-point scale.
- Number of ties remains unchanged for the converted scores also and thus fails to distinguish the respondents with same score. Occurrences of tied score in Likert items are common as different responses to different items can generate the same aggregate scores for more than one respondent.
- Mean and variance of raw scores of the test and also items get changed in the converted scale and shape of the original data gets distorted.
- Formula for conversion is not unique. Preston and Colman (2000) used the formula $\frac{(Score-1)}{(No.of\ response\ categpries-1)} \times 100$ to rescale to scores out of 100. Multiplication by 10 instead of 100 rescales all scale formats to a score out of 10. Cummins (1997, 2003) used linear transformation in the percentage of scale maximum.
- There could be other linear transformations based on frequencies of response categories of items being converted for transferring the levels to a new scale.
- The converted scores and raw scores will have equal mean and SD, if *Z*-scores are used. However, *Z*-scores involving mean and SD for Likert items are not meaningful since (i) mean and SD presume interval measurement which is not the case for Likert items. Thus, $\bar{X} > \bar{Y}$ is meaningless since the arithmetic mean is not defined for ordinal scales (Hand, 1996). (ii) distance between levels is unequal and unknown (Wu, 2007; Ferrando, 2003; Munshi, 1990) (iii) subjects do perceive Likert-type scales as non-equidistant,(Bendixen and Sandler,1995) (iv) addition of item scores assuming equal weight to items may not be right as factor analysis (FA), Principal Component Analysis (PCA) often comes out with different factor loadings for the items comprising that scale and thus reflect lack of justification of equal weights.
- Standardized scores making implicit assumptions about psychological equivalences between scales of different lengths cannot compare meaningfully rating scales of unequal lengths, (Colman, Norris and Preston, 1997).

Instead of mapping the numerical values attached to the end points, attempts can be made to map or equate raw scores of *k*-point item to say ($k \pm \Delta$)-point item where $k \geq 3$ and $\Delta \geq 1$ are positive integers, ensuring similarity in distributions of scores. However, initial transformations of raw scores of all *k*-point items are required since often skewed ordinal Likert data assumes wrongly equal importance to items, equal distance between two successive levels and often results in large number of tied scores. Moreover, assumptions of statistical techniques used in the parametric set up are not satisfied by ordinal Likert data. Thus, the Basic question is modified as follows:

Modified Basic Question: **“**If the raw scores of a *k*-point scale and a ($k \pm \Delta$)-point scale are transformed to continuous, equidistant scores with zero ties, normally distributed and are denoted by $\mathcal{X}^{(k)}$ *and* $\mathcal{X}^{(k\pm\Delta)}$, how to link a value of $\mathcal{X}_0^{(k)}$ to a value $\mathcal{X}_0^{(k\pm\Delta)}$ and facilitate meaningful combination of the responses to get test score of each respondent”?

Equivalent scores: A score X_0 in *k*-point scale is equivalent to Y_0 in ($k \pm 1$) -point scale if

$$\int_{-\infty}^{X_0} f(X)dx = \int_{-\infty}^{Y_0} g(Y)dy$$

The score combinations (X_0, Y_0) to be found for each pair of sub-tests and clubbed together to generate test/questionnaire scores. However, it may be difficult to approximate scores by integrable continuous function $f(X)$ or $g(Y)$ addressing the issue of goodness of fit.

A simpler and approximate way to find various combination of $(X_0,\ Y_0)$ could be to consider equal percentile scores (Equipercentile equating). Alternately, if each $f(X)$ and $g(Y)$ is the density function of Normal distribution with mean $\bar{X}$ and variance ${S_X}^2$, X_0 will be equivalent to $Y_0 \Leftrightarrow X_0 = Y_0$. In other words, finding of the equivalent combinations (X_0, Y_0) or linking of a score in k-point scale with a score in $(k\pm1)$-point scale is obvious. Normal Curve Equivalent (NCE) score by Mertler, (2002) considered $N(50, 21.06^2)$. Purpose of linking is to compare the respondents who have taken different versions of a scale or different subtests of a questionnaire, where the subtests may differ in length or content. Other methods can be worked out for linking test scores to make the distribution of test scores equivalent to make the distribution of test scores equivalent in terms of normality with equal mean and SD for all the k-point scales, using further transformation to achieve proposed mean and SD with positive scores.

Studies in equating scores involving several samples usually consider smoothing of raw data to avoid irregularities and equating design for data collection. Both can result into systematic errors. There are a number of smoothing methods and a number of equating designs, each having its advantages and disadvantages. However, in the present set up where all respondents responded to a number of k-point and $(k\pm\Delta)$-point Likert items (k is a positive integer ≥ 3 and $k-\Delta \geq 3$) of a questionnaire, issues relating to meaningful test scores as sum of item scores need to be answered realistically for one sample and may not deal with issues relating to smoothing or equating design.

Attempt to find meaningful test scores of respondents as sum of item scores or sum of sub-test scores where each sub-test contains a fixed value of k-point Likert items, may be undertaken keeping in mind the limitations of ordinal Likert data, non-availability of reference scores and following major observations regarding score equating by Livingston, (2004) and associated comments:

1. Equating is not Forecasting. Hence, equating method must be different from any forecasting methods including regression analysis.
2. Equating is symmetric i.e. interchangeable.
3. Equating demands deciding the range of scaled scores. However, if the observed scores are transformed to (μ_X, σ_X^2), $-\infty < \text{transformed score} < \infty$.
4. Equating may not consider similarity of dimensions/factors of raw scores and transformed scores.
5. Score X is equivalent to a score Y for a sample of respondents if X and Y represent the same relative position in the group (Equipercentile equating). However, equipercentile equating scores are not equal-interval and strictly speaking, not additive. Score linking by other methods like percentage of scores under Normal curve or NCE score may be explored.
6. Item response theory (IRT) equating (Lord, 1980) involving abstract definition of equated scores, strong assumptions about person's estimate (θ), item difficulty parameters (a), etc. may not fit well the reality of testing with Likert items.

The above motivates need to transfer first the raw scores of the Likert items with different number of response categories ensuring satisfaction of equidistant property along with other desirable properties like continuous data which is monotonic, etc. followed by normalizing the item scores and further rescaling to a desired range say [1, 5] and suggesting better method of combining such scores to obtain test scores satisfying the requirements from the angle of measurement theories, without making assumptions of distribution of observed/underlying variables, person's estimate, item parameters, etc.

Rest of the paper is organized as follows. Methodology of the proposed method and properties are described in the following section. Section 3 deals with empirical verifications to the suggested methods. The paper is rounded up in Section 4 by recalling the salient outcomes and emerging suggestions.

2. METHODOLOGY

Consider a Likert questionnaire with m- number of items, administered among n-number of respondents where $m_3, m_4, m_5, m_6, and\ m_7$ are the number of items with3, 4, 5, 6 and 7 response categories(levels) respectively and $\sum_{j=3}^{7} m_j = m$. Without loss of generality, assume that response categories of each item is ordered from low to high where the lowest level is marked as 1.

2.1 Proposed method

A multi-staged method is used to transfer raw scores to comparable test scores as shown in figure 1.

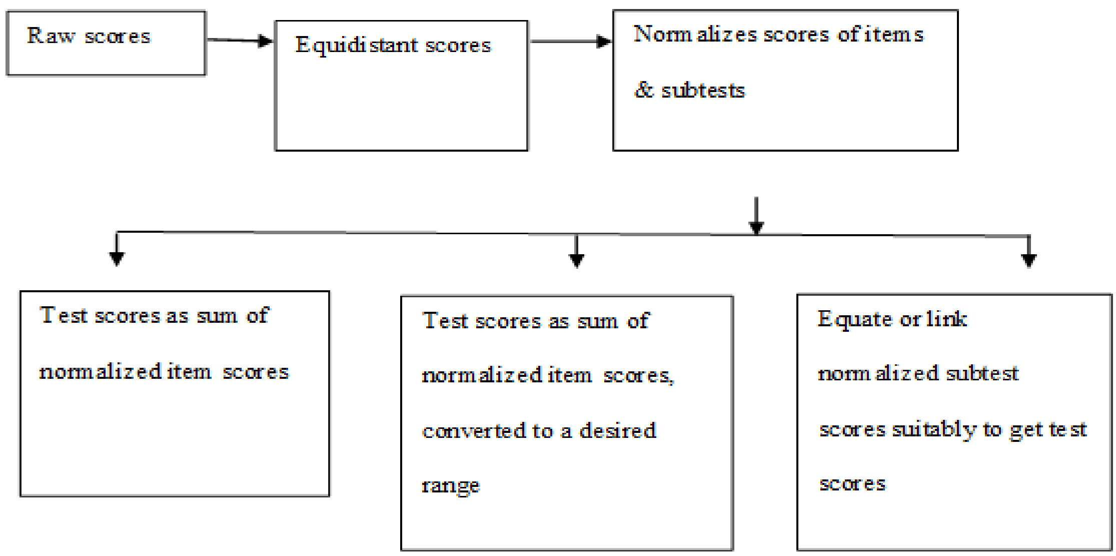

Figure 1. The Method

Stage 1: Sub-tests:

Consider all 3-point items in sub-test 1. Similarly, constitute sub-tests 2, 3, 4 and 5 by considering respectively all 4-point, 5-point, 6-point and 7-point items.

Stage 2: Equidistant scores by assigning different weights to response categories of different items:

This is described for 5-point items. However, it can be used for any k-point items, where $k \neq 5$. Let X_{ij} represents discrete raw score of the i-th respondent in the j-th response category, for $i = 1,2, \ldots, n$ and j=1 to 5. Problem is to find $W_{ij}'s$ such that $W_{ij} > 0, \sum_{j=1}^{5} W_{ij} = 1$ and $W_1, 2W_2, 3W_3, 4W_4, 5W_5$ forms an AP.

Such weights can be found using frequency of each level by following steps:

1. Find from the data, frequency of i-th item for the j-th level (f_{ij}). For each item, find maximum (f_{max}) and minimum frequency (f_{min}).
2. Find proportions $\omega_{ij} = \frac{f_{ij}}{n}$. Clearly, $\omega_{ij} > 0$ and $\sum_{j=1}^{5} \omega_{ij} = \frac{\sum_{j=1}^{5} f_{ij}}{n} = 1$.
3. Put initial weights $W_{i1} = \omega_{i1} = \frac{f_{min}}{n}$
4. Find the common difference α so that $W_{i1} + 4\alpha = 5W_{i5} \Longrightarrow \alpha = \frac{5f_{max} - f_{min}}{4n}$
5. Define $W_{i2} = \frac{\omega_{i1} + \alpha}{2}$, $W_{i3} = \frac{\omega_{i1} + 2\alpha}{3}$; $W_{i4} = \frac{\omega_{i1} + 3\alpha}{4}$; and $W_{i5} = \frac{\omega_{i1} + 4\alpha}{5}$
6. Here, $W_{ij} > 0$ and $\sum_{j=1}^{5} W_j \neq 1$.
7. 4: Get final weights $W_{ij(Final)} = \frac{W_{ij}}{\sum_{j=1}^{5} W_j}$ so that $\sum W_{ij(Final)} = 1$
8. Here, $0 < W_{j(Final)} < 1$ and
9. $j.W_{j(Final)} - (j-1).W_{(j-1)(Final)} = \frac{\alpha}{\sum_{j=1}^{5} W_j}\ \forall j =$ 2,3,4,5 (Equidistant scores)

Observations:

i) $W_{j(Final)}$ are based on empirical probabilities obtained from the basic Item score matrix.
ii) If $f_{ij} = 0$ for a particular j-th level of an item, the method fails and can be taken as zero value for scoring Likert items as weighted sum
iii) Mean, variance and range of weighted test scores and also weighted item scores will get reduced in comparison to the same from usual summative scores.
v) Different weights to different levels of different items may break ties of subject scores in usual summative scores and thus distinguish the same summative score on the basis of how the score was obtained.
vi) Generated scores are continuous.

Stage 3: Normalize scores:

From weighted scores described at Stage 2, take Z- scores for each item. For the i-th item, $Z_{ij} = \frac{X_{ij} - \bar{X_i}}{SD(X_i)} \sim N(0,1)$ where $-\infty < Z_{ij} < \infty$. But sub-test score as a sum of item scores will also be normal with mean zero and SD = $\sqrt{\sum Z_i^2 + 2\sum_{i \neq j} Cov(Z_i, Z_j)}$.

Stage 4: Covert the Z-scores to have a fixed range:

Find minimum and maximum values of Z obtained for each item. Convert the Z-score of an item to say [1, 5], as follows:

$$Y = \frac{(5-1)*(Z_{ij} - Min(Z_{ij}))}{Max\,(Z_{ij}) - Min(Z_{ij})} + 1 \qquad (1.1)$$

Linear transformation (1.1) will change range of Item score in the interval [1, 5]. Distributions of item scores for each k-point scale will be normal with similar values of means and variances. However, item scores will not satisfy definition of Equating Scores.

Range of sub-test scores as sum of converted item scores may vary. Variance of sub-test scores will also vary depending on correlations between pair of items.

Stage 5: Combining scores of items or subtests to obtain test scores:

NCE scores are equal interval scores following normal distribution with mean = 50 and SD = 21.06. NCE scores are additive. It need not assume that the scales have similar measurement errors, (Mâsse, et al., 2006).

For equipercentile scores, the converted scores on Form-X have the same distribution as scores on Form-Y implying that the cumulative distribution of equated scores on Form-X is equal to the cumulative distribution of scores on Form-Y. A percentile rank indicates the percentage of respondents below a certain score. Distance between say 10^{th} and 20^{th} percentile is different from the distance between say 40^{th} and 50^{th} percentile, Percentile scores range from 1 to 99 and are not additive. Percentile scores could be tied if such scores are not strictly increasing. If P_i is the *i*-th percentile score, then $P_i \ngtr P_{(i-1)}$ if the frequency in the interval containing the percentile point is zero. In other words, a particular P_i for a sub-test could be same for say 14% to 25% of scores with similar phenomenon in other sub-tests for say 20% to 30%. Thus, it is problematic to find equated score in (*k+1*)-point and (*K+2*)-point scales for a particular P_i in *k*-point scale. Hence, percentile scores may not be very suitable to equate individual scores of subtests with different number of response categories. However, scores of NCE's and percentiles are identical at 1, 50, and 99. Thus, NCE appears to have advantageous in comparison to Equipercentile scores at least for converting responses on different subtests with different response categories to attain comparable test scores.

Two disadvantages of NCE scores are as follows:

i) For$Z < (-)2.3738$, NCE score is negative.

ii) If Z-scores of five items of a subtest are $Z_{i1}, Z_{i2}, \ldots\ldots, Z_{i5}$ such that $\sum_{j=1}^{5} Z_{ij} = Z_{Subtest}$ then $\sum_{j=1}^{5} var(NCE_{ij}) > var(NCE_{Subtest})$. Similarly, sum of variance of NCE scores of the subtests may exceed the variance of NCE score of the test, if Z-value of test is taken as sum of Z-values of the subtests. This may make Cronbach alpha of NCE test score greater than unity. Cronbach alpha > 1 may occur even if $\frac{Sum\ of\ varNCE_{Subtests}}{var(NCE_{Test})} \approx 1$ say 0.99

To avoid the disadvantages of equipercentile scores and NCE scores and to make the distribution of different sub-test scores similar ensuring equality of mean, variance, further transformation may be used as follows:

$$Modified\ (Y_{k-point}) = \frac{(X_{k-point} - Mean_{k-point})}{SD_{k-point}} \times Proposed(SD) + Proposed(Mean) \qquad (1.2)$$

For example, proposed SD could be taken as ≥10 to increase variance of the distribution and the proposed mean could be taken to ensure all positive values of *Modified* $(Y_{k-point})$

Distribution of modified test scores for each *k*-point scale will be $N(\ Proposed(Mean),\ [Proposed(SD)]^2$. Thus, the *k*-point subtests for various values of *k*, could be considered as Equivalent Form having features of parallel tests.

2.2 Approaches to get transformed scores following $N(\mathbf{0},\ \mathbf{1})$

To get test scores of the respondents, the following approaches are proposed based on Normalized item scores following $N(0,\ 1)$

Approach 1.1 Test score as sum of *Z*-scores of all items will follow $N(0,\ \sigma^2)$ where $\sigma^2 \neq 1$.

Approach 1.2 To avoid negative values and to have a desired score range of items, item scores may be converted to say [1, 5] by (1.1) and added so as to maintain normality and unique rank.

Approach 1.3 To make the distribution of test scores equivalent in terms of normality with equal mean and variance for all the *k*-point scales, use transformation (1.2) with proposed mean and proposed SD and add the converted items scores to get test scores. Subtest scores obtained in this fashion will be equivalent. Finding equivalent scores will provide satisfactory answer to the Modified Basic Question ensuring similarity of mean, SD and shape of the subtest scores with added feature of high correlations between subtests.

Each of the above said approaches is likely to generate scores with zero ties, implying a unique rank for each respondent. Since the approaches are based on linear function of *Z*-scores of items, correlation between a pair of the Approaches will be ≈1.00. Such test scores will help in ranking or comparing or classifying the respondents, and also to undertake statistical analysis under parametric set up.

2.3 Comparison of test scores

Test scores obtained by each of the above said approaches are continuous, normally distributed with high correlations. Comparison of the approaches could be made considering advantages/limitations and statistics like Coefficient of variation (CV), reliability, etc. CV is a well known measure of relative precision which is independent of change of scale but not of origin. CV indicates the extent of variability in relation to the mean. Lower value of CV is desirable. It can be proved easily that square of CV of observed scores is equal to the ratio of square of CV of true scores and test reliability i.e. $CV_X{}^2 = \frac{CV_T{}^2}{r_{tt}}$. Thus, a negative relationship is there between test reliability and CV. In other words, lower the CV, higher is the reliability.

3. EMPIRICAL VERIFICATION

The method of obtaining equidistant Likert scores by assigning weights to different response categories of different items and method of linking those scores are shown empirically using hypothetical data involving 5 items in each of *k*- point scale, for *k*= 3, 4, 5 and 7 to 100 subjects.

3.1 Calculation of weights

Equidistant scores: Calculation of different weights to different response categories of different items as per the Stage 2 above is shown in Table 1.

Table 1. Calculation of weights to response categories of different Items

Item	Description	RC-1	RC-2	RC-3	RC-4	RC-5	Total
3-point items							
1	Frequency	15	45	40			100
	Proportions(ω_{1j})	0. 15	0.45	0.40			1.00
	Intermediate weights (W_{1j}) (α = 0.45)	0.15	0.30	0.35			0.80
	Final weights ($W_{1j(Final)}$)	0.1875	0.375	0.4375			1.00
2	Frequency	15	42	43			100
	Proportions(ω_{2j})	0.15	0.42	0.43			1.00
	Intermediate weights(W_{2j}) (α =0.42)	0.15	0.285	0.33			0.765
	Final weights ($W_{2j(Final)}$)	0.19608	0.37255	0.43137			1.00
3	Frequency	12	42	46			100
	Proportions(ω_{3j})	0.12	0.43	0.46			1.00
	Intermediate weights(W_{3j}) (α =0.51)	0.12	0.315	0.38			0.815
	Final weights ($W_{3j(Final)}$)	0.14724	0.38650	0.46626			1.00
4	Frequency	9	35	56			100
	Proportions (ω_{4j})	0.09	0.35	0.56			1.00
	Intermediate weights (W_{4j}) (α =0. 705)	0.09	0.3975	0.50			0.9875
	Final weights ($W_{4j(Final)}$)	0.09114	0.40253	0.50633			1.00
5	Frequency	8	48	44			100
	Proportions (ω_{5j})	0.08	0.48	0.44			1.00
	Intermediate weights (W_{5j}) (α =0.60)	0.08	0. 34	0.42667			0.84667
	Final weights ($W_{5j(Final)}$)	0.09449	0.40157	0.50394			1.00
4-point items							
1	Frequency	9	25	32	34		100
	Proportions (ω_{1j})	0.09	0.25	0.32	0.34		1.00
	Intermediate weights (W_{1j}) (α =0.33333)	0.09	0.21167	0.25222	0.2725		
	Final weights ($W_{1j(Final)}$)	0.10891	0.25613	0.30521	0.32975		1.00
2	Frequency	8	23	33	36	---	100
	Proportions (ω_{2j})	0.08	0.23	0.33	0.36	---	1.00
	Intermediate weights (W_{2j}) (α=0.37333)	0.08	0.22667	0.27556	0.30	---	
	Final weights ($W_{2j(Final)}$)	0.09068	0.25693	0.31234	0.34005	---	1.00
3	Frequency	9	21	34	36	---	100
	Proportions (ω_{3j})	0.09	0.21	0.34	0.36	---	1.00
	Intermediate weights (W_{3j}) (α =0.36)	0.09	0.225	0.27	0.2925		
	Final weights ($W_{3j(Final)}$)	0.10256	0.25641	0.30769	0.33333		1.00
4	Frequency	8	20	33	39		100
	Proportions (ω_{4j})	0.08	0.20	0.33	0.39		1.00
	Intermediate weights (W_{4j}) (α =0.41333)	0.08	0.24667	0.30222	0.33		
	Final weights ($W_{4j(Final)}$)	0.08343	0.25724	0.31518	0.34415		1.00
5	Frequency	7	29	26	38	---	100
	Proportions (ω_{5j})	0.07	0.29	0.26	0.38		1.00
	Intermediate weights (W_{5j}) (α =0.41333)	0.07	0.24167	0.29889	0.3275		
	Final weights ($W_{5j(Final)}$)	0.07462	0.25762	0.31863	0.34913		1.00

Table 1. Calculation of weights to response categories of different Items (Continued)

Item	Description	RC-1	RC-2	RC-3	RC-4	RC-5	Total
5-point items							
1	Frequency	3	9	21	28	39	100
	Proportions (ω_{1j})	0.03	0.09	0.21	0.28	0.39	1.00
	Intermediate weights (W_{1j}) (α =0.45)	0.03	0.24	0.31	0.345	0.366	
	Final weights ($W_{1j(Final)}$)	0.02324	0.18590	0.24012	0.26724	0.28350	1.00
2	Frequency	6	14	13	29	38	100
	Proportions (ω_{2j})	0.06	0.14	0.13	0.29	0.38	1.00
	Intermediate weights (W_{2j}) (α =0.40)	0.06	0.23	0.28667	0.315	0.332	
	Final weights ($W_{2j(Final)}$)	0.04903	0.18796	0.23427	0.25742	0.27132	1.00
3	Frequency	7	10	14	29	40	100
	Proportions (ω_{3j})	0.07	0.10	0.14	0.29	0.40	1.00
	Intermediate weights (W_{3j}) (α =0.4125)	0.07	0.24125	0.29833	0.32687	0.344	
	Final weights ($W_{3j(Final)}$)	0.05467	0.18841	0.23299	0.25528	0.26865	1.00
4	Frequency	8	8	16	30	38	100
	Proportions (ω_{4j})	0.08	0.08	0.16	0.30	0.38	
	Intermediate weights (W_{4j}) (α =0.375)	0.08	0.2275	0.27667	0.30125	0.316	
	Final weights ($W_{4j(Final)}$)	0.06659	0.18936	0.23028	0.25075	0.26302	1.00
5	Frequency	7	10	16	31	36	100
	Proportions (ω_{5j})	0.07	0.10	0.16	0.31	0.36	1.00
	Intermediate weights (W_{5j}) (α =0.3625)	0.07	0.21625	0.265	0.28937	0.304	
	Final weights ($W_{5j(Final)}$)	0.06115	0.18893	0.23152	0.25281	0.26559	1.00

Item	Description	RC-1	RC-2	RC-3	RC-4	RC-5	RC-6	RC-7
7-point items								
1	Frequency	5	6	7	17	18	22	25
	Proportions (ω_{1j})	0.05	0.06	0.07	0.17	0.18	0.22	0.25
	Intermediate weights (W_{1j}) (α=0.23333)	0.05	0.14167	0.17222	0.1875	0.19667	0.20278	0.20714
	Final weights ($W_{1j(Final)}$)	0.04318	0.12234	0.14873	0.16192	0.16984	0.17511	0.17888
2	Frequency	4	5 5	9 9	14 14	20 20	21 21	27 27
	Proportions (ω_{2j})	0.04	0.05	0.09	0.14	0.20	0.21	0.27
	Intermediate weights (W_{2j}) (α =0.60)	0.04	0.15417	0.19222	0.21125	0.22267	0.23028	0.23571
	Final weights ($W_{2j(Final)}$)	0.03110	0.11985	0.14944	0.16423	0.17311	0.17902	0.18325
3	Frequency	3	4	7	16	19	22	28
	Proportions (ω_{3j})	0.03	0.04	0.07	0.16	0.19	0.22	0.28
	Intermediate weights(W_{3j}) (α =0.60)	0.03	0.16667	0.21222	0.235	0.24867	0.25778	0.26429
	Final weights ($W_{3j(Final)}$)	0.02121	0.11782	0.15002	0.16612	0.17578	0.18222	0.18683
4	Frequency	4	4	6	19	21	20	26
	Proportions (ω_{4j})	0.04	0.04	0.06	0.19	0.21	0.20	0.26
	Intermediate weights(W_{4j}) (α =0.60)	0.04	0.14833	0.18444	0.2025	0.21333	0.22056	0.22571
	Final weights ($W_{4j(Final)}$)	0.03239	0.12012	0.14936	0.16398	0.17276	0.17861	0.18278
5	Frequency	5	3	7	30	20	13	22
	Proportions (ω_{5j})	0.05	0.03	0.07	0.30	0.20	0.13	0.22
	Intermediate weights(W_{5j}) (α =0.60)	0.05	0.1825	0.22667	0.24875	0.262	0.27083	0.27714
	Final weights ($W_{5j(Final)}$)	0.03294	0.12023	0.14933	0.16388	0.17261	0.17843	0.18258

*Legend: RC- j $\Longrightarrow$ j-th Response category for j=1, 2, 3, 4,5,6,7

The transformed scores are continuous and equidistant. For example, finally selected weights for Item 3 of 5-point scale were 0.05467, 0.18841, 0.23299, 0.25528 and 0.26865 respectively for response category 1, 2, 3, 4 and 5. Weighted scores of Item 3 of 5-point scale are equidistant, since $5W_{35(Final)} - 4W_{34(Final)} = 4W_{34(Final)} - 3W_{33(Final)} = 3W_{33(Final)} - 2W_{32(Final)} = 2W_{32(Final)} - W_{31(Final)} = 0.32215$

3.2 Breaking of ties

Raw scores resulted in large number of ties of subject scores. For 5-point sub-test, 96 out of 100 respondents got tied scores at various scores and length of tie ranged between 2 to 15. Tied raw score of 20 with 15 ties in 5-point sub-test and corresponding scores in equidistant scaling are shown in Table 2.

Table 2. Scores in the equidistant scaling corresponding to score of 20 in 5-point sub-test

Sl. no.	Corresponding score in equidistant scale	Sl. no.	Corresponding score in equidistant scale
1	5.10773	9	5.16128
2	5.11905	10	5.18149
3	5.15566	11	5.14873
4	5.15038	12	5.15875
5	5.08043	13	5.12324
6	5.13617	14	5.11686
7	5.06094	15	5.09754
8	5.15583		

Thus, the equidistant scores avoiding ties distinguish the same summative score on the basis of how the score was obtained.

3.3 Descriptive statistics

Raw scores of each item and sub-test were found to be negatively skewed. However, skewness of sub-tests ranged between -0.0026 (7-point scale) to -0.3865 (3-point scale) implying almost symmetric distributions. Some of item skewness ranged between -1 to -0.5 implying moderately skewed distributions. Mean variance and ranges of raw scores and scaled scores are shown in Table 3.

Table 3. Descriptive statistics

Description	Sub-test scores		Range of scores	
	Mean	Variance	Item	Sub-test
A. Raw scores				
3-point scale	11.70	3.08082	1 to 3	7 to 15
4-point scale	14.83	3.96071	1 to 4	10 to 19
5-point scale	19.16	6.82263	1 to 5	13 to 25
7-point scale	25.38	13.20768	1 to 7	14 to 34
B. Equidistant scores				
3-point scale	4.97611	1.23910	0.14724 to 1.51898	1.90471 to 7.03618
4-point scale	4.60777	0.70716	0.07462 to 1.39650	2.56414 to 6.34499
5-point scale	4.862951	0.712497	0.02323 to 1.41750	2.90749 to 6.76041
7-point scale	4.401098	0.57093	0.02121 to 1.30777	2.04788 to 6.19877
C. Standardized equidistant scores				
3-point scale	0.01	6.81222	-2.22866 to 1.06878	-7.05372 to 4.86195
4-point scale	0.0001	4.23828	-2.1184 to 1.11760	-4.99715 to 4.31887
5-point scale	0.00	4.50750	-2.61934 to 0.98113	-4.83517 to 4.78790
7-point scale	-0.00053	4.75793	-2.63255 to 1.32412	-6.85314 to 5.20309
D. Standardized equidistant scores converted to [1, 5]				
3-point scale	18.40581	12.52909	1 to 5	9 to 25
4-point scale	18.11001	7.052861	1 to 5	11.66668 to 23.66667
5-point scale	19.16009	6.822733	1 to 5	13 to 25
7-point scale	18.58606	5.866278	1 to 5	11 to 24.33333

Observations:

1. Mean and variance of raw scores differed for k-point sub-tests for k = 3, 4, 5 and 7. Values of mean and also variance tended to increase with increasing value of k, despite same number of items for each sub-test.
2. Distributions of raw scores for 3-point, 4-point, 5-point and 7-point sub-tests were different.
3. Equidistant scores reduced considerably mean and variance and made the data more homogeneous. Variance of sub-tests ranged between 0.57093 (for 7-point scale) to 1.23911(for 3-point scale). Item scores and sub-test scores were positive and continuous.
4. Standardized equidistant scores resulted in normality of scores of each item with zero mean and SD of one. Sub-test scores as sum of item scores also followed

normal distribution since it is well known that if X~ N(μ_X,σ_X^2) and Y ~ N(μ_Y,σ_Y^2) then (X + Y) ~ N(μ_X+ μ_Y,σ_X^2+σ_Y^2+2σ_(X,Y)). Variance of sub-tests varied between 4.23828 (4-point scale) to 6.81222 (3-point scale). However, items were not independent. Item scores and sub-test scores ranged from negative to positive.

5. When the Standardized equidistant scores of items were converted to [1, 5] by (1.1), means were found to be close for subtests, max difference being 0.05. But, variance of subtest varied between 5.87 to 12.53, due to different correlations between pair of items. Distribution of item scores and also sub-test scores were normal. Item scores ranged between 1 to 5. But range of sub-test scores differed to some extent for 3-point, 4-point, 5-point and 7-point scales.

3.4 Test scores using converted standardized scores

Test scores in each of the Approach 1.1, 1.2 and 1.3 for 3-point, 4-point, 5-point and 7-point sub-tests resulted in continuous, equidistant, normally distributed scores avoiding ties. Each respondent got unique integer valued rank. A rank correlation was maximum for ρ_1.1,1.3= 0.998 followed by ρ_1.1,1.2=0.997 and ρ_1.2,1.3=0.994. Test scores were all positive for Approach 1.2 and Approach 1.3. Mean, variance, CV and Score ranges of Test scores are shown in Table 4.

Table 4. Mean, variance, CV and Score ranges of Test scores

Description of Test scores	Mean	Variance	CV=$\frac{SD}{Mean}$	Score range
As sum of standardized equidistant scores i.e. *Z*-scores of all the items (Approach 1.1)	≈0	34.6329	0.16992	-14.6197 to 11.15762
As sum of standardized equidistant scores of items converted to [1,5] (Approach 1.2)	74.26196	56.04761	0.13357	55.66674 to 88.33346
As sum of *Z*- scores of all the items converted to $N(33,100)$ (Approach 1.3)	132	678.5452	0.19734	67.12237 to 182.2059

Lower value of CV in each of the approach and marginal difference in CV values indicate that number of response categories may not have much influence on variation about the mean. Approach 1.2 resulted in minimum value of CV.

Sum of variance of NCE scores of items was found to be higher than variance of NCE scores of the corresponding subtest. For example, $\sum_{j=1}^{5} var(NCE_{3-point,j})$ was 2218.249 against $var(NCE_{3-point\ subtest})$ = 443.65. CV of NCE test scores at the level of 0,273 was higher than CV of any of the approaches.

When test score as sum of standardized equidistant scores of items in [1,5] was further converted to equipercentile scores, large number of tied percentile scores emerged, which made it problematic to equate a particular percentile score in 3-point subtest with 4-point or 5-point or 7-point subtests.

3.5 Correlations

Correlations between subtests: When raw scores were considered, correlations between the subtests ranged between 0.0396 (4-point & 7-point) to 0.4797 (3-point & 4-point). Linear transformation used for converting scores at various stages like Raw scores → Equidistant scores as weighted sum → Z-scores → Z-scores converted to [1,5] or $N(33,100)$, did not change much the correlations between a pair of subtests.

Correlations between test scores: Test scores obtained by Approach 1.1, 1.2 and 1.3, based on linear function of Z-scores were highly correlated. $r_{(1.1,1.2)} = r_{(1.1,1.3)}$ = 0.998 and $r_{(1.2,1.3)}$ =0.996.

Correlations between subtest scores and Test scores: Correlation between subtest scores and Test scores are shown in Table 5.

Table 5. Correlation between subtest scores and Test scores

Subtests	Test scores		
	Approach 1.1	**Approach 1.2**	**Approach 1.3**
Subtest-1(3-point)	0.81225	0.83318	0.78286
Subtest-2(4-point)	0.63338	0.64887	0.64354
Subtest-3(5-point)	0.62721	0.62850	0.64505
Subtest-4(7-point)	0.51777	0.48407	0.53344

Correlation between subtest scores and Test scores ranged between 0.4841 to 0.8122. Subtest-1 was found to have maximum correlation with test scores under each approach. For the raw data, subtest 1(3-point) had correlation of 0.68 with the corresponding test scores. Raw scores → Equidistant scores by weighted sum modified r_(3-point,Test) to 0.88 followed by r_(4-point,Test)=0.65,r_(5-point,Test)=0.63 and r_(7-point,Test)=0.47. Similar trend was maintained in the three approaches because of use of linear transformation to obtain Z-scores from equidistant scores and subsequent stages.

3.6 Test reliability

Test reliability in terms of Cronbach alpha (without verification of assumptions) for raw scores and the three approaches are shown in Table 6.

For raw scores, α = 0.36. For Z-scores obtained after making the raw scores equidistant by assigning different weights to response categories of different items improved α to 0.44. Subsequent linear transformations to convert Z-scores to [1, 5] or to N(33,100) did not change much the Alpha. For NCE scores, α exceeded unity. For the Equivalent Forms resulting after converting the standardized equidistant scores to N (33,100), reliability of the subtests were found as correlation between a pair of such subtests. However, r_(Subtest i,Subtest j) differed for different values of i and j as can be seen from the Table – 6. Subtest reliability was found to be maximum for 3-point scale in line with maximum correlation with test scores under each approach.

Table 6. Reliability of test scores

Cronbach alpha (α)			
Raw scores	**Standardized equidistant scores(Z-Scores)**	**Standardized equidistant scores converted to [1, 5]**	**Standardized equidistant scores converted to $N(33, 100)$**
0.36294	0.43515	0.44655	0.432106

4. LIMITATIONS

Each Approach assumed equal importance to all items to facilitate addition.

5. CONCLUSIONS

Three approaches are described to convert and combine scores of Likert items with different response categories to obtain comparable test scores using only the permissible operations for a Likert scale i.e. considering the cell frequencies or empirical probabilities of Item – Response categories without making any assumptions of continuous nature or linearity or normality for the observed variables or the underlying variable being measured. A multi-staged method was followed to transfer Raw scores → Equidistant scores by data driven weights to response categories of different items → Z-scores → Z-scores converted to [1,5] or N(33,100) scores. Thus, these are assumption-free simple method which can be applied to a single sample.

Test scores generated by each of the three approaches are continuous, equidistant, normally distributed with zero ties. In case, frequency of a particular response category of an item = 0, it may be taken as zero value for scoring Likert items as weighted sum. Empirically, each respondent got a unique integer valued rank. Such test scores help in ranking or comparing or classifying the respondents, and also to undertake statistical analysis under parametric set up. The proposed approaches of combining Likert items of different response categories are critically relevant to practitioners and researchers in the social sciences in general and survey research in particular. Use of such methods of combining scores of Likert items is recommended for clear theoretical advantages and easiness in calculations.

Correlation between a pair of the Approaches was >0.99. Empirical results show marginal difference of Coefficient of variation (CV), indicating the extent of variability in relation to the mean and test reliability, in terms of Cronbach alpha across the Approaches. However, minimum CV and maximum alpha were found for Approach 1.2 where test scores were obtained as sum of standardized equidistant item scores converted to [1,5]. Because of a negative relationship between test reliability and CV, a lower value of CV is associated with higher reliability. Thus, the Approach 1.2 appears to have slight advantages.

Future studies may be undertaken (i) to assign weights to items after obtaining equidistant scores by assigning weights to different response categories of different items, so that item scores are equicorrelated with test scores i.e. equal importance to items – which may be followed by normalized scores (ii) to obtain measures of discriminating value of Likert item and also for Likert Scale using only the frequencies of Item – Response categories and simultaneously deal with reliability and discriminating value of combined scores of Likert items.

References:

Bendixen, M., & Sandler, M. (1995). Converting verbal scales to interval scales using

Correspondence analysis. *Management Dynamics: Contemporary Research*, *4*, 31-49.

Bentler, P. M. (1995). *EQS Structural Equations Program Manual.* Encino, CA: Multivariate Software Inc.

Colman, A. M., Norris, C. E., & Preston, C. C. (1997). Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. *Psychological Reports*, *80*, 355-362.

Cook, C. F. Heath, R., Thompson, L. ,& Thompson, B. (2001). Score reliability in web or internet-based surveys: unnumbered graphic rating scales versus Likert-type scales. *Educational and Psychological Measurement*, *61*, 697-706.

Cummins, R. A. (1997). *The Comprehensive Quality of Life Scale—intellectual/cognitive disability*, (ComQol-I5) (5th ed.). Melbourne: School of Psychology, Deakin University.

Cummins, R. A. (2003). Normative life satisfaction: Measurement issues and homeostatic model. *Social Indicators Research*, *64*, 225-240.

Dawes, John (2007). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research*, *50*(1), 61-77.

Dawes, J.G. (2002). Five point vs. eleven point scales: does it make a difference to data characteristics? *Australasian Journal of Market Research*, *10*(1), 39-47.

Ferrando, P. J. (2003). A Kernel density analysis of continuous typical-response scales. *Educational and Psychological Measurement*, *63*, 809-824

Finn, R.H. (1972). Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement*, *32*(7), 255–265.

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

Jabrayilov, R., Emons, W, H. M., & Sijtsma, K. (2016). Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment, *Applied Psychological Measurement*, *40*(8), 559-572. doi: 10.1177/0146621616664046

Johnson, S. M., Smith, P., & Tucker, S. (1982). Response format of the job descriptive index: assessment of reliability and validity by the multitrait-multimethod matrix. *Journal of Applied Psychology*, *67*(4), 500-505.

Lim, Hock-Eam (2008). The Use of Different Happiness Rating Scales: Bias and Comparison Problem? *Social Indicators Research*, *87*(2), 259-267

Livingston, S. A. (2004). *Equating test scores (without IRT).* Princeton, NJ: ETS.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*.

Hillsdale, NJ: Lawrence Erlbaum Associates

Mâsse,L. C., Allen,D., Wilson,M. & Williams, G. (2006): Introducing equating methodologies to compare test scores from two different self-regulation scales, *Health Education Research*, *21*(1), 110-120, https://doi.org/10.1093/her/cyl088

Mazaheri, M., & Theuns, P. (2009): Effects of varying response formats on self-ratings of life-satisfaction. *Social Indicators Research*, *90*, 381-395.

Mertler, C. A. (2002). *Using standardized test data to guide instruction and intervention.* College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Munshi, J. (2014). A Method for Constructing Likert Scales, *Social Science Research Network*. doi:10.2139/ ssrn.2419366.

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

Sangster, R. L., Willits, F. K., Saltiel, J., Lorenz, F. O., & Rockwood, T. H. (2001). *The effect of numerical labels on response scales*. Article presented at the annual meeting of the American Statistical Association, Atlanta, GA. http://www.bls.gov/osmr/pdf/st010120.pdf.

Saris, W. E., & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research.* Hoboken, New York, USA: Wiley-Interscience. Wiley series in survey methodology, ISBN 978-0-470-11495-7.

Schwarz, N., Knauper, B., Hippler, H. J., Noelle-Neumann, E., & Clark, W. (1991). Rating scales: Numeric values may change the meaning of scale labels. *The Public Opinion Quarterly*, *55*, 570-582. http://www.jstor.org/stable/2749407.

Wu, Chien-Ho (2007). An Empirical Study on the Transformation of Likert scale Data to Numerical Scores, *Applied Mathematical Sciences*, *1*(58), 2851-2862

Satyendra Nath Chakrabartty
Indian Ports Association
Snc12@rediffmail.com