Reflections on and test of the metrological properties of summated rating, Likert, and other scales based on sums of ordinal variables

Reference:
Kampen Jarl.- Reflections on and test of the metrological properties of summated rating, Likert, and other scales based on sums of ordinal variables
Measurement / International Measurement Confederation - ISSN 0263-2241 - 137(2019), p. 428-434
Full text (Publisher's DOI): https://doi.org/10.1016/J.MEASUREMENT.2019.01.083
To cite this reference: https://hdl.handle.net/10067/1593730151162165141

Reflections on and test of the metrological properties of summated rating, Likert, and other scales based on sums of ordinal variables

Kampen K. Jarl

PII: S0263-2241(19)30092-2
DOI: https://doi.org/10.1016/j.measurement.2019.01.083
Reference: MEASUR 6327

To appear in: *Measurement*

Please cite this article as: K.K. Jarl, Reflections on and test of the metrological properties of summated rating, Likert, and other scales based on sums of ordinal variables, *Measurement* (2019), doi: https://doi.org/10.1016/j.measurement.2019.01.083

Reflections on and test of the metrological properties of summated rating, Likert, and other scales based on sums of ordinal variables

January 25, 2019

Abstract

This study aims to contribute to the perpetual controversy on the parametric analysis of ordinal data, by giving a perchance long overdue examination of the widely held notion that sums of ordinal variables (e.g., Likert and summated rating scales) produce measures at ordinal level. In the present study, all 1,048,574 subscales of a well-known and widely applied sumscale, the 20-item CESD scale for depression, were assessed for their metrological properties. It was found that subscales consisting of less than 60% of the items of the original scale have lost all metrological properties of that scale, including ordinality as measured by Kendall's tau. This result justifies concern about the robustness of measurement scale properties of (shortened) sumscales, and by implication, of the empirical findings based on such scales.

Keywords: *Summated rating scale; Likert scale; ordinal; measurement level; metrology; CESD; depression*

1 Introduction

The summated rating scale (SRS) is one of the most frequently used data collection tools in the social sciences, psychology, and health studies. This type of measurement is highly likely to end up as part of a portfolio measurement in interdisciplinary research [1]. Its conception is attributed to Rensis Likert [2], which is why a scale consisting of the sum of ordinal scaled ("Likert-type") items is often called "Likert scale." Among the documented reasons for use of Likert scales and SRSs are that a well-developed scale can have good reliability and validity (i.e. psychometric properties), the scale is relatively cheap and easy to develop, is usually quick and easy for respondents to complete, and typically does not induce complaints from them [3]. This paper concerns all scales produced by the summation of ordinal variables, including Likert and summated rating scales, and refers to such scales as *sumscales.*

1.1 Building a sumscale

From standard textbooks, the procedure to produce sumscales that are called valid and reliable proceeds through five steps. In the first step, the construct to be measured must be defined. A construct is a theoretical abstraction, like an unobservable cognitive state, often with no known objective reality. Defining good constructs is in and by it self complicated [4], and merits its own discussion for which this paper is not the proper place. In the remainder, we shall inappropriately assume that we are dealing with well-defined constructs (i.e., they have familiarity, resonance, parsimony, coherence, etc.). This means that we have a set of indicators reflective of the construct of interest.

The second step is to design (Likert-type) items by formulating statements that reflect the attitude toward/opinion about/cognitive state of, each indicator. These items are typically measured at ordinal level. Using Kampen & Swyngedouw's [5] classification of types of ordinal variables, items for Likert scales are of Type 5 (unstandardized discrete variables with ordered categories). However, other types of ordinal variables are very common in building sumscales, and in this paper we will examine a sumscale composed of items of Type 1 (categorized metric variables with known thresholds), which have a material standard required for expressing metrological properties.

Ideally, in step 3 the formulation of items is followed by a first pilot study, when the items and the sumscale are tested with units purposively selected to stress the instrument. Items can be adjusted if necessary, such that the resulting scale is adequate with respect to the target population. This first pilot study must be followed by a second pilot, when the items are tested in a larger sample for consistency.

That is, the fourth step proceeds by performing an item analysis, a process of trial and error whereby 'item-remainder coefficients'(correlations of each item with the sum of the remaining items) are computed, items with (near) zero coefficients are deleted, and scores of items with negative coefficients will be reversed (recoded). The resulting scale's internal consistency is computed by Cronbach's alpha, which has it own problems outside the scope of this paper [6, 7].

In step number 5, the scale is validated (addressing the question: what does the sumscale measure if anything). First, validation is done at item level (face & content validity), and then at the level of the scale (criterion validity, discriminant validity, convergent validity).

After successful completion of these 5 steps, the sumscale can start its lifetime as a validated instrument for measuring the construct of interest.

1.2 Analysing a sumscale

While sumscales are perhaps easy to develop and present low burdens for respondents, the analysis of such scales is problematic because its metrological properties are subject to controversy. Controversy regards their level of measurement. Following Stevens [5, 8, 9], measurements can have different measurement scales. If we denote a given attribute by X, a measurement of the attribute by M, and consider two objects i and j, the following definitions apply:

1. Nominal scale: If $x_i = x_j$ then $m_i = m_j$ and when $x_i \neq x_j$ then $m_i \neq m_j$

2. Ordinal scale: At least nominal, and if $x_i < x_j$ then $m_i < m_j$

3. Interval scale: At least ordinal, and $x_i - x_j = \beta(m_i - m_j)$ for some $\beta > 0$

4. Ratio scale: At least interval, and $x_i \div x_j = m_i \div m_j$

Nominal and ordinal scales are also known as categorical or "qualitative" measurement scales, while interval and ratio scales are known as metric or "quantitative" measurement scales.

Different levels of measurement require different means of data analysis. For instance, association of variables measured at metric level is usually done by Pearson correlations (parametric analysis); analysis of associations of categorical variables is appropriately done by Pearson's chi-square for independence (non-parametric analysis). With respect to the analysis of sumscales, the controversy is between those claiming that parametric analysis of such data is "illegal" because the sumscale is measured at ordinal level, against so-called "liberals" taking a pragmatic approach to its analysis and simply treat it as measured at interval level [10, 11].

Neither the liberals nor the pragmatists offer any explanation about the scientific principle (or miracle) that is at work transforming sums of ordinal items into (ordinal or metric) measurement scales. Throughout literature, the assumption that the level of measurement of the sumscale is at least ordinal is hardly ever contested. However, taking one step back from the issue as to whether or not a sumscale can or cannot be meaningfully analysed by parametric means, one may question validity of the assumption that sumscales are ordinal scales. To give a simple example of the potential problem at work, imagine three drugs issued in three doses measured at ordinal scale: low (0), medium (1), lethal (2). The scores on the sumscale range between 0 and 6, which we can in turn classify as low (0−2), medium (3−4), and high (5−6). One can immediately spot the flaw in this measurement procedure, because having a low score on the sumscale can mean that a lethal dosis was issued (score 2 on one, score 0 on the other drugs). Such a sumscale cannot be considered to be useful.

Exactly when is a scale useful? Thurstone [12, 13] proposed 6 criteria:

1. Unidimensionality: A universal characteristic of measurement is that it describes only one attribute

2. Linearity: Measurement implies some linear continuum

3. Abstraction: The linear continuum is an abstraction (e.g., a measure is assumed to be valid for a given period in the past and the future)

4. Invariance: The procedure of measurement is always the same

5. Sample-free callibration: The scale must transcend the group measured and its function must be independent of the object of measurement

6. Test-free measurement: It should be possible to omit several items in a test and still obtain the same individual score (measure)

Applying these criteria to the sumscale will render it useless by definition. First, its unidimensionality won't be detected by Cronbach's alpha. Second, its measurement scale is obscure (but the present paper will try to shed some light on this issue). Third, its abstraction fails on lack of unidimensionality and linear continuum. The fourth criterion, invariance, can only be satisfied if everybody gets the same test in the same mode (in-person, self-administration), and in the same context (content of questionnaire, length of the questionaire, purpose of the questionaire, etc.) [14, 15]. Fifth, sample-free calibration is violated because in applied research, item selection and critical thresholds are based on samples under study. Finally, omitting items in a sumscale seriously affects resulting scores (no test-free measurement). But perhaps Thurstone's criteria are simply too rigorous in the realm of measuring cognitive states, and we had better stick to Steven's (1951) less demanding definition, where "measurement is the attaching of numbers to objects in a meaningful way" [5, 8, 16, 17].

1.3 Shortening Existing Sumcales

To add one further complication besides validation and analysis of sumscales, researchers are often tempted to use shortened sumscales in order to gather more information about a respondent in less time. Assume a sumscale constructed of P items conducted in a sample of N respondents. The number of subscales that can be constructed from this set of items equals $2^P - 2$ [18]. Each scale can be identified by a unique 2^P-bit binary code. For instance, assuming $P = 20$, the three item scale including items 4, 5 and 17 can be conveniently denoted in the P-vector **v** = (0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0). These code vectors can be collected in the $P \times (2^P - 2)$ matrix **V**, and together with the $N \times P$ data matrix **X**, the $N \times (2^P - 2)$ super matrix **Y** = **VX** can be computed that consists of all possible subscales. Please note that the last vector of **Y** contains the complete P-item sumscale. This complete sumscale can be considered as the gold standard. In order to make comparisons between the subscales convenient, a simple transformation can be applied such that each scale ranges between 0 and 100. Assume that all items take values between 0 and c, in which case the vector $c\mathbf{V1}$ contains the maxima of the subscales. The diagonal matrix **T** = $100 \times \mathrm{diag}(c\mathbf{V1})$ contains these maxima on the diagonal, and **M** = $\mathbf{VXT}^{-1}$ is the rescaled matrix of all subscales.

While researchers should keep validity and reliability on their mind when they shorten scales, it remains an issue as to what extent shortened sumscales preserve the metrological properties of the parent sumscale [19]. It was stated by Nunnally [20] that apparently, ". . . no one has made it clear what types of evidence would justify the assumption of a particular type of scale." Nevertheless, if a gold standard exists which one can safely assume to coincide with the attribute of interest, any measurement instrument can be tested for its capacity to correctly

1. Classify different objects as different and equal objects as equal (required for nominal scale measurement)

2. Rank the objects from low to high (required for ordinal scale measurement)

3. Quantify the relative differences between objects (required for interval scale measurement)
4. Quantify the ratio of objects (required for ratio scale measurement).

Like calibrating measurement, assessing measurement scale properties can only be done in the presence of a gold standard (that is decisive with respect to issues of equality and distance of objects). Precisely *because* the assumption that sumscales are measured at ordinal level is uncontested in the literature, shortened summated rating scales can be assessed for nominal or ordinal measurement level by using the (complete) sumscale as gold standard. Moreover, if we (unjustifiably) assume the sumscale is measured at interval scale, the shortened scales can be tested for interval and ratio measurement as well. Below, an example of such study is given, where all 1 million subscales of a well-known and widely applied sumscale, the 20-item CESD scale for depression, are assessed for their metrological properties. Data and method, an account of the results, and a discussion of the main findings are in the next three sections, followed by a conclusion acknowledging the limitations of the produced evidence.

2 An empirical assessment of the metrological properties of shortened sumscales

2.1 *Criteria for measurement scale assessment*

Metrology excludes from its definition nominal and ordinal variables because measurement units are neither defined nor supported by material standards. So when we assume metrological properties of (shortened) sumscales, we assume that these scales satisfy the requirements for interval and ratio measurement. This assumption must be verified by measurement scale assessment of each subscale, which requires evaluation of properties 1 through 4 listed above for all pairs of observations. A scale that preserves these properties is not only meaningful in Stevens' sense, but also obeys Thurstone's test-free measurement criterion for useful measurement.

Let m_i and m_{ip} denote the score of the ith individual on the complete sumscale and the pth subscale respectively, both rescaled such that they range from $0 - 100$ (strictly speaking, we have $m_{iP} = m_i$). The subscales in **M** have simple statistical properties such as mean, variance, standard error, percentiles, critical threshold (e.g., the highest p percent) that can be compared to the complete sumscale at aggregate level. Measurement scale assessment, however, requires analysis of the (sub)scales at individual level.

Nominal level of measurement means that individuals with equal values on the gold standard will have equal values on the subscale, and individuals with unequal values on the gold standard must have unequal values on the subscale. A pair of measurements for individuals i and j is defined to satisfy the requirement for nominal scale measurement if we have $m_{ip} = m_{jp}$ when $m_i = m_j$, and $m_{ip} \neq m_{jp}$ when $m_i \neq m_j$. In other words, the higher the percentage of pairs consistently classified as (un)equal, the more the subscale respects the nominal scale of measurement. Of special interest are the sensitivity and the specificity of the subscales [21, 22]. Assume a critical region of the gold standard defined by a cut point θ, with perc$(m > \theta)$ the percentage of respondents having scores higher than θ. The sensitivity of a subscale is defined as the percentage of cases ranked in the critical region both by the subscale and by the gold standard, and the specificity of a subscale is the percentage of cases ranked outside the critical region both by the subscale and by the gold standard (this definition of sensitivity appeals to category β, type 5, of Mencattini and Mari's classification [22]).

Ordinal level of measurement is respected when a subscale ranks individuals in the same way as the gold standard. A pair of measurements satisfies the requirement for ordinal scale when $(m_{ip} - m_{jp})(m_i - m_j) \geq 0$ for all possible pairs. In other words, concordant pairs in the subscale must be concordant pairs in the original (gold standard) scale, and similarly, disconcordant pairs in the subscale must be disconcordant pairs in the parent scale. The literature has produced several measures of concordance, of which Kendalls tau-b [23] is perhaps the most appropriate because it corrects for ties.

While testing subscales for nominal and ordinal scale measurement can be justified on the "Baron of Munchhausen" argument that the parent original scale is measured at ordinal level, such (however weak) justification is lacking when we assess for metric scale measurement. Nevertheless, the same "as if" logic applies: we will simply assume that the parent scale is metric and test the properties of the subscales accordingly. That is, a pair of measurements satisfies the requirement for interval scale when $m_i - m_j = \beta_p(m_{ip} - m_{jp})$ for some $\beta_p > 0$. Because this equality assumes a linear relationship of the gold standard and the subscale, a quick assessment for interval level measurement is Pearson's product moment correlation coefficient. Finally, a pair of measurements satisfies the requirement for ratio scale when $m_i \div m_j = m_{ip} \div m_{jp}$, or more easily, $m_i \div m_{ip} = m_j \div m_{jp} = 1$.

2.2 Data: The CES-D scale for depression

The original CES-D scale is a 20-item self-report scale developed by the Center for Epidemiologic Studies of the National Institute of Mental health to gauge depression in the general population [24]. The National Health and Nutrition Examination Survey (NHANES-I), collected in 1971-1975, administered the full 20-item CES-D to a subsample of 3,059 of adults aged between $25 - 74$ [25]. Respondents were asked to select how often they had experienced each of the described symptoms during the last week, on a 4-category response scale ($c = 3$, see Section 1.3), ranging from "rarely or none of the time (less than 1 day)" to "most or all of the time (5-7 days)." The CES-D, as measured on the NHANES-1, includes 20 items, for instance, "I was bothered by things that usually don't bother me," "I thought my life had been a failure," etc. Listwise deletion of missing data in NHANES-I resulted in $N = 2,414$ in the final data subjected to the analysis.

A commonly-used (though arbitrary) cut point for the CES-D is $\theta = 16$ (see Section 2.1), which according to Radloff [24] distinguishes the general population from psychiatric patients; 70% of psychiatric patients scored 16 or higher, compared to 21% of the general population. So the region containing the highest 21% of the scale is considered the critical region, which we adapt as criterion for computing sensitivity and specificity in the below analyses.

Several shortened versions of the full 20-item CES-D scale have been used in community, population, and clinical studies. For instance, the Health and Retirement Study, Wave 2 and beyond [26], and Round 3 of the European Social Survey (ESS-3) [27],

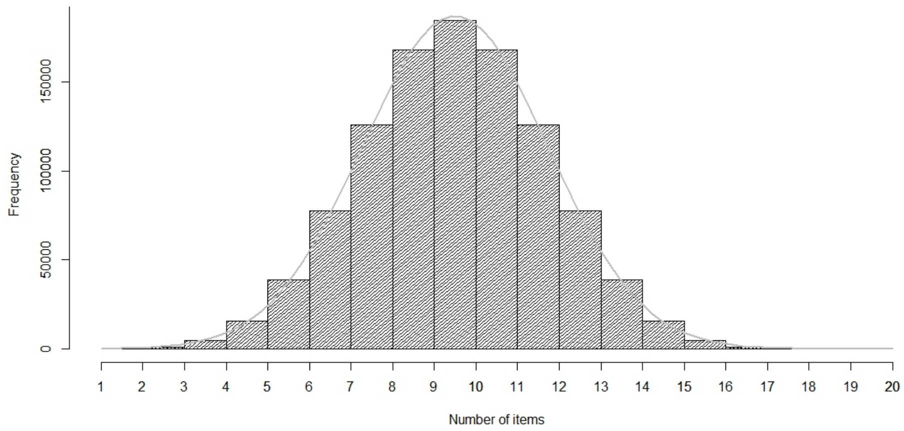

Figure 1: *Frequency histogram of the number items within each subscale of the 20-item CES-D sumscale*

adapted an 8-item shortened version of the CES-D scale consisting of the items “depressed,” “effort,” “sleep,” “happy,” “lonely,” “enjoy,” “cry,” “sad” and “get going.” Performance of this subscale, which we refer to as the “ESS scale,” will be compared to all other 8-item subscales. Admittedly, the common tendency has been to plunge into analysis of data without having a clear idea as to when a single dimension exists and when it does not [28]. However, a principal components analysis produced a four-factor solution reported by Radloff [24], who identified factors as

1. Depressed affect
2. Positive affect
3. Somatic and retarded activity
4. Interpersonal

This four factor solution was replicated in the data used in the present analysis. Because multidimensionality may affect the measurement properties of the scales, the dimension with the highest number of item loadings $>$ .35, “Depressed Affect,” was subjected to a similar analysis as the full scale (meaning that all subscales were examined for metrological properties). The .35 demarcation criterion was applied in the source article. The “Depressed Affect” scale included 8 items which were denoted as “bothered,” “blues,” “depressed,” “failure,” “fearful,” “lonely,” “cry,” and “sad.” This scientifically derived 8-item scale will, just like the full 20-item scale, be thoroughly studied in terms of all of possible subscales (in casu, 254).

2.3 Computations and analysis

On the basis of $P = 20$, the total number of proper subsets (excluding the empty subset) of the CESD scale equals 1,048,574. For the sake of clarity, the frequency histogram of the number of items in these subscales (NItems) is displayed in Figure 1. Computations of the one million odd subscales were done in a script written for the software package R (available at request). Whereas computations were fairly straightforward, the mere number of computations made the process a bit time consuming.

In the statistical analysis for measurement scale assessment, for each scale the following statistics were computed ("ideally" means, under the condition of *homomorphic mapping* [29]):

1. Nominal scale: Sensitivity and specificity (all ideally equal 100%)

2. Ordinal scale: Kendall's tau (ideally equals 1)

3. Interval scale: Pearson's r, explained variance $R^2 = r^2$ (ideally equals 1)

4. Ratio scale: The ratio $m_i \div m_{ip}$ (ideally equals 1).

Of course in practice, we have to allow for some measurement error that will lead to deviations from the ideal situation. In particular subscales consisting of only few items must be expected to suffer from a lack of resolution (they have only very few scale points).

In addition to the above statistics, also the mean, variance and root mean squared error (RMSE; comparing the subscale mean with the mean of the original 20-item 'gold standard' sumscale) of the subscales will be reported. These results will be aggregated to the number of items in each subscale.

Finally, in order to assess the relative importance of individual items on measurement scale preservation, two quality indicators QI(.8) and QI(.9) are introduced. Quality indicator QI(.8) equals 1 when sensitivity, Kendalls tau and R^2 (the square of the correlation r between the sub- and sumscale) of a subscale are all bigger than .8 and equals 0 otherwise, and QI(.9) is defined analogously. Item importance is measured by the probability that an item is included in a subscale satisfying the respective quality constraints: the higher this probability, the more important is the item.

2.4 Results

The main results of the analysis in function of the number of items in the subscale are in Table 1, with special attention for the 8-item ESS scale. Diagnostics with respect to the nominal scale of measurement include sensitivity and specificity of the subscales. Order preservation is determined by Kendall's tau. Other printed statistics include Pearson's correlation r, and the average ratio of individual observations. Descriptive statistics of the subscales are printed in the last four columns of this table (the scale mean, variance, standard error and mean root squared error).

Generally speaking, these results show that the quality of a subscale in terms of reflecting the metrological properties of the parent sumscale increases with the number of constituent items. The relationship between the number of items of the subscale and

Table 1. Measurement scale properties of shortened sumscales

# Items	*Sens.*	*Spec.*	*Kend. τ*	*r*	*Ratio*	*Mean*	*MVar*	*SE*	*MRSE*
1	0.38	0.92	0.40	0.53	0.97	13 .83	686 .6	44 .87	5 .58
2	0.43	0.94	0.50	0.66	1 .0	13 .97	418 .0	20 .44	3 .68
3	0.50	0.94	0.56	0.74	1 .0	13 .97	328 .8	12 .87	2 .90
4	0.58	0.94	0.61	0.79	1 .0	13 .97	284 .1	9 .08	2 .43
5	0.63	0.94	0.65	0.83	1 .0	13 .97	257 .3	6 .81	2 .11
6	0.66	0.95	0.68	0.86	1 .0	13 .97	239 .5	5 .29	1 .86
7	0.73	0.93	0.71	0.88	1 .0	13 .97	226 .7	4 .21	1 .65
8	0.72	0.95	0.74	0.90	1 .0	13 .97	217 .2	3 .40	1 .49
ESS scale	0.77	0.98	0.77	0.93	1.19	16.95	303.4	8.82	3.00
9	0.74	0.95	0.76	0.91	1 .0	13 .97	209 .7	2 .77	1 .34
10	0.76	0.96	0.78	0.93	1 .0	13 .97	203 .8	2 .27	1 .21
11	0.78	0.96	0.81	0.94	1 .0	13 .97	198 .9	1 .85	1 .10
12	0.80	0.97	0.83	0.95	1 .0	13 .97	194 .9	1 .51	0.99
13	0.82	0.97	0.85	0.96	1 .0	13 .97	191 .4	1 .22	0.89
14	0.84	0.97	0.86	0.96	1 .0	13 .97	188 .5	0.97	0.79
15	0.85	0.97	0.89	0.97	1 .0	13 .97	185 .9	0.75	0.7
16	0.86	0.98	0.91	0.98	1 .0	13 .97	183 .7	0.56	0.61
17	0.89	0.98	0.93	0.98	1 .0	13 .97	181 .7	0.4	0.51
18	0.91	0.98	0.94	0.99	1 .0	13 .97	179 .9	0.25	0.41
19	0.93	0.99	0.97	0.99	1 .0	13 .97	178 .4	0.11	0.29
20	1 .0	1 .0	1	1 .0	1 .0	13 .97	177 .0	0	0

the quality of that scale in terms of preserving nominal (sensitivity), ordinal (Kendall's tau) and metric (R^2) scale of measurement is graphically displayed in Figure 2. Sensitivity, Kendall's tau-b and squared correlation increase as the number of constituent items increases, reaching acceptable levels only when the number of items is 15 or more.

Table 2 presents the by-item analysis of measurement scale preservation. A scale satisfying QC(.8) consists on average of 12.3 items, and a scale satisfying QC(.9) consists of 15.7 items. The probability for a randomly selected item to be part of a scale satisfying QI(.8) therefore equals .61 and this inclusion probability equals .79 for QI(.9). A set of 12 items have higher inclusion probability then these two benchmarks (arranged on the left side of Table 2). With respect to the quality indicator QI(.9), there are 2,267 (.22%) subscales that satisfy the property that sensitivity, Kendall's tau and R^2 are greater than or equal to .90, while 227,220 (21.7%) subscales satisfy the (lenient!) constraints for QI(.8).

A summary of the main results for the 8-item scale tapping into the dimension of "Depressed Affect," analogous to that of the 20 item scale, is in Figure 3. It leads to the same conclusions as those based on the complete 20-item scale, and detailed discussion of its statistics will therefore be omitted in order to save space. No 8-item subscale satisfies QI(.9) while only 797 out of 125,970 (or .78%) satisfy QI(.8). The scale consisting of the top 8 items in Table 2 has fairly average values of sensitivity, Kendall's tau and R^2 (resp. .75, .80 and .83). While no 8 item subscale combines the maxima of all three parameters, one of the best performing 8 item subscales reaches values for specificity = .71, Kendall's tau-b = .83 and R^2 = .88. This 8-item subscale

Table 2. Measurement scale preservation by item: probability that an item is included in a scale satisfying the quality constraints*

Item	QI(.8)	QI(.9)	Item	QI(.8)	QI(.9)
Sleep	.66	.93	Bothered	.60	.78
Good	.58	.89	Blues	.61	.75
Effort	.68	.88	Fearful	.60	.75
Talk	.65	.86	*Enjoy*	.59	.71
Hopeful	.66	.85	Failure	.58	.70
Lonely	.64	.85	*Happy*	.61	.69
Appetite	.62	.85	Cry	.56	.68
Depressed	.66	.83	Unfriendly	.55	.66
Sad	.64	.81	Dislike	.55	.64
Get going	.64	.81			
Mind	.62	.80	(*Av. # Items*	12.3	15.7)

* The quality constraints QI(.8) and QI(.9) correspond to sensitivity, Kendall's tau and R^2 bigger than .8 and .9 respectively. The items that are used in the ESS 8-item subscale are italic. Items are ordered by QI(.9). So for instance, 93% of all subscales that satisfy QI(.9) contain the item "Sleep," while only 64% of these subscales contain "Dislike." Therefore, "Sleep" can be considered to be more important than "Dislike" in preserving the metrological properties of the parent sumscale, i.e. the 20-item CES-D scale.

differs from the "depressed affect" and "ESS" 8-item subscales.

2.5 Discussion

The mean of the original 20-item scale is equal to 14.0, while subscale means can be as low as 4.83 and as high as 29.06, which violates Thurstone's criterion of test-free measurement. This variability can also be inferred from the average root mean squared error (RMSE) comparing, the mean of the subscale over 2814 respondents with that of the original scale. If we allow a margin of error of 1 point of the scale or less, MRSE suggests that 12 (or 60% of total) items or more are required.

Except when the sumscale equals zero (which means all items equal zero), the nominal scale is not preserved by the subscales. Even the adequacy of the low-resolution Boolean classification determined by the subscale's sensitivity is disappointingly low for most of the subscales. That is, sensitivity of scales constructed of 3 items or less does not outperform a coin toss, subscales constructed with less than 12 items have at most 80% sensitivity, and scales built with less than 17 items have a sensitivity of at most 90% (see Table 1 and Figure 2). Specificity on the other hand, is always 90% or higher.

Order of observations is preserved fairly well in subscales with 16 items or more (Kendall's tau $>$ .90), and reasonably well in subscales made of 11 or more items (Kendall's tau $>$ .80). But the ordinal level of measurement quickly erodes in subscales built from 10 or less items. Turning to the correlations of the subscales and the original scale, subscales constructed of 7 items or less correlate .9 or less with the original scale, which corresponds to 81% or less explained variance. If 90% or more of variance of

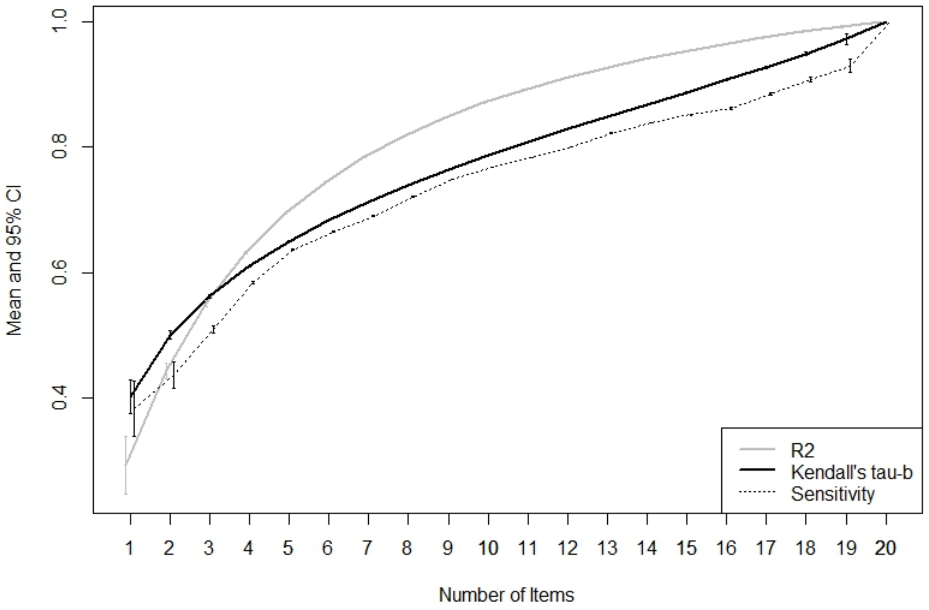

Figure 2: *Average sensitivity, Kendall's tau and square Pearson correlation with 95% confidence intervals of the subscales using the 20-item original CESD scale as gold standard.*

the original scale is to be explained by the subscale ($r > .95$), at least 12 items are needed. Considering ratio scale of measurement, on average the subscales reproduce the original scores as can be seen from the ratio the (individual) subscores and the original scores, which equals 1.0 for most subscales. However, individual scores can be easily under- or overestimated by 50% or more.

Conclusions on the effects of using shorter versions do not change drastically when we look into the 8-item scale of items about "Depressed Affect," see Figure 3. If we apply a benchmark of .8 for sensitivity, Kendall's tau and R^2, again a minimal number of 60% of the original constituent items is required to obtain mildly acceptable metrological properties of the subscales.

Finally, a word on the 8 item subscale used in, for instance, the ESS (see Section 2.2). The statistics for the ESS scale (see Table 1) reveal that within the subscales consisting of 8 items it performs fairly well by comparison, with sensitivity $= .77$, Kendall's tau $= .77$, and $R^2 = .86$, although its mean (and associated measures) is significantly higher (16.9). That said, even the best performing 8-item scale fails to satisfy the QI(.9) quality constraints.

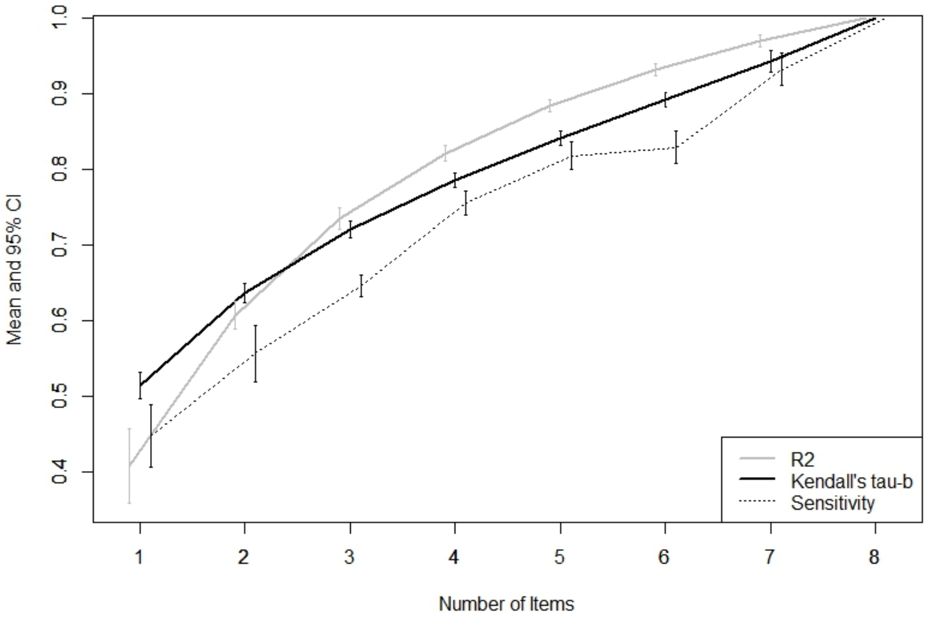

Figure 3: *Average sensitivity, Kendall's tau and square Pearson correlation with 95% confidence intervals of the subscales with 8-item subscale "Depressed Affect" as gold standard*

3 Conclusions

Summated rating, Likert, or sumscales don't obey Thurstone's criteria for useful measurement. What remains for critical assessment of their metrological properties is their meaningfulness in terms of Stevens' criteria regarding nominal, ordinal, interval and ratio scale measurement. A subscale can be assumed to be a valid measure of the construct measured by the parent scale *if and only if* it has high correlation with that parent scale. In the present study, applying a validated 20-item sumscale as the gold standard for all possible subscales, the analyses showed that measurement scale characteristics suffer from shortening. Scales of 60% of the total number of items or less have basically lost all metrological properties of the original scale. Uni- or multidimensionality doesn't appear to play a role in measurement scale preservation.

To illustrate the undesirable consequences of loss of correlation in applied research, consider two different sumscales tapping into two different dimensions while having a (conceptually relevant) correlation of .6 (or $R^2 = .36$). If shortened counterparts correlate .9 with their parent scales, which means the subscales obey the QI(.9) quality constraints, the correlation of the shortened scales will equal $.9 \times .6 \times .9 = .49$ (or $R^2 = .24$). This number drops to .38 (or $R^2 = .14$) when the shortened scales correlate only .8 i.e. obey the QI(.8) quality constraints. And perhaps easily overlooked, but this deflationary effect of correlations also occurs when empirical correlations found in shortened scales are to be replicated using full scales.

While it is tempting to conclude that a reliable sumscale should contain at least 60% of all items of the parent scale, the reader must be aware that these results are limited to the CESD scale under study and that other properties of the applied items, such as the number of response options, the negative-positive orientation of these options, the order and number of items, may have significant impact on the quality of subscales in other applications. In many cases, the constituent ordinal items also lack the material standard present in Type 1 ordinal variables, that is, categorized metric variables where classification is done with reference to the units of a metric variables (here, the days of experiencing a given affect). In other words, it may well be the case that in other applications of sumscales matters are even worse than in this particular data.

Finally, while finding that measurement scale properties are scarcely preserved in shortened sumscales is one thing, explaining this lack of robustness is another. At least two rival hypotheses can be formulated. The first is, that not even the parent sumscale is measured at ordinal measurement level, and therefore it is unreasonable to expect that random assemblies of constituent items will reproduce its metrological properties. The second hypothesis is, that the parent sumscale is valid but simply doesn't allow for (meaningful) shortening. Absent a gold standard that is statistically reliable and enjoys universal consensus regarding its validity, proving either hypothesis is impossible. Nevertheless, the results of this study justify concern about the robustness of measurement scale properties of shortened sumscales, and by implication, of the empirical findings based on such scales. The inescapable and highly inconvenient conclusion is that there is no miracle that transforms sums of ordinal items into an ordinal or interval scale.

References

[1.] H. Tobi, Measurement in interdisciplinary research: The contributions of widely-defined measurement and portfolio representations, Measurement 48 (2014) 228-231.
[2.] R. Likert, A technique for the measurement of attitudes, Archives of Psychology 22 (1932) 1-55.
[3.] P. E. Spector, Summated rating scale construction: an introduction, SAGE, London, 1996.
[4.] J. Gerring, What makes a concept good? A critical framework for understanding concept formation in the social sciences, Polity 31 (1999) 357-393.
[5.] J. K. Kampen & M. Swyngedouw, The ordinal controversy revisited, Quality & Quantity 34 (2000) 87-102.
[6.] L. J. Cronbach, Internal consistency of tests: analyses old and new, Psychometrika 53 (1988) 63-70.
[7.] K. Sijtsma On the use, the misuse, and very limited usefulness of Cronbach's alpha, Psychometrika 74 (2009) 107-120.
[8.] Stevens, S. S. (1951). *Handbook of experimental psychology*, Wiley, New York, 1951.

[9.] G. B. Rossi, Cross-disciplinary concepts and terms in measurement, Measurement 42 (2009) 88-96.
[10.] T. R. Knapp, Treating ordinal scales as interval scales: an attempt to resolve the controversy, Nursing Research 39 (1990) 121-123.
[11.] S. Jamieson, Likert scales: how to (ab)use them, Medical Education 38 (2004) 1212-18.
[12.] L. L. Thurstone, Measurement of social attitudes, Journal of Abnormal and Social Psychology, 26 (1931) 249-269.
[13.] B.D. Wright, A history of social science measurement, Educational Measurement: Issues and Practice 16 (1997) 33-45.
[14.] J. K. Kampen, The impact of survey methodology and context on central tendency, nonresponse and associations of subjective indicators of government performance, Quality & Quantity 41 (2007) 793-813.
[15.] N. M. Bradburn & S. Sudman, The current status of questionnaire research. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz and S. Sudman (Eds.), Measurement errors in surveys, John Wiley & Sons, Inc, Hoboken, NJ, 2004.
[16.] L. Finkelstein, Widely-defined measurement – An analysis of challenges, Measurement 42 (2009) 70-77.
[17.] L. Mari, A quest for the definition of measurement, Measurement 46 89-95.
[18.] J. M. Stanton, Empirical distributions of correlations as a tool for scale reduction, Behavior Research Methods, Instruments, & Computers 32 (2000) 403-406.
[19.] J. M. Stanton, E. F. Sinar, W. K. Baltzer & P. C. Smith, Issues and strategies for shortening the length of self-report scales, Personnel Psychology 55 (2002). 167-194.
[20.] J. C. Nunnally, Psychometric Theory, McGraw-Hill, New York, 1967.
[21.] D. G. Altman, Practical statistics for medical research, Chapman-Hall, London, 1991.
[22.] A. Mencattini & L. Mari, A conceptual framework for concept definition in measurement: The case of 'sensitivity', Measurement 72 (2015) 77-87.
[23.] A. Agresti, Categorical Data Analysis, Wiley, New York, 1990.
[24.] L. S. Radloff, A self-report depression scale for research in the general population, Applied Psychological Measurement 1 (1977) 385-401.
[25.] J. M. Zich, C. C. Attkisson & T. K. Greenfield, Screening for depression in primary care clinics: the CES-D and the BDI, The International Journal of Psychiatry in Medicine 20 (1990) 259-277.
[26.] D. E. Steffick, Documentation of affective functioning measures in the Health and Retirement Survey, HRS/AHEAD Documentation Report, Survey Research Center, University of Michigan, 2000.
[27.] F. A. Huppert, N. Marks, A. Clark, J. Siegrist, A. Stutzer, J. Viterso, & M. Wahrendorf, Measuring well-being across Europe: Description of the ESS Well-being Module and preliminary findings, Social Indicators Research 91 (2009) 301-315.
[28.] L. Guttman, The basis for scalogram analysis. In: S. A. Stouffer et al. (eds.), *Studies in social psychology in WOII. Vol 4: Measurement and prediction*, Wiley, New York, 1950.
[29.] S. V. Muravyov & V. Savolainen, Representation theory treatment of measurement semantics for ratio, ordinal and nominal scales, Measurement 22 (1997) 37-46.