AUCTORES
Globalize your Research
Psychology and Mental Health Care
Satyendra Nath Chakrabartty*
Open Access
Review Article

Measurement Issues in Mental Health

Satyendra Nath Chakrabartty *

Indian Ports Association, Indian Statistical Institute, Indian Maritime University

***Corresponding Author:** Satyendra Nath Chakrabartty, Indian Ports Association, Indian Statistical Institute, Indian Maritime University.

Received date: December 11, 2023; **Accepted date:** December 27, 2023; **Published date:** January 09, 2023

Citation: Satyendra N. Chakrabartty, (2024), Measurement Issues in Mental Health, *Psychology and Mental Health Care,* 8(1): **DOI:10.31579/2637-8892/241**

Abstract

The paper discusses limitations of existing self-reported measures of mental health and suggests transforming ordinal item scores to continuous, monotonic scores satisfying desired properties including meaningful aggregation. Here, an item scores with response-categories marked as 1, 2, 3, 4, 5, and so on are combined by [1] transforming to normally distributed scores, [2] angular similarity and [3] function of Geometric Mean (GM) for measuring mental health of individuals. Each proposed measure avoids selection of weights, scaling and considers all chosen domains, even if they have different correlations with scale scores. Each method can identify poorly performing domains and assess overall improvement/decline of a patient across time. However, changes over time need to be validated with clinical findings. Proposed mental health scores help better comparisons, ranking, classifying and testing of mean and variance for a sample. Approach [3] may be preferred for additional features like constant domain-elasticity, time-reversal test, chain indices.

Keywords: mental health; likert items; geometric mean; cosine similarity; monotonic; responsiveness

Introduction

Mental Health (MH), an integral part of general health and well-being is a basic human right (WHO, 2022). Measurement of multidimensional Mental Health is complex because of differences in cultures and social and psychological confounders, methodological limitations of measurement from scales/tests, etc. Psychological, social and behavioral features associated with mental illness vary in conceptualizations and are difficult to measure. Measurement of Mental Health is complex because of its multidimensional nature, differences in cultures and social and psychological confounders, methodological limitations of measurement from scales/tests, etc. Psychological, social and behavioral features associated with mental illness vary in conceptualizations and are difficult to measure. Need for community-based MH systems and services were outlined by World Health Organization (2021). Measurement issues of MH are important for accurate diagnosis, assessing severity; monitoring, tracking path of recovery of both individuals and sample. MH systems and services were outlined (WHO, 2022). Measurement issues of MH are important for accurate diagnosis, assessing severity; monitoring, tracking path of recovery of both individuals and sample. Attempts to measure outcomes of MH frequently using International standards for diagnosis of mental illnesses are: International Classification of Disease version 10 (ICD-10) and Diagnostic Statistical Manual version 5 (DSM-V). However, DCM–V contains heterogeneous diagnostic categories since pragmatic criteria undermine the diagnostic model (Kate et al. 2019). Structured Clinical Interview for DSM-IV and Best-estimate consensus diagnoses showed poor agreements and use of diagnostic interviews in clinical contexts are questionable (Kvig and Nilssen, 2023). Major purposes of such tools are to identify cases, screen those at risk of developing mental disorder and monitor the progress, classify, compare and rank individuals and also to track impact of interventions/ treatments.

While diagnostic tools assess clinical symptoms involving clinical interviews and multi-expert assessments, screening tools attempt to assess severity of a mental health disorder and track changes of one or a group of patients or response to treatments.

Commonly used tools to assess MH include:

- Self-reported Likert or Numerical rating scales (NRS) for psychiatric diagnoses or psychiatric medications
- Rigorous psychiatric interview by trained psychiatrist or clinical psychologist

Self-assessed scales and interviews cover a limited range of problems like anxiety and depression and may not capture isolated, short-lived cases or mild-cases requiring early treatments. Major concerns are scoring based on nature of data generated from such MH measuring tools and non-satisfaction of properties like monotonically increasing continuous scores along with their responsiveness, reliability, sensitivity, **specificity**, etc. Significant variations in the questionnaires for 16 common depression identification tools were found (Williams et al. 2002).

The paper aims at reviewing limitations of existing self-reported measures of mental health and suggesting transformations so that the transformed scores facilitate meaningful application of operations and satisfy desired properties of measurement.

Literature survey:

Tolls for assessing MH differ in terms of number and format of items, scoring methods, dimensions considered viz. clinical (depression, anxiety, schizophrenia, etc.) and social (social support, etc.), and are not comparable. For example, 109 different measurement tools in health literature were identified for social isolation only (Cordier, 2017). Illustrative assessment tools along with their uses, features and observations are given in Table-1.

Tool	**Uses**	**Features**	**Brief observations**
Clinical Interview Schedule – Revised (CIS-R) (Lewis, et al. 1992)	Diagnostic of specific Mental Health disorders (GAD, depression, panic disorders, phobias, OCDs and CMD-NOS)	Validation done in minority and ethnically diverse populations	Diagnosis of depressive disorder using CIS-R may not be practical in large surveys (Head et al. 2013). CIS-R is moderately valid and recommended much lower CIS-R cut-point (Jordanova et al. 2004)
General Health Questionnaire (GHQ) (Goldberg et al. 1997)	Measuring psychological distress and general mental wellbeing. Often used to assess severity of psychological distress of a person or population.	Self-completed 4-point questionnaire, available with 12, 28, 30 or the full 60 items. Four scoring methods: GHQ scoring (0-0-1-1); Likert scoring(0-1-2-3); Modified Likert scoring (0-0-1-2) and C-GHQ scoring (0-0-1-1) for positive items and 0-1-1-1 for negative items).	Factor analysis (FA) of GHQ 12 showed 2–3 factors against the claim of one-dimensional tool. Separate use of the factors has no practical advantages (Gao et al. 2004). Cronbach's alpha is used for reliability, interrater and intrarater reliability, even violating one-dimensional assumption of alpha (Montazeri et al. 2003). Response bias on the negative items exists (Hankins, 2008). GHQ 28 had negative correlation with QOL2: mental health (Alexopoulos et al. 2014) K 6 & K 10 had better psychometric properties than GHQ 12 (Cornelius et al. 2013)
Mini International Neuropsychiatric Interview (MINI) (Sheehan et al. 1998)	Diagnostic assessment of both ICD-10 MH and DSM-IV/V categories.	10 items, each of 4-point from 0 (do not agree at all) to 3 (agree fully) and a Visual Analog scale (VAS, 0 to 100).	Some questions are problematic and few are seen as extreme. Results could be biased by interpretation and the extent of guessing. Can be used as first step in outcome tracking in clinical settings.

36 item Short Form survey (SF-36) (Mishra et al. 2014)	Quality monitoring purposes, and Medicare assessments	- 36-items in 8 domains differ in format ("Yes-No" type, 3-point and 5-point Likert items) -Raw scores (X) are transformed to [0, 100] by Z = $\frac{X-Min_X}{Range\ of\ X}(100)$ where 0 score implies maximum disability.	Manual of SF 36 does not allow computation of $SF36_{Total}$ since several independent dimensions are being measured by SF-36. If Min_X is changed, ranking may be changed due to change in marginal rates of substitution (Seth and Villar, 2017). Negatively correlated with PHQ and GAD-7 (Johnson et al. 2019)
Kessler Psychological Distress Scales (K 6 and K 10) (Kessler et al. 2002)	For assessing Non-specific psychological distress (as a proxy for case or non-case of serious DSM-V mental illnesses). Six K-6 items, each in 5-point scale (0 – 4) assess patients' feelings, symptoms during last 30 days	- K-6 score is weighted sum of frequencies where weights are scale-values attached to the levels. Sum of weights $\neq 1$ implying deviation from convex-set. -Range of discrete score is [0, 24]. -Mental illness is severe if K-6 score ≥ 13	-Sensitivity of the tool questioned. - No consensus on dimensional structure and cut-off score for identification of moderate psychological distress. -Cronbach's alpha may not be valid since tau-equivalent property of all items is not established and the scale is not unidimensional. -Kappa and weighted kappa as reliability have limitations (Prochaska, et al. 2012)
Centre for Epidemiological Studies – Depression too(CES–D) (Lewinsohn et al. 1997)	-Depression specific screening assessment. -Correlates with DSM-V -Used as an indicator of symptom severity.	Questionnaire with 20 numbers of 4-point items (0 – 3). Subjects rate how frequently each item is applied to them over the past week. Higher scores⇒ more symptomatology	-Latent factor structure and item content are major areas of concern (Manea et al. 2014) -Validity and psychometric properties of several items have been questioned (Radloff, 1977)
Patient Health Questionnaire (PHQ) (Carleton et al. 2013)	For screening of depression defined by DSM-IV; diagnosis of major depressive disorder and monitoring of impact of treatment in terms of severity of symptoms.	- Self-reported Likert questionnaire; 2 items (PHQ 2) and 9 items with 4 levels (PHQ 9). Higher summative score imply higher depression severity	-*The algorithm scoring showed* low sensitivity for detecting major depressive disorder (MDD). - The PHQ was particularly limited in identifying depressed individuals with dysthymia(Cheng et al. 2006)

		Different methods for scoring PHQ include an algorithm based on Diagnostic and Statistical Manual of Mental Disorders.	
Geriatric Depression Scale (GDS) (Manea, et al. 2014)	Screening of risk of depression and assessing severity of depressive Symptoms in elderly populations	15 or 30 items (Yes – No types) are distributed over domains like Physical health. Mental health, Functional, Social and Environmental issues with equal importance.	Equal importance to the items and domains are not justified. Contribution of a domain to GDS may vary. Addition of domain scores assumes a higher score of Physical health can substitute lower score on Mental health.
General anxiety disorder questionnaire (GAD–7)	Used for screening of severity of anxiety symptomatology, and monitoring severity progress after diagnosis	- Self- reported Likert scales with 7 items, each having 4 levels -Higher score indicates more severe GAD symptoms.	Attempt to evaluate latent structure of GAD-7 through one-factor CFA failed as the model did not fit the data (Eack et al. 2006). No cut-off scores had adequately balanced sensitivity and specificity.
Community Screening Instrument for Dementia (CSI–D)	A 32-item cognitive test and a 26-item informant interview. Used for assessment of cognitive deficit. Score represents severity of cognitive impairment and dementia symptoms	Scores: 1.cognitive score (COGSCORE), item weighted total score for the cognitive test (lower score⇒ worse cognitive status) 2. Informant score (RELSCORE), Unweighted total score from the informant interview(higher score⇒ greater decline in cognitive and functional status) 3.Discriminant function score (DFSCORE), weighted score combining the COGSCORE and RELSCORE using an algorithm developed by its originators(high DFSCORE ⇒ cognitive impairment)	Influence of education was not eliminated by COGSCORE alone. But, RELSCORE was not affected by education. DFSCORE reduced effect of education and improved overall performance (increased areas under the ROC curves) (Yesavage et al. 1982)

Table 1: Illustrative Assessment Tools

Observations:

Most of the MH assessing tools use summative scores of Likert items/NRS suffer from following limitations:

- Unequal and unknown distance between levels (Rutter and Brown, 2017)
- Assumes equal weight to items and dimensions despite different item-total correlations, factor loadings, etc.
- Strictly speaking, arithmetic mean is not defined for ordinal scales and $\bar{X} > \bar{Y}$ is meaningless. Use of mean and SD for ordinal scales was disfavored (Liu et al. 2005).
- Different responses to different items can generate the same summative score for several respondents and cannot discriminate the respondents with tied score.
- Mean and variance tend to increase with increase in number of levels. Estimated mean is more influenced by number of response-categories, than the underlying variable (Wu, 2007).
- Likert scales with 2-point, 3-point, and 4-point items performed poorly on reliability, validity and discriminating power (Jamieson, 2004).
- Zero as an anchor value lowers mean, variance and distorts skew, kurtosis of scales and does not permit computation of expected value as product of value and probability of the value. Too many zero responses to an item artificially lower correlation with that item.

Possible solutions:

(i) Convert scores of Likert items with equal number of response-categories, to ratio scale using frequencies of levels to get continuous, equidistant and monotonic scores (Chakrabartty, 2020).

(ii) For items with different number of response-categories, transfer raw item-wise scores ensuring satisfaction of equidistant property, followed by normalizing and further rescaling to a desired range and combining such scores to obtain test scores which are normally distributed (Chakrabartty, 2020).

Major limitations of Kappa and weighted Kappa or kappa reliability coefficient as used in K 6 and K 10 to find degree of agreement among the raters are:

- A low kappa does not imply low agreement (Chakrabartty, 2019 ; Bajpai et al. 2015) Confidence interval for Kappa ≤ 0.60 may be surmised as large volume of incorrect evaluation of data (Simundic, 2008). For ordered categories, methods of deciding weights for weighted kappa vary and may give different values of weighted kappa.
- Concept of agreement in terms of Kappa or weighted kappa and concept of reliability of test/scale are different. No measure for inter-rater reliability is in line with definition of reliability as ratio of true score variance and observed score variance.

Other Limitations:

SF-36 was negatively correlated with GAD 7 and PHQ, presumably due to different domains measured by each of them. Multi-domains MH tools give equal importance to the domains. Such equal importance or no weights amounts to a compensatory approach, without differentiating essential and less important domains. As a result, low score of one domain gets countered by a high score of other domain. Theoretically, the domains may be given weights (considering relative importance of domains) and take MH score (*Y*) as a weighted sum. Here, 'trade-off' between a pair of domains since $\frac{W_1}{W_2}$ is the amount of domain-2 that needs to be sacrificed to gain an extra unit of domain-1. Weights from Principal Component Analysis (PCA) poorly weigh those items which do not have strong correlations with *Y*, even if they are theoretically and practically important. Thus, PCA ignores judgments as to what are important. Assumptions of PCA include relatively homogeneous large sample size, normality of item scores, etc. If one variable has a SD which far exceeds the rest the variable, it will dominate the first eigenvector. Moreover, PCA weights vary over time and space and thus comparisons become difficult. No weighting system is above criticism (Greco et al. 2019). Similarly, there is no perfect aggregation scheme.

Possible solution:

Multi-dimensional MH score (*Y*) may be defined by cosine similarity between the two vectors showing domain scores of the current period and base period or by geometric mean (GM) of ratios of current domain scores and respective domain score for the base period, to accommodate all relevant domains and facilitate computation of *Y* for an individual and also for a group of individuals.

Proposed methods:

Ignoring the issues of selection of indicators, following methods are proposed for measurement of multi-dimensional MH score (*Y*) avoiding scaling of raw data and choosing weights.

Pre-processing of data:

- Assign 1, 2, 3, 4, 5, .. to the response-categories of items avoiding zero which keeps invariant the nature of generated data .
- Ensure each item is positively related to MH. Take reciprocal of each item whose lower value implies higher MH value.
- Convert Likert scores to Ratio scale.

Method 1: For Likert items with equal number of response-categories, method suggested by [33] is described below:

Let X_{ij} be the raw score of the *i*-th individual in the *j*-th item, for $i = 1,2,\dots,n$ and $j = 1,2,\dots\dots,m$. For a 5-point item, $X_{ij} = 1, 2, 3, 4$ and 5.

I: For *i*-th item find positive weights (W_{ij}) which are different for different levels and $\sum_{j=1}^{5} W_{ij} = 1$ satisfying the equidistant condition i.e. $W_1, 2W_2, 3W_3, 4W_4, 5W_5$ forms an Arithmetic Progression. A positive value of the common difference will ensure $5W_5 > 4W_4 > 3W_3 > 2W_2 > W_1$

One way to find such weights are:

i) Let f_{ij} be the frequency of *i*-th item for the *j*-th level. For each item, find maximum (f_{max}) and minimum frequency (f_{min}).

ii) Find proportions $\omega_{ij} = \frac{f_{ij}}{n}$. Note, $\omega_{ij} > 0$ and $\sum_{j=1}^{5} \omega_{ij} = \frac{\sum_{j=1}^{5} f_{ij}}{n} = 1$.

iii) Put initial weights $W_{i1} = \omega_{i1} = \frac{f_{min}}{n}$. Find the common difference $\alpha = \frac{5f_{max} - f_{min}}{4n}$.

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 $W_{i5} = \frac{\omega_{i1} + 4\alpha}{5}$

Here, $W_{ij} > 0$ and $\sum_{j=1}^{5} W_j \neq 1$.

iv) Get final weights $W_{ij(Final)} = \frac{W_{ij}}{\sum_{j=1}^{5} W_j}$ so that $\sum W_{ij(Final)} = 1$

Weighted sum of raw scores gives equidistant scores (*E*) and provides meaningful arithmetic aggregations.

II: Normalize the scores obtained at I by $Z = \frac{E - \bar{E}}{SD(E)} \sim N(0,1)$.

III: Take further weights to items to satisfy additional property of making the test scores equi-correlated with the items i.e. equal item reliability and thus justify addition of such converted item scores.

Method 2: For Likert scale consisting of subtests consisting of 3-point, 4-point, 5-point, 6-points items.

I: 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 and repeat Stage I of Method 1separately for each sub-test.

II: Take *Z*- scores for each item. For the *i*-th item, $Z_{ij} = \frac{E_{ij} - \bar{E}_i}{SD(E_i)} \sim N(0,1)$. Sub-test score as a sum of item scores will also follow $N\left(0, \sqrt{\sum Z_i^2 + 2\sum_{i\neq j} Cov(Z_i, Z_j)}\right)$

III: Convert *Z*-score of an item to Y_i in the range say [1, 100], by:

$$Y_i = \frac{(99)*(Z_i - Min(Z_i))}{Max(Z_i) - Min(Z_i)} + 1$$

Distributions of item scores for each *K*-point scale will be normal. However, 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.

IV: To have same distribution of different sub-test scores, further transformation may be used as follows:

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

Modified test scores for each *K*-point scale will be $N(Proposed\ mean, Proposed\ SD)$. Thus, the *K*-point subtests for various values of *K* could be considered as Equivalent Forms having features of parallel tests.

Methodology:

For one-dimensional tools:

Use Method 1 to find *Y* for a tool which is one-dimensional. For multi-dimensional tool, find scores of a domain consisting of Likert items by the above said method. Such scores of one-dimensional tools or domain scores are continuous satisfying equidistant property with a fixed zero point and has the following advantages:

i. Higher value indicates higher value of MH or domain score

ii. Generate monotonic scores since choice of *j*-th level will result in higher score than the choice of (*j*-1)-th level for any item for *j* =2, 3, 4, 5

iii. Rank a group of patients uniquely avoiding ties unlike the usual summative scores.

iv. Possible to find sample mean and SD for a group of patients.

v. If X_{it} denotes severity of the *i*-th patient in *t*-th time period, then $\frac{X_{it} - X_{i(t-1)}}{X_{i(t-1)}} \times 100$ will indicate percentage of progress/deterioration registered by the *i*-th patient in *t*-th time in comparison to (*t-1*)-th time period i.e. responsiveness of the scale.

For multi-dimensional tools:

Let $\boldsymbol{X_{m\times n}}$ be the matrix for *m*-persons and *n*-domains where each row vector $\boldsymbol{X_C} = (X_{1c}, X_{2c}, \ldots \ldots, X_{nc})^T$ represents scores of *n*- domains in the current period of a person. Here, $X_{ic} > 0\ \forall i = 1,2,\ldots,n$ have been obtained after the data pre-processing presented above. Let corresponding base period vector is $\boldsymbol{X_0} = (X_{10}, X_{20}, \ldots \ldots, X_{n0})^T$. Let θ be the angle between $\boldsymbol{X_C}$ and $\boldsymbol{X_0}$. The domains may be independent or correlated with varying degrees.

MH score (*Y*) of an individual combining the domain scores without considering correlations among the domains and avoiding selection of weights and normalization of domain scores are proposed as follows:

a) Cosine similarity approach:

Angular similarity approach proposed by (Chakrabartty, 2019) is adopted to combine the domain scores to get MH score (*Y*) as follows:

$$Y_{C0} = Cos\theta = \frac{X_C^T X_0}{\|X_C\| \|X_0\|} \quad (1)$$

where $\|X_C\|$ and $\|X_0\|$ are length of $\boldsymbol{X_C}$ and $\boldsymbol{X_0}$ respectively.

Here $0 \leq Cos\theta \leq 1$.

The equation (1) reflects overall achievement made by a person over the base period. It can also be taken as a disability intensity of a person at current period which is a continuous variable and offers a uni-variate platform for parametric analysis. Higher value of $Cos\theta_i$ implies the patient is close to the “No symptoms” status and lower value implies the patient is away from the “No symptoms” status. Lower values of $Cos\theta$ make the data more homogeneous. Patients can be ranked with respect to $Cos\theta_i$. The measure also helps to classify the patients into two or more non-overlapping classes. $Cos\theta_{i1} > Cos\theta_{i0} \Longrightarrow$ the *i*-th patient has improved in period 1 from the 0-th period. The ratio $\frac{Cos\theta_{i1}}{Cos\theta_{i0}} > 1$ quantifies progress made by the *i*-th patient and $\frac{Cos\theta_{i1}}{Cos\theta_{i0}} < 1 \Longrightarrow$ the patient has deteriorated and treatment plans, cares need to be looked into. Thus, the ratio reflects responsiveness of the tool to evaluate effect of interventions on a patient when disability is measured by $Cos\theta_i$. Norms of such ratio or difference may be determined statistically that is clinically important. Association between *i*-th and *j*-th person can be evaluated by $Cos\theta_{ij} = \frac{X_i^T X_j}{\|X_i\| \|X_j\|}$ for $i \neq j$.

Averaging of $Cos\theta_i$ for a group of persons is not meaningful as $Cos\theta_i$ does not obey triangle inequality. Mean and dispersion of angles $\emptyset_1, \emptyset_2, \emptyset_3, \ldots \ldots \emptyset_k$, can be obtained for vectors of unit length (Rao, 1973).

Example of computation of overall MH by $Cos\theta$ using hypothetical data involving 6 individuals and 4 domains, where value of each indicator was improved by 1 unit in the current period is given in Table-2.

Individual	Base period or previous period				Current period				$Y = Cos\theta \times 100$	$Y = \frac{X_{ic}}{X_{i0}} \times 100$
	D-1	D–2$	D-3	D-4	D-1	D–2 $	D-3	D-4		
1	114	0.033003	32	25.7	115	0.031949	33	26.7	99.9967	104.6245
2	120	0.038462	28	17.3	121	0.037037	29	18.3	99.9961	106.3797
3	104	0.045455	32	76.3	105	0.043478	33	77.3	99.99876	100.895

4	123	0.037736	20	16.4	124	0.036364	21	17.4	99.99568	108.2242

Legend: D_i denotes the i-th Domain. $: Reciprocals of the negatively related domain

Table 2: Computation of proposed measures for aggregating domain scores

The table shows similarity in ranks of individuals with respect to values of Y by $Cos\theta(100)$ and $\frac{X_{ic}}{X_{i0}}(100)$. This tends to indicate linear relationship between the two proposed methods.

Mean or most preferred direction is estimated by $\bar{\emptyset} = Cot^{-1}\frac{\sum \cos\emptyset_i}{\sum \sin\emptyset_i}$ and the dispersion by $\sqrt{1-r^2}$ where $r^2 = (\frac{\sum Cos\emptyset i}{k})^2 + (\frac{\sum Sin\emptyset_i}{k})^2$

Convert $\boldsymbol{X_C}$ and $\boldsymbol{X_0}$ to $\boldsymbol{\pi_C}$ and $\boldsymbol{\pi_0}$ where $\boldsymbol{\pi_C} = \sqrt{\frac{X_i}{\|X_C\|}}$ and $\boldsymbol{\pi}_{\mathbf{0}} = \sqrt{\frac{1}{\|X_0\|}}$ so that $\|\pi_C\|^2 = \|\pi_0\|^2 = 1$ and compute sample mean by $Cos(\bar{\theta}) = Cos\,(Cot^{-1}\frac{\sum \cos\theta_i}{\sum \sin\theta_i})$ and sample dispersion as $\sqrt{1-r^2}$

b) Geometric Mean approach:

A function of Geometric mean of the unit-free positive ratios $\frac{X_{ic}}{X_{i0}}$ for i=1, 2, …., n is considered to combine domain scores to get Y_{c0} (XXX anonymized for peer review) as follows:

$$Y_{c0} = \frac{X_{1c}, X_{2c}, \ldots\ldots, X_{nc}}{X_{10}\, X_{20}\, \ldots\ldots X_{n0}} \quad (2)$$

$Y_{c0} > 1 \implies$ Overall improvement of a person from the base-period. Quantification of progress of the i-th person in period t over $(t\text{-}1)$ th period is given by $Y_{i_t} - Y_{i_{(t-1)}} > 0$ or $\frac{Y_{i_t}}{Y_{i_{(t-1)}}} > 1$. Progress and decline of the i-th domain at c-th time period over the base-period are indicated respectively by $\frac{X_{ic}}{X_{i0}} > 1$ and $\frac{X_{ic}}{X_{i0}} < 1$. The domains where deterioration took place can be easily observed by observing the values of $\frac{X_{ic}}{X_{i0}}$. The proposed index Y_{c0} for the i-th person can be taken as intensity of mental disorder of the person and thus helps to rank a group of patients in terms of mental disorder intensity.

GM approach is applicable for data in ratio scale or ordinal scale or in percentages and even for skewed longitudinal data and snap-shot data. It is not affected much by extreme values (outliers) and produces no bias for measuring disease intensity of a patient. Level of substitutability among the variables is reduced significantly since low value of one variable does not get linearly compensated by high values in another variable.

It may be noted that (2) is the GM^n. Considering distribution of GM which approaches lognormal, computation of mean and variance of MH for a sample suggested as $e^{\mu_X + \frac{\sigma_X{}^2}{2}}$ and $e^{2\mu_X + \sigma_X{}^2}(e^{\sigma_X{}^2} - 1)$ respectively where ln(Y) = X ~ $N(\mu_X, \sigma_X)$ (Alf and Grossberg, 1979).

Discussion:

(1) and (2) are simple, avoid scaling and selection of weights. Each of (1) and (2) may be multiplied by 100 for general convention. Each measure satisfies the following:

- Reflects overall improvement/decline of a person across time by a continuous function which increases monotonically showing responsiveness of measurement of MH
- Independent of change of scale
- Reduced substitutability among the domains; not affected much by outliers and satisfies the principle of population replication (Herrero et al. 2010)
- Can be computed for properly defined sub-groups say gender, socially and economically backward groups, elderly people with specific morbidity, etc.
- Possible to compute mean and variance of MH score for a group of individuals.
- Individuals may also be compared in terms of progress made from base period or on Year-to-Year basis

(2) has additional features like:

- 1% increase in $X_{ic} \implies$ increase in Y if all others remain unchanged.
- Critical domains are those for which $\frac{X_{ic}}{X_{i0}} < 1$ or $\frac{X_{it}}{X_{i(t-1)}} < 1$
- Relative contribution of the domains to Y can be quantified easily.
- Satisfies Time-reversal test since $Y_{t0}.Y_{0t} = 1$.
- Possible to form chain-indices since $Y_{20} = Y_{21}.Y_{10}$. Chain-indices help to draw path of improvement/decline since the base period.

Thus, the proposed method in terms of (2) with higher desirable properties is an improvement over the existent measures.

Conclusions:

After reviewing major limitations of measuring mental health, the paper proposed methods of converting item-wise ordinal Likert scores to normally distributed scales, with equal and different number of response-categories for arithmetic aggregation of item scores. For combining domain scores, the paper proposed two indices in terms of angular similarity and function of Geometric Mean (*GM*) for measuring mental health. Each measure is non-parametric, simple, avoids scaling or finding weights or reduction of dimensionality and considers all chosen domains and indicators. Scores generated by each of the method were continuous, monotonic and assess progress/deterioration of a patient across time. Each depicts overall improvement or decline of a patient or a sample of patients in the current year with respect to base year or on Year-to-Year basis and facilitates better comparison, ranking, classification and assessing paths of progress. However, changes over time need to be validated with clinical findings. Measure based on of angular similarity and function of *GM* reduce level of substitutability among the indicators, not affected much by outliers and satisfies the principle of population replication. Both satisfy desired properties like monotonically increasing continuous function, assessment of responsiveness, which in turn helps drawing of path of improvement/decline over time. It is possible to compute mean and variance of mental health for a group of persons. Normality helps in estimating/testing population parameters.

Each proposed measure can be used to find mental health scores (Y) of a uni-dimensional tool or domain scores for multi-dimensional tool. GM approach is preferred for its additional features like linearity between gain in a domain and gain in mental health, time-reversal test, easy identification of critical areas requiring attention and contribution of the domains/indicators to the mental health.

Simulation studies with multi dataset to explore issues relating to dimensionality and rank robustness of tools and to find distribution of

$Y_{C0} = Cos\theta$ and relationship between $Cos\theta$ and $\prod_{i=1}^{n} \frac{X_{ic}}{X_{i0}}$ are suggested for future studies.

Declarations:

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

Conflicts of interest/Competing interests: The author has no conflicts of interest to declare

Ethical approval: Not applicable since the paper does not involve human participants.

Consent of the participants: Not applicable since the paper does not involve data from human participants

Data Availability statement: The paper did not use any datasets

Code availability: No application of software package or custom code

Authors' contributions: Sole Author

Reference:

1. Alexopoulos EC, Palatsidi V, Tigani X, Darviri C. (2014). Exploring Stress Levels, Job Satisfaction and Quality of Life in a Sample of Police Officers in Greece, *Safety and Health at Work*, 5(4), 210–215.
2. Alf, EF and Grossberg, JM (1979). The geometric mean: Confidence limits and significance tests, *Perception &Psychophysics*, 26 (5), 419-421
3. Bajpai, S., Bajpai, R. and Chaturvedi, HK.(2015). Evaluation of Inter-Rater Agreement and Inter-Rater Reliability for Observational Data: An Overview of Concepts and Methods, *Journal of the Indian Academy of Applied Psychology*,41(3), 20-27
4. Carleton RN, Thibodeau MA, Teale MJ, Welch PG, Abrams MP, Robinson T, *et al.*(2013). The center for epidemiologic studies depression scale: a review with a theoretical and empirical examination of item content and factor structure. *PLoS One*; 8(3):e58067.
5. Chakrabartty, Satyendra Nath (2020). Better Use of Scales as Measuring Instruments in Mental Disorders. *Journal of Neurology Research Reviews & Reports*. SRC/JNRRR-128.
6. Chakrabartty, Satyendra Nath (2019). Alternate measure of disability intensity: Modified Rankin Scale, *Journal of Stroke Medicine*,1-8.
7. Chakrabartty SN and Talukdar, GC. (2022).Statistics in Cancer: diagnosis, intensity, treatment efficacy and patient survival studies (473 – 488) in *Cancer diagnostics and therapeutics - Current trends, challenges and future* (Ed.) Basu, SK; Panda, CK and Goswami, S. eBook ISBN 9811647518, Springer Nature,
8. Cheng, ST, Chan AC, Fung HH.(2006). Factorial structure of a short version of the Center for Epidemiologic Studies Depression scale. *Int J Geriatr Psychiatry*, 21: 333–336
9. Cordier, R.(2017). A systematic review evaluating the psychometric properties of measures of social inclusion. *PLoS One*, 9; 12(6).
10. Cornelius, BL, Groothoff, JW, van der Klink, JJ, Brouwer, S.(2013).The performance of the K10, K6, and GHQ-12 to screen for present state DSM-IV disorders among disability claimants. *BMC Public Health.* 13:128.
11. Eack SM, Greeno CG, Lee BJ.(2006). Limitations of the patient health questionnaire in identifying anxiety and depression: Many cases are undetected. *Res Soc Work Pract.*;16:625–631.
12. Gao F, Luo N, Thumboo J, Fones C, Li SC and Cheung YB. (2004). Does the 12-item General Health Questionnaire contain multiple factors and do we need them? *Health Qual Life Outcomes*, 2, 63.
13. Goldberg DP, Gater R, Sartorius N, Ustun TB et al. (1997). The validity of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychological medicine*, 27,191-197
14. Greco, S., Ishizaka, A., Tasiou, M. and Torrisi, G. (2019). On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness, *Soc Indic Res.*141: 61–94.
15. Hankins M. (2008). The reliability of the twelve-item general health questionnaire (GHQ-12) under realistic assumptions. *BMC Public Health*; 8:355.
16. Head J, Stansfeld SA, Ebmeier KP, Geddes JR, et al. (2013). Use of self-administered instruments to assess psychiatric disorders in older people: validity of the General Health Questionnaire, Center for epidemiological studies Depression Scale and the self-completion version of the revised Clinical Interview Schedule. *Psychological Medicine*, 43(12), 2649-2656
17. Herrero, C., Martinez, R., & Villar, A. (2010). Multidimensional social evaluation: An application to the measurement of human development. *Review of Income and Wealth*,56, 483–497.
18. Jamieson, S.(2004). Likert scales: how to (ab) use them. *Medical Education*; 38, 1212-1218
19. Johnson SU, Ulvenes PG, Øktedalen T, Hoffart A. (2019). Psychometric Properties of the General Anxiety Disorder 7-Item (GAD-7) Scale in a Heterogeneous Psychiatric Sample. *Front Psychol*; 10:1713.
20. Jordanova V., Wickramesinghe C., Gerada C. Prince, M.(2004). Validation of two survey diagnostic interviews among primary care attendees: a comparison of CIS-R and CIDI with SCAN ICD-10 diagnostic categories. *Psychological Medicine*, 34(6):1013-1024
21. Kate, A., John, R**.**, Rhiannon, C. and Peter, K.(2019). Heterogeneity in psychiatric diagnostic classification, *Psychiatry Research*,279, 15-22
22. Kessler R, Andrews G, Colpe, L, Hiripi EE, Mroczek D, Normand Sl. *et al.* (2002). Short screening scales to monitor population prevalence and trends in non-specific psychological distress. *Psychological Medicine*, 32, 959-956.
23. Kvig EI, Nilssen S. (2023). Does method matter? Assessing the validity and clinical utility of structured diagnostic interviews among a clinical sample of first-admitted patients with psychosis: A replication study. *Front Psychiatry*.14:1076299.
24. Lewinsohn, PM; Seeley, JR; Roberts, RE & Allen, NB (1997).Center for Epidemiological Studies-Depression Scale (CES-D) as a screening instrument for depression among community-residing older adults. *Psychology and Aging*,12, 277- 287
25. Lewis G, Pelosi AJ, Araya R, Dunn G. (1992). Measuring psychiatric disorder in the community: a standardized assessment for use by lay-health workers. *Psychological Medicine*,22, 465-486
26. Liu SI, Prince M, Chiu MJ, Chen TF, Sun YW, Yip PK. (2005). Validity and reliability of a Taiwan Chinese version of the community screening instrument for dementia. *Am J Geriatr Psychiatry*.13(7):581-588.
27. Manea. L, Gilbody S, McMillan D. (2014). A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression, *Gen Hosp Psychiatry*, 37(1):67-75.
28. Mishra, GD, Hockey R. and Dobson AJ. (2014). A comparison of SF-36 summary measures of physical and mental health for women across the life course, *Qual Life Res.*;23(5):1515-21.
29. Montazeri A, Harirchi A, Shariati M, Garmaroudi G, Ebadi M. et al. (2003). The 12-item general health questionnaire (GHQ-

12): translation and validation study of the Iranian version. *Health Qual Life Outcomes*; 1:66.
30. Prochaska JJ, Sung H, Max W, Shi Y, Ong M. (2012). Validity Study of the K6 Scale as a Measure of Moderate Mental Distress based on Mental Health Treatment Need and Utilization. *Int J Methods Psychiatr Res.*21(2): 88–97.
31. Radloff, LS. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied psychological measurements*, 1(3), 385-401.
32. Rao, CR. (1973). *Linear Statistical Inference and its Application*. 2nd Edition, Wiley Eastern Private Limited, New Delhi.
33. Rutter LA and Brown TA. (2017). Psychometric properties of the generalized anxiety disorder scale-7 (GAD-7) in outpatients with anxiety and mood disorders. *J. Psychopathol. Behav. Assess.*39 140–146.
34. Seth, S. and Villar, A. (2017). Measuring human development and human deprivations. OPHI Working Paper 110, University of Oxford
35. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E. *et al.* (1998). The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry*;59 Suppl 20:22-33
36. Simundic AM. (2008). Confidence interval. *Biochem Med*,18:154–161.
37. Williams JW, Pignone M, Ramirez G, Perez Stellato C. (2002). Identifying depression in primary care: a literature synthesis of case-finding instruments. *Gen Hosp Psych*, 24(4), 225-237
38. WHO (2022). *World mental health report: Transforming mental health for all*. Geneva: World Health Organization; Licence: CC BY-NC-SA 3.0 IGO.
39. WHO (2021). *World Health Report 2021, Mental Health: New Understanding, New Hope*, World Health Organization, Geneva
40. Wu, Chien-Ho (2007). An Empirical Study on the Transformation of Likert scale Data to Numerical Scores, *Applied Mathematical Sciences*, (58),2851–2862
41. Yesavage JA, Brink TL, Rose TL, Lum O, Huang V. et al. (1982). Development and validation of a geriatric depression screening scale: a preliminary report. *J Psychiatr Res*. 17(1):37-49.

DOI:10.31579/2637-8892/241