BMJ Open

# Psychometric evaluation of an interview-administered version of the Kessler 10-item questionnaire (K10) for measuring psychological distress in rural Bangladesh



| | |
|---|---|
| Journal: | *BMJ Open* |
| Manuscript ID | bmjopen-2018-022967 |
| Article Type: | Research |
| Date Submitted by the Author: | 15-Mar-2018 |
| Complete List of Authors: | Uddin, Mohammed ; Swinburne University of Technology, Department of Statistcs, Data science and Epidemiology<br>Islam, Fakir; Swinburne University of Technology, Statistics, Data Science and Epidemiology<br>Al-Mahmud , Abdullah ; Swinburne University of Technology, 3. School of Design; Faculty of Health, Arts and Design |
| Keywords: | Phychological Distress, Rasch Analysis, Validation, Kessler Psychological Distress Scale, Rural Bangladesh, Item Response Theory |
| | |

# Title: Psychometric evaluation of an interview-administered version of the Kessler 10-item questionnaire (K10) for measuring psychological distress in rural Bangladesh

Mohammed Nazim Uddin[1], Fakir M Amirul Islam[1,2], Abdullah Al Mahmud[3]

1. Department of Statistics, Data Science and Epidemiology; Faculty of Health, Arts and Design; Swinburne University of Technology, Hawthorn VIC 3122, Australia
2. Organisation for Rural Community Development, Dariapur, Narail, Bangladesh
3. Centre for Design Innovation, School of Design, Swinburne University of Technology, Hawthorn, VIC 3122, Australia

Email: Mohammed Nazim Uddin – mnuddin@swin.edu.au; Fakir M Amirul Islam – fislam@swin.edu.au; Abdullah Al Mahmud – aalmahmud@swin.edu.au;

Corresponding author:

Email: fislam@swin.edu.au

**Title: Psychometric evaluation of an interview-administered version of the Kessler 10-item questionnaire (K10) for measuring psychological distress in rural Bangladesh**

## ABSTRACT

**Objective** The aim of this study was to translate, adapt and validate the Kessler 10-item questionnaire (K10) for measuring psychological distress in rural Bangladesh.

**Design** Cohort study.

**Setting** Narail district, Bangladesh.

**Participants** A random sample of 2425 adults of age 18–90 years were recruited.

**Outcome measure** Validation of the K10 was the major outcome. Socio-demographic factors were measured to assess if the K10 needed adjustment for factors such as age or gender. The Rasch measurement model was used for the validation, and RUMM2030 and SPSS24 software were used for analyses.

**Results** Initial inspection of the total sample showed poor overall fit. A sample size of 300, which is more satiated for Rasch analysis, also showed poor overall fit, as indicated by a significant item-trait interaction ($\chi^2$ = 262.27, df = 40, p < 0.001) and item fit residual values (mean = –0.25, SD = 2.49). Of 10 items, five items were disordered thresholds, and seven items showed misfit, suggesting problems with the response format and items. After removing three items ("feel tired", "depressed" and "worthless") and changing the Likert scale categories from five to four categories, the remaining seven items showed ordered threshold. A revised seven-item scale has shown adequate internal consistency, with no evidence of multidimensionality, no differential item functioning (DIF) on age and gender, and no signs of local dependency.

**Conclusions** Analysis of the psychometric validity of K10 using the Rasch model showed that 10 items are not appropriate for measuring psychological distress in rural Bangladesh. A modified version of seven items (K7) with four-response category would provide a psychometrically more robust scale than the original K10. The study findings suggest repeating the K7 version in other remote areas for further validation can substantiate an efficient screening tool for measuring psychological distress among the general Bangladeshi population.

**Strengths and limitations of the study**

- This study provides the first reliable data on the K10 questionnaire from a general population of a typical rural district in Bangladesh.
- This study used numerous primary data on K10 and associated covariates.
- The data were collected through face-to-face interviews of people from a typical rural district that generally represents Bangladesh.
- The sophisticated Rasch analysis technique was applied to validate as well as identify a suitable unidimensional structure of the K10. The study provides a unique opportunity to assess psychological distress in a rural population of Bangladesh.
- The potential drawback of this study is that it is based on a single-occasion collection of data from a rural district in Bangladesh. While we have attempted to capture the situation in the Narail district, the study needs to be repeated in a random sample of other rural districts to be truly representative of the national population.

## Introduction

A high prevalence of physiological distress is recognised worldwide.[1] Psychological distress is associated with chronic diseases and other health related problems,[2] and early diagnosis is seen as an important measure to ensure effective and targeted intervention.[3] In recent years, epidemiological studies have attempted to employ short dimensional scales to effectively measure and monitor the extent of psychological distress in the general community for the purposes of early diagnosis.[4] The Kessler 10-item questionnaire (K10) is one such scale among similar tools such as the Beck Depression Inventory (BDI),[5] the Hospital Anxiety and Depression Scale (HADS)[6] and the Depression Anxiety Stress Scales (DASS)[7] which are designed to assess non-specific psychological distress and screen for common psychiatric disorders.[8-11]

The K10 was developed in 1992 by Professor Kessler and Mroczek[12] to be used in the United States National Health Interview Survey as a brief measure of non-specific psychological distress along the anxiety-depression spectrum. The K10 comprises ten questions (rated on five-point Likert-type scales, where 1 = none of the time to 5 = all of the time) about psychological distress.

Although K10 is not a diagnostic tool, it does indicate psychological distress and is used to identify people in need of further assessment for anxiety and depression. The K10 measurement of a client's psychological distress levels can also be used as an outcome measure and assist treatment planning and monitoring.[13] In the context of the general population, there is often a shortage of space for the inclusion of more items in the scale. The BDI (21 items),[5] HADS (14 items)[6] and the DASS (42 items)[7] are limited as screening tools because of their long list of items. Moreover, studies confirm that well-constructed short scales can be as strong predictors as the more lengthy instruments or interviews.[12 14] Because of its small number of items, the K10 has, since its development, been widely used in many countries, including the USA, Canada and Australia. The tool has also being adopted in the World Health Organization's World Mental Health Survey.[8-11 15] Moreover, another advantage of the K10 is that it was developed using methods associated with the item response theory.[15]

Although the K10 was originally developed to identify levels of non-specific psychological distress in the general population, the tool has also demonstrated a strong relationship with severe mental illnesses as defined by structured diagnostic interviews.[16] As such, clinicians have been encouraged to use the K10 to screen for psychiatric illness.[17 18] Further, the K10 has been used as a routine outcome measure in specialist public mental health services in multiple Australian states and territories.[4] A recent review of the literature suggests that the K10 is an effective and reliable assessment tool applicable to a variety of settings and cultures for detecting the risk of clinical psychological disorders.[19 20] However, a major limitation of the K10 is the lack of consistency across studies about its factor structure. Although it was initially designed to yield a single score indicating the level of psychological distress,[15] one study demonstrated a four-factor model with acceptable fit in large community samples;[21] another study proposed a two-factor solution, one factor for depression and another for anxiety;[4] while another study did not find an adequate fit.[22]

Bangladesh is a country of 163 million people[23] where mental health complaints are a major public health concern, especially in rural areas.[24-26] The prevalence of mental disorders in such areas varies between 6.5% and 31%, possibly due to the use of different protocols and definitions of mental disorders.[27] A culturally validated tool is needed for quick screening of psychological

distress in Bangladesh, as well as in other countries with similar socio-economic conditions. Due to lack of published research on the K10 in rural settings, and uncertainties surrounding the scale noted above, we need to develop a valid measurement scale of psychological distress in Bangladesh.

The present study pursues an update of Rasch analysis technique to evaluate the suitability of the K10 for measuring psychological distress in rural Bangladesh, and to provide guidance on suitable modification to the instrument to improve its performance. Accuracy and precision of K10 scores can lead to a more efficient allocation of health care resources as well as more efficient screening of psychological distress among the rural population.

## Materials and Methods

### Study Population

Participants were recruited from the Narail district, located approximately 200 km south-west of Dhaka, the capital city of Bangladesh. We recruited a total of 2425 adults aged 18–90 years, from May to July 2017. The study protocol, including its geographic location and population density, is described in detail elsewhere.[28]

### Sample Size and Statistical Power

A sample of approximately 300 is more suitable for a Rasch analysis, because large sample sizes can result in type 1 errors that falsely reject an item for not fitting in the Rasch model.[29] A sample size of 300 is considered large enough for 99% confidence that the estimated item difficulty would be within ±½ logit of its stable value.[30] We did the analysis five times with five different random sample sizes of 300 each, from the total sample of 2425, to check the robustness of the models using different subsamples. For the initial test of the model, we also used the total sample.

### Sampling Frame

A multilevel cluster random sampling technique was used for this cohort study. Three unions (smallest rural administrative unit) out of 13 and 1 Pourashava (smallest urban administrative unit) of Narail Upazilla (the third largest type of administrative division in Bangladesh) were

randomly selected at level 1. Two to three villages (a smallest territorial and social unit for administrative and representative purposes), from each selected union and two wards (an electoral district, for administrative and representative purposes) were randomly chosen from selected Pourashava at the second level. In total, 150 adults (18–59 year old) and 120 older adults (60–90 year old) from each of the villages/wards were interviewed. Recruitment strategy and quality assurance in data collection are described previously.[28]

**Kessler Psychological Distress Scale**

The K10 measures how often participants have experienced symptoms of anxiety and depressive disorders in the previous four weeks prior to screening.[12] Respondents were asked, 'During the past four weeks, how often did you feel: 1) tired out for no good reason; 2) nervous; 3) so nervous that nothing could calm you down; 4) hopeless; 5) restless or fidgety; 6) so restless you could not sit still; 7) sad or depressed; 8) so depressed that nothing could cheer you up; 9) everything was an effort; 10) worthless.' Items are rated on a five-point ordinal scale: all of the time (score 5), most of the time (score 4), some of the time (score 3), a little of the time (score 2) and none of the time (score 1). Questions 3 and 6 are not asked if the preceding question was answered 'none of the time', in which case questions 3 and 6 would automatically receive a score of one. Scores for the ten questions are summed: the maximum score is 50, indicating severe distress; the minimum score is ten indicating no distress. Low scores indicate low levels of psychological distress and high scores indicate higher levels of psychological distress.[12]

**Outcome Variables**

The main outcome measure was the validation of the K10.

**Factor Variables of Differential Item Functioning (DIF)**

Participants were categorised as either adults (18 to 59 year old) or older adults (60 to 90 year old), and by gender (male or female).

**Scale Validation:** Item response theory (IRT) and Classical Test Theory (CTT)

**Item Response Theory (IRT)**

IRT is a paradigm for the design, analysis and scoring of tests, questionnaires and similar instruments measuring abilities, attitudes or other variables.[31] It is based on the relationship between individuals' performances on a test item and their personal performance on an overall measure of the ability that the item seeks to quantify.[32] All IRT models attempt to explain observed (actual) item performance as a function of an underlying ability (unobserved) or latent trait.

**Classical Test Theory (CTT)**

Classical test theory is a quantitative approach to testing the reliability and validity of a scale based on its items. CTT is a simple linear model which links the observable score (X) to the sum of two unobservable (often called latent) variables, true score (T) and error score (E); i.e., X = T + E. Because of, each examinee there are two unknowns, without simplified assumption the equation will not be solved. The assumptions in the classical test model are that (a) true scores and error scores are uncorrelated, (b) the average error score in the population of examinees is zero, and (c) error scores on parallel tests are uncorrelated.[33] The true score (T) is defined as the expected value of the observed score over an infinite number of repeat administrations of the same instrument.[33 34]

**Rationale for using the Rasch analysis instead of the CTT**

Similar to the IRT, the CTT is another fundamental measurement theory that researchers employ to construct measures of latent traits. Both IRT and CTT can be used to construct measures of latent traits, but the two measurement systems are entirely dissimilar. A more in-depth explanation of the literature on CTT[35-37] and IRT.[38-41] So far, the K10 was validated mostly using CTT in which the items and the latent trait being measured are considered separately and, therefore, cannot be meaningfully and systematically compared.[42 43] These limitations can be solved rationally using Rasch modelling.[38 39 44-46]

**The Rasch Model**

The Rasch model was named after the Danish mathematician Georg Rasch.[47] The model shows what should be expected of responses to items if measurement (at the metric level) is to be achieved. Two versions of the Rasch model are available:

dichotomous, $P\{X_{ni} = x\} = \frac{e^{x(\beta_n - \delta_i)}}{1+e^{x(\beta_n - \delta_i)}}$;[47]

and polytomous, $P\{X_{ni} = x\} = \frac{e^{-\tau_{1i}-\tau_{2i}\cdots\cdots-\tau_{xi}+x(\beta_n-\delta_i)}}{\sum_{x'=0}^{m_i} e^{-\tau_{1i}-\tau_{2i}\cdots\cdots-\tau_{xi}+x(\beta_n-\delta_i)}}$;[48]

where $\beta_n$ is the location of person n and $\delta_i$ is the location of item i.$\tau_{xi}$, $x = 1, 2, \ldots, m_i$ are thresholds which partitioned the latent continuum of item i into $m_i + 1$ ordered categories. X is the response value that qualifies the expression by $\beta_n - \delta_i$.

The Rasch analysis in this study was conducted using the RUMM 2030 package.[49] In the assessment of K10, respondents were presented with the ten-item questionnaire regarding psychological distress. The purpose of the Rasch analysis was to maximise the homogeneity of the trait and to allow more significant reduction of redundancy without sacrificing the measurement of information by decreasing items and scoring levels to yield a more valid and straightforward measure. The Rasch model requires some assumptions that need to be evaluated to ensure that an instrument has Rasch properties. The Rasch assumptions most commonly assessed are a) unidimensionality, b) local independence and c) invariability.

Chi-square item-trait interaction statistics define the overall fit of the model for the scale.[50] A non-significant chi-square probability value indicated that the hierarchical ordering of the items is consistent across all levels of the underlying trait. A Bonferroni adjustment[51] is typical of the alpha value used to assess statistical significance, by dividing the alpha value of 0.05 by the number of items in the scale. Item-person interaction statistics distributed as z-statistic with a mean of zero and SD of 1 (indicating perfect fit with the model). Values of SD above 1.5 for either items or person suggest a problem. Individual item fit statistics are presented as residuals (acceptable within the range ±2.5) and chi-square statistic (require a non-significant chi-square value).

The Rasch model can be extended to analyse items with more than two response categories, which involves a 'threshold' parameter, represented by the two response categories where either

response is probable. Common sources of item misfit occur with 'disorder thresholds' failure of the respondents to use the response category in a manner consistent with the level of the trait being measured.

Unidimensionality occurs when a set of items measures just one thing in common.[52] To establish this, the first step is to run a Principal Component Analysis (PCA) on the residuals to identify two subsets of the items having the most difference. Second, the items loading on the first factor are extracted, items having positive and negative loadings are defined, and estimates for these two sets are derived. Applying an independent t-test to both sets, which conduct t-tests for each person in the sample comparing their score on the Set 1 items and Set 2 items. If less than 5% of the estimates are outside the range of ±1.96, the scale is considered unidimensional.

In case of local independence,[53] the items in a test are expected to be unrelated to each other; i.e. the response on each item should not be associated with that of another items. To test for local independence, we need to check the residuals correlation matrix, and any correlation coefficient value greater than 0.3 suggests the two items are locally dependent. In a situation where the correlation value is greater than 0.3, the two items need to be merged into one, called subtest analysis, to achieve a significant improvement on PSI value. If so, it is a sign of local dependency and a violation of one of the Rasch assumptions.

Invariability indicates that 'items are not dependent on the distribution of persons' abilities and the persons' abilities are not dependent on the test items.[54] In Rasch measurement theory, the scale should work in the same way, irrespective of which group (e.g., gender or age) is being assessed. If for some reason one gender does not display equal likelihood of confirming the item, then the items would display DIF and would violate the requirement of unidimensionality.[55] DIF is an analysis of variance of the person-item deviation residuals with the person's factors (e.g., age, gender).

The reliability and internal consistency of the model are defined by the Person Separation Index (PSI).[56] In addition to item fit, examination of person fit is essential. A few responses with unusual response pattern (identified by high positive residuals) may seriously affect the fit at the

item level. Such aberrant response patterns occur due to unrecorded co-morbidity or respondents with cognitive defects. Therefore, if some response pattern showed high positive fit residuals, removal from the analysis may make a significant difference to the scale internal construct validity.

## Results

### *Overview of the respondents*

Table 1 shows the summary statistics of both the validation and the total data sets by gender (male and female). The mean (SD, range) age of the total participant sample was 52.0 years (17, 18–90). Of the total sample, 48.5% were men, 27.6% had no formal education, 4% had at least a bachelor's degree level of education.

### *Primary analysis of the original set of ten items and five response categories*

K10 scores ranged from 10 to 50 with a mean of 16.7 (SD = 11.3). Initial inspection of the scale with the total 2425 participants showed poor overall fit with the Rasch model, as indicated by a significant item-trait interaction ($\chi^2$ = 1729.89, df = 40, $p < 0.001$) and item fit residual values (mean = –0.25, SD = 6.75) outside the acceptable range. Eight items were found to be misfit based on the overall fit residual values outside the range of ±2.5. Five items were found to have disordered thresholds, signifying problems with the 5-point response format used for the scale. A check found multidimensionality: the model fit statistics for the five separate random subsamples of 300 each from the total participant sample produced almost identical results, indicating the results and sample selections were robust (Table 2).

Initial inspection of scores in the random sample of 300 participants showed poor overall fit to the Rasch model ($\chi^2$ = 262.27, df = 40, $p < 0.001$) and items fit residual values (mean = –0.25, SD = 2.49). However, the person fit residuals (mean = 0.18, SD = 1.24) were within the acceptable range (Table 2, sample 1). Five items were found to have disordered thresholds, and seven of the individuals' item fit statistics showed misfit, suggesting problems with the 5-point

response format used for the questionnaire. The value of the PSI (analogous to Cronbach's alpha) for the original set of ten items with five response categories was 0.84, indicating that the scale worked well to separate persons. The frequency distribution of the items showed (data not shown) mistargeting. Across all five items, the distribution was skewed towards the lower values, indicating low psychological distress among the respondents in the sample. Seven items (items 1, 2, 3, 4, 7, 8 and 9) showed misfit (Table 3: initial solution) while five items showed disorder thresholds (1, 4, 7, 8, 9) (Figure 1: initial solution). A visual examination of the threshold map shows that the estimates of the thresholds defining the categories in item 1 (tired) (Figure 2: category probability curve), item 4 (feel hopeless), item 7 (depressed), item 8 (an effort) and item 9 (so sad) do not form distinctive regions of the continuum. We have examined the category probability curve of each disorder threshold item, and found response 1 and 2 adjacent category were not the same (Figure 2, category probability curve).

To address the issue of disordered categories, Rasch analysis was conducted on only the disordered items, by merging the two middle categories ('a little of the time' and 'some of the time'). This reduced the scoring to a 4-point format from 01234 to 01123, and made the overall score range 0 to 40. Following this, eight misfit items were identified with significant chi-square probability values, or high positive or high negative residual values (± 2.5), and found only item 5 to be disordered. (Table 3: only disorder items were rescored as 01123). Then we carried out all items Likert scale categories from five to four categories and found all items were ordered thresholds. (Figure 1: rescore all items to 01123). However, five items were still misfit in the model (Table 3: rescore all items to 01123).

***Proposed final analysis of the seven items and four response categories***

Misfit items were removed one at a time iteratively, based on positive or negative residual values as well as the degree of the significant chi-square probability values. The total model fit and individual item fit statistics were checked after each iteration, until the remaining items were shown to fit Rasch model's expectations. The three removed items were items 1, 7 and 10.

The final solution, retaining seven items, showed overall fit with the model (Table 4). The PSI was found to be high (PSI = 0.84), making the model suitable for individual use. The items of the

K7 scale were assessed for DIF across gender (male/female) and age (adults: 18–59 year old) and older adults (60–90 year old) (Table 5). A significant DIF was found on item 9 (feel so sad); however, using a Bonferroni-adjusted alpha value (.05/7=.007), the value became non-significant. In the final model, seven items with four response categories showed all items to have ordered thresholds (Figure 3). There was no indication of item or person misfit (Table 4: Individuals' items fit statistics of final K7). Unidimensionality of the K7 scale was tested using PCA (3.34%, 95% CI 0.9% to 5.8%), and from a binomial distribution was found non-significant, which supports unidimensionality of the K7 (Table 4, final solution of K10 and Figure 4, final solution of K7).

**Discussion**

The purpose of the paper was to evaluate the suitability of the Kessler 10-item questionnaire for measuring psychological distress in rural Bangladesh. This article examines the potential contribution of Rasch analysis in exploring several issues concerning the K10. This includes an assessment of the appropriateness of using all K10 items to represent the underlying dimension of psychological distress. In addition, the article includes an evaluation of the validity of the category scoring system, the fit of individual items and an assessment of the potential bias of items by gender and age, from the perspective of the Rasch model. The initial descriptive analysis of the frequency distributions indicated that the 10-item scale with five response categories mistargeted the current sample of the rural Bangladeshi population. Non-responses or very few responses in the categories may manifested to the mistargeting. Two items ('tired' and 'depressed') showed misfit, and two items ('so nervous' and 'so restless') showed redundancy (i.e., little impact on the scale). Moreover, items with disordered thresholds indicating problems with the categorisation of the items and scale showed evidence of multidimensionality. Since the K10 scale has not previously undergone a rigorous psychometric analysis in rural Bangladesh and even in neighbouring countries, the detection of problems was not surprising, even though attention had been paid to targeting when the scale was constructed. In these circumstances, the analysis elaborated on taking advantage of the Rasch model.

One response category was warped, which resulted in four instead of five response categories for each item. Moreover, those items showing misfit were removed from the model gradually after going through all possible steps to improve the model. Item 1 ('how often did you feel tired out

for no good reason') was removed because it showed high fit residuals value and DIF for age (adults and older adults). Although techniques exist for solving uniform DIF by allowing the item difficulty to vary by group, we believe that option is inappropriate because it is not useful as an everyday screening environment. Therefore, we decided to delete the biased item, which also had a large chi-square value. On the other hand, the item may not play the concepts of psychological distress in Bangladesh. This could be one reason why the item works differently according to age (adults and older adults). The removal of this item from the scale improved the overall fit of the model, supporting this decision. Moreover, the item removed was one of the four items that Kessler[15] had earlier used to reduce 10 to 6 items. Item 7 ('how often did you feel depressed') was also removed from the scale due to misfit with the model. The large positive residual value indicates misfit in that it contributed little or no information additional to other items, as well as having a large chi-square value. However, the item showed no DIF on age and gender. Removal of the item from the model significantly improved the fit of remaining items. Moreover, the item removed was one of the four-item that Kessler[15] earlier used to reduce 10 to six items. Item 10 ('how often did you feel worthless') has been removed from the scale due to high chi-square value and significant chi-square probability, as well as high positive residuals which contribute to an overall model misfit. The high chi-square value indicates that it adds nothing to the information gained by other items, and this item is the only one, which increased the overall chi-square value and made the overall model misfit. The study results support the retention of item 10.

Removal of items from the scale would eliminate at least some redundancy.[57-59] However, our analysis identified that Cronbach's alpha for the K7 (0.88) was equivalent to the original K10 Cronbach's alpha (0.87); in addition, the PSI of K7 (0.84) was the same as that of the original K10's PSI (0.84). A study reported by Fassaert et al,[19] showed that some redundancy happens in Cronbach's alpha, when comparing K10 (0.93) and K6 (0.89). However, our model showed superior value of Cronbach's alpha K7 (0.88) compared to the original K10 (0.87) model, and confirms adequate fit of the model in the rural settings in Bangladesh. Although we have proposed seven validated items (K7), a previous study proposed six (K6) items[17] was more robust than the K10. Of K7, five items were common in K6. We only tested K6 items using Rasch analysis and found a poor overall fit. In particular, the presence of the item "feel

worthless" showed a large positive fit residual and significantly large chi-square value, which influenced the overall model misfit under Rasch assumptions. Therefore, the current study found that the K7 model is more robust in our sample compared to K6.[17 20]

Gender differences in psychology are ubiquitous,[60] so it is essential to verify whether the model is affected by gender or not. Our revised seven-item model showed no DIF on gender, i.e., there is no gender bias in the revised K7 scale. The K7 scale is equally valid for men and women, which supports the previous findings reported in Australia.[61] Another important factor is age, and there is inconsistency in the literature on the relationship between age and psychological distress.[62] The study conducted by Kessler et al. documented a good deal of inequality in the relationship between age and screening scales of depressive symptoms.[63] However, other studies showed a stable nonlinear association between age and psychological distress in several cross-sectional epidemiologic surveys.[62 64 65] Our revised model of K7 confirmed that there is no age bias (adults and older adults), and the model is equally applicable to any one between the age of 18–90 years.

Application of the Rasch measurement model in this study has supported the viability of a seven-item version of the K10 scale for measuring psychological distress in rural Bangladesh. The scale shows high reliability, with no disordering of thresholds and no evidence of DIF. The model also showed high PSI (0.84) and reliability (0.87), which indicated the power of the test of fit. Furthermore, there is good evidence from this sample that a single total score of psychological distress is viable. Thus, the seven-item scale appears robust when tested against the strict assumptions of the Rasch measurement model.

This paper shows how the Rasch model can be used for rigorous examination and development of measurement instruments such as the K10 psychological distress scale. The Rasch model simplifies measurement problems such as lack of invariance, which was overlooked in traditional analysis.[66] The Rasch analysis of the K10 scale indicates that the psychometric properties of the original scale most likely would have been much better if scale developmental had been guided by IRT (Rasch analyses). In future, importance should be given to improving the targeting of person and items. Reducing the number of response categories as well as the number of items

might also improve the properties of the scale.[67] Therefore, data on the general rural population regarding psychological distress based on the revised seven-item scale from the K10, with four-response category, is superior to the original scale.

This study provides the first reliable data on levels of psychological distress among the general population of rural Bangladesh. The analysis was based on a large data set of adults and older adults across a wide range of age, from whom data were collected directly in a face-to-face interview. The Rasch analysis in this study guided a detailed examination of the structure of the scale. The response category orderings (threshold ordering) were not examined earlier, and evidence from the current study does not support the response format or the validity of the original 10-item scale.

The potential drawback of this study is that it is based on single-occasion collection of data from people in a rural district of Bangladesh. While we have attempted to capture the situation in the Narail district, the study would obviously need to be repeated in a random sample of other rural districts for the results to be truly representative of a national population.

**Conclusion**

Overall, the authors favours the use of K10 in rural Bangladesh, as has been used elsewhere. However, this study acknowledges that due to cultural variations and strict adherence to Rasch properties, modification is needed to measure psychological distress in rural Bangladesh. The results of this study suggest that a revised seven-item version of the K10, with four-response category, would provide a more robust psychometric scale than the original K10. The modified seven-item scale fulfils all the assumptions of the Rasch model, and the model has shown no differential item functioning (DIF) on age and sex as well as no local dependency. The study findings can be repeated using a random sample of other remote areas in Bangladesh to further validate the revised scale, as well as to better establish the level of psychological distress nationwide. The tool can be applied in clinical settings at the national level, where psychological distress has yet to be diagnosed.

**List of abbreviations**

CTT, Classical Test Theory; IRT, Item Response Theory; DIF, Differential Item Functioning; PSI; Person Separation Index, K10, Kessler Psychological Distress Scale

**Ethical aspects:** Human Ethics Approval was received from the Swinburne University of Technology Human Ethics Committee (SHR Project 2015/065) in accordance with the tenets of the Declaration of Helsinki. Study participants provided written informed consent.

**Authors' contributions:** MNU and FMAI jointly designed the study. MNU analysed the data and drafted the manuscript. FMAI contributed to writing the manuscript. FMAI supervised the overall analyses and preparation of the manuscript. AAM reviewed the manuscript. All authors contributed to the development of the manuscript and read and approved its final version.

**Acknowledgements:** We particularly acknowledge the contribution of Md Rafiqul Islam, Md Sajibul Islam, Saburan Nesa and Arzan Hosen for their hard work in door-to-door data collection. Finally, we would like to express our gratitude to the study participants for their voluntary participation.

**Competing interests:** None declared

**Funding:** The Faculty of Health, Arts and Design of the Swinburne University Technology under the Research Development Grant Scheme, funded data collection for this research project. The funder had no role in the design of the study, data collection or analysis, interpretation of data or writing of the manuscript.

**Data sharing statement:** Data will be available upon request.

**Table 1. Characteristics of participants who were included and who were not in the current study, by gender**

| | Total N=2425 | | | In validation N=300 | | |
|---|---|---|---|---|---|---|
| Characteristic | Total (2425) | Male (1176) | Female (1249) | Total (300) | Male (143) | Female (153) |
| Age groups (in years) | | | | | | |
| 18–59 | 1278 (52.7) | 603 (51.3) | 675 (54.0) | 172 (57.3) | 73 (51.0) | 99 (63.1) |
| 60–90 | 1147 (47.3) | 573 (48.7) | 574 (46.0) | 128 (42.7) | 70 (49.0) | 58 (36.9) |
| Education | | | | | | |
| No education | 671 (27.7) | 289 (24.6) | 382 (30.6) | 76 (25.3) | 37 (25.9) | 39 (24.8) |
| Primary (1–5) | 946 (39.0) | 447 (38.0) | 499 (40.0) | 124 (41.3) | 58 (40.6) | 66 (42.0) |
| Secondary (6–9) | 327 (13.5) | 146 (12.4) | 181 (14.5) | 38 (12.7) | 13 (9.1) | 25 (15.9) |
| SSC or HSC Pass (10–12) | 385 (15.9) | 224 (19.0) | 161 (12.9) | 50 (16.7) | 26 (18.2) | 24 (15.3) |
| Degree or equivalent (13–16) | 96 (4.0) | 70 (6.0) | 26 (2.1) | 12 (4.0) | 9 (6.3) | 3 (1.9) |

SSC – Secondary School Certificate, HSC- Higher Secondary Certificate

**Table 2. Model Fit Statistics for total sample and five random samples of 300 with all 10 items**

| Initial solution | Total sample N=2425 | Sample 1 n=300 | Sample 2 n=300 | Sample 3 n=300 | Sample 4 n=300 | Sample 5 n=300 |
|---|---|---|---|---|---|---|
| Overall model fit, Chi-square value | 1727.89 | 262.27 | 212.30 | 204.07 | 194.37 | 282.14 |
| Degree of freedom (DF) | 40 | 40 | 40 | 40 | 40 | 40 |
| P | 0.00000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Item fit residuals (mean (SD)) | –0.25 (6.75) | 0.13 (2.49) | 0.05 (2.40) | –0.23 (2.12) | 0.11 (2.38) | –0.16 (2.64) |
| Person fit residuals (mean (SD)) | –0.29 (1.32) | –0.18 (1.24) | –0.28 (1.33) | –0.34 (1.32) | –0.30 (1.37) | –0.27 (1.32) |
| Person separation index (PSI) | 0.84 | 0.84 | 0.83 | 0.85 | 0.82 | 0.83 |
| Coefficient alpha | 0.86 | 0.87 | 0.86 | 0.87 | 0.86 | 0.87 |
| Unidimensionality test (% that goes beyond 95% CI) | 10.3 (9.6–11.2) | 9.3 (6.9–11.8) | 11.7 (9.2–14.1) | 8.3 (5.9–10.8) | 9.0 (6.5–11.5) | 10.33 (7.9–12.8) |

**Table 3. Fit statistics (location, residuals and P values) of the 10 items for the first random sample of 300**

| | Initial solution | | | Rescore only disordered items to 01123* | | | Rescore all items to 01123 | | |
|---|---|---|---|---|---|---|---|---|---|
| **Items** | **Location** | **Residuals** | **P value** | **Location** | **Residuals** | **P value** | **Location** | **Residual** | **P value** |
| Feel tired (1) | –0.42 | 4.28 | 0.000*^ | 0.00 | 1.35 | 0.005 | –0.51 | 1.22 | 0.000^ |
| Feel nervous (2) | –0.11 | –0.85 | 0.001^ | –0.56 | –1.19 | 0.004^ | –0.12 | –3.26 | 0.020 |
| Feel so nervous (3) | 0.13 | –3.16 | 0.000^ | –0.32 | –3.65 | 0.002^ | 0.05 | –4.13 | 0.002^ |
| Feel hopeless (4) | –0.06 | -0.62 | 0.008* | 0.34 | –1.54 | 0.001^ | –0.03 | –1.77 | 0.104 |
| Feel restless or fidgety (5) | –0.22 | 0.46 | 0.002^ | –0.69 | 0.11 | 0.001*^ | –0.26 | –1.93 | 0.302 |
| Feel so restless (6) | 0.08 | –3.11 | 0.000*^ | –0.38 | –3.39 | 0.007 | 0.04 | –3.73 | 0.003^ |
| Feel depressed (7) | 0.26 | 3.87 | 0.000*^ | 0.74 | 3.00 | 0.000^ | 0.35 | 3.18 | 0.000^ |
| Everything was an effort (8) | –0.15 | –0.33 | 0.125* | 0.28 | –1.90 | 0.000^ | –0.16 | –2.36 | 0.301 |
| Feel so sad (9) | 0.16 | –0.48 | 0.058 | 0.65 | –2.32 | 0.001^ | 0.25 | –2.64 | 0.003^ |
| Feel worthless (10) | 0.34 | 1.33 | 0.001^ | –0.06 | 2.41 | 0.003^ | 0.39 | 0.60 | 0.247 |

*Disordered items; ^ P values depend on chi-square values (Bonferroni correction (p value/number of items)) =.05/10=.005)

**Table 4. Individuals' item fit statistics of original Kessler K10 and final seven-items model**

| | Individuals' items fit statistics of original K10 | | | | | Individuals' items fit statistics of Final K7 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Items** | **Location** | **SE** | **Residual** | **$\chi^2$** | **P value** | **Location** | **SE** | **Residual** | **$\chi^2$** | **P value** |
| Feel tired (1) | –0.42 | 0.08 | 4.28 | 46.76 | 0.000 | | | | | |
| Feel nervous (2) | –0.11 | 0.09 | –0.85 | 19.94 | 0.001 | –0.20 | 0.15 | –1.40 | 3.99 | 0.41 |
| Feel so nervous (3) | 0.13 | 0.09 | –3.16 | 30.36 | 0.000 | 0.10 | 0.15 | –2.66 | 11.01 | 0.03 |
| Feel hopeless (4) | –0.06 | 0.08 | –0.62 | 13.66 | 0.008 | 0.03 | 0.15 | 0.62 | 3.35 | 0.50 |
| Feel restless or fidgety (5) | –0.22 | 0.09 | 0.46 | 16.88 | 0.002 | –0.28 | 0.16 | –0.81 | 3.98 | 0.41 |
| Feel so restless (6) | 0.08 | 0.09 | –3.11 | 30.53 | 0.000 | 0.09 | 0.15 | –2.78 | 8.04 | 0.09 |
| Feel depressed (7) | 0.26 | 0.09 | 3.87 | 70.15 | 0.000 | | | | | |
| Everything was an effort (8) | –0.15 | 0.08 | –0.33 | 7.21 | 0.125 | –0.09 | 0.15 | –0.86 | 7.03 | 0.13 |
| Feel so sad (9) | 0.16 | 0.09 | –0.48 | 9.11 | 0.058 | 0.34 | 0.16 | –0.56 | 2.42 | 0.65 |
| Feel worthless (10) | 0.34 | 0.09 | 1.33 | 17.69 | 0.001 | | | | | |
| | **Initial solution of K10** | | | | | **Final solution of K7** | | | | |
| Overall model fit | 262.27 | | | | | 39.82 | | | | |
| Degree of freedom (DF) | 40 | | | | | 28 | | | | |
| P | 0.000 | | | | | 0.068 | | | | |
| Item fit residuals (mean (SD)) | 0.13 (2.49) | | | | | –0.20 (1.20) | | | | |
| Person fit residuals (mean (SD)) | –0.18 (1.24) | | | | | –0.63 (1.40) | | | | |
| Person separation index (PSI) | 0.84 | | | | | 0.84 | | | | |
| Coefficient alpha | 0.87 | | | | | 0.88 | | | | |
| Unidimensionality test (% that goes beyond 95% CI) | 9.33% CI (6.9–11.8) | | | | | 3.34% CI (0.9–5.8) | | | | |

**Table 5. DIF on age (adults and older adults) and gender (male and female)**

| Items | DIF on Age | | | | DIF on Gender | | | |
|---|---|---|---|---|---|---|---|---|
| | MS | F | DF | Prob | MS | F | DF | Prob |
| Feel nervous (2) | 0.58 | 0.88 | 1 | 0.35 | 0.59 | 0.91 | 1 | 0.34 |
| Feel so nervous (3) | 1.00 | 1.86 | 1 | 0.17 | 0.06 | 0.11 | 1 | 0.74 |
| Feel hopeless (4) | 0.07 | 0.08 | 1 | 0.78 | 2.41 | 2.59 | 1 | 0.11 |
| Feel restless or fidgety (5) | 0.49 | 0.67 | 1 | 0.41 | 0.66 | 0.89 | 1 | 0.35 |
| Feel so restless (6) | 0.50 | 0.92 | 1 | 0.34 | 0.00 | 0.00 | 1 | 0.98 |
| Everything was an effort (8) | 0.12 | 0.17 | 1 | 0.68 | 0.26 | 0.36 | 1 | 0.55 |
| Feel so sad (9) | 5.29 | 6.86 | 1 | 0.01 | 0.80 | 1.04 | 1 | 0.31 |

## Appendix

### Table 1. Analysis history

| | Analysis | Item fit residuals mean (SD) | Person fit residuals mean (SD) | PSI (CF) | Overall model fit $\chi^2$ (p value) | Status of disorder items | % Significant t-test CI |
|---|---|---|---|---|---|---|---|
| K10 Original scale | 1 | 0.14 (2.50) | –0.19 (1.24) | 0.85 (0.87) | 262.28 (0.0000) | Five items (1,4,7,8,9) | 9.33% CI (6.9–11.8) |
| Rescore only disorder items to 01123 | 2 | –0.71 (2.34) | –0.30 (1.13) | 0.84 (0.87) | 202.53 (0.0000) | One item (5) | 7.3% CI (4.9–9.8) |
| Rescore all to 01123 | 3 | –1.48 (2.38) | –0.55 (1.44) | 0.83 (0.87) | 166.67 (0.0000) | No items | 4.7% CI (2.2–7.1) |
| Positive worded items (1,2,3,5,6) | 4 | –0.15 (1.92) | –0.30 (1.04) | 0.85 (0.83) | 107.05 (0.0000) | One item (1) | 6.3% CI (3.9–8.8) |
| Negative worded items (4,7,8,9,10) | 5 | 0.51 (1.90) | –0.21 (1.03) | 0.59 (0.80) | 98.51 (0.0000) | Three items (4, 7 ,8) | 3.0% CI (0.5–5.5) |
| Remove only tired from the model | 6 | –1.07 (2.39) | –0.57 (1.55) | 0.81 (0.87) | 107.13 (0.0000) | No item | 6.7% CI (4.2–9.1) |
| Remove only depressed from the model | 7 | –1.50 (1.97) | –0.67 (1.49) | 0.85 (0.88) | 96.61 (0.0000) | No item | 10.3% CI (7.9–12.8) |
| Remove only worthless from the model | 8 | –1.58 (2.59) | –0.53 (1.40) | 0.82 (0.86) | 164.14 (0.0000) | No item | 8.3% CI (5.9–10.8) |
| Remove tired and depressed together from the model | 9 | –1.27 (1.66) | –0.73 (1.62) | 0.83 (0.88) | 58.73 (0.0027) | No item | 16.0% CI (13.5–18.5) |
| Remove tired and worthless together from the model | 10 | –1.20 (2.61) | –0.55 (1.48) | 0.80 (0.86) | 126.24 (0.0000) | No item | 12.0% CI (9.5–14.5) |
| Remove depressed and worthless together from the model | 11 | –1.55 (1.90) | –0.64 (1.42) | 0.85 (0.87) | 76.64 (0.0000) | No item | 9.3% CI (6.9–11.8) |
| Remove tired, depressed and worthless from the model | 12 | –1.34 (1.27) | –0.70 (1.53) | 0.84 (0.88) | 40.11 (0.0647) | No item | 6.7% CI (4.2–9.2) |
| Remove Further from tired, depressed and worthless from the model+ one person id 164 | 13 | –0.20 (1.20) | –0.06 (1.41) | 0.84 (0.88) | 39.83 (0.0685) | No item | 3.3% CI (0.9–5.8) |

The history of statistical analysis of the K10 using Rasch analysis has been mentioned in Table 1 (supplementary file). First, we ran the Rasch analysis with original ten items. Out of the ten items, five items had disordered thresholds, and overall chi-square values as well as item fit residuals that were high and significant. We rescored only the disordered items by following the pattern of categorical probability curve, which suggested the combination of the middle two response categories into one, but one item still had a disordered threshold. Next, we rescored all items to 0,1,1,2, 3 from 0, 1, 2, 3, 4. The problem of disordered items was solved, but overall chi-square values and item fit residuals SD were high. Then, we tried to use the PCA technique to check whether the scale was more than one dimension. To achieve this, we used PCA technique to separate positively and negatively worded items. We found that items 1, 2, 3, 5 and 6 were positively worded items, and 4, 7, 8, 9 and 10 were negatively worded items. We applied Rasch analysis technique to positively and negatively worded items and found one disordered item among the positively worded items and three among the negatively worded items, and overall model fits were poor for both models. We confirmed that the K10 was not a two-factor solution. Then we revisited the model where we rescored all items to 01123.

We checked the individual items fit chi-square value, which might influence the overall chi-square value. We found that item 1 (feel tired) had a high chi-square value followed by item 7 (feel depressed) and item 10 (worthless). First, we removed the item 'feel tired' from the model and then 'depressed' and finally 'worthless.' Removing one item at a time in the following sequential order, 'tired', 'depressed' and 'worthless' resulted in chi-square values (SD) of 107.13 (SD = 2.34), 96.61 (SD=1.97) and 164.14 (2.59), respectively, indicating the models were poorly fit. Removing two items at a time in the following sequential order, 'tired and depressed', 'depressed and worthless', and 'worthless and tired' did not improve the model significantly. Going through different iteration process in removing items, removing three items together produced the desired model except the individual's person fit statistics SD (1.53). Further investigation showed that one person was misfit. Removing the misfit person, Rasch analysis produced a perfect fit model with seven items, with four categories for each item (Appendix Table 1). All the assumptions of the Rasch analysis have been met in our model.

## References

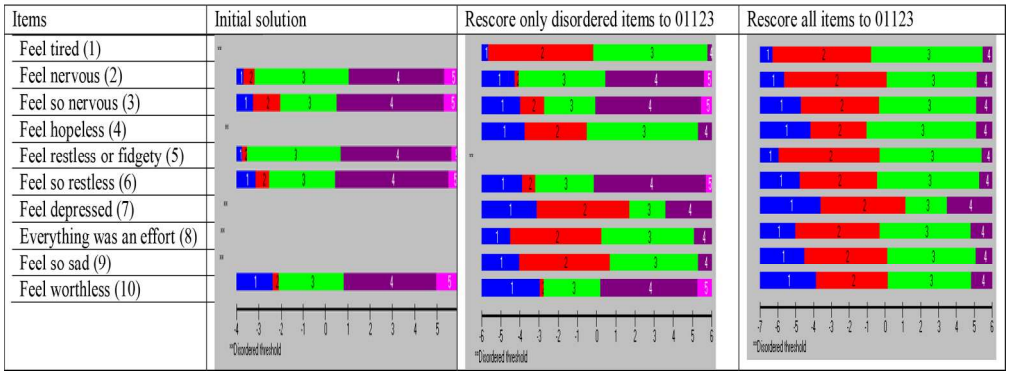
1. Lopez AD, Murray CCJL. The global burden of disease, 1990-2020. Nat Med 1998;**4**(11):1241-43.
2. Egede LE. Effect of comorbid chronic diseases on prevalence and odds of depression in adults with diabetes. Psychosomatic Medicine 2005;**67**(1):46-51.
3. Costello EJ. Early Detection and Prevention of Mental Health Problems: Developmental Epidemiology and Systems of Support. J Clin Child Adolesc 2016;**45**(6):710-17.
4. Sunderland M, Mahoney A, Andrews G. Investigating the Factor Structure of the Kessler Psychological Distress Scale in Community and Clinical Samples of the Australian Population. J Psychopathol Behav 2012;**34**(2):253-59.
5. Beck AT, Erbaugh J, Ward CH, et al. An Inventory for Measuring Depression. Arch Gen Psychiat 1961;**4**(6):561-&.
6. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiat Scand 1983;**67**(6):361-70.
7. Parkitny L, McAuley J. The Depression Anxiety Stress Scale (DASS). J Physiother 2010;**56**(3):204-04.
8. Andrews G, Peters L. The psychometric properties of the composite international diagnostic interview. Soc Psych Psych Epid 1998;**33**(2):80-88.
9. Kessler RC, Ustun TB. The World Mental Health (WMH) Survey Initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). Int J Meth Psych Res 2004;**13**(2):93-121.
10. Kessler R MD. Final Versions of our Non-Specific Psychological Distress Scale Ann Arbor(MI): Survey Research Center of the Institute for Social Research, University of Michigan, 1994. 1994.
11. Slade T, Johnston A, Browne MAO, et al. 2007 National Survey of Mental Health and Wellbeing: methods and key findings. Aust Nz J Psychiat 2009;**43**(7):594-605.
12. Kessler RC, Barker PR, Colpe LJ, et al. Screening for serious mental illness in the general population. Arch Gen Psychiat 2003;**60**(2):184-89.
13. Slade T, Grove R, Burgess P. Kessler Psychological Distress Scale: normative data from the 2007 Australian National Survey of Mental Health and Wellbeing. Aust Nz J Psychiat 2011;**45**(4):308-16.
14. Smith PC, Schmidt SM, Allensworth-Davies D, et al. Primary Care Validation of a Single-Question Alcohol Screening Test. J Gen Intern Med 2009;**24**(7):783-88.
15. Kessler RC, Andrews G, Colpe LJ, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002;**32**(6):959-76.
16. Andrews G, Slade T. Interpreting scores on the Kessler Psychological Distress Scale (K10). Aust Nz J Publ Heal 2001;**25**(6):494-97.
17. Furukawa TA, Kessler RC, Slade T, et al. The performance of the K6 and K10 screening scales for psychological distress in the Australian National Survey of Mental Health and Well-Being. Psychol Med 2003;**33**(2):357-62.
18. Kessler RC, Green JG, Gruber MJ, et al. Screening for serious mental illness in the general population with the K6 screening scale: results from the WHO World Mental Health (WMH) survey initiative. Int J Meth Psych Res 2010;**19**:4-22.

19. Fassaert T, de Wit MAS, Tuinebreijer WC, et al. Psychometric properties of an interviewer-administered version of the Kessler Psychological Distress scale (K10) among Dutch, Moroccan and Turkish respondents. Int J Meth Psych Res 2009;**18**(3):159-68.
20. Furukawa TA, Kawakami N, Saitoh M, et al. The performance of the Japanese version of the K6 and K10 in the World Mental Health Survey Japan. Int J Meth Psych Res 2008;**17**(3):152-58.
21. Brooks RT, Beard J, Steel Z. Factor structure and interpretation of the K10. Psychol Assessment 2006;**18**(1):62-70.
22. Berle D, Starcevic V, Milicevic D, et al. The Factor Structure of the Kessler-10 Questionnaire in a Treatment-Seeking Sample. J Nerv Ment Dis 2010;**198**(9):660-64.
23. Bank W. Bangladesh Current Population. Secondary Bangladesh Current Population 2016. https://data.worldbank.org/country/bangladesh.
24. Hosain GMM, Chatterjee N, Ara N, et al. Prevalence, pattern and determinants of mental disorders in rural Bangladesh. Public Health 2007;**121**(1):18-24.
25. Islam MM, Ali M, Ferroni P, et al. Prevalence of psychiatric disorders in an urban community in Bangladesh. Gen Hosp Psychiat 2003;**25**(5):353-57.
26. Moussavi S, Chatterji S, Verdes E, et al. Depression, chronic diseases, and decrements in health: results from the World Health Surveys. Lancet 2007;**370**(9590):851-58.
27. Hossain MD, Ahmed HU, Chowdhury WA, et al. Mental disorders in Bangladesh: a systematic review. Bmc Psychiatry 2014;**14**.
28. Uddin MN, Bhar S, Al Mahmud A, et al. Psychological distress and quality of life: Rationale and protocol of a prospective cohort study in a rural district in Bangladesh. BMJ Open 2017.
29. Smith AB, Rush R, Fallowfield LJ, et al. Rasch fit statistics and sample size considerations for polytomous data. Bmc Med Res Methodol 2008;**8**.
30. Linacre JM. Sample Size and Item Calibration Stability, 7, p. 328. 1994.
31. May K. Fundamentals of Item Response Theory - Hambleton,Rk, Swaminathan,H, Rogers,Hj. Appl Psych Meas 1993;**17**(3):293-94.
32. Wise PS. Handbook of educational psychology - Berliner,DC, Calfee,RC. Contemp Psychol 1997;**42**(11):983-85.
33. Lord FM, Novick MR, Birnbaum A. Statistical theories of mental test scores. 1968.
34. Ndalichako JL, Rogers WT. Comparison of finite state score theory, classical test theory, and item response theory in scoring multiple-choice items. Educ Psychol Meas 1997;**57**(4):580-89.
35. Lord FM NM. Demonstration of formulae for true measurement of correlation. Am J Psychol 1907;18:161–1968.
36. Cappelleri JC, Lundy JJ, Hays RD. Overview of Classical Test Theory and Item Response Theory for the Quantitative Assessment of Items in Developing Patient-Reported Outcomes Measures Comment. Clin Ther 2014;**36**(5):648-62.
37. Gregory RJ. *Psychological testing: History, principles, and applications*: Allyn & Bacon, 2004.
38. Birnbaum A. Some latent trait models and their use in inferring an examinee's ability. Statistical theories of mental test scores 1968.
39. Bock RD. A brief history of item theory response. Educational measurement: issues and practice 1997;**16**(4):21-33.
40. Chang C-H, Reeve BB. Item response theory and its applications to patient-reported outcomes measurement. Evaluation & the Health Professions 2005;**28**(3):264-82.

41. Nguyen TH, Han H-R, Kim MT, et al. An introduction to item response theory for patient-reported outcome measurement. The Patient-Patient-Centered Outcomes Research 2014;**7**(1):23-35.
42. Bartholomew D. Fundamentals of Item Response Theory - Hambleton,Rk, Swaminathan,H, Rogers,Hj. Brit J Math Stat Psy 1993;**46**:184-85.
43. Raykov T, Marcoulides GA. Fundamentals and Models of Item Response Theory. Introduction to Psychometric Theory 2011:269-304.
44. Andrich D. Controversy and the Rasch model: a characteristic of incompatible paradigms? Med Care 2004;**42**(1):I-7.
45. Andrich D. Rating scales and Rasch measurement. Expert review of pharmacoeconomics & outcomes research 2011;**11**(5):571-85.
46. Andrich D. The Legacies of RA Fisher and K. Pearson in the Application of the Polytomous Rasch Model for Assessing the Empirical Ordering of Categories. Educ Psychol Meas 2013;**73**(4):553-80.
47. Rasch G. An Item Analysis Which Takes Individual Differences into Account. Brit J Math Stat Psy 1966;**19**:49-&.
48. Andrich D. Rating Formulation for Ordered Response Categories. Psychometrika 1978;**43**(4):561-73.
49. RUMM2030 For analysing assessment and attitude questionnaire data [program], 2017.
50. Engelhard G. Rasch Models for Measurement - Andrich,D. Appl Psych Meas 1988;**12**(4):435-36.
51. Leon AC. Multiplicity-adjusted sample size requirements: A strategy to maintain statistical power with Bonferroni adjustments. J Clin Psychiat 2004;**65**(11):1511-14.
52. Gerbing DW, Anderson JC. An Updated Paradigm for Scale Development Incorporating Unidimensionality and Its Assessment. J Marketing Res 1988;**25**(2):186-92.
53. Alan Tennant, Conaghan PG. The Rasch Measurement Model in Rheumatology:What Is It and Why Use It? When Should It Be Applied, and What Should One Look for in a Rasch Paper? 2007.
54. . Psychological Testing in the Service of Disability Determination. Washington (DC), 2015.
55. Smith RM. Fit analysis in latent trait measurement models. J Appl Meas 2000;**1**(2):199-218.
56. Andrich D, Sheridan B, Lyne A, et al. RUMM: A windows-based item analysis program employing Rasch unidimensional measurement models. Perth, Australia: Murdoch University 2000.
57. Dickens GL, Rudd B, Hallett N, et al. Factor validation and Rasch analysis of the individual recovery outcomes counter. Disabil Rehabil 2017:1-12.
58. Jones PW, Harding G, Berry P, et al. Development and first validation of the COPD Assessment Test. Eur Respir J 2009;**34**(3):648-54.
59. McDowell J, Courtney M, Edwards H, et al. Validation of the Australian/English version of the Diabetes Management Self-Efficacy Scale. Int J Nurs Pract 2005;**11**(4):177-84.
60. Eaton NR, Keyes KM, Krueger RF, et al. An invariant dimensional liability model of gender differences in mental disorder prevalence: evidence from a national sample. J Abnorm Psychol 2012;**121**(1):282-8.
61. Baillie AJ. Predictive gender and education bias in Kessler's psychological distress Scale (K10). Soc Psych Psych Epid 2005;**40**(9):743-48.
62. Kessler RC, Foster, C., Webster, P. S., & House, J. S. The relationship between age and depressive symptoms in two national surveys. Psychol Aging 1992; **7**(1):119-26.

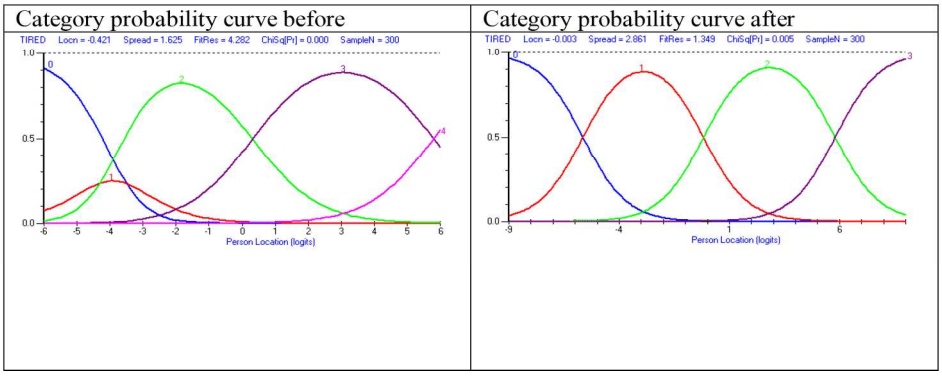
63. Feinson MC. Are Psychological Disorders Most Prevalent among Older Adults - Examining the Evidence. Soc Sci Med 1989;**29**(10):1175-81.
64. Newmann JP. Aging and Depression. Psychol Aging 1989;**4**(2):150-65.
65. Roberts RE. Reliability of the CES-D Scale in different ethnic contexts. Psychiatry Res 1980;**2**(2):125-34.
66. Fan XT. Item response theory and classical test theory: An empirical comparison of their item/person statistics. Educ Psychol Meas 1998;**58**(3):357-81.
67. Hagquist C, Bruce M, Gustavsson JP. Using the Rasch model in nursing research: An introduction and illustrative example. Int J Nurs Stud 2009;**46**(3):380-93.

**Figure 1. Threshold maps of the original Kessler K10 items**



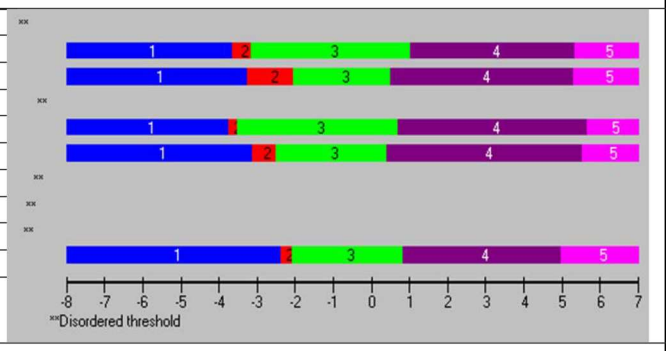
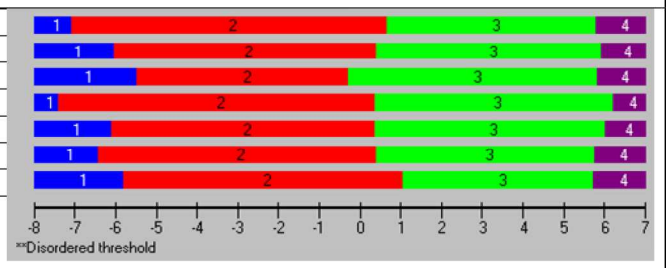

173x73mm (300 x 300 DPI)

**Figure 2. Category probability curve of item 'feel tired' before and after rescoring**



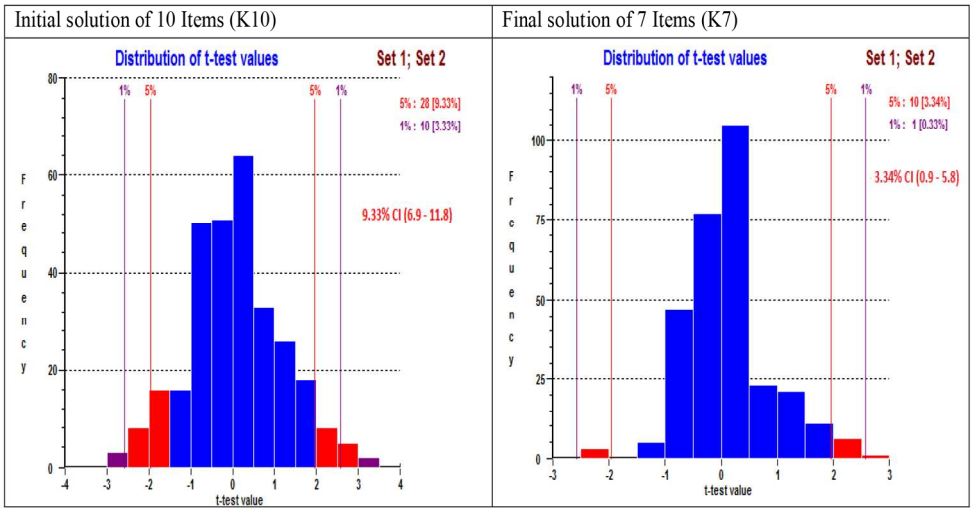

173x74mm (300 x 300 DPI)

**Figure 3. Threshold maps of the original 10-items (Kessler 10) vs. the final 7-items model**

| Initial solution (10 items) | |
|---|---|
| Feel tired (1)<br>Feel nervous (2)<br>Feel so nervous (3)<br>Feel hopeless (4)<br>Feel restless or fidgety (5)<br>Feel so restless (6)<br>Feel depressed (7)<br>Everything was an effort (8)<br>Feel so sad (9)<br>Feel worthless (10) |  |
| Final solution (7 items) | |
| Feel nervous (2)<br>Feel so nervous (3)<br>Feel hopeless (4)<br>Feel restless or fidgety (5)<br>Feel so restless (6)<br>Everything was an effort (8)<br>Feel so sad (9)<br>Rescoring all to 01123 |  |

173x138mm (300 x 300 DPI)

**Figure 4. Dimensionality testing original 10-items (Kessler 10) vs the final 7-items model**




173x98mm (300 x 300 DPI)

**STROBE 2007 (v4) Statement—Checklist of items that should be included in reports of *cohort studies***

| Section/Topic | Item # | Recommendation | Reported on page # |
|---|---|---|---|
| **Title and abstract** | 1 | (*a*) Indicate the study's design with a commonly used term in the title or the abstract | Pages 1 and 2 |
| | | (*b*) Provide in the abstract an informative and balanced summary of what was done and what was found | Page 2 |
| **Introduction** | | | |
| Background/rationale | 2 | Explain the scientific background and rationale for the investigation being reported | Page 3 |
| Objectives | 3 | State specific objectives, including any prespecified hypotheses | Page 5 |
| **Methods** | | | |
| Study design | 4 | Present key elements of study design early in the paper | Page 5 |
| Setting | 5 | Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection | Page 5 (in protocol paper)(ref 28) |
| Participants | 6 | (*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up | Page 5 (in protocol paper)(ref 28) |
| | | (*b*) For matched studies, give matching criteria and number of exposed and unexposed | |
| Variables | 7 | Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable | Page 6 |
| Data sources/ measurement | 8* | For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group | *Page 6* |
| Bias | 9 | Describe any efforts to address potential sources of bias | |
| Study size | 10 | Explain how the study size was arrived at | Page 5 (in protocol paper)(ref 28) |
| Quantitative variables | 11 | Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why | |
| Statistical methods | 12 | (*a*) Describe all statistical methods, including those used to control for confounding | Pages 7-9 |
| | | (*b*) Describe any methods used to examine subgroups and interactions | |
| | | (*c*) Explain how missing data were addressed | |
| | | (*d*) If applicable, explain how loss to follow-up was addressed | In protocol paper (ref 28) |
| | | (*e*) Describe any sensitivity analyses | |

| | | | |
|---|---|---|---|
| **Results** | | | |
| Participants | 13* | (a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed | Page 5 |
| | | (b) Give reasons for non-participation at each stage | |
| | | (c) Consider use of a flow diagram | |
| Descriptive data | 14* | (a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders | Table 1 on page 17 |
| | | (b) Indicate number of participants with missing data for each variable of interest | |
| | | (c) Summarise follow-up time (eg, average and total amount) | |
| Outcome data | 15* | Report numbers of outcome events or summary measures over time | |
| Main results | 16 | (*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included | |
| | | (*b*) Report category boundaries when continuous variables were categorized | |
| | | (*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period | |
| Other analyses | 17 | Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses | Appendix pages 21-22 |
| **Discussion** | | | |
| Key results | 18 | Summarise key results with reference to study objectives | Page 12 |
| **Limitations** | | | page 15 |
| Interpretation | 20 | Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence | Pages 12-15 |
| Generalisability | 21 | Discuss the generalisability (external validity) of the study results | |
| **Other information** | | | |
| Funding | 22 | Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based | Page 16 |

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

**Note:** An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.

BMJ Open

# Psychometric evaluation of an interview-administered version of the Kessler 10-item questionnaire (K10) for measuring psychological distress in rural Bangladesh



| | |
|---|---|
| Journal: | *BMJ Open* |
| Manuscript ID | bmjopen-2018-022967.R1 |
| Article Type: | Research |
| Date Submitted by the Author: | 29-Apr-2018 |
| Complete List of Authors: | Uddin, Mohammed ; Swinburne University of Technology, Department of Statistcs, Data science and Epidemiology<br>Islam, Fakir; Swinburne University of Technology, Statistics, Data Science and Epidemiology<br>Al-Mahmud , Abdullah ; Swinburne University of Technology, 3. School of Design; Faculty of Health, Arts and Design |
| Primary Subject Heading: | Health services research |
| Secondary Subject Heading: | Mental health |
| Keywords: | Phychological Distress, Rasch Analysis, Validation, Kessler Psychological Distress Scale, Rural Bangladesh, Item Response Theory |
| | |

# Title: Psychometric evaluation of an interview-administered version of the Kessler 10-item questionnaire (K10) for measuring psychological distress in rural Bangladesh

Mohammed Nazim Uddin[1], Fakir M Amirul Islam[1,2*] , Abdullah Al Mahmud[3]

1. Department of Statistics, Data Science and Epidemiology; Faculty of Health, Arts and Design; Swinburne University of Technology, Hawthorn VIC 3122, Australia
2. Organisation for Rural Community Development, Dariapur, Narail, Bangladesh
3. Centre for Design Innovation, School of Design, Swinburne University of Technology, Hawthorn, VIC 3122, Australia

Email: Mohammed Nazim Uddin – mnuddin@swin.edu.au; Fakir M Amirul Islam – fislam@swin.edu.au; Abdullah Al Mahmud – aalmahmud@swin.edu.au;

*Corresponding author:

Email: fislam@swin.edu.au

**Title: Psychometric evaluation of an interview-administered version of the Kessler 10-item questionnaire (K10) for measuring psychological distress in rural Bangladesh**

## ABSTRACT

**Objective** The aim of this study was to translate, adapt and validate the Kessler 10-item questionnaire (K10) for measuring psychological distress in rural Bangladesh.

**Design** Cohort study.

**Setting** Narail district, Bangladesh.

**Participants** A random sample of 2425 adults of age 18–90 years were recruited.

**Outcome measure** Validation of the K10 was the major outcome. Socio-demographic factors were measured to assess if the K10 needed adjustment for factors such as age or gender. The Rasch measurement model was used for the validation, and RUMM2030 and SPSS24 software were used for analyses.

**Results** Initial inspection of the total sample showed poor overall fit. A sample size of 300, which is more satiated for Rasch analysis, also showed poor overall fit, as indicated by a significant item-trait interaction ($\chi^2$ = 262.27, df = 40, p < 0.001) and item fit residual values (mean = –0.25, SD = 2.49). Of 10 items, five items were disordered thresholds, and seven items showed misfit, suggesting problems with the response format and items. After removing three items ("feel tired", "depressed" and "worthless") and changing the Likert scale categories from five to four categories, the remaining seven items showed ordered threshold. A revised seven-item scale has shown adequate internal consistency, with no evidence of multidimensionality, no differential item functioning (DIF) on age and gender, and no signs of local dependency.

**Conclusions** Analysis of the psychometric validity of K10 using the Rasch model showed that 10 items are not appropriate for measuring psychological distress in rural Bangladesh. A modified version of seven items (K7) with four-response category would provide a psychometrically more robust scale than the original K10. The study findings suggest repeating the K7 version in other remote areas for further validation can substantiate an efficient screening tool for measuring psychological distress among the general Bangladeshi population.

**Strengths and limitations of the study**

- This study provides the first reliable data on the K10 questionnaire from a general population of a typical rural district in Bangladesh.
- This study used numerous primary data on K10 and associated covariates.
- The data were collected through face-to-face interviews of people from a typical rural district that generally represents Bangladesh.
- The sophisticated Rasch analysis technique was applied to validate as well as identify a suitable unidimensional structure of the K10. The study provides a unique opportunity to assess psychological distress in a rural population of Bangladesh.
- The potential drawback of this study is that it is based on a single-occasion collection of data from a rural district in Bangladesh. While we have attempted to capture the situation in the Narail district, the study needs to be repeated in a random sample of other rural districts to be truly representative of the national population.

## Introduction

A high prevalence of psychological distress is recognised worldwide.[1] Psychological distress is associated with chronic diseases and other health related problems,[2] and early diagnosis is seen as an important measure to ensure effective and targeted intervention.[3] In recent years, epidemiological studies have attempted to employ short dimensional scales to effectively measure and monitor the extent of psychological distress in the general community for the purposes of early diagnosis.[4] The Kessler 10-item questionnaire (K10) is one such scale among similar tools such as the Beck Depression Inventory (BDI),[5] the Hospital Anxiety and Depression Scale (HADS)[6] and the Depression Anxiety Stress Scales (DASS)[7] which are designed to assess non-specific psychological distress and screen for common psychiatric disorders.[8-11]

The K10 was developed in 1992 by Professor Kessler and Mroczek[12] to be used in the United States National Health Interview Survey as a brief measure of non-specific psychological distress along the anxiety-depression spectrum. The K10 comprises ten questions (rated on five-point Likert-type scales, where 1 = none of the time to 5 = all of the time) about psychological distress.

Although K10 is not a diagnostic tool, it does indicate psychological distress and is used to identify people in need of further assessment for anxiety and depression. The K10 measurement of a client's psychological distress levels can also be used as an outcome measure and assist treatment planning and monitoring.[13] In the context of the general population, there is often a shortage of space for the inclusion of more items in the scale. The BDI (21 items),[5] HADS (14 items)[6] and the DASS (42 items)[7] are limited as screening tools because of their long list of items. Moreover, studies confirm that well-constructed short scales can be as strong predictors as the more lengthy instruments or interviews.[12 14] Because of its small number of items, the K10 has, since its development, been widely used in many countries, including the USA, Canada and Australia. The tool has also being adopted in the World Health Organization's World Mental Health Survey.[8-11 15] Moreover, another advantage of the K10 is that it was developed using methods associated with the item response theory.[15]

Although the K10 was originally developed to identify levels of non-specific psychological distress in the general population, the tool has also demonstrated a strong relationship with severe mental illnesses as defined by structured diagnostic interviews.[16] As such, clinicians have been encouraged to use the K10 to screen for psychiatric illness.[17 18] Further, the K10 has been used as a routine outcome measure in specialist public mental health services in multiple Australian states and territories.[4] A recent review of the literature suggests that the K10 is an effective and reliable assessment tool applicable to a variety of settings and cultures for detecting the risk of clinical psychological disorders.[19 20] However, a major limitation of the K10 is the lack of consistency across studies about its factor structure. Although it was initially designed to yield a single score indicating the level of psychological distress,[15] one study demonstrated a four-factor model with acceptable fit in large community samples;[21] another study proposed a two-factor solution, one factor for depression and another for anxiety;[4] while another study did not find an adequate fit.[22]

Bangladesh is a country of 163 million people[23] where mental health complaints are a major public health concern, especially in rural areas.[24-26] The prevalence of mental disorders in such areas varies between 6.5% and 31%, possibly due to the use of different protocols and definitions of mental disorders.[27] A culturally validated tool is needed for quick screening of psychological

distress in Bangladesh, as well as in other countries with similar socio-economic conditions. Due to lack of published research on the K10 in rural settings, and uncertainties surrounding the scale noted above, we need to develop a valid measurement scale of psychological distress in Bangladesh.

The present study pursues an update of Rasch analysis technique to evaluate the suitability of the K10 for measuring psychological distress in rural Bangladesh, and to provide guidance on suitable modification to the instrument to improve its performance. Accuracy and precision of K10 scores can lead to a more efficient allocation of health care resources as well as more efficient screening of psychological distress among the rural population.

## Materials and Methods

### Study Population

Participants were recruited from the Narail district, located approximately 200 km south-west of Dhaka, the capital city of Bangladesh. We recruited a total of 2425 adults aged 18–90 years, from May to July 2017. The study protocol, including its geographic location and population density, is described in detail elsewhere.[28]

### Sample Size and Statistical Power

A sample of approximately 300 is more suitable for a Rasch analysis, because large sample sizes can result in type 1 errors that falsely reject an item for not fitting in the Rasch model.[29] A sample size of 300 is considered large enough for 99% confidence that the estimated item difficulty would be within ±½ logit of its stable value.[30] We did the analysis five times with five different random sample sizes of 300 each, from the total sample of 2425, to check the robustness of the models using different subsamples. For the initial test of the model, we also used the total sample.

### Sampling Frame

A multilevel cluster random sampling technique was used for this cohort study. Three unions (smallest rural administrative unit) out of 13 and 1 Pourashava (smallest urban administrative unit) of Narail Upazilla (the third largest type of administrative division in Bangladesh) were

randomly selected at level 1. Two to three villages (a smallest territorial and social unit for administrative and representative purposes), from each selected union and two wards (an electoral district, for administrative and representative purposes) were randomly chosen from selected Pourashava at the second level. In total, 150 adults (18–59 year old) and 120 older adults (60–90 year old) from each of the villages/wards were interviewed. Recruitment strategy and quality assurance in data collection are described previously.[28]

**Patient and public involvement**

Our study participants are the general people with or without any particular disease. There was a public involvement in conducting the research including informing the district commissioner, district police super, civil surgeon, and the public representatives such as the Chairman of the union parishad. We conducted a pilot survey and arranged a focus group discussion regarding the understanding of the questionnaire by the general people.

Recruitment strategy was reported in the protocol paper.[28] To maintain an approximately equal number of male and female participants, one female was interviewed immediately after a male participant. Participants did not involve in the recruitment to and conduct of the study. Although the results are being published in peer-reviewed journals, the results will be disseminated via community briefs and presentations at national and international conferences. However, the participants those will be identified with severe psychological depressed, the Organisation for Rural Community Development (ORCD) intends to refer them to the psychologists for their treatment. This is also plan to use the modified version of the questionnaire for mass scale screening program for measuring psychological distress.

**Kessler Psychological Distress Scale**

The K10 measures how often participants have experienced symptoms of anxiety and depressive disorders in the previous four weeks prior to screening.[12] Respondents were asked, 'During the past four weeks, how often did you feel: 1) tired out for no good reason; 2) nervous; 3) so nervous that nothing could calm you down; 4) hopeless; 5) restless or fidgety; 6) so restless you could not sit still; 7) sad or depressed; 8) so depressed that nothing could cheer you up; 9) everything was an effort; 10) worthless.' Items are rated on a five-point ordinal scale: all of the

time (score 5), most of the time (score 4), some of the time (score 3), a little of the time (score 2) and none of the time (score 1). Questions 3 and 6 are not asked if the preceding question was answered 'none of the time', in which case questions 3 and 6 would automatically receive a score of one. Scores for the ten questions are summed: the maximum score is 50, indicating severe distress; the minimum score is ten indicating no distress. Low scores indicate low levels of psychological distress and high scores indicate higher levels of psychological distress.[12]

**Outcome Variables**

The main outcome measure was the validation of the K10.

**Factor Variables of Differential Item Functioning (DIF)**

Participants were categorised as either adults (18 to 59 year old) or older adults (60 to 90 year old), and by gender (male or female).

**Scale Validation:** Item response theory (IRT) and Classical Test Theory (CTT)

**Item Response Theory (IRT)**

IRT is a paradigm for the design, analysis and scoring of tests, questionnaires and similar instruments measuring abilities, attitudes or other variables.[31] It is based on the relationship between individuals' performances on a test item and their personal performance on an overall measure of the ability that the item seeks to quantify.[32] All IRT models attempt to explain observed (actual) item performance as a function of an underlying ability (unobserved) or latent trait.

**Classical Test Theory (CTT)**

Classical test theory is a quantitative approach to testing the reliability and validity of a scale based on its items. CTT is a simple linear model which links the observable score (X) to the sum of two unobservable (often called latent) variables, true score (T) and error score (E); i.e., X = T + E. Because of, each examinee there are two unknowns, without simplified assumption the equation will not be solved. The assumptions in the classical test model are that (a) true scores and error scores are uncorrelated, (b) the average error score in the population of examinees is

zero, and (c) error scores on parallel tests are uncorrelated.[33] The true score (T) is defined as the expected value of the observed score over an infinite number of repeat administrations of the same instrument.[33 34]

**Rationale for using the Rasch analysis instead of the CTT**

Similar to the IRT, the CTT is another fundamental measurement theory that researchers employ to construct measures of latent traits. Both IRT and CTT can be used to construct measures of latent traits, but the two measurement systems are entirely dissimilar. A more in-depth explanation of the literature on CTT[35-37] and IRT.[38-41] So far, the K10 was validated mostly using CTT in which the items and the latent trait being measured are considered separately and, therefore, cannot be meaningfully and systematically compared.[42 43] These limitations can be solved rationally using Rasch modelling.[38 39 44-46]

**The Rasch Model**

The Rasch model was named after the Danish mathematician Georg Rasch.[47] The model shows what should be expected of responses to items if measurement (at the metric level) is to be achieved. Two versions of the Rasch model are available:

dichotomous, $P\{X_{ni} = x\} = \frac{e^{x(\beta_n - \delta_i)}}{1+e^{x(\beta_n - \delta_i)}}$ ;[47]

and polytomous, $P\{X_{ni} = x\} = \frac{e^{-\tau_{1i}-\tau_{2i}\cdots\cdots-\tau_{xi}+x(\beta_n-\delta_i)}}{\sum_{x'=0}^{m_i} e^{-\tau_{1i}-\tau_{2i}\cdots\cdots-\tau_{xi}+x(\beta_n-\delta_i)}}$ ;[48]

where $\beta_n$ is the location of person n and $\delta_i$ is the location of item i.$\tau_{xi}$, $x = 1, 2, \ldots, m_i$ are thresholds which partitioned the latent continuum of item i into $m_i + 1$ ordered categories. X is the response value that qualifies the expression by $\beta_n - \delta_i$.

The Rasch analysis in this study was conducted using the RUMM 2030 package.[49] In the assessment of K10, respondents were presented with the ten-item questionnaire regarding psychological distress. The purpose of the Rasch analysis was to maximise the homogeneity of the trait and to allow more significant reduction of redundancy without sacrificing the measurement of information by decreasing items and scoring levels to yield a more valid and straightforward measure. The Rasch model requires some assumptions that need to be evaluated

to ensure that an instrument has Rasch properties. The Rasch assumptions most commonly assessed are a) unidimensionality, b) local independence and c) invariability.

Chi-square item-trait interaction statistics define the overall fit of the model for the scale.[50] A non-significant chi-square probability value indicated that the hierarchical ordering of the items is consistent across all levels of the underlying trait. A Bonferroni adjustment[51] is typical of the alpha value used to assess statistical significance, by dividing the alpha value of 0.05 by the number of items in the scale. Item-person interaction statistics distributed as z-statistic with a mean of zero and SD of 1 (indicating perfect fit with the model). Values of SD above 1.5 for either items or person suggest a problem. Individual item fit statistics are presented as residuals (acceptable within the range ±2.5) and chi-square statistic (require a non-significant chi-square value).

The Rasch model can be extended to analyse items with more than two response categories, which involves a 'threshold' parameter, represented by the two response categories where either response is probable. Common sources of item misfit occur with 'disorder thresholds' failure of the respondents to use the response category in a manner consistent with the level of the trait being measured.

Unidimensionality occurs when a set of items measures just one thing in common.[52] To establish this, the first step is to run a Principal Component Analysis (PCA) on the residuals to identify two subsets of the items having the most difference. Second, the items loading on the first factor are extracted, items having positive and negative loadings are defined, and estimates for these two sets are derived. Applying an independent t-test to both sets, which conduct t-tests for each person in the sample comparing their score on the Set 1 items and Set 2 items. If less than 5% of the estimates are outside the range of ±1.96, the scale is considered unidimensional.

In case of local independence,[53] the items in a test are expected to be unrelated to each other; i.e. the response on each item should not be associated with that of another items. To test for local independence, we need to check the residuals correlation matrix, and any correlation coefficient value greater than 0.3 suggests the two items are locally dependent. In a situation where the

correlation value is greater than 0.3, the two items need to be merged into one, called subtest analysis, to achieve a significant improvement on PSI value. If so, it is a sign of local dependency and a violation of one of the Rasch assumptions.

Invariability indicates that 'items are not dependent on the distribution of persons' abilities and the persons' abilities are not dependent on the test items.[54] In Rasch measurement theory, the scale should work in the same way, irrespective of which group (e.g., gender or age) is being assessed. If for some reason one gender does not display equal likelihood of confirming the item, then the items would display DIF and would violate the requirement of unidimensionality.[55] DIF is an analysis of variance of the person-item deviation residuals with the person's factors (e.g., age, gender).

The reliability and internal consistency of the model are defined by the Person Separation Index (PSI).[56] In addition to item fit, examination of person fit is essential. A few responses with unusual response pattern (identified by high positive residuals) may seriously affect the fit at the item level. Such aberrant response patterns occur due to unrecorded co-morbidity or respondents with cognitive defects. Therefore, if some response pattern showed high positive fit residuals, removal from the analysis may make a significant difference to the scale internal construct validity.

## Results

### *Overview of the respondents*

Table 1 shows the summary statistics of both the validation and the total data sets by gender (male and female). The mean (SD, range) age of the total participant sample was 52.0 years (17, 18–90). Of the total sample, 48.5% were men, 27.6% had no formal education, 4% had at least a bachelor's degree level of education.

### *Primary analysis of the original set of ten items and five response categories*

K10 scores ranged from 10 to 50 with a mean of 16.7 (SD = 11.3). Initial inspection of the scale with the total 2425 participants showed poor overall fit with the Rasch model, as indicated by a significant item-trait interaction ($\chi^2$ = 1729.89, df = 40, $p < 0.001$) and item fit residual values

(mean = –0.25, SD = 6.75) outside the acceptable range. Eight items were found to be misfit based on the overall fit residual values outside the range of ±2.5. Five items were found to have disordered thresholds, signifying problems with the 5-point response format used for the scale. A check found multidimensionality: the model fit statistics for the five separate random subsamples of 300 each from the total participant sample produced almost identical results, indicating the results and sample selections were robust (Table 2).

Initial inspection of scores in the random sample of 300 participants showed poor overall fit to the Rasch model ($\chi^2$ = 262.27, df = 40, p < 0.001) and items fit residual values (mean = –0.25, SD = 2.49). However, the person fit residuals (mean = 0.18, SD = 1.24) were within the acceptable range (Table 2, sample 1). Five items were found to have disordered thresholds, and seven of the individuals' item fit statistics showed misfit, suggesting problems with the 5-point response format used for the questionnaire. The value of the PSI (analogous to Cronbach's alpha) for the original set of ten items with five response categories was 0.84, indicating that the scale worked well to separate persons. The frequency distribution of the items showed (data not shown) mistargeting. Across all five items, the distribution was skewed towards the lower values, indicating low psychological distress among the respondents in the sample. Seven items (items 1, 2, 3, 4, 7, 8 and 9) showed misfit (Table 3: initial solution) while five items showed disorder thresholds (1, 4, 7, 8, 9) (Figure 1: initial solution). A visual examination of the threshold map shows that the estimates of the thresholds defining the categories in item 1 (tired) (Figure 2: category probability curve), item 4 (feel hopeless), item 7 (depressed), item 8 (an effort) and item 9 (so sad) do not form distinctive regions of the continuum. We have examined the category probability curve of each disorder threshold item, and found response 1 and 2 adjacent category were not the same (Figure 2, category probability curve).

To address the issue of disordered categories, Rasch analysis was conducted on only the disordered items, by merging the two middle categories ('a little of the time' and 'some of the time'). This reduced the scoring to a 4-point format from 01234 to 01123, and made the overall score range 0 to 40. Following this, eight misfit items were identified with significant chi-square probability values, or high positive or high negative residual values (± 2.5), and found only item 5 to be disordered. (Table 3: only disorder items were rescored as 01123). Then we carried out

all items Likert scale categories from five to four categories and found all items were ordered thresholds. (Figure 1: rescore all items to 01123). However, five items were still misfit in the model (Table 3: rescore all items to 01123).

***Proposed final analysis of the seven items and four response categories***

Misfit items were removed one at a time iteratively, based on positive or negative residual values as well as the degree of the significant chi-square probability values. The total model fit and individual item fit statistics were checked after each iteration, until the remaining items were shown to fit Rasch model's expectations. The three removed items were items 1, 7 and 10.

The final solution, retaining seven items, showed overall fit with the model (Table 4). The PSI was found to be high (PSI = 0.84), making the model suitable for individual use. The items of the K7 scale were assessed for DIF across gender (male/female) and age (adults: 18–59 year old) and older adults (60–90 year old) (Table 5). A significant DIF was found on item 9 (feel so sad); however, using a Bonferroni-adjusted alpha value (.05/7=.007), the value became non-significant. In the final model, seven items with four response categories showed all items to have ordered thresholds (Figure 3). There was no indication of item or person misfit (Table 4: Individuals' items fit statistics of final K7). Unidimensionality of the K7 scale was tested using PCA (3.34%, 95% CI 0.9% to 5.8%), and from a binomial distribution was found non-significant, which supports unidimensionality of the K7 (Table 4, final solution of K10 and Figure 4, final solution of K7). The details statistical analysis history of the K10 using Rasch analysis is shown in (Appendix).

## Discussion

The purpose of the paper was to evaluate the suitability of the Kessler 10-item questionnaire for measuring psychological distress in rural Bangladesh. This article examines the potential contribution of Rasch analysis in exploring several issues concerning the K10. This includes an assessment of the appropriateness of using all K10 items to represent the underlying dimension of psychological distress. In addition, the article includes an evaluation of the validity of the category scoring system, the fit of individual items and an assessment of the potential bias of items by gender and age, from the perspective of the Rasch model. The initial descriptive

analysis of the frequency distributions indicated that the 10-item scale with five response categories mistargeted the current sample of the rural Bangladeshi population. Non-responses or very few responses in the categories may manifested to the mistargeting. Two items ('tired' and 'depressed') showed misfit, and two items ('so nervous' and 'so restless') showed redundancy (i.e., little impact on the scale). Moreover, items with disordered thresholds indicating problems with the categorisation of the items and scale showed evidence of multidimensionality. Since the K10 scale has not previously undergone a rigorous psychometric analysis in rural Bangladesh and even in neighbouring countries, the detection of problems was not surprising, even though attention had been paid to targeting when the scale was constructed. In these circumstances, the analysis elaborated on taking advantage of the Rasch model.

One response category was warped, which resulted in four instead of five response categories for each item. Moreover, those items showing misfit were removed from the model gradually after going through all possible steps to improve the model. Item 1 ('how often did you feel tired out for no good reason') was removed because it showed high fit residuals value and DIF for age (adults and older adults). Although techniques exist for solving uniform DIF by allowing the item difficulty to vary by group, we believe that option is inappropriate because it is not useful as an everyday screening environment. Therefore, we decided to delete the biased item, which also had a large chi-square value. On the other hand, the item may not play the concepts of psychological distress in Bangladesh. This could be one reason why the item works differently according to age (adults and older adults). The removal of this item from the scale improved the overall fit of the model, supporting this decision. Moreover, the item removed was one of the four items that Kessler[15] had earlier used to reduce 10 to 6 items. Item 7 ('how often did you feel depressed') was also removed from the scale due to misfit with the model. The large positive residual value indicates misfit in that it contributed little or no information additional to other items, as well as having a large chi-square value. However, the item showed no DIF on age and gender. Removal of the item from the model significantly improved the fit of remaining items. Moreover, the item removed was one of the four-item that Kessler[15] earlier used to reduce 10 to six items. Item 10 ('how often did you feel worthless') has been removed from the scale due to high chi-square value and significant chi-square probability, as well as high positive residuals which contribute to an overall model misfit. The high chi-square value indicates that it adds

nothing to the information gained by other items, and this item is the only one, which increased the overall chi-square value and made the overall model misfit. The study results support the retention of item 10.

Removal of items from the scale would eliminate at least some redundancy.[57-59] However, our analysis identified that Cronbach's alpha for the K7 (0.88) was equivalent to the original K10 Cronbach's alpha (0.87); in addition, the PSI of K7 (0.84) was the same as that of the original K10's PSI (0.84). A study reported by Fassaert et al,[19] showed that some redundancy happens in Cronbach's alpha, when comparing K10 (0.93) and K6 (0.89). However, our model showed superior value of Cronbach's alpha K7 (0.88) compared to the original K10 (0.87) model, and confirms adequate fit of the model in the rural settings in Bangladesh. Although we have proposed seven validated items (K7), a previous study proposed six (K6) items[17] was more robust than the K10. Of K7, five items were common in K6. We only tested K6 items using Rasch analysis and found a poor overall fit. In particular, the presence of the item "feel worthless" showed a large positive fit residual and significantly large chi-square value, which influenced the overall model misfit under Rasch assumptions. Therefore, the current study found that the K7 model is more robust in our sample compared to K6.[17 20]

Gender differences in psychology are ubiquitous,[60] so it is essential to verify whether the model is affected by gender or not. Our revised seven-item model showed no DIF on gender, i.e., there is no gender bias in the revised K7 scale. The K7 scale is equally valid for men and women, which supports the previous findings reported in Australia.[61] Another important factor is age, and there is inconsistency in the literature on the relationship between age and psychological distress.[62] The study conducted by Kessler et al. documented a good deal of inequality in the relationship between age and screening scales of depressive symptoms.[63] However, other studies showed a stable nonlinear association between age and psychological distress in several cross-sectional epidemiologic surveys.[62 64 65] Our revised model of K7 confirmed that there is no age bias (adults and older adults), and the model is equally applicable to any one between the age of 18–90 years.

Application of the Rasch measurement model in this study has supported the viability of a seven-item version of the K10 scale for measuring psychological distress in rural Bangladesh. The scale shows high reliability, with no disordering of thresholds and no evidence of DIF. The model also showed high PSI (0.84) and reliability (0.87), which indicated the power of the test of fit. Furthermore, there is good evidence from this sample that a single total score of psychological distress is viable. Thus, the seven-item scale appears robust when tested against the strict assumptions of the Rasch measurement model.

This paper shows how the Rasch model can be used for rigorous examination and development of measurement instruments such as the K10 psychological distress scale. The Rasch model simplifies measurement problems such as lack of invariance, which was overlooked in traditional analysis.[66] The Rasch analysis of the K10 scale indicates that the psychometric properties of the original scale most likely would have been much better if scale developmental had been guided by IRT (Rasch analyses). In future, importance should be given to improving the targeting of person and items. Reducing the number of response categories as well as the number of items might also improve the properties of the scale.[67] Therefore, data on the general rural population regarding psychological distress based on the revised seven-item scale from the K10, with four-response category, is superior to the original scale.

This study provides the first reliable data on levels of psychological distress among the general population of rural Bangladesh. The analysis was based on a large data set of adults and older adults across a wide range of age, from whom data were collected directly in a face-to-face interview. The Rasch analysis in this study guided a detailed examination of the structure of the scale. The response category orderings (threshold ordering) were not examined earlier, and evidence from the current study does not support the response format or the validity of the original 10-item scale.

The potential drawback of this study is that it is based on single-occasion collection of data from people in a rural district of Bangladesh. While we have attempted to capture the situation in the Narail district, the study would obviously need to be repeated in a random sample of other rural districts for the results to be truly representative of a national population.

## Conclusion

Overall, the authors favours the use of K10 in rural Bangladesh, as has been used elsewhere. However, this study acknowledges that due to cultural variations and strict adherence to Rasch properties, modification is needed to measure psychological distress in rural Bangladesh. The results of this study suggest that a revised seven-item version of the K10, with four-response category, would provide a more robust psychometric scale than the original K10. The modified seven-item scale fulfils all the assumptions of the Rasch model, and the model has shown no differential item functioning (DIF) on age and sex as well as no local dependency. The study findings can be repeated using a random sample of other remote areas in Bangladesh to further validate the revised scale, as well as to better establish the level of psychological distress nationwide. The tool can be applied in clinical settings at the national level, where psychological distress has yet to be diagnosed.

## List of abbreviations

CTT, Classical Test Theory; IRT, Item Response Theory; DIF, Differential Item Functioning; PSI; Person Separation Index, K10, Kessler Psychological Distress Scale

**Ethical aspects:** Human Ethics Approval was received from the Swinburne University of Technology Human Ethics Committee (SHR Project 2015/065) in accordance with the tenets of the Declaration of Helsinki. Study participants provided written informed consent.

**Authors' contributions:** MNU and FMAI jointly designed the study. MNU analysed the data and drafted the manuscript. FMAI contributed to writing the manuscript. FMAI supervised the overall analyses and preparation of the manuscript. AAM reviewed the manuscript. All authors contributed to the development of the manuscript and read and approved its final version.

**Acknowledgements:** We particularly acknowledge the contribution of Md Rafiqul Islam, Md Sajibul Islam, Saburan Nesa and Arzan Hosen for their hard work in door-to-door data collection. Finally, we would like to express our gratitude to the study participants for their voluntary participation.

**Competing interests:** None declared

**Funding:** The Faculty of Health, Arts and Design of the Swinburne University Technology under the Research Development Grant Scheme, funded data collection for this research project. The funder had no role in the design of the study, data collection or analysis, interpretation of data or writing of the manuscript.

**Data sharing statement:** Data will be available upon request.

**Table 1. Characteristics of participants who were included and who were not in the current study, by gender**

| Characteristic | Total N=2425 | | | In validation N=300 | | |
|---|---|---|---|---|---|---|
| | Total (2425) | Male (1176) | Female (1249) | Total (300) | Male (143) | Female (153) |
| Age groups (in years) | | | | | | |
| 18–59 | 1278 (52.7) | 603 (51.3) | 675 (54.0) | 172 (57.3) | 73 (51.0) | 99 (63.1) |
| 60–90 | 1147 (47.3) | 573 (48.7) | 574 (46.0) | 128 (42.7) | 70 (49.0) | 58 (36.9) |
| Education | | | | | | |
| No education | 671 (27.7) | 289 (24.6) | 382 (30.6) | 76 (25.3) | 37 (25.9) | 39 (24.8) |
| Primary (1–5) | 946 (39.0) | 447 (38.0) | 499 (40.0) | 124 (41.3) | 58 (40.6) | 66 (42.0) |
| Secondary (6–9) | 327 (13.5) | 146 (12.4) | 181 (14.5) | 38 (12.7) | 13 (9.1) | 25 (15.9) |
| SSC or HSC Pass (10–12) | 385 (15.9) | 224 (19.0) | 161 (12.9) | 50 (16.7) | 26 (18.2) | 24 (15.3) |
| Degree or equivalent (13–16) | 96 (4.0) | 70 (6.0) | 26 (2.1) | 12 (4.0) | 9 (6.3) | 3 (1.9) |

SSC – Secondary School Certificate, HSC- Higher Secondary Certificate

**Table 2. Model Fit Statistics for total sample and five random samples of 300 with all 10 items**

| Initial solution | Total sample N=2425 | Sample 1 n=300 | Sample 2 n=300 | Sample 3 n=300 | Sample 4 n=300 | Sample 5 n=300 |
|---|---|---|---|---|---|---|
| Overall model fit, Chi-square value | 1727.89 | 262.27 | 212.30 | 204.07 | 194.37 | 282.14 |
| Degree of freedom (DF) | 40 | 40 | 40 | 40 | 40 | 40 |
| P | 0.00000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Item fit residuals (mean (SD)) | –0.25 (6.75) | 0.13 (2.49) | 0.05 (2.40) | –0.23 (2.12) | 0.11 (2.38) | –0.16 (2.64) |
| Person fit residuals (mean (SD)) | –0.29 (1.32) | –0.18 (1.24) | –0.28 (1.33) | –0.34 (1.32) | –0.30 (1.37) | –0.27 (1.32) |
| Person separation index (PSI) | 0.84 | 0.84 | 0.83 | 0.85 | 0.82 | 0.83 |
| Coefficient alpha | 0.86 | 0.87 | 0.86 | 0.87 | 0.86 | 0.87 |
| Unidimensionality test (% that goes beyond 95% CI) | 10.3 (9.6–11.2) | 9.3 (6.9–11.8) | 11.7 (9.2–14.1) | 8.3 (5.9–10.8) | 9.0 (6.5–11.5) | 10.33 (7.9–12.8) |

**Table 3. Fit statistics (location, residuals and P values) of the 10 items for the first random sample of 300**

| | Initial solution | | | Rescore only disordered items to 01123* | | | Rescore all items to 01123 | | |
|---|---|---|---|---|---|---|---|---|---|
| **Items** | **Location** | **Residuals** | **P value** | **Location** | **Residuals** | **P value** | **Location** | **Residual** | **P value** |
| Feel tired (1) | –0.42 | 4.28 | 0.000*^ | 0.00 | 1.35 | 0.005 | –0.51 | 1.22 | 0.000^ |
| Feel nervous (2) | –0.11 | –0.85 | 0.001^ | –0.56 | –1.19 | 0.004^ | –0.12 | –3.26 | 0.020 |
| Feel so nervous (3) | 0.13 | –3.16 | 0.000^ | –0.32 | –3.65 | 0.002^ | 0.05 | –4.13 | 0.002^ |
| Feel hopeless (4) | –0.06 | -0.62 | 0.008* | 0.34 | –1.54 | 0.001^ | –0.03 | –1.77 | 0.104 |
| Feel restless or fidgety (5) | –0.22 | 0.46 | 0.002^ | –0.69 | 0.11 | 0.001*^ | –0.26 | –1.93 | 0.302 |
| Feel so restless (6) | 0.08 | –3.11 | 0.000*^ | –0.38 | –3.39 | 0.007 | 0.04 | –3.73 | 0.003^ |
| Feel depressed (7) | 0.26 | 3.87 | 0.000*^ | 0.74 | 3.00 | 0.000^ | 0.35 | 3.18 | 0.000^ |
| Everything was an effort (8) | –0.15 | –0.33 | 0.125* | 0.28 | –1.90 | 0.000^ | –0.16 | –2.36 | 0.301 |
| Feel so sad (9) | 0.16 | –0.48 | 0.058 | 0.65 | –2.32 | 0.001^ | 0.25 | –2.64 | 0.003^ |
| Feel worthless (10) | 0.34 | 1.33 | 0.001^ | –0.06 | 2.41 | 0.003^ | 0.39 | 0.60 | 0.247 |

*Disordered items; ^ P values depend on chi-square values (Bonferroni correction (p value/number of items)) =.05/10=.005)

**Table 4. Individuals' item fit statistics of original Kessler K10 and final seven-items model**

| | Individuals' items fit statistics of original K10 | | | | | Individuals' items fit statistics of Final K7 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Items** | **Location** | **SE** | **Residual** | **$\chi^2$** | **P value** | **Location** | **SE** | **Residual** | **$\chi^2$** | **P value** |
| Feel tired (1) | –0.42 | 0.08 | 4.28 | 46.76 | 0.000 | | | | | |
| Feel nervous (2) | –0.11 | 0.09 | –0.85 | 19.94 | 0.001 | –0.20 | 0.15 | –1.40 | 3.99 | 0.41 |
| Feel so nervous (3) | 0.13 | 0.09 | –3.16 | 30.36 | 0.000 | 0.10 | 0.15 | –2.66 | 11.01 | 0.03 |
| Feel hopeless (4) | –0.06 | 0.08 | –0.62 | 13.66 | 0.008 | 0.03 | 0.15 | 0.62 | 3.35 | 0.50 |
| Feel restless or fidgety (5) | –0.22 | 0.09 | 0.46 | 16.88 | 0.002 | –0.28 | 0.16 | –0.81 | 3.98 | 0.41 |
| Feel so restless (6) | 0.08 | 0.09 | –3.11 | 30.53 | 0.000 | 0.09 | 0.15 | –2.78 | 8.04 | 0.09 |
| Feel depressed (7) | 0.26 | 0.09 | 3.87 | 70.15 | 0.000 | | | | | |
| Everything was an effort (8) | –0.15 | 0.08 | –0.33 | 7.21 | 0.125 | –0.09 | 0.15 | –0.86 | 7.03 | 0.13 |
| Feel so sad (9) | 0.16 | 0.09 | –0.48 | 9.11 | 0.058 | 0.34 | 0.16 | –0.56 | 2.42 | 0.65 |
| Feel worthless (10) | 0.34 | 0.09 | 1.33 | 17.69 | 0.001 | | | | | |
| | **Initial solution of K10** | | | | | **Final solution of K7** | | | | |
| Overall model fit | | | | 262.27 | | | | | 39.82 | |
| Degree of freedom (DF) | | | | 40 | | | | | 28 | |
| P | | | | 0.000 | | | | | 0.068 | |
| Item fit residuals (mean (SD)) | | | | 0.13 (2.49) | | | | | –0.20 (1.20) | |
| Person fit residuals (mean (SD)) | | | | –0.18 (1.24) | | | | | –0.63 (1.40) | |
| Person separation index (PSI) | | | | 0.84 | | | | | 0.84 | |
| Coefficient alpha | | | | 0.87 | | | | | 0.88 | |
| Unidimensionality test (% that goes beyond 95% CI) | | | | 9.33% CI (6.9–11.8) | | | | | 3.34% CI (0.9–5.8) | |

**Table 5. DIF on age (adults and older adults) and gender (male and female)**

| Items | DIF on Age | | | | DIF on Gender | | | |
|---|---|---|---|---|---|---|---|---|
| | MS | F | DF | Prob | MS | F | DF | Prob |
| Feel nervous (2) | 0.58 | 0.88 | 1 | 0.35 | 0.59 | 0.91 | 1 | 0.34 |
| Feel so nervous (3) | 1.00 | 1.86 | 1 | 0.17 | 0.06 | 0.11 | 1 | 0.74 |
| Feel hopeless (4) | 0.07 | 0.08 | 1 | 0.78 | 2.41 | 2.59 | 1 | 0.11 |
| Feel restless or fidgety (5) | 0.49 | 0.67 | 1 | 0.41 | 0.66 | 0.89 | 1 | 0.35 |
| Feel so restless (6) | 0.50 | 0.92 | 1 | 0.34 | 0.00 | 0.00 | 1 | 0.98 |
| Everything was an effort (8) | 0.12 | 0.17 | 1 | 0.68 | 0.26 | 0.36 | 1 | 0.55 |
| Feel so sad (9) | 5.29 | 6.86 | 1 | 0.01 | 0.80 | 1.04 | 1 | 0.31 |

**Figure legends**

**Figure 1: Threshold maps of the original Kessler K10 items**

**Figure 2: Category probability curve of item 'feel tired' before and after rescoring**

**Figure 3: Threshold maps of the original 10-items (Kessler 10) vs. the final 7-items model**

**Figure 4: Dimensionality testing original 10-items (Kessler 10) vs the final 7-items model**

## References


1. Lopez AD, Murray CCJL. The global burden of disease, 1990-2020. Nat Med 1998;**4**(11):1241-43.
2. Egede LE. Effect of comorbid chronic diseases on prevalence and odds of depression in adults with diabetes. Psychosomatic Medicine 2005;**67**(1):46-51.
3. Costello EJ. Early Detection and Prevention of Mental Health Problems: Developmental Epidemiology and Systems of Support. J Clin Child Adolesc 2016;**45**(6):710-17.
4. Sunderland M, Mahoney A, Andrews G. Investigating the Factor Structure of the Kessler Psychological Distress Scale in Community and Clinical Samples of the Australian Population. J Psychopathol Behav 2012;**34**(2):253-59.
5. Beck AT, Erbaugh J, Ward CH, et al. An Inventory for Measuring Depression. Arch Gen Psychiat 1961;**4**(6):561-&.
6. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiat Scand 1983;**67**(6):361-70.
7. Parkitny L, McAuley J. The Depression Anxiety Stress Scale (DASS). J Physiother 2010;**56**(3):204-04.
8. Andrews G, Peters L. The psychometric properties of the composite international diagnostic interview. Soc Psych Psych Epid 1998;**33**(2):80-88.
9. Kessler RC, Ustun TB. The World Mental Health (WMH) Survey Initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). Int J Meth Psych Res 2004;**13**(2):93-121.
10. Kessler R MD. Final Versions of our Non-Specific Psychological Distress Scale Ann Arbor(MI): Survey Research Center of the Institute for Social Research, University of Michigan, 1994. 1994.
11. Slade T, Johnston A, Browne MAO, et al. 2007 National Survey of Mental Health and Wellbeing: methods and key findings. Aust Nz J Psychiat 2009;**43**(7):594-605.
12. Kessler RC, Barker PR, Colpe LJ, et al. Screening for serious mental illness in the general population. Arch Gen Psychiat 2003;**60**(2):184-89.
13. Slade T, Grove R, Burgess P. Kessler Psychological Distress Scale: normative data from the 2007 Australian National Survey of Mental Health and Wellbeing. Aust Nz J Psychiat 2011;**45**(4):308-16.
14. Smith PC, Schmidt SM, Allensworth-Davies D, et al. Primary Care Validation of a Single-Question Alcohol Screening Test. J Gen Intern Med 2009;**24**(7):783-88.
15. Kessler RC, Andrews G, Colpe LJ, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002;**32**(6):959-76.
16. Andrews G, Slade T. Interpreting scores on the Kessler Psychological Distress Scale (K10). Aust Nz J Publ Heal 2001;**25**(6):494-97.
17. Furukawa TA, Kessler RC, Slade T, et al. The performance of the K6 and K10 screening scales for psychological distress in the Australian National Survey of Mental Health and Well-Being. Psychol Med 2003;**33**(2):357-62.
18. Kessler RC, Green JG, Gruber MJ, et al. Screening for serious mental illness in the general population with the K6 screening scale: results from the WHO World Mental Health (WMH) survey initiative. Int J Meth Psych Res 2010;**19**:4-22.

19. Fassaert T, de Wit MAS, Tuinebreijer WC, et al. Psychometric properties of an interviewer-administered version of the Kessler Psychological Distress scale (K10) among Dutch, Moroccan and Turkish respondents. Int J Meth Psych Res 2009;**18**(3):159-68.
20. Furukawa TA, Kawakami N, Saitoh M, et al. The performance of the Japanese version of the K6 and K10 in the World Mental Health Survey Japan. Int J Meth Psych Res 2008;**17**(3):152-58.
21. Brooks RT, Beard J, Steel Z. Factor structure and interpretation of the K10. Psychol Assessment 2006;**18**(1):62-70.
22. Berle D, Starcevic V, Milicevic D, et al. The Factor Structure of the Kessler-10 Questionnaire in a Treatment-Seeking Sample. J Nerv Ment Dis 2010;**198**(9):660-64.
23. Bank W. Bangladesh Current Population. Secondary Bangladesh Current Population 2016. https://data.worldbank.org/country/bangladesh.
24. Hosain GMM, Chatterjee N, Ara N, et al. Prevalence, pattern and determinants of mental disorders in rural Bangladesh. Public Health 2007;**121**(1):18-24.
25. Islam MM, Ali M, Ferroni P, et al. Prevalence of psychiatric disorders in an urban community in Bangladesh. Gen Hosp Psychiat 2003;**25**(5):353-57.
26. Moussavi S, Chatterji S, Verdes E, et al. Depression, chronic diseases, and decrements in health: results from the World Health Surveys. Lancet 2007;**370**(9590):851-58.
27. Hossain MD, Ahmed HU, Chowdhury WA, et al. Mental disorders in Bangladesh: a systematic review. Bmc Psychiatry 2014;**14**.
28. Uddin MN, Bhar S, Al Mahmud A, et al. Psychological distress and quality of life: Rationale and protocol of a prospective cohort study in a rural district in Bangladesh. BMJ Open 2017.
29. Smith AB, Rush R, Fallowfield LJ, et al. Rasch fit statistics and sample size considerations for polytomous data. Bmc Med Res Methodol 2008;**8**.
30. Linacre JM. Sample Size and Item Calibration Stability, 7, p. 328. 1994.
31. May K. Fundamentals of Item Response Theory - Hambleton,Rk, Swaminathan,H, Rogers,Hj. Appl Psych Meas 1993;**17**(3):293-94.
32. Wise PS. Handbook of educational psychology - Berliner,DC, Calfee,RC. Contemp Psychol 1997;**42**(11):983-85.
33. Lord FM, Novick MR, Birnbaum A. Statistical theories of mental test scores. 1968.
34. Ndalichako JL, Rogers WT. Comparison of finite state score theory, classical test theory, and item response theory in scoring multiple-choice items. Educ Psychol Meas 1997;**57**(4):580-89.
35. Lord FM NM. Demonstration of formulae for true measurement of correlation. Am J Psychol 1907;18:161–1968.
36. Cappelleri JC, Lundy JJ, Hays RD. Overview of Classical Test Theory and Item Response Theory for the Quantitative Assessment of Items in Developing Patient-Reported Outcomes Measures Comment. Clin Ther 2014;**36**(5):648-62.
37. Gregory RJ. *Psychological testing: History, principles, and applications*: Allyn & Bacon, 2004.
38. Birnbaum A. Some latent trait models and their use in inferring an examinee's ability. Statistical theories of mental test scores 1968.
39. Bock RD. A brief history of item theory response. Educational measurement: issues and practice 1997;**16**(4):21-33.
40. Chang C-H, Reeve BB. Item response theory and its applications to patient-reported outcomes measurement. Evaluation & the Health Professions 2005;**28**(3):264-82.

41. Nguyen TH, Han H-R, Kim MT, et al. An introduction to item response theory for patient-reported outcome measurement. The Patient-Patient-Centered Outcomes Research 2014;**7**(1):23-35.
42. Bartholomew D. Fundamentals of Item Response Theory - Hambleton,Rk, Swaminathan,H, Rogers,Hj. Brit J Math Stat Psy 1993;**46**:184-85.
43. Raykov T, Marcoulides GA. Fundamentals and Models of Item Response Theory. Introduction to Psychometric Theory 2011:269-304.
44. Andrich D. Controversy and the Rasch model: a characteristic of incompatible paradigms? Med Care 2004;**42**(1):I-7.
45. Andrich D. Rating scales and Rasch measurement. Expert review of pharmacoeconomics & outcomes research 2011;**11**(5):571-85.
46. Andrich D. The Legacies of RA Fisher and K. Pearson in the Application of the Polytomous Rasch Model for Assessing the Empirical Ordering of Categories. Educ Psychol Meas 2013;**73**(4):553-80.
47. Rasch G. An Item Analysis Which Takes Individual Differences into Account. Brit J Math Stat Psy 1966;**19**:49-&.
48. Andrich D. Rating Formulation for Ordered Response Categories. Psychometrika 1978;**43**(4):561-73.
49. RUMM2030 For analysing assessment and attitude questionnaire data [program], 2017.
50. Engelhard G. Rasch Models for Measurement - Andrich,D. Appl Psych Meas 1988;**12**(4):435-36.
51. Leon AC. Multiplicity-adjusted sample size requirements: A strategy to maintain statistical power with Bonferroni adjustments. J Clin Psychiat 2004;**65**(11):1511-14.
52. Gerbing DW, Anderson JC. An Updated Paradigm for Scale Development Incorporating Unidimensionality and Its Assessment. J Marketing Res 1988;**25**(2):186-92.
53. Alan Tennant, Conaghan PG. The Rasch Measurement Model in Rheumatology:What Is It and Why Use It? When Should It Be Applied, and What Should One Look for in a Rasch Paper? 2007.
54. . Psychological Testing in the Service of Disability Determination. Washington (DC), 2015.
55. Smith RM. Fit analysis in latent trait measurement models. J Appl Meas 2000;**1**(2):199-218.
56. Andrich D, Sheridan B, Lyne A, et al. RUMM: A windows-based item analysis program employing Rasch unidimensional measurement models. Perth, Australia: Murdoch University 2000.
57. Dickens GL, Rudd B, Hallett N, et al. Factor validation and Rasch analysis of the individual recovery outcomes counter. Disabil Rehabil 2017:1-12.
58. Jones PW, Harding G, Berry P, et al. Development and first validation of the COPD Assessment Test. Eur Respir J 2009;**34**(3):648-54.
59. McDowell J, Courtney M, Edwards H, et al. Validation of the Australian/English version of the Diabetes Management Self-Efficacy Scale. Int J Nurs Pract 2005;**11**(4):177-84.
60. Eaton NR, Keyes KM, Krueger RF, et al. An invariant dimensional liability model of gender differences in mental disorder prevalence: evidence from a national sample. J Abnorm Psychol 2012;**121**(1):282-8.
61. Baillie AJ. Predictive gender and education bias in Kessler's psychological distress Scale (K10). Soc Psych Psych Epid 2005;**40**(9):743-48.
62. Kessler RC, Foster, C., Webster, P. S., & House, J. S. The relationship between age and depressive symptoms in two national surveys. Psychol Aging 1992; **7**(1):119-26.

63. Feinson MC. Are Psychological Disorders Most Prevalent among Older Adults - Examining the Evidence. Soc Sci Med 1989;**29**(10):1175-81.
64. Newmann JP. Aging and Depression. Psychol Aging 1989;**4**(2):150-65.
65. Roberts RE. Reliability of the CES-D Scale in different ethnic contexts. Psychiatry Res 1980;**2**(2):125-34.
66. Fan XT. Item response theory and classical test theory: An empirical comparison of their item/person statistics. Educ Psychol Meas 1998;**58**(3):357-81.
67. Hagquist C, Bruce M, Gustavsson JP. Using the Rasch model in nursing research: An introduction and illustrative example. Int J Nurs Stud 2009;**46**(3):380-93.

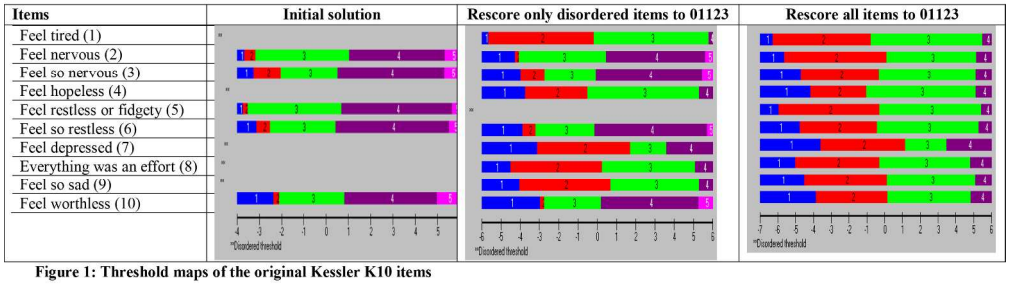



**Figure 1: Threshold maps of the original Kessler K10 items**

247x77mm (300 x 300 DPI)

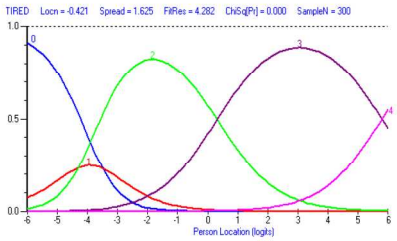
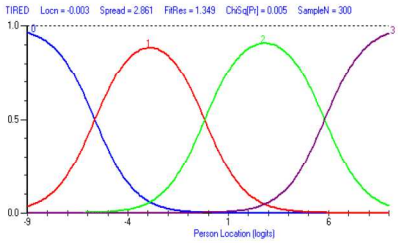
| Category probability curve before | Category probability curve after |
|---|---|
|  |  |



**Figure 1: Category probability curve of item 'feel tired' before and after rescoring**

176x78mm (300 x 300 DPI)

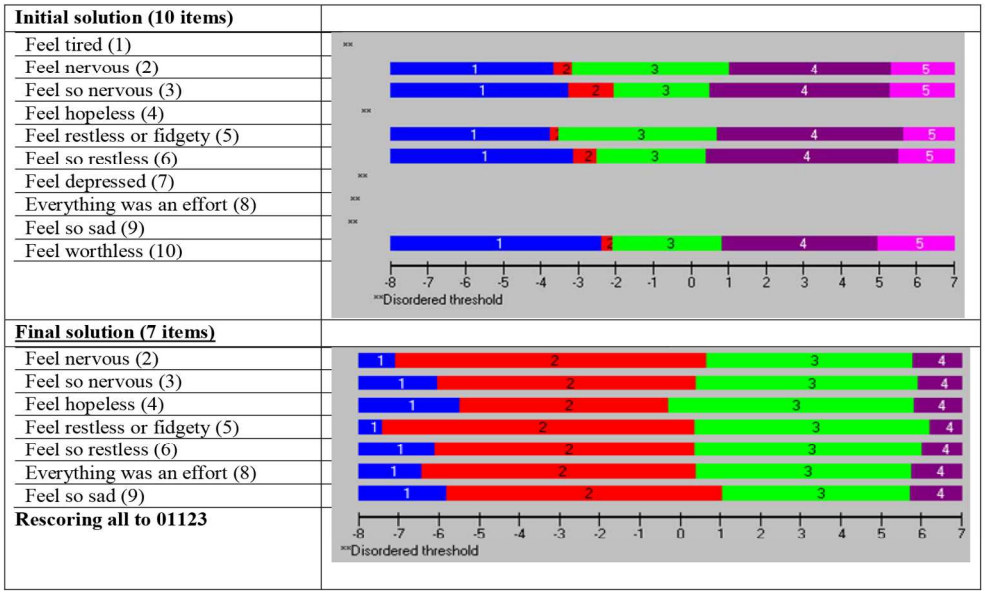



**Figure 3: Threshold maps of the original 10-items (Kessler 10) vs. the final 7-items model**

156x102mm (300 x 300 DPI)

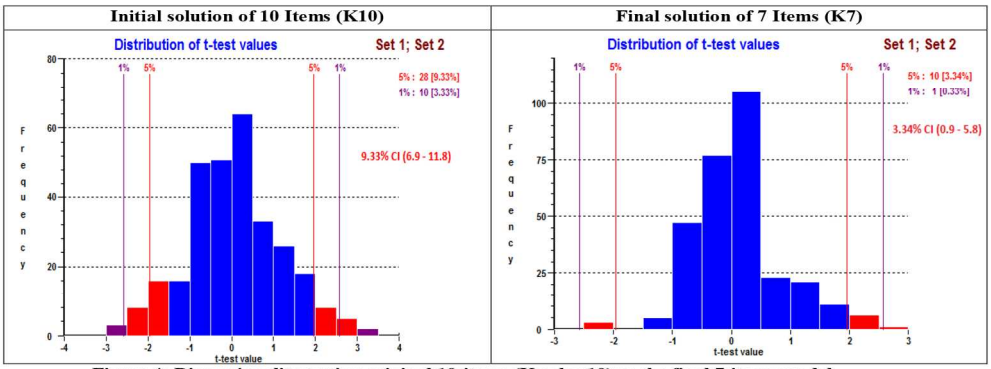



**Figure 4: Dimensionality testing original 10-items (Kessler 10) vs the final 7-items model**

163x67mm (300 x 300 DPI)

## Appendix

### Table 1. Analysis history

| | Analysis | Item fit residuals mean (SD) | Person fit residuals mean (SD) | PSI (CF) | Overall model fit $\chi^2$ (p value) | Status of disorder items | % Significant t-test CI |
|---|---|---|---|---|---|---|---|
| K10 Original scale | 1 | 0.14 (2.50) | –0.19 (1.24) | 0.85 (0.87) | 262.28 (0.0000) | Five items (1,4,7,8,9) | 9.33% CI (6.9–11.8) |
| Rescore only disorder items to 01123 | 2 | –0.71 (2.34) | –0.30 (1.13) | 0.84 (0.87) | 202.53 (0.0000) | One item (5) | 7.3% CI (4.9–9.8) |
| Rescore all to 01123 | 3 | –1.48 (2.38) | –0.55 (1.44) | 0.83 (0.87) | 166.67 (0.0000) | No items | 4.7% CI (2.2–7.1) |
| Positive worded items (1,2,3,5,6) | 4 | –0.15 (1.92) | –0.30 (1.04) | 0.85 (0.83) | 107.05 (0.0000) | One item (1) | 6.3% CI (3.9–8.8) |
| Negative worded items (4,7,8,9,10) | 5 | 0.51 (1.90) | –0.21 (1.03) | 0.59 (0.80) | 98.51 (0.0000) | Three items (4, 7 ,8) | 3.0% CI (0.5–5.5) |
| Remove only tired from the model | 6 | –1.07 (2.39) | –0.57 (1.55) | 0.81 (0.87) | 107.13 (0.0000) | No item | 6.7% CI (4.2–9.1) |
| Remove only depressed from the model | 7 | –1.50 (1.97) | –0.67 (1.49) | 0.85 (0.88) | 96.61 (0.0000) | No item | 10.3% CI (7.9–12.8) |
| Remove only worthless from the model | 8 | –1.58 (2.59) | –0.53 (1.40) | 0.82 (0.86) | 164.14 (0.0000) | No item | 8.3% CI (5.9–10.8) |
| Remove tired and depressed together from the model | 9 | –1.27 (1.66) | –0.73 (1.62) | 0.83 (0.88) | 58.73 (0.0027) | No item | 16.0% CI (13.5–18.5) |
| Remove tired and worthless together from the model | 10 | –1.20 (2.61) | –0.55 (1.48) | 0.80 (0.86) | 126.24 (0.0000) | No item | 12.0% CI (9.5–14.5) |
| Remove depressed and worthless together from the model | 11 | –1.55 (1.90) | –0.64 (1.42) | 0.85 (0.87) | 76.64 (0.0000) | No item | 9.3% CI (6.9–11.8) |
| Remove tired, depressed and worthless from the model | 12 | –1.34 (1.27) | –0.70 (1.53) | 0.84 (0.88) | 40.11 (0.0647) | No item | 6.7% CI (4.2–9.2) |
| Remove Further from tired, depressed and worthless from the model+ one person id 164 | 13 | –0.20 (1.20) | –0.06 (1.41) | 0.84 (0.88) | 39.83 (0.0685) | No item | 3.3% CI (0.9–5.8) |

The history of statistical analysis of the K10 using Rasch analysis has been mentioned in Table 1 (supplementary file). First, we ran the Rasch analysis with original ten items. Out of the ten items, five items had disordered thresholds, and overall chi-square values as well as item fit residuals that were high and significant. We rescored only the disordered items by following the pattern of categorical probability curve, which suggested the combination of the middle two response categories into one, but one item still had a disordered threshold. Next, we rescored all items to 0,1,1,2, 3 from 0, 1, 2, 3, 4. The problem of disordered items was solved, but overall chi-square values and item fit residuals SD were high. Then, we tried to use the PCA technique to check whether the scale was more than one dimension. To achieve this, we used PCA technique to separate positively and negatively worded items. We found that items 1, 2, 3, 5 and 6 were positively worded items, and 4, 7, 8, 9 and 10 were negatively worded items. We applied Rasch analysis technique to positively and negatively worded items and found one disordered item among the positively worded items and three among the negatively worded items, and overall model fits were poor for both models. We confirmed that the K10 was not a two-factor solution. Then we revisited the model where we rescored all items to 01123.

We checked the individual items fit chi-square value, which might influence the overall chi-square value. We found that item 1 (feel tired) had a high chi-square value followed by item 7 (feel depressed) and item 10 (worthless). First, we removed the item ‘feel tired’ from the model and then ‘depressed’ and finally ‘worthless.’ Removing one item at a time in the following sequential order, ‘tired’, ‘depressed’ and ‘worthless’ resulted in chi-square values (SD) of 107.13 (SD = 2.34), 96.61 (SD=1.97) and 164.14 (2.59), respectively, indicating the models were poorly fit. Removing two items at a time in the following sequential order, ‘tired and depressed’, ‘depressed and worthless’, and ‘worthless and tired’ did not improve the model significantly. Going through different iteration process in removing items, removing three items together produced the desired model except the individual's person fit statistics SD (1.53). Further investigation showed that one person was misfit. Removing the misfit person, Rasch analysis produced a perfect fit model with seven items, with four categories for each item (Appendix Table 1). All the assumptions of the Rasch analysis have been met in our model.

**STROBE 2007 (v4) Statement—Checklist of items that should be included in reports of *cohort studies***

| Section/Topic | Item # | Recommendation | Reported on page # |
|---|---|---|---|
| **Title and abstract** | 1 | (*a*) Indicate the study's design with a commonly used term in the title or the abstract | Pages 1 and 2 |
| | | (*b*) Provide in the abstract an informative and balanced summary of what was done and what was found | Page 2 |
| **Introduction** | | | |
| Background/rationale | 2 | Explain the scientific background and rationale for the investigation being reported | Page 3 |
| Objectives | 3 | State specific objectives, including any prespecified hypotheses | Page 5 |
| **Methods** | | | |
| Study design | 4 | Present key elements of study design early in the paper | Page 5 |
| Setting | 5 | Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection | Page 5 (in protocol paper)(ref 28) |
| Participants | 6 | (*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up | Page 5 (in protocol paper)(ref 28) |
| | | (*b*) For matched studies, give matching criteria and number of exposed and unexposed | |
| Variables | 7 | Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable | Page 6 |
| Data sources/ measurement | 8* | For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group | *Page 6* |
| Bias | 9 | Describe any efforts to address potential sources of bias | |
| Study size | 10 | Explain how the study size was arrived at | Page 5 (in protocol paper)(ref 28) |
| Quantitative variables | 11 | Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why | |
| Statistical methods | 12 | (*a*) Describe all statistical methods, including those used to control for confounding | Pages 7-9 |
| | | (*b*) Describe any methods used to examine subgroups and interactions | |
| | | (*c*) Explain how missing data were addressed | |
| | | (*d*) If applicable, explain how loss to follow-up was addressed | In protocol paper (ref 28) |
| | | (*e*) Describe any sensitivity analyses | |

| | | | |
|---|---|---|---|
| **Results** | | | |
| Participants | 13* | (a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed | Page 5 |
| | | (b) Give reasons for non-participation at each stage | |
| | | (c) Consider use of a flow diagram | |
| Descriptive data | 14* | (a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders | Table 1 on page 17 |
| | | (b) Indicate number of participants with missing data for each variable of interest | |
| | | (c) Summarise follow-up time (eg, average and total amount) | |
| Outcome data | 15* | Report numbers of outcome events or summary measures over time | |
| Main results | 16 | (*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included | |
| | | (*b*) Report category boundaries when continuous variables were categorized | |
| | | (*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period | |
| Other analyses | 17 | Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses | Appendix pages 21-22 |
| **Discussion** | | | |
| Key results | 18 | Summarise key results with reference to study objectives | Page 12 |
| **Limitations** | | | page 15 |
| Interpretation | 20 | Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence | Pages 12-15 |
| Generalisability | 21 | Discuss the generalisability (external validity) of the study results | |
| **Other information** | | | |
| Funding | 22 | Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based | Page 16 |

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

**Note:** An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.