International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 23(1): 1–8 (2014)
Published online 16 January 2014 in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1431

A new look at the construct validity of the K6 using Rasch analysis

ASADUZZAMAN KHAN,[1] CHI-WEN CHIEN[1] & NICOLA W. BURTON[2]

1 School of Health and Rehabilitation Sciences, The University of Queensland, Brisbane, Qld, Australia
2 School of Human Movement Studies, The University of Queensland, Brisbane, Qld, Australia

Key words
Kessler, K6, psychological distress, construct validity, Rasch analysis, mental health

Correspondence
Asaduzzaman Khan, School of Health and Rehabilitation Sciences, The University of Queensland, St Lucia, Brisbane, Qld 4072, Australia.
Telephone (+61) 7-3346-7456
Fax (+61) 7-3365-1877
Email: a.khan2@uq.edu.au

Received 13 July 2012;
accepted 29 August 2012

Abstract

The Kessler six-item psychological distress (K6) scale is widely used to screen for mental disorders; however, information is lacking on the rating scale performance or dimensionality structure of the scale. This study used a population based sample ($n = 7596$) to evaluate the construct validity of the K6 scale using Rasch partial credit analysis. The analysis showed that almost all of the five-point rating scales in the K6 items were used appropriately to differentiate psychological distress of the study participants. The analysis provided evidence of unidimensionality of the scale, although items 1 (so sad) and 3 (restless or fidgety) might offer a potential second off-dimensional component. All items appeared to fit the Rasch model's expectation as demonstrated by the acceptable item fit statistics. The study participants demonstrated valid response patterns when answering K6 items, except for some who were younger or had higher psychological distress. This study using Rasch analysis confirms the construct validity of the K6 scale and suggests that the K6 is a useful and valid instrument for assessing psychological distress in the mid-aged general population. Further research can facilitate better understanding about the unidimensionality of the scale. *Copyright © 2014 John Wiley & Sons, Ltd.*

Introduction

Assessment of non-specific psychological distress can be used to screen for potential mental health disorders in both individuals and populations in order to monitor prevalence, guide clinical decision-making and evaluate service provision. The Kessler Psychological Distress scales have become increasingly popular as a short screening tool to measure the level of distress associated with non-specific psychological symptoms in the general population (Kessler *et al.*, 2002; Andersen *et al.*, 2011). Taking items from 18 commonly used screening scales, Kessler *et al.* (2002) used modern item-response theory (IRT) to evaluate the contribution of each selected item to the sensitivity of the total scale in the severity range of the population distribution. Two versions of the Kessler scales, K6 (six-item) and K10 (10-item), were thus derived from identifying items that were best able to discriminate respondents for current psychological distress at the 90th–99th percentile range of the population distribution. Internal consistency was demonstrated with high Cronbach's alpha values of 0.89 and 0.93 for the K6 and K10, respectively. In addition, the K6 and K10 have been validated against commonly recognized diagnostic tools such as the Composite International Diagnostic Interview (CIDI) as well as the Diagnostic and Statistical Manual for Mental Disorders, Fourth Edition (DSM-IV) mood and anxiety disorders in various large community samples (Andrews and Slade, 2001; Kessler *et al.*, 2002, 2003; Rodgers *et al.*,

Copyright © 2014 John Wiley & Sons, Ltd.

2007). Clinical studies have also endorsed the validity of the scales in a variety of patient populations (Baggaley *et al.*, 2007; Hides *et al.*, 2007; Haller *et al.*, 2009). The K6 and K10 scales were shown to outperform the widely used 12-question General Health Questionnaire (GHQ-12) and were comparable in detecting depressive and anxiety disorders in terms of receiver operating characteristic (ROC) curve performances (Furukawa *et al.*, 2003; Kessler *et al.*, 2003).

While the K6 scale is a subset of the K10 scale, this abbreviated version of six items has been found to demonstrate similar sensitivity to the K10 in discriminating between cases and non-cases of serious mental illness (SMI) (Arnaud *et al.*, 2010). The K6 scale also has comparable performance to the K10 for screening DSM-IV mental disorders (Furukawa *et al.*, 2003) and good concordance with independent clinical ratings of SMI (Kessler *et al.*, 2010). Because of the consistency across sub-samples of different characteristics and the brevity of its contents with only six items, the K6 scale has been recommended as a preferred choice for screening SMI in the general population (Furukawa *et al.*, 2003; Kessler *et al.*, 2010).

Although research evidence on the concurrent validity of the K6 scale in screening for psychological distress is promising, information on some specific aspects of construct validity such as the rating scale performance or dimensionality structure of the scale is limited. In the original study, the five-point rating scales of each item were re-coded as dichotomous (e.g. 0 versus 1–4, 0–1 versus 2–4, 0–2 versus 3–4, 0–3 versus 4) and submitted to IRT analyses as individual items (Kessler *et al.*, 2002). It remained unclear whether the five-point rating scale can be used appropriately as a whole within individual items to differentiate people's psychological distress as clearly as the five-point rating scale allows. Furthermore, Kessler *et al.* (2002) used principal axis factor analysis to examine the dimensionality of the K6 scale based on eigenvalues. Their results extracted more than one factor structure with eigenvalues greater than one, which was viewed as the conventionally accepted criterion to determine the number of viable factors (Cliff, 1988). A recent study examined the factor structure of the K6 (Arnaud *et al.*, 2010) using factor analysis, and demonstrated a possible two-dimensional solution for the scale with the first factor explaining 46.5% of the total variance and the second 17.3%. However, another recent methodological study using confirmatory factor analysis found that the K6 was unidimensional across gender or age groups (Drapeau *et al.*, 2010). This inconsistency in the dimensionality and/or factor structure of the K6 scale has been attributed to the use of factor analysis which tends to leave some noise in the scores (residual noise) and as such the main factor does not often explain the variance that it is meant to explain (Waugh, 2002). In addition, the mathematics behind factor analysis does not allow item difficulties and person measures to be estimated separately on the same scale. Even if the items are indicative of measuring the same trait, items which are difficult to endorse may not be highly correlated with the items that are easy to endorse, and this could lead to misleading results in factor analysis (Waugh, 2002).

Rasch analysis, one type of IRT based approaches, has been increasingly applied to validate construct validity of assessment tools. This analysis is sensitive to the full matrix of residuals and thus the residual noise can be avoided in estimating the model parameters (Wright, 1996). In addition to producing objective linear measures, Rasch analysis performs well to detect unidimensionality more accurately than the principal component factor analysis (Smith, 1996; Wright, 1996; Waugh and Chapman, 2005). It has advantages of examining the appropriateness of the rating scale when used as a whole to differentiate between people with different levels of phenomenon of interest. In addition, it can deal with items with varying difficulty levels (e.g. high or low difficulty). Simultaneously, Rasch analysis considers each respondent's abilities (e.g. psychological distress), which allows detection of certain respondents who have inconsistent response patterns in relation to the modeled expectations based on Rasch analysis and serves as one evidence source of person response validity (Fisher *et al.*, 2000). Nevertheless the performance of K6 using Rasch analysis has not yet been examined.

Given the inconsistency of results from studies using factor analysis and the added advantages of Rasch analysis in examining construct validity, the present study was designed to use Rasch analysis to evaluate three aspects of construct validity of the K6. In particular, two validity elements (i.e. unidimensionality and rating scale performance) were focused on to establish cumulative evidence for K6. An examination of person response validity was also included because the K6 is reported as suitable for use in the general population but there is little empirical evidence about whether the respondents could exhibit valid or expected response patterns when answering the K6 items.

Methods

Participants

Data for the present study were from a mail survey conducted in Brisbane, Australia. This HABITAT study was designed to investigate physical activity and health

Int. J. Methods Psychiatr. Res. 23(1): 1–8 (2014). DOI: 10.1002/mpr
Copyright © 2014 John Wiley & Sons, Ltd.

over a period of five years in a randomly selected sample of adults aged 40–65 years in 2007. Details on the study design, sampling and data collection for HABITAT have been published elsewhere (Burton *et al.*, 2009). The second wave of data collected in 2009 was the first to include the K6 and hence used in the present study. Of the 10,844 surveys sent in 2009, 322 participants declined participation, 161 were deceased or unable to respond, and 2765 did not respond to the variable of interest (K6). The remaining 7596 participants form the basis for the present study. The socio-demographic and health profile of respondents are provided in Table 1.

Measures

The K6 scale consisted of six items that ask participants to rate how often during the past four weeks they felt: (1) so sad that nothing could cheer them up; (2) nervous; (3) restless or fidgety; (4) hopeless; (5) worthless; and (6) that everything was an effort. A five-point rating scale was used as the response options, indicating 0 (none of the time), 1 (a little of the time), 2 (some of the time), 3 (most of the time), and 4 (all of the time).

Data analysis

Construct validity of the K6 was assessed using Rasch analysis that included four components. Each component was used to examine different aspects of construct validity for the K6, including (a) the appropriateness of the rating scale such as whether the five-point frequency rating scales can be used appropriately by participants to differentiate their psychological distress severity, (b) the unidimensionality or the extent to which the K6 items measure a single construct defining psychological distress, (c) the person response validity or the extent to which the responses of participants demonstrate logical hierarchical ordering, and (d) test targeting or the extent to which the K6 items are of appropriate severity for the sample. The Rasch partial credit model was implemented using the WINSTEPS software version 3.73 (Linacre, 2011). Details of each part of Rasch analysis were described as follows.

Table 1. Socio-demographic and health characteristics of the 2009 HABITAT respondent sample ($n = 7596$)

Characteristics	*n*	%
Sex		
Male	3370	44.37
Female	4226	55.63
Age group (years)		
42–49	2362	31.10
50–59	3099	40.80
60–67	2135	28.10
Highest educational qualification		
School only (up to 12 years)	3342	44.00
Certificate/diploma	2310	30.41
University degree	1944	25.59
Employment status		
Full-time	3691	51.11
Part-time	1572	21.77
No paid work/retired	1959	27.13
Self-rated health status		
Excellent/very good	3104	41.85
Good	3001	40.46
Fair/poor	1312	17.69

Rating scale analysis

The appropriateness of the five-point rating scales of the K6 items was analyzed with regards to basic rating scale assumptions (Linacre, 2002). Specific K6 items were identified as having problematic rating categories if they did not meet the criteria: (1) at least 10 cases per category, (2) monotonically increasing average measures across categories, (3) category outfit mean square (MnSq) values less than 2.0, and (4) monotonically increasing step calibrations indicating that the difficulty level of a lower step (e.g. 0–1 between none and a little of the time) was lower than that of its adjacent higher step (e.g. 1–2 between a little and some of the time). The rating categories were reorganized (i.e. some were collapsed) if most of those criteria were not satisfied.

Unidimensionality

To examine the unidimensional construct of the K6 items, a Rasch principal component analysis (PCA) of residuals (Linacre, 1998; Smith, 2002) was conducted. In the PCA of residuals, it is expected that the Rasch identified construct should account for $> 60\%$ of the total variance and, after removal of the principal component, the residuals for the item/person interactions are likely to be randomly distributed and show no structure (Linacre, 2011). Also, the eigenvalue size of < 2 that is explained by the first contrast (which is the largest secondary dimension after the Rasch-derived construct is removed) has been proposed as a cutoff value to indicate that there is unidimensionality in the test items Raîche, 2005). However, if the first contrast has an eigenvalue value > 2, it indicates that another dimension may be present, provided that the dimension has at least three items (Linacre, 2011).

Rasch analysis also provides goodness-of-fit statistics to examine how well test items fit with the model's

Copyright © 2014 John Wiley & Sons, Ltd.

expectations through using infit and outfit statistics. The (weighted) infit statistic is most sensitive to ratings on the K6 items located close to the participants' psychological distress statuses, while the (unweighted) outfit statistic is more influenced by the ratings on the off-target items (i.e. those much lower or higher than the participants' psychological distress statuses). The fit statistics are usually reported as the MnSq; infit and outfit MnSq values of 1.4 have been suggested as acceptable criteria for clinical assessments (Bond and Fox, 2007; Krumlinde-Sundholm *et al.*, 2007; Chien and Bond, 2009) and were adopted for this study. When more than 95% of the items demonstrated acceptable fit, the unidimensionality of the K6 was further supported.

Person response validity

The goodness-of-fit statistics can also be used to examine the response validity of participants. Infit and outfit MnSq values of < 1.4 were considered as acceptable fit of the participants' response patterns with the Rasch model's expectations. An overall rate of $> 95\%$ of participants exhibiting acceptable fit (Fisher *et al.*, 2000) indicated person response validity of the K6 items that can be used appropriately in the community-dwelling population with a range of characteristics.

Test targeting

The unidimensional acceptable-fit K6 items were eventually expressed as logits in the Rasch analysis output and were calibrated along a hierarchical order from mild to severe psychological distress to be endorsed. The well-fitting participants' level of stress were also calibrated from less-distress to more-distress and were placed together with the items' severity calibrations on the same linear interval-level measurement continuum (also referred to as an item-person map). The Rasch-generated item-person map provides a visual inspection method to examine which range of psychological distress in the participants the K6 items targeted (or encompassed).

Results

Rating scale analysis

Rasch analysis results showed that almost all of the five-point rating scales in the K6 items were used appropriately to differentiate psychological distress of the participants (Table 2). Two items exhibited poor fit with Rasch model's expectations (i.e. outfit MnSq > 2.0) only in the highest category representing "*All of the time*"; thus, no rating scale reorganization was made.

Unidimensionality

The initial PCA of the K6 revealed that 60.4% of total variance was explained by the Rasch-derived measures, and a small eigenvalue of 1.6 was accounted for by the second major component. The item fit analysis identified no items as misfitting (Table 3). These results provided evidence of unidimensionality in the K6 items. However it was noted that, in the principal component plot of item loading for the first contrast, the Items 1 (so sad) and 3 (restless or fidgety) had substantive, positive loading values of > 0.4 in contrast to other items substantial loading on the negative values. This indicates that these two items may produce a potential but not sizable off-dimensional component, since in unexplained variance a secondary dimension must have the strength of at least three items with eigenvalues > 2.0 (Raîche, 2005; Linacre, 2011).

Person response validity

The Rasch goodness-of-fit person-response analyses revealed that 814 (10.7%) of the participants misfit the K6 items. Given that this rate was higher than the accepted criterion of 5%, a detailed investigation was conducted to examine the sources of the disturbance in the K6 measurement. By comparing the K6 data between the 814 misfitting participants and the rest of the sample ($n = 6782$), the participants who had misfitting responses exhibited significantly higher mean psychological distress (mean $= -1.23$ logits compared to -5.06 logits). Furthermore, significantly more misfitting participants were included in the youngest group aged 42–49 years ($P < 0.05$ using Chi-square), but they did not differ significantly from those who demonstrated acceptable-fit responses in relation to gender and educational levels ($P < 0.05$ using Chi-square).

The examination of the source of disturbance also focused on the unexpected misfitting responses on individual K6 items. We found that a total of 301 (0.9%) out of 34,026 valid responses were identified as extremely unexpected (e.g. standardized z values beyond ± 3.0). Particularly the 112 (37.2%) misfitting ratings were related to the Item 4 (hopelessness), which was scored higher than Rasch model's expectations. Large proportions of the Item 4 misfitting ratings were produced by the participants who received secondary school education (44.6%) and/or were female (63.4%).

Considering these secondary analyses, younger participants (42–49 years) or those with higher psychological distress, and the most unexpected variations in the

Int. J. Methods Psychiatr. Res. 23(1): 1–8 (2014). DOI: 10.1002/mpr
Copyright © 2014 John Wiley & Sons, Ltd.

Table 2. Rating scale analysis for the K6 items

Item/category label	Observed count	Average measure	Outfit MnSq	Step calibration
1. So sad nothing can cheer				
None of the time (0)	3585	−4.13	1.43	—
A little of the time (1)	2558	−3.43	1.15	−3.91
Some of the time (2)	1235	−1.77	1.02	−1.38
Most of the time (3)	186	0.31	1.04	1.62
All of the time (4)	32	1.63	1.52	3.67
2. Nervous				
None of the time (0)	5606	−3.98	0.84	—
A little of the time (1)	1203	−2.32	0.56	−2.46
Some of the time (2)	622	−0.75	0.58	−1.40
Most of the time (3)	134	1.03	0.66	1.04
All of the time (4)	31	2.41	0.98	2.81
3. Restless				
None of the time (0)	3464	−4.22	1.34	—
A little of the time (1)	2534	−3.48	1.09	−3.90
Some of the time (2)	1359	−1.79	0.99	−1.45
Most of the time (3)	204	0.02	1.18	1.65
All of the time (4)[1]	35	0.93	2.33	3.70
4. Hopeless				
None of the time (0)	5845	−3.87	0.87	-
A little of the time (1)	1112	−2.20	0.82	−2.52
Some of the time (2)	514	−1.50	0.67	−1.40
Most of the time (3)	102	1.16	0.80	1.05
All of the time (4)	23	2.53	1.02	2.86
5. Everything was an effort				
None of the time (0)	3927	−4.35	0.95	—
A little of the time (1)	2406	−3.15	0.91	−3.29
Some of the time (2)	939	−1.52	0.93	−0.91
Most of the time (3)	256	−0.06	1.24	1.16
All of the time (4)[1]	68	0.58	2.84	2.93
6. Worthless				
None of the time (0)	6218	−3.77	0.90	—
A little of the time (1)	840	−1.87	0.51	−1.88
Some of the time (2)	387	−0.40	0.60	−1.10
Most of the time (3)	113	1.10	0.81	0.78
All of the time (4)	38	2.19	1.24	2.19

[1]Indicates that the category demonstrated misfit such as outfit MnSq values of >2.0.

Table 3. Measures and fit statistics for the K6 items

Items	Measure	SE	Infit MnSq	Outfit MnSq
1. So sad nothing can cheer	−0.42	0.02	1.22	1.21
2. Nervous	0.37	0.03	0.74	0.63
3. Restless	−0.55	0.02	1.02	1.01
4. Hopeless	0.70	0.03	0.81	0.80
5. Everything was an effort	−0.68	0.02	1.02	1.01
6. Worthless	0.59	0.03	0.79	0.63

Copyright © 2014 John Wiley & Sons, Ltd.

Item 4 (hopelessness) were likely to be the sources of the disturbance to the measurement for K6. Therefore, a decision was made to eliminate those 814 participants who demonstrated misfit due to the potential threat for subsequent Rasch analyses. Based on the reduced group of 6782 participants with acceptable fit, the unidimensionality of the K6 items were re-examined. Higher total variance of 69.4% (originally 60.4%) explained by the Rasch-derived measures was found with a smaller eigenvalue of 1.4 loaded on in the first contrast. No misfitting items were identified. The reduced sample was thus used in the subsequent analyses.

Test targeting

Figure 1 shows the distribution of the item and person logit measures plotted along the same axis in Rasch item-person maps for the K6 items. The mean logit measures of the participants' psychological distress status in the K6 scale were found to have a large deviation (i.e. −5.06 logits) from the mean of the K6 item severity measures. Furthermore, the K6 revealed a floor effect with 28.3% of the participants with minimum scores, while no ceiling effect (0.1%) was found. Thus, both results represented an inadequate item to person targeting, particularly for those participants who had lower or no distress and were not addressed by any K6 items located at the bottom of Figure 1. Nevertheless, this would be conceptually correspondent with the original intention of the K6 scale for screening use in the 90th − 99th percentile range of the general population distribution.

In addition, Rasch analysis reported a person reliability coefficient of 0.80 for the K6 with a separation index of 2.02. This person reliability coefficient can be interpreted as the replicability of the person ordering, and the separation index can be used to identify the number of statistically distinct ability strata of the individuals in the sample (Bond and Fox, 2007). The study results indicated that the participants' scores can be reliably estimated at the acceptable level of > 0.70 (Bond and Fox, 2007) and be differentiated into at least three statistically distinct strata. Similarly, Rasch analysis also generated the item reliability coefficient for the K6 that was found to be 1.00 in association with the separation index of 23.9. These results indicated that the K6 items defining the construct of non-specific psychological distress yielded precise estimates and were well separated into at least 32 strata of severity for the participants in the study sample.

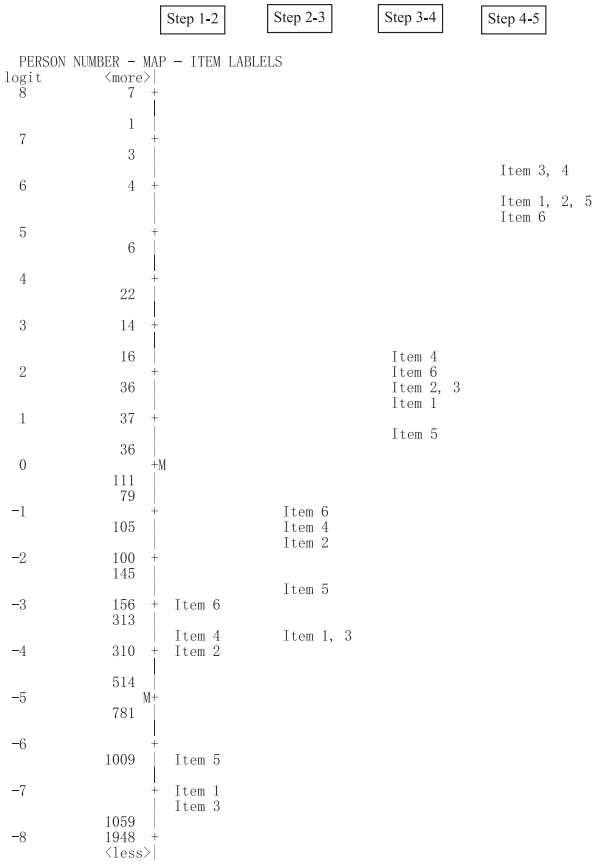

Figure 1. The distribution of the item and person logit measures. Note: The item–person map in Figure 1 presents person ability measures in relation to item difficulty calibrations. The number of participants with the same ability measures are shown in the right side, whereas the item difficulty measures are presented according to step calibration (1–2, 2–3, 3–4, and 4–5) for the five-point rating scale. Higher logit measures indicate higher item difficulty (i.e. the items with the psychological problems that people suffered rarely). It also indicates higher person ability (i.e. experience higher psychological distress).

Discussion

Applying Rasch analysis to the K6 scale in the present study confirmed several aspects of construct validity of this scale when used in a large community sample of mid-aged adults. In examining the five-point rating scale performance of the scale in earlier non-Rasch analysis, the response categories were converted into four ordered dichotomies as individual items and were found to exhibit progressing severities (Kessler *et al.*, 2002). However, Kessler *et al.* (2002) have acknowledged that the generated

Int. J. Methods Psychiatr. Res. 23(1): 1–8 (2014). DOI: 10.1002/mpr
Copyright © 2014 John Wiley & Sons, Ltd.

ordering of severity estimates may be misleading due to possibly biased standard errors resulted from the dependence among those dichotomous related items. Therefore, we alternatively treated the K6 items with the five-point rating scale as polychotomous items to examine the rating scale appropriateness within each item. Rasch analysis is a specialized IRT model in estimating one parameter (e.g. severity) and can estimate severity values for each of four step thresholds (i.e. 0–1, 1–2, 2–3, and 3–4) between the five rating categories in K6 items. The finding of this Rasch analysis provides clearer evidence supporting that all of the five-point rating scales in the K6 items were used appropriately as a whole by demonstrating increased severities to differentiate different levels of non-specific psychological distress of the study participants. While the two items exhibited misfit in the highest category of "all of the time", the category misfit did not substantially affect goodness-of-fit of these two items and no collapsing reorganization was made.

All of the K6 items appeared to constitute a unidimensional scale as confirmed by the Rasch PCA of residuals and could be placed hierarchically as demonstrated by acceptable item fit statistics. However, Item 1 (so sad) and Item 3 (restless or fidgety) may produce a potential second off-dimensional component, while an earlier study on a small clinical sample demonstrated the possibility of a second dimension with Items 2 (nervous) and 3 (restless or fidgety) (Arnaud *et al.*, 2010). Considering that there are only two items with no obvious clinical explanations and decisively statistical patterns for an additional construct, we tentatively concluded that it is unlikely to pose a substantial threat to the unidimensionality of the K6 scale. More research is needed to confirm the unidimensionality of the scale.

This study provides preliminary evidence of person response validity for the K6 scale and confirms its clinical use in a large community sample. While a number of participants were found to detract from the construct validity of the K6 scale, it appeared to be clinically expected. For example, some participants could report atypical patterns of psychological distress depending on their personal experience. In addition, the Item 4 (hopelessness) was likely to cause the disturbance to the measurement for K6 in comparison with other items. However, this item demonstrated acceptable goodness-of-fit values (Table 3) and, according to Kessler *et al.* (2002), its severity calibration remained invariance across gender and education. It is thus suggested as a minor issue that women or those with secondary school education may show a different psychological distress pattern specifically on this item, but overall the construct validity of the K6 is not substantially affected.

In addition, the K6 scores did not appear to suffer from ceiling effects in this large community sample of the study according to the Rasch generated item–person map (Figure 1). This finding also confirmed the K6 developers' postulation that those items could cover with 90th–99th percentile range of the population. Approximately 10% of the participants in this study sample were located above −2 logit (inclusive) in the map, implying that each K6 item has at least three rating categories (two, three, and four) with sufficient threshold severities in the target range of the 90th–99th percentile.

There are a few potential limitations to this study. The study sample was comprised only of mid-aged adults aged 42–67 years. Future studies are warranted to examine whether the findings of the present study remain stable in younger adults or elderly populations. Non-responses of the HABITAT study were higher for people from non-English speaking backgrounds, those with socio-economic disadvantage or in poor health. Each of these groups may be more vulnerable to psychological distress (Gill *et al.*, 2009), which poses a threat to the external validity of the study findings. In addition, the present study has found a considerable amount of floor effect (28%) for K6 items, which nonetheless is not unexpected for a short screening tool like the K6.

In conclusion, application of Rasch analysis to the K6 scale provides additional evidence confirming the construct validity of the K6. Appropriate rating scale performance of the K6 along with lack of ceiling effects suggests that K6 is suitable for measuring non-specific psychological distress of mid-aged adults at a population level and can be considered as a valid measure of mental health in epidemiological research.

Acknowledgements

The HABITAT study was supported by project grants from the (Australian) National Health and Medical Research Council (NHMRC) (ID339718, 497236). The authors gratefully acknowledge Mr Martin O'Flaherty for his assistance with data coding and cleaning, and Ms Robyn Baguley and Ms Sophie Miller for project management. NB is supported by an NHMRC Program grant (ID 569940: Owen, Bauman, Brown).

Declaration of interest statement

The authors have no completing interest.

Copyright © 2014 John Wiley & Sons, Ltd.

References

Andersen L., Grimsrud A., Myer L., Williams D., Stein D., Seedat S. (2011) The psychometric properties of the K10 and K6 scales in screening for mood and anxiety disorders in the South African stress and health study. *International Journal of Methods in Psychiatric Research*, **20**(4), 215–223.

Andrews G., Slade T. (2001) Interpreting scores on the Kessler Psychological Distress Scale (K10). *Australian and New Zealand Journal of Psychiatry*, **25**(6), 494–497.

Arnaud B., Malet F., Teissedre M., Izaute F., Moustafa J., Geneste J., Schmidt P., Llorca M., Brousse G. (2010) Validity study of Kessler's psychological distress scales conducted among patients admitted to French emergency department for alcohol consumption-related disorders. *Alcoholism, Clinical and Experimental Research*, **34**(7), 1235–1245.

Baggaley R., Ganaba R., Filippi V., Kere M., Marshall T., Sombie I., Storeng K., Patel V. (2007) Detecting depression after pregnancy: the validity of the K10 and K6 in Burkina Faso. *Tropical Medicine and International Health*, **12** (10), 1225–1229.

Bond T.G., Fox C.M. (2007) Applying the Rasch Model: Fundamental Measurement in the Human Sciences, 2nd edition, Mahwah, NJ, Lawrence Erlbaum Associates.

Burton N., Haynes M., Wilson L., Giles-Corti B., Oldenburg B., Brown W., Giskes K., Turrell G. (2009) HABITAT: a longitudinal multilevel study of physical activity change in mid-aged adults. *BMC Pubic Health*, **9**, 76. DOI: 10.1186/1471-2458-9-76

Chien C.W., Bond T.G. (2009) Measurement properties of fine motor scale of Peabody developmental motor scales-second edition: a Rasch analysis. *American Journal of Physical Medicine & Rehabilitation*, **88**(5), 376–386.

Cliff N. (1988) The eigenvalues-greater-than-one rule. *Psychological Bulletin*, **103**(2), 276–279.

Drapeau A., Beaulieu-Prevost D., Marchand A., Boyer R., Préville M., Kairouz S. (2010) A life-course and time perspective on the construct validity of psychological distress in women and men. Measurement invariance of the K6 across gender. *BMC Medical Research Methodology*, **10**, 68. DOI: 10.1186/1471-2288-10-68

Fisher A.G., Bryze K., Atchison B.T. (2000) Naturalistic assessment of functional performance in school settings: reliability and validity of the School AMPS scales. *Journal of Outcome Measurement*, **4**(1), 491–512.

Furukawa T., Kessler R., Slade T., Andrews G. (2003) The performance of the K6 and K10 screening scales for psychological distress in the Australian National Survey of Mental Health and Well-Being. *Psychological Medicine*, **33**(2), 357–362.

Gill T., Taylor A., Winefield H., Chittleborough C., Koster C. (2009) Determinants of Psychological Distress (PD) in South Australia – Final Report, Adelaide, SA, Discipline of Psychiatry, University of Adelaide.

Haller D., Sanci L., Sawyer S., Patton G. (2009) The identification of young people's emotional distress: a study in primary care. *British Journal of General Practice*, **59**(560), e61–e70.

Hides L., Lubman D., Devlin H., Cotton S., Aitken C., Gibbie T., Hellard M. (2007) Reliability and validity of the Kessler 10 and Patient Health Questionnaire among injecting drug users. *Australian and New Zealand Journal of Psychiatry*, **41**(2), 166–168.

Kessler R., Andrews G., Colpe L., Hiripi E., Mroczek D., Norand S., Walters E., Zaslavsky A. (2002) Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*, **32**(6), 959–976.

Kessler R., Barker P., Colpe L., Epstein J., Gfroerer J., Hiripi E., Howes M., Normand S., Manderscheid R., Walters E., Zaslavsky A. (2003) Screening for serious mental illness in the general population. *Archives of General Psychiatry*, **60**(2), 184–189.

Kessler R., Green J., Gruber M., Sampson N., Bromet E., Cuitan M., Furukawa T., Gureje O., Hinkov H., Hu C., Lara C., Lee S., Mneimneh Z., Myer L., Oakley-Browne M., Posada-Villa J., Sagar R., Viana M., Zaslavsky A.M. (2010) Screening for serious mental illness in the general population with the K6 screening scale: results from the WHO World Mental Health (WMH) Survey Initiative. *International Journal of Methods in Psychiatric Research*, **19**(Suppl 1), 4–22.

Krumlinde-Sundholm L., Holmefur M., Kottorp A., Eliasson A.C. (2007) The assisting hand assessment: current evidence of validity, reliability, and responsiveness to change. *Developmental Medicine & Child Neurology*, **49**(4), 259–264.

Linacre J.M. (1998) Detecting multidimensionality: which residual data-type works best? *Journal of Outcome Measurement*, **2**(3), 266–283.

Linacre J.M. (2002) Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, **3**(1), 85–106.

Linacre J.M. (2011) User's Guide to WINSTEPS, Chicago, IL, MESA Press.

Raîche G. (2005) Critical eigenvalue sizes in standardized residual principal components analysis (PCA). *Rasch Measurement Transactions*, **19**(1), 1012.

Rodgers B., Parslow R., Degenhardt L. (2007) Affective disorders, anxiety disorders and psychological distress in non-drinkers. *Journal of Affective Disorders*, **99**(1-3), 165–172.

Smith E.V. (2002) Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, **3**(2), 205–231.

Smith R. (1996) A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling*, **3**(1), 25–40.

Waugh R., Chapman E. (2005) An analysis of dimensionality using factor analysis (true-score theory) and Rasch measurement: what is the difference? Which method is better? *Journal of Applied Measurement*, **6**(1), 80–99.

Waugh R. (2002) Creating a scale to measure motivation to achieve academically: linking attitudes and behaviours using Rasch measurement. *British Journal of Educational Psychology*, **72**(1), 65–86.

Wright B. (1996) Comparing Rasch measurement and factor analysis. *Structural Equation Modeling*, **3**(1), 3–24.

Copyright © 2014 John Wiley & Sons, Ltd.