Psychological distress across the lifespan: examining age-related item bias in the Kessler-6 Psychological Distress Scale

Author:
Sunderland, Matthew; Hobbs, Megan J.; Anderson, Tracy M.; Andrews, Gavin

Publication details:
International Psychogeriatrics
v. 24
Chapter No. 2
pp. 231-242
1041-6102 (ISSN)

Publication Date:
2012

Publisher DOI:
http://dx.doi.org/10.1017/S1041610211001852

Word Count: 5,158 (excluding abstract, references, tables and figures)

Psychological distress across the lifespan: examining age related item bias in the Kessler 6 Psychological Distress Scale

Matthew Sunderland[1]

Megan J Hobbs[1]

Tracy M Anderson[1]

Gavin Andrews[1]

1. Clinical Research Unit for Anxiety and Depression, School of Psychiatry, UNSW at St Vincent's Hospital, Sydney, Australia.

Corresponding Author:

Dr Matthew Sunderland
CRUfAD, Level 4, O'Brien Centre, St Vincent's Hospital, 394-404 Victoria Street,
Darlinghurst, NSW 2010, Australia.

Email: matthews@unsw.edu.au. Ph: +612 8382 1437.

Running Title: Psychological Distress across the lifespan

ABSTRACT

Background: Old age respondents may systemically differ in their responses to measures of psychological distress over and above their actual latent distress levels when compared to younger respondents. The current study aimed to investigate the potential for age-related bias(es) in the Kessler 6 psychological distress scale (K6) items.

Methods: Data from the 2007 Australian National Survey of Mental Health and Wellbeing were analysed using Item Response Theory to detect the presence of item bias in each of the K6 items. The potential for item bias was assessed by systematically comparing young (16-34), middle age (35-64), and old age (65-85) respondents. The significance and magnitude of the item bias between the age groups was assessed using the log-likelihood ratio method of differential item functioning.

Results: After statistical adjustment, there were no biases of significant magnitude influencing the endorsement of K6 items between young and middle age respondents or between middle age and old age respondents. There was a bias of significant magnitude present in the endorsement of the K6 item addressing levels of fatigue between young and old age respondents.

Conclusions: Despite the identification of significant item bias in the endorsement of K6 items between the age groups, the magnitude and influence of the bias on total K6 scores is likely to have little influence on the overall interpretation of group data when comparing psychological distress across the lifespan. Researchers should take some caution however when examining individual levels of fatigue related to psychological distress in older individuals.

Key Words: Kessler 6, item bias, differential item functioning, age comparisons, validity, item response theory.

INTRODUCTION

The Kessler Psychological Distress Scale measures psychological distress experienced in the previous 30 days. The scale has ten item and six item versions, commonly referred to as the K10 and the K6 respectively (Kessler *et al.*, 2002). Since the initial development of the K6/K10 both scales have been widely implemented in epidemiologic surveys and clinical settings. Indeed, epidemiological studies have demonstrated that the K10 and K6 are strongly related to DSM-IV mood and anxiety disorders (Andrews and Slade, 2001; Furukawa *et al.*, 2003; Kessler *et al.*, 2010a; Slade *et al.*, 2011; Sunderland *et al.*, in press). The Kessler scale's screening ability, coupled with its brevity, make it an attractive tool for facilitating comparisons between the psychopathology that is associated with various sociodemographic groups. Such comparisons have substantial heuristic value, in that these analyses could reveal group- and/or setting-specific determinants of mental illness that may warrant further research or treatment.

When examining the distribution of psychopathology as a function of age, epidemiological data have consistently indicated that aging is associated with decreases in prevalence of DSM-IV mood and anxiety disorders (Kessler *et al.*, 2010b; Troller *et al.*, 2007), and this pattern of results has been replicated using the K6/K10 (Slade *et al.*, 2011). Sceptics of the finding that the prevalence of syndromal psychopathology decreases across the lifespan have argued that this pattern goes against what practitioners frequently observe in old age settings (Snowdon, 2001). In contrast, proponents of this finding have argued that a decrease in emotional responsiveness, an increase in emotional control, and an increase in coping skills associated with increasing age contributes to the finding of lower psychopathology (Ernst and Angst, 1995; Henderson *et al.*, 1998; Jorm, 2000).

An empirically robust way of informing the broad debate regarding whether the observed decline in prevalence of psychopathology in old age is a legitimate trend or whether the downward trend in prevalence is artefactually influenced by some form of bias involves the formal identification of measurement invariance. Essentially, measurement invariance confirms that the likelihood of endorsing a questionnaire item is related solely to the respondent's clinical presentation rather than some other aspect of the individual such as their age or sex. During the development of the K6/K10 measures, detailed item analyses were performed to select questionnaire items that were invariant across age and sex (Kessler *et al.*, 2002). Since this initial item selection phase of developing the K6/K10, few studies have specifically investigated the measurement invariance of the K6/K10 across the lifespan. The most robust findings occur once a measure has been in circulation for some time and a large body of data exists across a variety of subpopulations.

Despite scant data with respect to the measurement invariance of the K6/K10, there has been recent speculation that some ways of assessing internalizing disorder in old age respondents are biased. For example, Grayson *et al.* (2000) and Tsang *et al.* (2008) have suggested items related to fatigue are problematic when assessing older individuals because of the association between fatigue and chronic physical conditions that become more prevalent with increase age. Likewise, measures of worthlessness may be associated with a decline or change in social roles and/or social relationships as well as a decline in overall physical and cognitive functioning, both of which have demonstrated a strong relationship in old age (Hong *et al.*, 2009; Vink *et al.*, 2008). Although, these studies did not focus on the potential biases of the K6/K10 specifically, the K6/K10 measures contain similar items to those in question. Consequently, assumptions about the applicability of the K6/K10 in old age or psychogeriatric settings and the measures' suitability for age-banded comparisons in epidemiological studies may be premature.

One approach that can be utilised for examining the presence of bias in questionnaire items is known as the identification of Differential Item Functioning (DIF) using procedures grounded in the field of Item Response Theory (IRT). The field of IRT is too large to adequately describe here, however a brief description is provided and the interested reader is encouraged to gain a greater understanding from the references provided. Briefly, IRT includes a series of models that can be fitted to individual data as a means of describing the probability of responding to a particular questionnaire or scale item in relation to a normally distributed latent dimension under examination (Baker, 2001; Embretson and Reise, 2000; Van den Linden and Hambleton, 1997). For example, IRT assumes that certain questionnaires or scales, such as the K6/K10, are designed to indirectly measure a latent dimension of interest, in this case psychological distress. Each item of the K6/K10 may index a different area along the normal distribution of the latent dimension and therefore modelling this relationship using IRT enables a further understanding of an individual's latent level of psychological distress based on the specific items that they endorse and vice versa. One model, known as the graded response model (Cohen *et al.,* 1993; Samejima, 1969), is particularly suitable for the investigation of ordinal items, such as those in the K6/K10 scales, since it models the relationship between different response options from each item with the normally distributed latent dimension.

The finding of DIF using IRT models implies that the probability of responding to a certain item cannot be solely attributed to the latent dimension but instead may be influenced by an additional feature related to group membership, i.e. gender, age, ethnicity, etc. This is achieved by first defining two groups based on a characteristic of interest and then equating the scale of the latent dimension using common invariant items (known as anchor items). Once all individuals from each group are placed on a common scale, the IRT models are fit separately to each group and the item characteristics of each questionnaire item are compared (Teresi and Fleishman, 2007). If there are no differences between the item characteristics of the two groups then it can be concluded that the relationship between the item and the latent dimension is the same for all individuals regardless of group membership. On the other hand, if there are significant differences between the item characteristics then it can be concluded that an additional feature related to group membership is influencing the endorsement of the item over and above the latent dimension. Therefore, the identification of DIF implies that a certain group of individuals may be biased towards responding to the items in a certain manner when compared to individuals belonging to the other group. Consequently, the aim of this study is to assess potential age-related bias(es) in the assessment of psychological distress, as measured by the Kessler scales, by investigating the possible presence of DIF in each item between individuals belonging to young, middle age, and old age groups of the Australian general population.

METHODS

Sample

The data for the current study were from the 2007 Australian NSMHWB, a cross-sectional household survey of the Australian general population (excluding very remote areas). Private households were randomly selected in each state and territory using a stratified, multistage area design. In total, there were 8,841 households that participated in the survey (a response rate of 60%). Over-sampling (a greater probability of being selected for the interview) of young (16-24 years) and old (65-85 years) age groups ensured that a sufficient sample size was selected to provide reliable estimates for these age groups.

Measures

The Composite International Diagnostic Interview (CIDI) 3.0 was used as the base instrument to derive psychiatric diagnoses in the 2007 NSMHWB. The CIDI 3.0 provided prevalence estimates across the person’s lifetime based on the DSM-IV diagnostic criteria. Prevalence in the past 30 days was estimated by determining if symptoms relating to a lifetime diagnosis were experienced in the 30 days prior to the interview. The diagnoses assessed by the CIDI 3.0 in the Australian survey included: depression, dysthymia, bipolar disorder (manic episode), agoraphobia, social phobia, panic disorder, generalised anxiety disorder (GAD), obsessive-compulsive disorder (OCD), post-traumatic stress disorder (PTSD), substance use, and substance dependence. Furthermore, the survey collected information on socio-demographics and physical functioning/activity limitation as assessed by the WHO Disability Assessment Scale 2.0 (WHODAS 2.0).

Non-specific psychological distress was measured by the Kessler Psychological Distress Scale (Kessler *et al.*, 2002). The current study examined the K6 rather than the K10, as the latter version did not meet the local independence assumption required by IRT. Local independence means that each item is independently related to the latent factor and the response to one item is not directly influenced by a response from another item. The K10 contains skip instructions meaning that some items are skipped depending on the responses to preceding questions, therefore creating a violation of the local independence assumption. That K6 was used as an alternative since the K6 contains items that exhibit local independence. The K6 has strong psychometric properties and is a valid substitute for the K10 when predicting DSM-IV diagnostic cases and serious mental illness (Furukawa *et al.*, 2003; Kessler *et al.*, 2010a; Sunderland *et al.,* in press). Due to its desirable psychometric properties and brevity, the K6 has been utilised in numerous population-based epidemiological studies around the world as part of the World Mental Health Survey initiative to measure and monitor trends in psychological distress (Kessler and Ustun, 2008). Likewise, the K6 has been routinely used in numerous Australian states and territories to measure distress and monitor treatment outcomes. The K6 includes items that measure the presence of nervousness, hopelessness, irritability, negative affect, fatigue, and worthlessness experienced over the past 30 days. Items are rated on a five point scale, with 0 indicating an absence of the symptom and 4 indicating the symptom was present all of the time in the past

30 days. As a result, the final score on the K6 can range from 0 to 24, with higher scores indicating higher levels of psychological distress.

Data analysis

Sample characteristics

The invariance of the K6 was examined across three age groups that defined young adulthood (16-34 years), middle age (35-64 years) and old age (65-85 years). The decision regarding the cut-points for the age groups is rather arbitrary but was informed by epidemiological studies that traditionally define old age as 65+. The young group (<65) was further defined to investigate any significant differences between those that could be considered young and those that could be considered middle aged. Frequencies and standard errors for socio-demographic characteristics as well as clinical characteristics were calculated for each age group. Diagnosis groups were categorised as any affective disorder (depression, dysthymia, bipolar disorder), any anxiety disorder (agoraphobia, social phobia, panic disorder, GAD, OCD, PTSD), and finally any mental disorder (affective, anxiety, substance use, and substance dependence). K6 mean scores and standard errors were calculated for each age group in the total population.

Age groups were compared on socio-demographic and clinical characteristics using χ^2 analysis whereas functional limitation (measured by WHODAS 2.0 scores) and the K6 mean scores were compared using Poisson regression where K6 scores and WHODAS 2.0 scores formed the dependent variable and age group formed independent variable. Frequencies and means were weighted for the sex and age distribution of the Australian general population and standard errors were adjusted to account for the complex survey design using the balanced repeated replication technique (Wolter, 2007). All descriptive analyses were conducted using the SUDAAN software package.

Dimensionality

The type of IRT model that was used in this study requires that that the K6 is unidimensional (i.e. the total variance between K6 items can be explained by one underlying latent dimension). Previous studies have demonstrated that the Kessler scales possess a single factor structure of non-specific psychological distress (Kessler *et al.*, 2002; Fassaert *et al.*, 2009). To further test this assumption, the pairwise polychoric correlations between K6 items were estimated and submitted to a single factor confirmatory factor analysis (CFA) using a robust weighted least squares mean and variance adjusted (WLSMV) method of estimation (Muthén and Muthén, 2010). Separate CFAs were conducted using the total sample and for each age group under investigation.

The fit of each CFA was based on several statistical indices. The Comparative Fit Index (CFI), the Tucker–Lewis Index (TLI), and the root mean square error of approximation (RMSEA) provide a complementary perspective of model fit. Current recommendations suggest that TLI and CFI values >=0.90 and values >=0.95 indicate acceptable and very good fit, respectively (Hu and Bentler, 1998). RMSEA values less than 0.05 indicate close mode fit, values up to 0.08 indicate reasonable model fit, and values exceeding 0.10 indicate poor fit (Browne and Cudeck, 1993).

Differential Item Functioning

The graded response model was used to model the probability of endorsing each response option (none of the time, a little of the time, some of the time, most of the time, all of the time) on each K6 item according to the latent level of psychological distress, which in this instance is measured as a normally distributed latent variable. As mentioned previously, the IRT model accounts for the relationship between individual distress levels on the latent dimension and the probability of endorsing a particular item response. This relationship can be described in IRT through the use of several item parameters, known as the discrimination or *a*-parameter and the difficulty or *b*-parameters. The *a*-parameter describes the relationship between the item and the latent dimension, for example the larger the value of the *a*-parameter the stronger the item discriminates between individuals along a specific point on the latent dimension. The *b*-parameters describe the location on the latent dimension where examinees have a 50% probability of endorsing a particular response option (Embretson and Reise, 2000). In the graded response model, each item can be described using one *a*-parameter and several *b*-parameters, the number of *b*-parameters is determined by one minus the total number of response options, in this case four.

The Item Response Theory Loglikelihood Ratio test for Differential Item Functioning (IRTLR-DIF) procedure was utilised to test for significant DIF in the *a*- and *b*-parameters of each of the K6 items and is described in further detail by Teresi *et al.*, (2007). By testing for DIF in the *a*- and *b*-parameters separately, the analysis could examine whether the degree of latent psychological distress was associated with respondents endorsement each of the items differently as a function of age (e.g., *a*-DIF) and whether age influenced the ability of the K6 items to distinguish between continuous degrees of psychological distress (e.g., *b*-DIF).

As mentioned previously, to examine whether the *a*- or *b*-parameters exhibited DIF across age groups, a set of invariant items, or 'anchor' set, is required to equate the separate age groups to a common latent scale. Several selection methods have been developed to 'purify' items to assist with the selection of an appropriate anchor set when there is no theoretical basis for the choice of items (Woods, 2009). A simulation study found that the identification of a pure anchor set (i.e. items that clearly do not exhibit DIF between the examined group) prior to DIF detection significantly improved the accuracy rate and reduced false positive results, particularly in IRT-based methods of DIF detection (Navas-Ara and Gomez-Benito, 2002).

This study used an iterative procedure recommended by Kim and Cohen (1995) to identify an appropriate anchor set. First, all K6 items were tested for DIF using IRTLR-DIF software (Thissen, 2001). At this stage no specific anchor set is specified therefore each item was tested using the remaining items as a temporary anchor set. For each item the likelihood-ratio test statistic (G^2 change), which compares a model with the parameter estimates constrained to be equal between the two examined groups with a model that frees the parameters to be estimated separately between the two groups, was produced. The item with the largest likelihood-ratio statistic, indicating the largest level of DIF, was removed from the analysis and the remaining items were re-examined. The procedure continued in this manner until items no longer displayed significant DIF at the 0.05 significance level. These items were selected as the anchor set to equate the two groups on a common latent scale. The alpha level of 0.05 was chosen at this stage, regardless of multiple comparisons, to protect against the possibility of Type II error and therefore ensuring a conservative selection of anchor items.

Once the anchor set was selected, the IRTLR-DIF procedure was again utilised for the final test for significant DIF in the remaining items. The log-likelihood ratio test statistic was again used as an omnibus test for significant DIF measured simultaneously in the *a*- and *b*-parameters of the graded item response model. Assuming a theoretical probability that DIF may be located in one parameter, the statistical significance of the omnibus test statistic was calculated using 1 degree of freedom as suggested by Thissen (2001). If the omnibus test was significant, follow-up tests were then conducted on the individual parameters to identify if the DIF was present in the *a*-parameter or *b*-parameters, or both. It was at this stage that the Benjamini-Hochberg method was used to adjust the critical *p*-value for multiple comparisons to reduce the possibility of Type I error (Benjamini and Hochberg, 1995; Thissen *et al.*, 2002). The final IRT parameters were estimated in the MULTILOG 7.03 software package.

Practically, it is possible for significant levels of DIF to be identified by the IRTLR-DIF procedure but the magnitude of the DIF may have a negligible effect on the overall performance of the scale. Graphs of the total expected score functions for each age group comparison were produced to examine the magnitude of DIF and the impact on the overall expected scale scores. These graphs were calculated by summing each of the individual items' expected score function for the K6 and graphing that as a function of the latent dimension. Briefly, these graphs represent the level of latent psychological distress (x-axis) required to generate an expected total score on the K6 scale (y-axis) for each age group. In addition, a method for quantifying the difference between the expected score functions for each item is known as the non-compensatory DIF (NCDIF) index. This index is calculated by averaging the squared difference in expected scale score for each item between members of the focal group and members of the reference group along the actual distribution of the latent trait (Raju *et al.*, 1995). Cut off values based on simulation studies indicate that NCDIF scores greater than 0.096 indicate that the DIF is of a sufficient magnitude to have a substantial impact on the expected scale score (Teresi *et al.*, 2007).

RESULTS

Sample characteristics

Only respondents with complete data for the K6 were included in the current study. The total sample size was 8,840 as one respondent had one or more missing values on the K6 items. The total sample size included 2,761 respondents aged between 16 and 34, 4,174 respondents aged between 35 and 64, and finally 1,905 respondents aged between 65 and 85. Socio-demographic characteristics for each of the age groups are presented in Table 1. The results revealed significant differences between the three age groups with respect to all socio-demographic characteristics. On average, the old age group had a slightly higher percentage of female, widowed/separated/divorced, not in the labour force, and lower educated individuals in comparison to the younger age groups.

Clinical characteristics

The percentages and standard errors of respondents with any affective, any anxiety, and any mental health disorder across the lifespan and in the 30 days priors to interview for each age group are presented in Table 2. χ^2 analyses revealed that old age respondents are less likely to have a diagnosis of any affective, any anxiety, and any mental health disorder over their lifetime and in the past 30 days. Indeed, the lifetime prevalence of any mental health disorder in old age respondents is approximately half that of both the younger age groups. Functional limitation, however, is significantly higher in the old age group in comparison to both younger age groups as measured by the WHODAS 2.0. Finally, K6 mean scores were significantly lower for the old age group in comparison to both younger age groups.

K6 item analysis

Dimensionality

One factor CFA of polychoric correlations in the total sample and separately for each age group indicate that a one factor model satisfactorily fit the K6 data as evidenced by the fit indices. CFI, TLI, and RMSEA statistics indicated good model fit according to our a priori cut points in the total sample with values of 0.99, 0.98, and 0.06, respectively. Likewise, CFA conducted separately for each age group indicates that a one factor model provides reasonably good fit in the young (CFI = 0.99, TLI = 0.98, RMSEA = 0.07), middle aged (CFI = 0.99, TLI = 0.99, RMSEA = 0.05), and old aged (CFI = 0.99, TLI = 0.98, RMSEA = 0.05). The individual standardised factor loadings for all the models were significantly high (>= 0.6). These results support the assumption of unidimensionality required for the IRT analysis.

Differential Item Functioning

The results of the DIF analyses and the *a*- and *b*-parameters comparing the young with the old, the middle age with the old, and the young with the middle age are presented in Tables 3, 4, and 5, respectively. The item addressing negative affect was identified as a suitable anchor item comparing young with old and young with middle age whilst the item addressing hopelessness was identified as a suitable anchor comparing middle age with old. No other items were found to be suitable for the anchor set. After Benjamini-Hochberg adjustments, there were no significant levels of DIF identified in the items between young and middle age respondents. There were two items, worthlessness and fatigue, that displayed significant levels of DIF between young and old respondents and middle age and old respondents. Additionally, the item addressing negative affect displayed significant levels of *a*-DIF whilst the item addressing nervousness displayed significant levels of *b*-DIF between middle age and old respondents.

Inspection of the total expected score functions between young and old respondents and middle age and old respondents, as shown in Figure 1, reveals that old age respondents tend to endorse slightly higher K6 scores when both groups are matched in terms of their underlying psychological distress level. Although, the magnitude of the DIF is relatively low, as evidenced by individual item NCDIF values below the a priori cut points. The one item with the greatest magnitude of DIF (NCDIF = 0.20), and above our a priori cut point of 0.09, was found in the item addressing fatigue between old and young respondents. This indicates that old respondents are more likely to endorse higher response options in the fatigue item in comparison to young respondents even though both groups are matched in terms of their underlying psychological distress level.

DISCUSSION

The series of IRT likelihood ratio tests of differential item functioning comparing old, middle aged and young respondents revealed that not all K6 items are invariant across age. Although worthlessness, nervousness, fatigue and negative affect items were found to be non-invariant, an assessment of their magnitude and therefore likely impact on scores was explored and only one item, fatigue, remained significant. In the old age group, for the item of fatigue, biases of sufficient magnitude artefactually *increased* the older groups overall mean K6 scores in comparison to the young age group.

The finding of item bias in the assessment of fatigue indicates that psychological distress is slightly over-estimated in the old age group. Inspection of the expected score functions presented in Figure 1 indicates that if we theoretically assume that 99% of the population lie within -3 and +3 on the latent dimension of psychological distress, then on average, old age respondents equated in terms of latent distress, receive an expected score on the K6 that is approximately 0.61 points higher than the young age group, reflecting an effect size of approximately 0.22. Subtracting 0.61 from the overall total K6 mean score of old age respondents brings their total score to 1.19 compared to 3.00 for the young group. This correction, at the group level, results in an apparent widening of the disparity of mean K6 scores between young and old.

Using K6 scores not corrected for item bias, the current results indicate that across the lifespan levels of psychological distress decrease with increasing age. As above, when correction for the fatigue item is made with the older group, the pattern remains and is more apparent. Given the strong association demonstrated between DSM-IV disorders and psychological distress (Andrews and Slade, 2001; Furukawa *et al.*, 2003; Kessler *et al.*, 2010a; Slade *et al.*, 2011; Sunderland *et al.*, in press), evidence for age related decline using formal diagnostic classification as operationalised by the CIDI was also assessed. The observed prevalence estimates of DSM-IV affective and anxiety disorders across the lifespan presented in the current study confirm the K6 age-related decline and are consistent with previous evidence that the prevalence of psychiatric disorders, particularly 30 day prevalence of affective and anxiety disorders declines with increasing age (Kessler *et al.*, 2010b; Trollor *et al.*, 2007).

The current finding, that the correction for age-related item bias in a measure of psychological distress served to heighten rather than minimise the age dependent disparity, goes some way in contributing to the broader debate about the legitimacy of a seemingly pervasive finding of improved mental health during old age. Psychological distress, is however only one measure of psychopathology and although the expected relationship with our formal diagnostic instrument (i.e. the CIDI) was observed, our finding would be complemented by systematic research into item bias in the CIDI instrument. Indeed, O'Connor and Parslow (2009) have argued that the due to the lengthy nature and item complexity associated with the CIDI, it may be subject to greater age dependent bias than screening measures such as the Kessler scales. Using prevalence ratios between young and old, they demonstrated that levels of psychological distress, as measured by the K10, did not decrease as dramatically in old age as the DSM-IV prevalence estimates would suggest (O'Connor and Parslow, 2009). Using the current data, we replicated this ratio pattern. For example, the decrease in K6 mean scores from young to old age in the current study was 3.0 to 1.8 respectively, a ratio of 1:0.6, whilst the decrease in 30 day prevalence for any mental disorder across young (11.2%) and old (3.3%) results in a smaller ratio of 1:0.3. Interestingly, when using K6 scores were corrected for the fatigue item, the K6 young to old ratio was reduced to 1:0.4, a result closer to that observed for the formal diagnostic instrument, although if the CIDI modules were subject to similar item bias, investigation and then subsequent corrections may cause the ratios to differ again.

This study is not alone in identifying that the clinical feature of fatigue may be interpreted differently in old age respondents filling out screening instruments. Indeed using the Centre for Epidemiologic Studies - Depression scale, Grayson *et al.* (2000) demonstrated that individuals with a disability, bone and joint disease, and stroke were more likely to report higher levels of fatigue above and beyond levels of depression. They concluded that bias may be present in the assessment of depression not due to aging per se, but due to increased physical comorbidities and functional impairment associated with age. In the current study, old age respondents had significantly higher levels of functional impairment than both the middle and younger groups and it could be that the age-related bias observed for the fatigue item is related to the higher functional impairment associated with old age. Although our analyses were not designed to confirm if physical comorbidity or any other particular factor mediates the endorsement of the fatigue item in old age, it may be prudent with individual patients for clinicians to consider if fatigue is presenting as a clinical feature of psychological distress or a subsequent symptom of physical disorders that occur with the normal aging process.

Overall, greater confidence in prevalence estimates, policy planning, and the setting of clinical benchmarks for the old will be achieved when there is further evidence regarding the impact of age-related biases across lengthy diagnostic instruments used to derive national prevalence estimates, as well as across popular screening instruments used in research and clinical settings. Until this is achieved, our findings suggest that clinicians and researchers can use the overall K6 scale to compare psychological distress across various age groups with some degree of confidence. Researchers and clinicians may wish to be mindful that the fatigue item was associated with bias of sufficient magnitude and view this finding in context.

The results of the current study should be interpreted against several limitations. Most notably, simulation studies have indicated that, whilst the use of one item as an anchor set provides an acceptable level of power, the pure anchor set used to equate the measurement scale between the two groups should ideally contain four or more items (Wang and Yeh, 2003). However, the use of small pure anchor sets appears to be favourable over the use of anchor sets that may contain DIF items, which could distort the results if used to equate the measurement scale (Woods, 2009). Given that the purification process in the current study indicated that only one item was free from DIF, it was decided to maintain the use of a one item anchor instead of increasing the chance of including additional items in the anchor that have some indication of DIF. Secondly, the current study utilised the IRT-LR method to detect DIF however there are many procedures to identify measurement invariance, including multigroup confirmatory factor analysis or multiple indicators multiple causes (MIMIC) models, each with their own pros and cons. The current results would benefit from future research replicating the findings using various samples and methods to detect age-related bias. Finally, due to statistical constraints our study examined the K6, rather than the K10. However, due to the similarity in items between the two Kessler measures, we suggest that it is likely that the K10 measure can also be used with confidence with old age individuals and across age groups when comparisons are desired.

In summary, the K6 when subjected to DIF analyses revealed only one item with age-related bias of sufficient magnitude to artefactually increase scores in old age. The relative impact of the fatigue item when comparing scores of respondents aged 65 years and older with respondents aged 16 to 34 is likely to be minor. Overall the measure can be used with confidence in old age respondents and when comparing young and middle age group data. Given the relative small impact of the fatigue item on the overall old age distress scores is up to the individual researcher or clinician to judge if this warrants consideration when interpreting results. Those interested in the broader debate about the legitimacy of the apparent age related decline in psychopathology will be assured that data based on the Kessler scales that have been used to inform the debate are unlikely to be subject to marked measurement bias.

CONFLICT OF INTEREST

None.

DESCRIPTION OF AUTHORS' ROLES

M Sunderland formulated the research question, was responsible for the statistical design and analyses and wrote the manuscript. M Hobbs formulated the research question, supervised the statistical analyses and assisted with writing the manuscript. T Anderson formulated the research question and assisted with writing the manuscript. G Andrews provided critical revisions to the draft manuscript.

ACKNOWLEDGEMENTS

This research was funded by a NHMRC Capacity Building Grant ID 568940.

REFERNCES

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

Baker, F. B. (2001). *The Basics of Item Response Theory*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Benjamini, Y. and Hochberg, Y. (1995). Controlling for false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B,* 57, 289-300.

Browne, M. W. and Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen and J. S. Long (Eds.), *Testing Structural Equation Models*. Newbury Park, CA: Sage.

Cohen, A. S., Kim, S-H. and Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement,* 17, 335-350.

Embretson, S. E. and Reise, S. P. (2000). *Item Response Theory for Psychologists.* New York, NY: Lawrence Erlbaum Associates, Inc.

Ernst, C. and Angst, J. (1995). Depression in old age. Is there a real decrease in prevalence? *European Archives of Psychiatry and Clinical Neuroscience*, 245, 272-287.

Fassaert, T., et al. (2009). Psychometric properties of an interviewer-administered version of the Kessler Psychological Distress scale (K10) among Dutch, Moroccan and Turkish respondents. *International Journal of Methods in Psychiatric Research,* 18, 159-169.

Furukawa, T.A., Kessler, R.C., Slade,T. and 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, 357-362.

Grayson, D.A., Mackinnon, A., Jorm, A.F., Creasey, H. and Broe, G. A. (2000) Item bias in the Center for Epidemiologic Studies Depression Scale: effects of physical disorders and disability in an elderly community sample. *Journal of Gerontology - Series B Psychological Sciences and Social Sciences*, 55B, 273-282.

Henderson, A.S., Jorm, A.F., Korten, A.E., Jacomb, P., Christensen, H. and Rodgers, B. (1998). Symptoms of depression and anxiety during adult life: evidence for a decline in prevalence with age. *Psychological Medicine*, 28, 1321-1328.

Hong, S-I., Hasche, L. and Bowland, S. (2009). Structural relationships between social activities and longitudinal trajectories of depression among older adults. *The Gerontologist,* 49, 1-11.

Hu, L. T. and Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods,* 3, 424-453.

Jorm A.F. (2000). Does old age reduce risk of anxiety and depression? A review of epidemiological studies across the adult life span. *Psychological Medicine*, 20, 11-22.

Kessler, R.C., Birnbaum, H., Bromet, E., Hwang, I., Sampson, N., and Shahly, V. (2010b). Age differences in major depression: results from the National Comorbidity Survey Replication (NCS-R). *Psychological Medicine*, 40, 225-237.

Kessler, R.C., et al. (2002). Short screening scales to monitor population prevalence and trends in non-specific psychological distress. *Psychological Medicine,* 32, 959-976.

Kessler, R.C., et al. (2010a). 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 Journals of Methods in Psychiatric Research*, 19, 4-22.

Kessler, R. C. and Ustun, T. B. (2008). *The WHO World Mental Health Surveys.* New York, NY: Cambridge University Press.

Kim, S.-H. and Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education,* 8, 291-312.

Muthén, L. K. and Muthén, B. O. (2010). *Mplus Users' Guide Sixth Edition*. Los Angeles, CA: Muthén & Muthén.

Navas-Ara, M.J. and Gomez-Benito, J. (2002). Effects of ability scale purification on the identification of DIF. *European Journal of Psychological Assessment,* 18, 9-15.

O'Connor, D.W. and Parslow, R.A. (2009). Different responses to K-10 and CIDI suggest that complex structured psychiatric interviews underestimate rates of mental disorder in old people. *Psychological Medicine,* 39, 1527-1531.

Raju, N. S., van der Linden, W. J. and Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement,* 19, 353-368.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement,* 34, 100.

Slade, T., Grove, R. and Burgess, P. (2011). Kessler Psychological Distress Scale: normative data from the 2007 Australian National Survey of Mental Health and Wellbeing. *Australian and New Zealand Journal of Psychiatry*, 45, 308-316.

Snowdon, J. (2001). Is depression more prevalent in old age? *Australian and New Zealand Journal of Psychiatry*, 35, 782-787.

Sunderland, M., Slade, T., Stewart, G. and Andrews, G. (in press). Estimating the prevalence of DSM-IV mental illness in the Australian general population using the Kessler psychological distress scale. *Australian and New Zealand Journal of Psychiatry.*

Teresi, J. A. and Fleishman, J. A. (2007). Differential item functioning and health assessment. *Quality Life Research,* 16, 33-42.

Teresi, J. A., et al. (2007). Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): applications to measures of physical functioning ability and general distress. *Quality Life Research*, 16, 43-68.

Thissen, D. (2001). *IRTLRDIF v 2.0b: Software for the computation of the statistics involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning*. Chapel Hill, NC: University of New Carolina. Software and user's manual available at: http://www.unc.edu/~dthissen/dl.html.

Thissen, D., Steinberg, L. and Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics,* 27, 77-83.

Trollor, J.N., Sachdev, P.S., Anderson, T.M., Andrews, G. and Brodaty, H. (2007). Age shall not weary them: Mental health in the middle-aged and the elderly. *Australian and New Zealand Journal of Psychiatry*, 41, 581-589.

Tsang, A., et al. (2008). Common chronic pain conditions in developed and developing countries: gender and age differences and comorbidity with depression-anxiety disorders *Journal of Pain*, 9, 883-891.

Van den Linden, W.J., and Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*. New York, NY: Springer-Verlag.

Vink, D., Aartsen, M.J. and Schoevers, R.A. (2008). Risk factors for anxiety and depression in the elderly: a review. *Journal of Affective Disorders*, 106, 29-44.

Wolter, K. M. (2007). *Introduction to Variance Estimation*. New York, NY: Springer.

Wang, W-C., and Yeh, Y-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement,* 27, 479-498.

Woods, C. (2009). Empirical selection of anchor for tests of differential item functioning. *Applied Psychological Measurement,* 33, 42-57.

TABLES

Table 1: Sample characteristics of young, middle aged, and old age respondents in the 2007 Australian National Survey of Mental Health and Well-being

Demographic		16-34 (n=2761)	35-64 (n=4174)	65-85 (n=1905)	χ2	p
Sex % (SE)	Male	50.6 (0.1)	49.7 (0.1)	47.4 (0.2)	89.2	<0.001
	Female	49.4 (0.1)	50.3 (0.1)	52.6 (0.2)		
Marital Status % (SE)	Never Married	73.9 (1.1)	13.9 (0.7)	3.9 (0.4)	1075.3	<0.001
	Widowed/Divorced/Separated	2.6 (0.3)	17.2 (0.8)	31.7 (1.0)		
	Married	23.5 (1.1)	68.8 (1.0)	64.4 (1.2)		
Labour Force % (SE)	Employed	74.6 (0.8)	75.2 (0.7)	10.3 (0.7)	1684.8	<0.001
	Unemployed	4.9 (0.5)	1.8 (0.3)	0.1 (0.5)		
	Not in labour force	20.5 (0.8)	23.0 (0.7)	89.7 (0.8)		
Education % (SE)	Post-School qualification	52.6 (1.0)	60.1 (1.0)	41.4 (1.4)	65.9	<0.001
	No post-school qualification	47.4 (1.0)	39.9 (1.0)	58.6 (1.4)		
Country of Birth % (SE)	Australia	77.9 (1.2)	70.7 (1.1)	69.2 (1.1)	19.9	<0.001
	English-speaking country	6.4 (0.6)	13.5 (0.7)	14.7 (1.0)		
	Other	15.7 (1.1)	15.8 (1.0)	16.1 (1.1)		

Table 2: Clinical characteristics of young, middle aged, and old age respondents in the 2007 Australian National Survey of Mental Health and Well-being

Disorder	16-34 (n=2761)	35-64 (n=4174)	65-85 (n=1905)	χ2	p
Lifetime Affective Disorder % (SE)	14.4 (0.8)	17.8 (0.9)	8.1 (0.7)	46.2	<0.001
Lifetime Anxiety disorder % (SE)	19.1 (0.9)	24.2 (1.0)	10.4 (0.8)	70.9	<0.001
Lifetime Any disorder % (SE)	42.0 (1.3)	46.2 (1.3)	24.6 (1.1)	103.6	<0.001
30 day Affective disorder % (SE)	2.5 (0.4)	3.0 (0.4)	0.9 (0.2)	20.8	<0.001
30 day Anxiety disorder % (SE)	7.7 (0.6)	7.8 (0.7)	2.5 (0.4)	35.5	<0.001
30 day Any disorder % (SE)	11.2 (0.7)	9.9 (0.7)	3.3 (0.5)	45.6	<0.001
WHODAS Mean (SE)	6.7 (0.6)	9.2 (0.9)	16.3 (2.2)	14.0*	<0.001
K6 mean score (SE)	3.0 (0.1)	2.8 (0.1)	1.8 (0.1)	53.4*	<0.001

Note: WHODAS = World Health Organization Disability Assessment Schedule, Any disorder classified as any affective, anxiety, and substance use disorder measured in the 2007 Australian survey, * significance generated using Wald F statistic with Poisson regression

Table 3: IRT parameters (SE) and differential item functioning of the K6 between young and old age respondents in the 2007 Australian National Survey of Mental Health and Well-being

Item	**Group**	**a**	**b1**	**b2**	**b3**	**b4**	**aDIF (p)**	**bDIF (p)**	**NCDIF**
Nervous	young	1.43(0.05)	-0.16(0.03)	1.19(0.05)	2.56(0.11)	3.59(0.19)	0.1 (0.75)	11.0 (0.03)	0.00
	old	1.43(0.05)	-0.16(0.03)	1.19(0.05)	2.56(0.11)	3.59(0.19)			
Hopeless	young	2.53(0.10)	0.70(0.03)	1.51(0.05)	2.28(0.08)	2.95(0.14)	5.6 (0.02)	12.0 (0.02)	0.00
	old	2.53(0.10)	0.70(0.03)	1.51(0.05)	2.28(0.08)	2.95(0.14)			
Restless/Fidgety	young	1.44(0.05)	-0.24(0.03)	0.91(0.05)	2.17(0.09)	3.32(0.16)	NS		
	old	-	-	-	-	-			
Negative Affect	young	2.96(0.14)	1.18(0.04)	1.72(0.05)	2.35(0.09)	3.31(0.21)	Anchor		
	old	-	-	-	-	-			
Fatigue	young	1.95(0.05)	0.13(0.03)	1.19(0.04)	2.11(0.07)	2.87(0.12)	2.4 (0.12)	***46.9 (<0.01)***	*0.20*
	old	1.95(0.05)	-0.25(0.04)	0.50(0.06)	1.32(0.08)	1.97(0.12)			
Worthless	young	4.48(0.25)	0.96(0.03)	1.54(0.04)	2.09(0.06)	2.70(0.11)	**10.0 (<0.01)**	**21.7 (<0.01)**	0.05
	old	2.79(0.23)	0.79(0.07)	1.22(0.09)	1.83(0.14)	2.31(0.11)			

Note: a, b1, b2, b3, and b4 represent the a-parameter and b-parameters for each item estimated by the graded response model. aDIF and bDIF represents the change in G^2 statistics between constrained and less constrained models. Bold indicates significant after Benjamini-Hochberg adjustments. Italics indicates the magnitude meets the required cut-off point

Table 4: IRT parameters (SE) and differential item functioning of the K6 between middle age and old age respondents in the 2007 Australian National Survey of Mental Health and Well-being

Item	Group	a	b1	b2	b3	b4	aDIF (p)	bDIF (p)	NCDIF
Nervous	middle	1.53(0.04)	0.16(0.03)	1.30(0.04)	2.50(0.08)	3.57(0.14)	3.5 (0.06)	**39.6 (<0.01)**	0.03
	old	1.53(0.04)	0.26(0.05)	1.17(0.08)	2.26(0.14)	2.75(0.19)			
Hopeless	middle	2.89(0.09)	0.83(0.02)	1.47(0.04)	2.15(0.06)	2.70(0.09)	Anchor		
	old	-	-	-	-	-			
Restless/Fidgety	middle	1.55(0.05)	0.10(0.03)	1.08(0.04)	2.21(0.08)	3.08(0.12)	NS		
	old	-	-	-	-	-			
Negative Affect	middle	3.76(0.17)	1.13(0.03)	1.62(0.04)	2.23(0.06)	3.04(0.14)	**6.6 (0.01)**	8.3 (0.08)	0.01
	old	3.15(0.36)	1.10(0.07)	1.46(0.10)	2.10(0.15)	3.16(0.48)			
Fatigue	middle	2.34(0.08)	0.23(0.02)	1.04(0.03)	1.83(0.06)	2.54(0.09)	**14.4 (<0.01)**	**17.8 (<0.01)**	0.03
	old	1.73(0.10)	0.02(0.05)	0.79(0.08)	1.62(0.13)	2.30(0.18)			
Worthless	middle	3.70(0.16)	1.01(0.02)	1.51(0.04)	2.14(0.06)	2.68(0.10)	**7.1 (<0.01)**	7.9 (0.10)	0.06
	old	3.32(0.25)	0.89(0.06)	1.26(0.07)	1.79(0.12)	2.20(0.15)			

Note: a, b1, b2, b3, and b4 represent the a-parameter and b-parameters for each item estimated by the graded response model. aDIF and bDIF represents the change in G^2 statistics between constrained and less constrained models. Bold indicates significant after Benjamini-Hochberg adjustments.

Table 5: IRT parameters (SE) and differential item functioning of the K6 between young and middle aged respondents in the 2007 Australian National Survey of Mental Health and Well-being

Item	Group	a	b1	b2	b3	b4	aDIF (p)	bDIF (p)	NCDIF
Nervous	young	1.50(0.05)	-0.07(0.03)	1.21(0.04)	2.49(0.08)	3.68(0.14)	0.5 (0.48)	13.8 (<0.01)	0.00
	middle	1.50(0.05)	-0.07(0.03)	1.21(0.04)	2.49(0.08)	3.68(0.14)			
Hopeless	young	3.11(0.09)	0.72(0.02)	1.42(0.03)	2.09(0.05)	2.73(0.09)	0.7 (0.40)	13.3 (0.01)	0.00
	middle	3.11(0.09)	0.72(0.02)	1.42(0.03)	2.09(0.05)	2.73(0.09)			
Restless/Fidgety	young	1.52(0.05)	-0.13(0.03)	0.94(0.04)	2.10(0.06)	3.05(0.11)	0.7 (0.40)	11.8 (0.02)	0.00
	middle	1.52(0.05)	-0.13(0.03)	0.94(0.04)	2.10(0.06)	3.05(0.11)			
Negative Affect	young	3.61(0.13)	1.03(0.02)	1.60(0.030	2.19(0.05)	3.01(0.11)	Anchor		
	middle	-	-	-	-	-			
Fatigue	young	2.28(0.06)	0.14(0.02)	1.02(0.03)	1.83(0.05)	2.53(0.07)	1.1 (0.29)	11.6 (0.02)	0.00
	middle	2.28(0.06)	0.14(0.02)	1.02(0.03)	1.83(0.05)	2.53(0.07)			
Worthless	young	3.96(0.14)	0.93(0.02)	1.46(0.03)	2.04(0.04)	2.60(0.07)	2.2 (0.14)	8.3 (0.08)	0.00
	middle	3.96(0.14)	0.93(0.02)	1.46(0.03)	2.04(0.04)	2.60(0.07)			

Note: a, b1, b2, b3, and b4 represent the a-parameter and b-parameters for each item estimated by the graded response model. aDIF and bDIF represents the change in G^2 statistics between constrained and less constrained models. Bold indicates significant after Benjamini-Hochberg adjustments.

FIGURE LEGEND

Figure 1: Expected Score Functions for K6 in young vs. old and middle age vs. old.