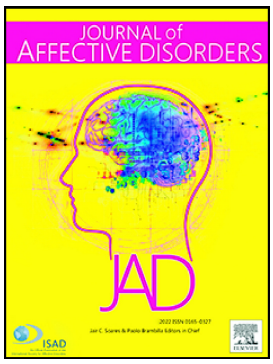

The performance of long vs. short questionnaire-based measures of depression, anxiety, and psychological distress among UK adults: A comparison of the patient health questionnaires, generalized anxiety disorder scales, malaise inventory, and Kessler scales

Dorottya Lantos, Darío Moreno-Agostino, Lasana T. Harris, George Ploubidis, Lucy Haselden, Emla Fitzsimons

PII: S0165-0327(23)00800-5

DOI: https://doi.org/10.1016/j.jad.2023.06.033

Reference: JAD 16219

To appear in:

Received date: 14 February 2023

Revised date: 12 June 2023

Accepted date: 16 June 2023

Please cite this article as: D. Lantos, D. Moreno-Agostino, L.T. Harris, et al., The performance of long vs. short questionnaire-based measures of depression, anxiety, and psychological distress among UK adults: A comparison of the patient health questionnaires, generalized anxiety disorder scales, malaise inventory, and Kessler scales, (2023), https://doi.org/10.1016/j.jad.2023.06.033

The Performance of Long vs. Short Questionnaire-Based Measures of Depression, Anxiety, and Psychological Distress Among UK Adults: A Comparison of the Patient Health Questionnaires, Generalized Anxiety Disorder Scales, Malaise Inventory, and Kessler Scales

Dorottya Lantos,[1,2]

Darío Moreno-Agostino,[1,3]

Lasana T. Harris,[1,2]

George Ploubidis,[1]

Lucy Haselden[1]

and

Emla Fitzsimons[1]

[1]Centre for Longitudinal Studies, Social Research Institute, UCL

[2]Department of Experimental Psychology, UCL

[3]ESRC Centre for Society and Mental Health, King's College London

Author credit statement: DL was responsible for developing the online survey, collecting and analysing data, and preparing the original draft of this manuscript. DM-A and GP contributed to the interpretation of data analyses. LH contributed to the selection of instruments and sampling. EF, LTH, and GP conceptualized the project and supervised data collection and analyses. All authors contributed to the writing process by providing critical comments and through editing.

Abstract

It is often important to minimise the time participants in social science studies spend on completing questionnaire-based measures, reducing response burden, and increasing data quality. Here, we investigated the performance of the short versions of some widely used depression, anxiety, and psychological distress scales and compared them to the performance of longer versions of these scales (PHQ-2 vs PHQ-9, GAD-2 vs GAD-7, Malaise-3 vs Malaise-9, K6 vs K10). Across a sample of UK adults (N = 987, ages 18-86), we tested the existing factor structure and accuracy of the scales through confirmatory factor analyses and exploration of the total information functions, observing adequate model fit indices across the measures. Measurement invariance was tested across birth sex and age groups to explore whether any differences in measurement properties or measurement bias may exist, finding support for the invariance of most measures. We conducted bivariate correlations across the measures as a way of obtaining evidence of the equivalence in the rank-ordering of short vs long scales. The results followed a similar pattern across the young adult subsample (N = 375, ages 18-39) as in the overall sample. Overall, these results indicate that the short forms of the tested scales may perform similarly to the full versions. Where brevity is important, researchers may opt to use the shorter versions of the scales based on these data.

Keywords: measurement, depression, psychological distress, anxiety, questionnaire optimisation

The Performance of Long vs. Short Questionnaire-Based Measures of Depression, Anxiety, and Psychological Distress Among UK Adults: A Comparison of the Patient Health Questionnaires, Generalized Anxiety Disorder Scales, Malaise Inventory, and Kessler Scales

Questionnaire-based measures are among the most frequently used methods for data collection in the psychological and social sciences (Fernández-Ballesteros, 2004; Stone et al., 2000). This is not only because of the ease with which such measures can be administered (i.e., using only a paper and pen or a computer-based survey instead of more complex instruments required for other types of measures, including behavioural, physiological, or reaction time-based measures), but also because of the assumption that it is the individual who can most accurately respond to questions about their thoughts and feelings. Yet questionnaire-based measures also come with limitations. For example, certain scales may become outdated as time passes (e.g.. Torsvik et al., 2021), while scales that work reliably in one cultural context may not do so in a different cultural context (Beaton et al., 2000; Paulhus & Vazire, 2007). Moreover, people may not have access to all of the thoughts and feelings that drive their behaviour (Scheuler & Schreiber, 2004). The time which participants spend completing social science research is invaluable, yet the available time in each data collection session and thus the amount of data that can be collected are both restricted.

For the above reasons, it is important to understand which, among the numerous validated and widely used measures, are most optimal for inclusion in a given research project. It is further often vital and desirable to keep such measures as short as possible since shorter measures reduce response burden and as a result increase data quality (Rolstad et al., 2011). Ensuring that assessments are as short as possible whilst remaining as valid and reliable is possible is likewise increasingly important in clinical settings where outpatient encounters may be brief and competing demands are continuously present (Levell, 2022).

How short can measures be while still capturing the variance in a construct of interest? Here, we focus on psychometric measures developed for capturing psychological distress, depression, and anxiety. We explore and compare the properties of short and long versions of such measures among a non-clinical sample of UK adults. To do so, we investigate the fit of the previously established factor structure as well as the measurement invariance of seven scales. Relying on item response theory, we further introduce a shorter version of the Malaise Inventory, as the shortest currently available measure includes 9 items (Ploubidis et al., 2019). We devote additional attention to the younger adult subsample (ages 18-39). This age group is of special interest as these analyses are also being used to inform the upcoming sweep (age 22) of the Millennium Cohort Study (MCS), an observational cohort study which has been following the lives of nearly 19,000 individuals born in the UK at the turn of the century (Connelly & Platt, 2014; Joshi & Fitzsimons, 2016).

Selecting Self-Report Measures: The Unique Case of Cohort Studies

Longitudinal birth cohort studies follow a cohort of individuals sharing a similar birth date. They help researchers understand how, why and when inequalities evolve over time, and how social, economic and environmental factors influence various life outcomes (e.g., mental health, Hunt & White, 1998; Samet & Muñoz, 1998). The UK is home to a unique set of birth cohort studies, all still running to this day, and including generations from 1946, 1958, 1970 and 2000/01. A key challenge of longitudinal studies is keeping participants engaged and minimising attrition. Another important challenge is including the best measures to ensure scientific rigour and relevance over time alongside minimising respondent burden.

Some measures may have as little as only two items, such as the two item versions of the Patient Health Questionnaire (PHQ-2, Kroenke et al., 2001, 2003; Kroenke & Spitzer, 2002) or that of the Generalized Anxiety Disorder Questionnaire (GAD-2, Kroenke et al., 2007; Spitzer et al., 2006), explored here, while single-item measures also exist in the

literature (Elo et al., 2003; Gardner et al., 1998; Postmes et al., 2013; Wanous & Reichers, 1996). In preparation for the upcoming MCS data sweep (2023), we test among a sample of UK adults complete and shortened versions of existing measures of psychological distress, depression, and anxiety to inform the selection of scales to be included in the data sweep. We aim to establish the performance of the short scales in comparison to the longer versions of the scales. Therefore, we anticipate that the results of the present analyses will indicate whether the short versions are valid and reliable, and whether their properties and performance are similar enough to the long scales for inclusion in this assessment. Furthermore, as cohort studies aim to facilitate cross-cohort comparisons, we explore the properties of the scales among not only among young adults in the UK, but also among UK adults of all ages.

Overview of the Study

In the present study, we explored the properties of the complete and short versions of various mental health measures using an online survey: the K10 and K6 measures (Kessler et al., 2002), the 9-item (PHQ-9, Kroenke et al., 2001; Kroenke & Spitzer, 2002) and 2-item (PHQ-2, Kroenke et al., 2003) versions of the Patient Health Questionnaire (PHQ), and the 7-item (GAD-7, Spitzer et al., 2006) and 2-item (GAD-2, Kroenke et al., 2007) versions of the Generalized Anxiety Disorder scale (GAD). Doing so, our aim was to clarify whether the short versions are comparable in a sample of UK adults to their full version. We further examined these characteristics among only the young adult subsample (18-39 years) in preparation for the next MCS data sweep (Connelly & Platt, 2014; Joshi & Fitzsimons, 2016). To gather data of the highest possible quality, keeping in mind the limited available time for the completion of survey-type measures, we aim to inform the selection of self-report questionnaires for use in the upcoming data sweep (age 22, 2023) with the results presented here. We additionally aim to inform researchers facing similar challenges who may

be looking at specific age groups in their work. The study was preregistered (https://osf.io/bk9xs)[1]. Ethical approval was obtained from the Ethics Committee of University College London. All data and syntax files are available via OSF (https://osf.io/vg4a9/).

Method

Participants

A sample of 1,068 UK adults started the survey. The sample was recruited to closely mimic one that is representative of the population. We removed the data of 8 participants who gave consent to partaking but did not consent to the storage of their data, as well as 40 participants who only filled in the consent form and nothing else. We excluded a further 33 participants from data analysis due to incorrect responses to (one or both) attention check questions (e.g., Please select agree). The final sample consisted of 987 participants (463 males, 505 females, 2 participants indicated that they did not wish to share their birth sex), ages 18-86, M = 45.21, SD = 15.61. Seventeen participants only partially completed the survey, and their demographic details were thus missing. Participants were recruited via Prolific Academic (https://www.prolific.co/), an online platform where participants may voluntarily register and complete surveys and studies in return for monetary rewards. Prolific Academic allows researchers to set specific demographic parameters during the recruitment process, which allowed us to recruit a sample which closely resembled the UK population with regards to sex, age and ethnicity. Participants were reimbursed £7.50 for their time. Across some of the analyses we were interested primarily in the responses of young adults, and hence completed them by including only the 375 participants who were aged 18-39 (M = 28.56, SD = 6.39, 184 males, 191 females).

Procedure

[1] Note that the preregistration did not include the plan to test measurement invariance.

Data was collected as part of a larger project. We created an online survey using Qualtrics software. Participants were first presented with an informed consent form and information sheet detailing their tasks throughout the study. They next completed several psychometric questionnaires. Among the measures, we implemented two attention check questions (e.g., Please select agree) to filter out participants who were not reading the items of the questionnaires with care. All scales were presented in a randomized order across participants. Finally, participants responded to demographic questions (birth sex, gender identity, age, ethnicity), were debriefed and thanked for their time.

Measures

Psychological distress was assessed using the 10-item K10 and the 6-item K6 scale (Kessler et al., 2002), along the 9-item version of the Malaise Inventory (Ploubidis et al., 2019; Rutter et al., 1970). The K10 and, embedded within it, the K6 (Kessler et al., 2002) were completed by 971 participants. Participants responded to the items (e.g., During the last 30 days, about how often did you feel hopeless?) on a 5-point Likert scale (1 = none of the time, 5 = all of the time). Participants' responses were summed, with higher scores indicating greater psychological distress.

The 9-item version of the Malaise Inventory (Ploubidis et al., 2019; Rutter et al., 1970) was completed by 974 participants. Participants completed the items of the questionnaire (e.g., Do you often feel miserable or depressed?) using binary yes/no responses. We scored 'yes' responses as 1 and 'no' responses as 0, and summed participants' overall answers, with higher scores indicating greater psychological distress.

Depression was assessed using the PHQ-9 (Kroenke et al., 2001; Kroenke & Spitzer, 2002) and, embedded within it, the PHQ-2 (Kroenke et al., 2003). These measures were completed by 976 participants. Participants responded to the items (e.g., 'Over the last 2 weeks, how often have you been bothered by any of the following problems? – Little interest

or pleasure in doing things') on a 4-point Likert scale (0 = not at all, 3 = nearly every day). Participants' responses were summed, with higher scores indicating increased experiences of depressive symptomatology.

Anxiety was assessed using the GAD-7 (Spitzer et al., 2006) and, embedded within it, the GAD-2 (Kroenke et al., 2007). These measures were completed by 974 participants. Participants responded to the items (e.g., 'Over the last 2 weeks, how often have you been bothered by the following problems? – Feeling nervous, anxious, or on edge') on a 4-point Likert scale (0 = not at all, 3 = nearly every day). Participants' responses were summed, with higher scores indicating increased experiences of anxiety.

Data Analyses

Measurement properties were investigated with a latent variable modelling approach using MPlus version 8.7 (Muthén & Muthén, 1998-2017). We conducted confirmatory factor analyses with a robust mean and variance adjusted weighted least squares (WLSMV) estimator to explore the latent structure of each self-report measure. We employed either a model for binary or for ordered categorical data depending on response options used for each scale (i.e., Yes/No binary responses vs. Likert-scales). As the scales included in the present manuscript all have well-established factor structures, we relied on confirmatory factor analyses. We used the root mean square error of approximation (RMSEA, Steiger, 1990), the comparative fit index (CFI, Bentler, 1990), and the Tucker-Lewis Index (TLI, Tucker & Lewis, 1973) to determine model fit. We interpreted RMSEA values up to .05 as indicating good fit, and values up to .08 as indicating adequate fit (Hu & Bentler, 1998). In the cases of CFI and TLI, we interpreted values greater than .90 as indicating adequate, and those greater than .95 as indicating good model fit (Barrett, 2007).

Drawing on item response theory, we additionally evaluated the precision of measurement of the self-report questionnaires by plotting the test information functions (TIF)

using MPlus version 8.7 (Muthén & Muthén, 1998-2017). TIF plots depict the Fischer information (a measure of the precision or reliability of the measure due to its inverse relationship with the standard error of measurement) at different levels of the underlying latent variable (Betz & Turner, 2011). All analyses exploring the properties of the self-report questionnaires were conducted on the complete sample as well as on the young adult subsample. The young adult sample was of special interest to our research group whilst preparing for the upcoming MCS data sweep, whereas the data of the complete sample with greater variance in age may be of interest to other researchers.

Item reduction. A 9-item Malaise Inventory is currently the shortest available version of this measure. We aimed to optimise this scale by shortening it. We first conducted a factor analysis to examine the general properties of the scale. Next, we selected the items with the highest discrimination parameters to create the short scale. We aimed to keep the TIF as similar as possible to that of the original scale and to ensure that internal consistency also remained optimal. We considered the item thresholds when making decisions about the items to retain. Where item thresholds were very high, thus resulting in low item endorsement and, subsequently, low variability in a general (not clinical) population like that of MCS, lower loadings but thresholds closer to the centre of the distribution of latent factor scores were preferred.

Measurement invariance was tested to explore whether the measurement properties of the questionnaires were equivalent across birth sex and age groups (Armstrong, 1998; Little, 2013; van de Schoot et al., 2013, 2015). This type of strategy could not be implemented in scales with three or less items, since in those cases the configural model is just-identified at best, leading to non-meaningful goodness-of-fit indices that cannot be compared to those from models with invariance constraints. As a result, it was not possible to test measurement invariance in most of the shorter versions of the scales. The analyses were

performed in the cases of the K10, K6, PHQ-9, GAD-9, and 9-item Malaise scales to detect potential differences in the measurement properties of the larger scales that may impact the shorter versions.

We conducted the analyses across four groups (birth sex * age): younger males, older males, younger females, and older females. As in the previous factor analyses, we used a WLSMV estimator and tested two levels of invariance: configural (where no measurement parameters were constrained to be equal across groups) and scalar invariance (where both the loadings and thresholds of the items were constrained to be equal across groups). We compared the goodness-of-fit indices of the two models. The chi-square difference test is very sensitive to sample size, which in this case is large enough to influence the results of the test. Models where the loss of fit was less than 0.01 for CFI and 0.015 for RMSEA met the criteria for invariance (Chen, 2007; Cheung & Rensvold, 2002). These analyses were conducted using MPlus version 8.7 (Muthén & Muthén, 1998-2017).

Scale properties were explored by looking at the descriptive statistics of each scale. We additionally conducted independent samples *t*-tests on the sum scores of all scales to test whether any differences existed between birth sexes or age groups (i.e., 18–39-year-olds compared to 40+ year-olds), and 2 x 2 ANOVAs to test for interactions. The two participants who did not disclose their birth sex were excluded from the analyses where splitting across sexes was meaningful. These analyses were conducted using SPSS 27.0. Internal consistency of the scales was assessed with McDonald's (1999) ω_t coefficient, estimated with the Omega macro for SPSS (Hayes & Coutts, 2020). McDonald's ω_t is the "proportion of test variance due to all common factors" (p. 152), and is equivalent to Cronbach's α when a scale is unidimensional (Revelle & Zinbarg, 2009). McDonald's ω_t indicates the internal consistency of the scales, where coefficients .70 or greater are considered adequate, with values closer to one indicating higher levels of internal consistency (Cicchetti, 1990, 1994).

Correlations. We computed the correlation matrix of longer and shorter versions of the psychological distress, depression, and anxiety scales. This allowed us to explore the equivalence in the rank ordering across the measures, convergent validity, and discriminant validity.

Results

Measurement Properties

We checked the fit of the established factor structures of each scale. The fit statistics of all administered scales (Table SM1), the item loadings (Figure SM1), and the TIFs of the configural models (Figure SM2) are presented in the Supplementary Materials. Only the Malaise Inventory showed adequate fit based on the RMSEA. However, the CFI and TLI showed a good model fit across all measures.

Shortening the Malaise Inventory

We aimed to optimise the 9-item Malaise Inventory by selecting only the 3 items with the highest loadings on the underlying latent variable. The three items matched across the full sample and the young adult subsample (Figure SM1, Appendix A). The analyses revealed that among young females, the responses were always the same to items 4 and 6, suggesting that including both items does not contribute to the variance in this age group. A similar finding was observed among the responses of older females to items 3, 5, and 7, and among those of older males to items 3 and 5. The final 3 items ('*Do you often get worried about things?', 'Are you easily upset or irritated?'*, '*Does every little thing get on your nerves and wear you out?*') do not contain any of those that could introduce redundant information in either age group.

Measurement Invariance

We first explored the results of the K10. No women over 40 responded with 'all of the time' to the question 'During the last 30 days, about how often did you feel so nervous that

nothing could calm you down?' (item 3), while only 3 men over 40 did, and no women over 40 responded with 'all of the time' to the question 'During the last 30 days, about how often did you feel so restless you could not sit still?' (item 6), while only 2 men over 40 did. This could be dealt with by grouping the two most extreme categories together and thus creating an overall cluster with existing responses. However, to form meaningful comparisons, we would in this case have to cluster the responses of the young age group together as well. As the younger age group provided responses across all scales in all categories, this would lead to the loss of information. For the sake of retaining such information, we did not compare the sample across ages, and instead we only explored sex differences within the young adult sample.

Across all remaining measures, we tested measurement invariance across birth sexes and age groups (i.e., total 4 groups: males ages 18-39, females ages 18-39, males ages 40+, females ages 40+). For the sake of consistency, we also conducted all analyses only among the young adult group, comparing the responses of males and females. The results of the measurement invariance testing procedure are presented in Table 1. Although the changes in RMSEA only indicated an adequate fit in the case of the Malaise Inventory, the changes in CFI and TLI indicated a good fit across models. These results are in line with the baseline RMSEA, which are higher than desirable (Table SM1).

Table 1. *Measurement Invariance Testing*

	Configural Model					*Scalar Model*					*Differences*		
	χ^2	RMSEA	CFI	TLI	SRMR	χ^2	RMSEA	CFI	TLI	SRMR	χ^2	RMSEA	CFI
K10Y	675.11***	0.22***	0.96	0.95	0.08	740.70***	0.18***	0.96	0.97	0.08	69.43***	0.04	< 0.01
K6	145.15***	0.11***	0.99	0.99	0.03	219.11***	0.07**	0.99	1.00	0.03	109.43***	0.04	< 0.01
K6Y	47.88***	0.09*	1.00	0.99	0.03	66.89**	0.06	1.00	1.00	0.03	26.09	0.03	< 0.01
Malaise	216.86***	0.07*	0.98	0.97	0.09	242.38***	0.06	0.97	0.97	0.10	32.45+	0.01	-0.01
MalaiseY	75.87*	0.05	0.98	0.98	0.09	85.19*	0.05	0.98	0.98	0.09	9.84	< 0.01	< 0.01
PHQ-9	404.77***	0.11***	0.98	0.97	0.05	435.21***	0.08***	0.98	0.99	0.05	113.32**	0.03	< 0.01
PHQ-9Y	159.80***	0.10***	0.98	0.98	0.05	176.69***	0.08***	0.99	0.99	0.05	34.92	0.02	0.01
GAD-7	250.13***	0.12***	0.99	0.99	0.04	293.52***	0.08***	0.99	1.00	0.04	81.09***	0.04	< 0.01
GAD-7Y	131.43***	0.14***	0.99	0.99	0.04	128.79***	0.10***	0.99	0.99	0.04	20.76	0.04	< 0.01

Note. PHQ = Patient health questionnaire. GAD = Generalized anxiety disorder scale. The letter Y denotes results reflecting only on the young adult (ages 18-39) subsample. RMSEA = Root mean square error of approximation. CFI = Comparative fit index. TLI = Tucker-Lewis index. SRMR = Standardized root mean squared residual. We interpreted RMSEA values up to .05 as indicating good fit, and values up to .08 as indicating adequate fit. In the cases of CFI and TLI, we interpreted values greater than .90 as indicating adequate, and those greater than .95 as indicating good model fit. Models where the loss of fit was less than 0.01 for CFI and 0.015 for RMSEA met the criteria for invariance. This type of strategy could not be implemented in scales with three or less items, since in those cases the configural model is just-identified at best,

leading to non-meaningful goodness-of-fit indices that cannot be compared to those from models with invariance constraints. Measurement invariance was not tested on the K10 scale in the full sample as adults aged 40+ did not endorse extreme categories of some items of this scale. $^{***}p \leq .001$. $^{**}p < .01$. $^{*}p < .05$. $^{+}p = .053$.

Scale Properties

Descriptive statistics of the sum scores of all scales are presented in Table 2. McDonald's ω_t suggests that internal consistency remained comparable after shortening the Malaise Inventory (Table 2). Independent samples *t*-tests revealed that females had worse mental health sum scores than males in all measures in both the overall sample (Table 3A), and in the young adult subsample (Table 3B). Younger adults' (ages 18-39) sum scores were also significantly worse on all measures compared to older adults (ages 40+, Table 3C). 2x2 ANOVAs further revealed a significant interaction across birth sex and age on the PHQ scales, and the same pattern of results was present in the case of the Malaise scales as well. These interactions showed that the difference between males and females was larger in the younger age groups then in the older age groups. The measures of effect size are presented along the results of the *t*-tests (Cohen's *d*; Table 3) and ANOVAs (η_p^2; Table 4), i.e., standardized measures of the magnitude of the observed effects. A Cohen's *d* value of (absolute) .02 is interpreted as small, that of (absolute) .05 is interpreted as medium, and that of (absolute) .08 or greater is interpreted as large; whereas a η_p^2 of .01 is considered small, that of .06 is interpreted as medium, and that of .14 or greater is interpreted as large, as suggested by commonly used guidelines (Cohen, 1988, 1992). The analyses including the 3-item version of the Malaise Inventory yielded a similar pattern of results as those including the 9-item Malaise Inventory (Tables 3, 4), suggesting that the shortening of the inventory was successful.

Correlations

The sum scores of the depression, anxiety, and psychological distress measures were strongly positively correlated with each other in the overall sample as well as the young adult subsample (Table 5). The correlations were strongest across the long and short versions of

each measure (*r*s = .87-.99). These results support the convergent and divergent validity of the scales.

Table 2. *Descriptive Statistics of the Scales Included*

	Full Sample				*Young Adult Subsample (Ages 18-39)*			
	M	*SD*	Range	McDonald's ω_t	*M*	*SD*	Range	McDonald's ω_t
K10	20.26	8.54	10-50	.94	23.34	9.25	10-50	.94
K6	12.24	5.44	6-30	.92	14.10	5.91	6-30	.92
Malaise9	3.03	2.50	0-9	.83	3.70	2.50	0-9	.81
Malaise3	1.40	1.13	0-3	.72	1.70	1.11	0-3	.69
PHQ-9	7.01	6.07	0-26	.91	8.81	6.62	0-25	.91
PHQ-2	1.62	1.67	0-6	-	2.07	1.81	0-6	-
GAD-7	5.42	5.24	0-21	.93	7.00	5.66	0-21	.93
GAD-2	1.65	1.76	0-6		2.16	1.90	0-6	-

Note. PHQ = Patient health questionnaire. GAD = Generalized anxiety disorder scale. Malaise9 refers to the 9-item Malaise Inventory, whereas Malaise3 refers to the 3-item version of the scale introduced in this manuscript. McDonald's ω_t indicates the internal consistency of the scales, where coefficients .70 or greater are considered adequate, with values closer to one indicating higher levels of internal consistency.

Table 3. *Mean Comparisons on All Sum Scores Across (A) Sexes in the Overall Sample, (B) Sexes in the Young Adult Sample, and (C) Age Groups*

A.	**Mean Comparisons Across Sexes in the Overall Sample**						
	Females: *M (SD)*	Males: *M (SD)*	*t*	*M* difference *(SE)*	95% CI of difference	*p*	Cohen's *d*
K10	21.00 (8.46)	19.40 (8.55)	-2.91	-1.59 (0.55)	[-2.67, -0.52]	.004	-.19
K6	12.63 (5.37)	11.78 (5.49)	-2.42	-0.85 (0.35)	[-1.53, -0.16]	.02	-.16
Malaise9	3.53 (2.51)	2.47 (2.38)	-6.76	-1.06 (0.16)	[-1.37, -0.76]	< .001	-.44
Malaise3	1.64 (1.11)	1.14 (1.10)	-6.94	-0.49 (0.07)	[-0.63, -0.36]	< .001	-.45
PHQ-9	7.77 (6.06)	6.13 (5.94)	-4.25	-1.64 (0.39)	[-2.40, -0.88]	< .001	-.27
PHQ-2	1.73 (1.71)	1.48 (1.60)	-2.31	-0.25 (0.11)	[-0.46, -0.04]	.02	-.15
GAD-7*	6.17 (5.32)	4.55 (5.02)	-4.83	-1.63 (0.33)	[-2.28, -0.97]	< .001	-.31
GAD-2*	1.94 (1.82)	1.32 (1.64)	-5.60	-0.63 (0.11)	[-0.85, -0.41]	< .001	-.36
B.	**Mean Comparisons Across Sexes in the Young Adult Subsample**						
	Females: *M (SD)*	Males: *M (SD)*	*t*	*M* difference *(SE)*	95% CI of difference	*p*	Cohen's *d*
K10	24.63 (9.29)	21.99 (9.12)	-2.78	-2.64 (0.95)	[-4.51, -0.77]	.01	-.29
K6	14.83 (5.90)	13.35 (5.83)	-2.45	-1.48 (0.61)	[-2.68, -0.29]	.02	-.25
Malaise9	4.41 (2.29)	2.95 (2.49)	-5.92	-1.46 (0.25)	[-1.95, -0.98]	< .001	-.61

Malaise3*	2.06 (0.99)	1.32 (1.10)	-6.92	-0.75 (0.11)	[-0.96, -0.54]	< .001	-.72
PHQ-9	10.12 (6.56)	7.44 (6.41)	-4.00	-2.68 (0.67)	[-4.00, -1.36]	< .001	-.41
PHQ-2	2.34 (1.84)	1.79 (1.73)	-2.96	-0.55 (0.19)	[-0.91, -0.18]	.003	-.31
GAD-7	8.17 (5.64)	5.78 (5.44)	-4.16	-2.38 (0.57)	[-3.51, -1.26]	< .001	-.43
GAD-2*	2.56 (1.94)	1.74 (1.77)	-4.28	-0.82 (0.19)	[-1.20, -0.44]	< .001	-.44

C. Mean Comparisons Across Younger (Ages 18-39) and Older Adults (Ages 40+)

	YA: *M (SD)*	OA: *M (SD)*	*t*	*M* difference *(SE)*	95% CI of difference	*p*	Cohen's *d*
K10*	23.34 (9.29)	18.31 (7.40)	8.87	5.03 (0.57)	[3.92, 6.15]	< .001	.62
K6*	14.10 (5.61)	11.05 (4.77)	8.43	3.05 (0.36)	[2.34, 3.76]	< .001	.58
Malaise9	3.70 (2.50)	2.60 (2.41)	6.78	1.09 (0.16)	[0.78, 1.41]	< .001	.45
Malaise3	1.70 (1.11)	1.22 (1.11)	6.49	0.48 (0.07)	[0.33, 0.62]	< .001	.43
PHQ-9*	8.81 (6.62)	5.88 (5.40)	7.19	2.93 (0.41)	[2.13, 3.73]	< .001	.50
PHQ-2*	2.07 (1.81)	1.34 (1.51)	6.50	0.73 (0.11)	[0.51, 0.95]	< .001	.45
GAD-7*	7.00 (5.66)	4.40 (4.69)	7.41	2.59 (0.35)	[1.91, 3.28]	< .001	.51
GAD-2*	2.16 (1.90)	1.33 (1.59)	7.05	0.83 (0.12)	[0.60, 1.06]	< .001	.48

Note. YA = Younger adults. OA = Older adults. CI = Confidence interval. PHQ = Patient health questionnaire. GAD = Generalized anxiety disorder scale. Malaise9 refers to the 9-item Malaise Inventory, whereas Malaise3 refers to the 3-item version of the scale introduced in this manuscript. The *t*-statistic is an index value that compares two means. The larger the absolute value of the *t*-statistic is, the more likely it is that the two means are statistically different. The effect size Cohen's *d* of (absolute) .02 is interpreted as small, that of (absolute) .05 is interpreted as medium, and that of (absolute) .08 or greater is interpreted as large.

* denotes analyses where Levene's test for equality of variances was significant, so the presented results are adjusted for equal variances not being assumed.

Table 4. *Interactions Between Birth Sex and Age on All Assessed Measures*

	F	*df*	*p*	η_p^2
K10	2.08	1	.15	.002
K6	1.89	1	.17	.002
Malaise9	3.84	1	.05	.004
Malaise3	7.79	1	.01	.01
PHQ-9	4.37	1	.04	.01
PHQ-2	4.77	1	.03	.01
GAD-7	3.06	1	.08	.003
GAD-2	1.66	1	.20	.002

Note. PHQ = Patient health questionnaire. GAD = Generalized anxiety disorder scale. Malaise9 refers to the 9-item Malaise Inventory, whereas Malaise3 refers to the 3-item version of the scale introduced in this manuscript. The effect size η_p^2 of .01 is considered small, that of .06 is interpreted as medium, and that of .14 or greater is interpreted as large.

Table 5. *Correlations Among the Sum Scores of the Scales in the Overall Sample and Young Adult Sample*

	1. K10	2. K6	3. Malaise9	4. Malaise3	5. PHQ-9	5. PHQ-9	7. GAD-7
1. K10	-						
2. K6	**.99***** / *.99****	-					
3. Malaise9	**.79***** / *.77****	**.77***** / *.74****	-				
4. Malaise3	**.67***** / *.63****	**.65***** / *.61****	**.88***** / *.87****	-			
5. PHQ-9	**.87***** / *.86****	**.86***** / *.86****	**.74***** / *.71****	**.63***** / *.60****	-		
6. PHQ-2	**.81***** / *.78****	**.81***** / *.79****	**.65***** / *.62****	**.57***** / *.52****	**.88***** / *.88****		
7. GAD-7	**.83***** / *.81****	**.81***** / *.79****	**.76***** / *.74****	**.67***** / *.65****	**.77***** / *.75****	**.67***** / *.63****	-
8. GAD-2	**.76***** / *.74****	**.75***** / *.73****	**.71***** / *.69****	**.62***** / *.59****	**.69***** / *.67****	**.63***** / *.59****	**.94***** / *.93****

Note. $^{***}p < .001$. PHQ = Patient health questionnaire. GAD = Generalized anxiety disorder scale. Malaise9 refers to the 9-item Malaise Inventory, whereas Malaise3 refers to the 3-item version of the scale introduced in this manuscript.
Correlations among the full sample are presented in **bold.**
Correlations among the young adult sample are presented in *italics.*

Discussion

In this manuscript, our aim was to test whether short versions of scales designed to assess psychological distress, depression, and anxiety may be used in a comparable manner to their full versions among a nonclinical sample of UK adults. Specifically, we investigated the K6 and K10 scales (Kessler et al., 2002), the PHQ-2 and PHQ-9 scales (Kroenke et al., 2001, 2003; Kroenke & Spitzer, 2002), and the GAD-2 and GAD-7 scales (Kroenke et al., 2007; Spitzer et al., 2006). We additionally tested the Malaise Inventory's 9-item version (Ploubidis et al., 2019; Rutter et al., 1970), which was the shortest available version of this measure, and developed a 3-item version. We relied on item response theory to do so.

The analyses revealed that the short scales were highly correlated with the full versions of the scales. In addition, the short scales performed comparably to their full versions across additional analyses. The results of these analyses corroborated previous findings suggesting that psychological distress, depression, and anxiety are more common among women than men and among younger rather than older adults (Brummer et al., 2014; Grenier et al., 2019; Jalnapurkar et al., 2018; Jorm et al., 2005). A significant interaction across age and birth sex on the Malaise and PHQ scales indicated that the differences in psychological distress and depression among males and females is more pronounced in the younger than the older age group.

Evidence of measurement invariance across age and sex groups was found in those cases in which this could be tested: The K6 scale, PHQ-9, GAD-7, and 9-item Malaise Inventory. Evidence of measurement invariance was also found across age groups among the subsample of young adults on the K10 scale. Although this could not be formally tested across the scales with 3 or less items, while we also refrained from testing it across the full sample on the K10 scale in attempt to ensure we do not lose information by merging across response categories, the lack of issues in at least one version of each scale may suggest that

the corresponding alternative versions may also have invariant measurement properties across the same groups.

The analyses presented here were limited by the nature of the short scales. Some of the analyses presented on the full scales could not be conducted on the short scales due to the number of items included. For example, measurement invariance testing cannot be implemented in scales with three or less items. In addition, the present analyses were conducted in a sample of UK adults. Based on the results presented here, we cannot be certain whether they would replicate in different cultural or national contexts. It should also be noted that the results presented here were collected from the general population. We thus cannot make any conclusions based on these results about the performance of the scales in clinical populations. More research is needed to examine the suitability of the short scales in clinical setting. For example, while they may be useful in detecting the presence of anxiety or depression, the assessment of the severity of such conditions may be more limited with the short measures than with the longer versions of the same measures.

It is important to note that administering the short versions of the measures in the present study as part of the long versions of each measure may bias the results somewhat. This is because the position of each item, as well as participants' responses to all other items of the long scales may influence their responses to the items of the short scales. Future studies should aim to examine the relationship between the short measures presented here, without any additional items included in each scale, and alternative related measures of affective disorders. Such analyses would serve to further confirm the validity of the measures.

Finally, as the shortening of the Malaise Inventory was primarily driven by the factor analyses, the final three items seem to capture anxiety more so than the longer scale's more diverse items pertaining to psychological distress. Nevertheless, the results of bivariate

correlations indicate that out of all measures included here, the 3-item Malaise was most strongly correlated to its 9-item version rather than to any of the other measures, including those specifically developed to assess anxiety. While future research may investigate this further (e.g., by testing the 3-item version of the Malaise Inventory along the 24-item full version of the scale), it is reasonable to conclude that the short version of the scale continues to assess a construct that overlaps with that assessed by the 9-item version.

Overall, these analyses indicate that the short scales may provide a good approximation of the full scales. This may be especially important in research or clinical settings where the time available for the completion of measures has strict constraints. In such cases, relying on the short scales tested here may save time and increase data quality whilst also maintaining the reliability and validity of the scales as close to their longer versions as possible.

References

Armstrong, B. G. (1998). Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occupational and Environmental Medicine*, *55*(10), 651–656. https://doi.org/10.1136/oem.55.10.651

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815–824. https://doi.org/10.1016/j.paid.2006.09.018

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures. *Spine*, *25*(24), 3186–3191.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Betz, N. E., & Turner, B. M. (2011). Using Item Response Theory and Adaptive Testing in Online Career Assessment. *Journal of Career Assessment*, *19*(3), 274–286. https://doi.org/10.1177/1069072710395534

Brummer, L., Stopa, L., & Bucks, R. (2014). The Influence of Age on Emotion Regulation Strategies and Psychological Distress. *Behavioural and Cognitive Psychotherapy*, *42*, 668–681.

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Cicchetti, D. V. (1990). Assessment of adaptive behavior in young children. In J. H. Johnson & J. Goldman (Eds.), *Developmental assessment in clinical child psychology: A handbook* (pp. 173–196).

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge Academic.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Connelly, R., & Platt, L. (2014). Cohort Profile: UK Millennium Cohort Study (MCS). *International Journal of Epidemiology*, *43*(6), 1719–1725. https://doi.org/10.1093/ije/dyu00

Elo, A.-L., Leppänen, A., & Jahkola, A. (2003). Validity of a single-item measure of stress symptoms. *Scandinavian Journal of Work, Environment & Health*, *29*(6), 444–451.

Fernández-Ballesteros, R. (2004). Self-report questionnaires. In S. N. Haynes & E. M. Heiby (Eds.), *Comprehensive handbook of psychological assessment: Vol. 3. Behavioural assessment* (pp. 194–221). Wiley.

Gardner, D. G., Cummings, L. L., Dunham, R. B., & Pierce, J. L. (1998). Single-Item Versus Multiple-Item Measurement Scales: An Empirical Comparison. *Educational and Psychological Measurement*, *58*(6), 898–915. https://doi.org/10.1177/0013164498058006003

Grenier, S., Payette, M., Gunther, B., Askari, S., Desjardins, F. F., Raymond, B., & Berbiche, D. (2019). Association of age and gender with anxiety disorders in older adults: A

systematic review and meta-analysis. *International Journal of Geriatric Psychiatry*, *34*(3), 397–407. https://doi.org/10.1002/gps.5035

Hayes, A. F., & Coutts, J. J. (2020). Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But…. *Communication Methods and Measures*, *14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453. https://doi.org/10.1037/1082-989X.3.4.424

Hunt, J. R., & White, E. (1998). Retaining and Tracking Cohort Study Members. *Epidemiologic Reviews*, *20*(1), 57–70. https://doi.org/10.1093/oxfordjournals.epirev.a017972

Jalnapurkar, I., Allen, M., & Pigott, T. (2018). Sex Differences in Anxiety Disorders: A Review. *HSOA Journal of Psychiatry, Depression & Anxiety*, *4*(12).

Jorm, A. F., Windsor, T. D., Dear, K. B. G., Anstey, K. J., Christensen, H., & Rodgers, B. (2005). Age group differences in psychological distress: The role of psychosocial risk factors that vary with age. *Psychological Medicine*, *35*(9), 1253–1263. https://doi.org/10.1017/S0033291705004976

Joshi, H., & Fitzsimons, E. (2016). The Millennium Cohort Study: The making of a multi-purpose resource for social science and policy. *Longitudinal and Life Course Studies*, *7*(4). https://doi.org/10.14301/llcs.v7i4.410

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S.-L. T., Walters, E. E., & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*, *32*(6), 959–976. https://doi.org/10.1017/S0033291702006074

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A New Depression Diagnostic and Severity Measure. *Psychiatric Annals*, *32*(9), 509–515. https://doi.org/10.3928/0048-5713-20020901-06

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2003). The Patient Health Questionnaire-2: Validity of a Two-Item Depression Screener. *Medical Care*, *41*(11), 1284–1292. https://doi.org/10.1097/01.MLR.0000093487.78664.3C

Kroenke, K., Spitzer, R. L., Williams, J. B. W., Monahan, P. O., & Löwe, B. (2007). Anxiety Disorders in Primary Care: Prevalence, Impairment, Comorbidity, and Detection. *Annals of Internal Medicine*, *146*(5), 317. https://doi.org/10.7326/0003-4819-146-5-200703060-00004

Levell, N. J. (2022). NHS outpatient secondary care: A time of challenges and opportunities. *Future Healthcare Journal*, *9*(2), 106–112. https://doi.org/10.7861/fhj.2022-0044

Little, T. D. (2013). *Longitudinal structural equation modeling*. The Guilford Press.

McDonald, R. P. (1999). *Test Theory: A Unified Treatment* (0 ed.). Psychology Press. https://doi.org/10.4324/9781410601087

Muthén, L. K., & Muthén, B. O. (1998). *MPlus User's Guide.* (8th ed.). Muthén & Muthén.

Paulhus, D. L., & Vazire, S. (2007). The Self-Report Method. In *Handbook of research methods in personality psychology* (pp. 224–239). Guilford.

Ploubidis, G. B., McElroy, E., & Moreira, H. C. (2019). A longitudinal examination of the measurement equivalence of mental health assessments in two British birth cohorts. *Longitudinal and Life Course Studies*, *10*(4), 471–489. https://doi.org/10.1332/175795919X15683588979486

Postmes, T., Haslam, S. A., & Jans, L. (2013). A single-item measure of social identification: Reliability, validity, and utility. *British Journal of Social Psychology*, *52*(4), 597–617. https://doi.org/10.1111/bjso.12006

Revelle, W., & Zinbarg, R. E. (2009). Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145–154. https://doi.org/10.1007/s11336-008-9102-z

Rolstad, S., Adler, J., & Rydén, A. (2011). Response Burden and Questionnaire Length: Is Shorter Better? A Review and Meta-analysis. *Value in Health*, *14*(8), 1101–1108. https://doi.org/10.1016/j.jval.2011.06.003

Rutter, M., Tizard, J., & Whitmore, K. (1970). *Education, health and behaviour*. Longman.

Samet, J. M., & Muñoz, A. (1998). Evolution of the Cohort Study. *Epidemiologic Reviews*, *20*(1), 1–14. https://doi.org/10.1093/oxfordjournals.epirev.a017964

Schooler, J. W., & Schreiber, C. A. (2004). Experience, meta-consciousness, and the paradox if introspection. *Journal of Consciousness Studies*, *11*(7–8), 17–39.

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092. https://doi.org/10.1001/archinte.166.10.1092

Steiger, J. H. (1990). Structural Model Evaluation and Modification: An Interval Estimation Approach. *Multivariate Behavioral Research*, *25*(2), 173–180. https://doi.org/10.1207/s15327906mbr2502_4

Stone, A. A., Turkkan, J. S., Bachrach, C. A., Jobe, J. B., Kurtzman, H. S., & Cain, V. S. (Eds.). (2000). *The science of self-report: Implications for research and practice*. Lawrence Erlbaum.

Torsvik, M., Johnsen, H. C., Lillebo, B., Reinaas, L. O., & Vaag, J. R. (2021). Has "The Ceiling" Rendered the Readiness for Interprofessional Learning Scale (RIPLS)

Outdated? *Journal of Multidisciplinary Healthcare*, *Volume 14*, 523–531. https://doi.org/10.2147/JMDH.S296418

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10. https://doi.org/10.1007/BF02291170

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00770

van de Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement Invariance. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01064

Wanous, J. P., & Reichers, A. E. (1996). Estimating the Reliability of a Single-Item Measure. *Psychological Reports*, *78*(2), 631–634. https://doi.org/10.2466/pr0.1996.78.2.631

Appendix A

Malaise Inventory – 9 items (Ploubidis et al., 2019; Rutter et al., 1970)

9-item scale:

1. Do you feel tired most of the time?
2. Do you feel miserable or depressed?
3. **Do you often get worried about things?**
4. Do you often get in a violent rage?
5. Do you often suddenly become scared for no good reason?
6. **Are you easily upset or irritated?**
7. Are you constantly keyed up and jittery?
8. **Does every little thing get on your nerves and wear you out?**
9. Does your heart often race like mad?

Responses: Yes/No

3-item version: Items 3, 6, 8

Conflict of interest
The authors disclose no conflict of interest.

Highlights

- Data quality may become poor as participants experience fatigue and boredom
- Short measures are thus vital for sound research
- We compared the performance of long and short versions of widely used measures
- These assessed depression, anxiety, and psychological distress
- The results suggest that short measures may perform similarly to longer scales