Psychological distress across adulthood: test-equating in three British birth cohorts

Hannah E. Jongsma[1*], Vanessa G. Moulton[1], George B. Ploubidis[1], Emily Gilbert[1], Marcus Richards[2], Praveetha Patalay[1,2]

[1] Centre for Longitudinal Studies, UCL Institute of Education

[2] MRC Unit of Lifelong Health and Ageing, UCL

[*] Hannah E. Jongsma is the corresponding author (h.jongsma@ucl.ac.uk)

Abstract

Background

Life-course and cross-cohort investigations of psychological distress are limited by differences in measures used across time within- and between- cohorts.

Aims

We aimed to examine adulthood distribution of symptoms and cross-cohort trends by test-equating mental health measures administered in the 1946, 1958 and 1970 British birth cohorts.

Methods

We used data from the above three birth cohorts (N=32,242) and an independently recruited calibration sample (n=5,800) where all measures of psychological distress that were used in at least one sweep of the cohorts were administered. We used two approaches to test-equating (equipercentile linking and multiple imputation) and two index-measured (General Health Questionnaire [GHQ]-12 and Malaise-9). We presented and compared means and prevalence of mental distress across adulthood in each cohort.

Results

While broad patterns of the shape of mental distress were similar across adulthood (inverse-Ushape) for all methods used, both test-equating method and index measure resulted in slightly different estimates, most notably for cross-cohort comparisons. Cross-cohort comparisons using GHQ-12 suggested that psychological distress is higher in younger cohorts, whereas using Malaise-9 there were inconsistent differences between cohorts. Sensitivity analysis (using incidents where both measures were simultaneously available in the cohorts) indicated that multiple imputation led to more accurate estimates compared to equipercentile linking.

Conclusion

When estimating life course trajectories of psychological distress we observe an inverse-U shaped trajectory across adulthood. Differences in point estimates between measures and methods do not allow for clear conclusions about consistent trends between cohorts.

Introduction

Despite the fact that common mental disorders are a leading cause of disease burden[1], with one in six adults in England meeting the threshold for a clinical diagnosis in 2014[2] our ability to assess psychological distress reliability across time, person and place is limited. Limitations to comparability result from a lack of 'gold standard' and plethora of instruments used as well as from differences in mode of administration and response options. Recently, studies have used Item Response Theory-based approaches[3,4], equipercentile linking[5,6], and multiple imputation[7] to investigate comparability of different mental health measures. Despite this recent interest in applying such test-equating methods to mental health measures, little is known about the effects of using different approaches on resulting estimates of distribution and above threshold prevalence.

A life-course and cross-cohort perspective

The adulthood distribution of psychological distress is expected to follow an inverse-U shape with symptoms increasing from early adulthood to mid-life and then decreasing from mid-life to old age[8,9], although there is limited evidence for other life-course distributions as well, including an increase in distress in the over-75s[10]. The limitations of relying on cross-sectional data across ages to determine the life-course shape of psychological distress are well recognised conflating of age and cohort effects [8,11,12]). One of the limitations to drawing stronger conclusions about the adulthood trajectory of symptoms in longitudinal data has been the use of different measures across time in population based cohort studies (for instance, in the 1946 birth cohort the Present State Examination (PSE) is used at age 36, the Psychiatric Symptom Frequency (PSF) scale at age 43, and the GHQ-28 at ages 53, 63 and 69). Similarly, cross-cohort comparisons are also limited by different cohorts using different measures. Where identical measures have been used to compare cohorts there is some evidence that mid-life psychological distress is higher in more recent mid-20th century cohorts[13], although other studies suggest recent cohorts have better mental health[11] or U-shaped cohort effects with psychological distress highest in the oldest and most recently born cohorts[14].

The present study

The availability of three successive national cohort studies with mental health measures through adulthood (the 1946, 1958, 1970 cohorts) makes possible the investigation of both life-course progression in the same individuals longitudinally and cross-cohort comparisons if measures can be successfully equated or harmonised. The aims of the present study are two-fold: to use two different test-equating approaches (equipercentile linking and multiple imputation) based on an external test-equating (hereafter calibration) sample with two index measures (Malaise-9 and General Health Questionnaire [GHQ]-12) to equate all adult psychological distress measures in the three British birth

cohorts and to compare the results of these different measures and methods on the estimation of the distribution and prevalence of psychological distress across the life-course and between cohorts.

Methods

Participants

We used data from four different datasets: a calibration sample, and the 1946, 1958 and 1970 British birth cohorts.

Calibration sample

Our calibration sample (target *n*=5,000) was recruited independently from the birth cohorts between 7 June and 18 July 2019. We used quota-sampling to ensure at least 1,000 participants in each of five age groups (23-30, 31-40, 41-50, 51-60, 61-70) and representativeness of the general population in terms of sex, ethnicity, country of residence (England, Scotland, Wales, Northern Ireland) and highest level of education achieved.

Cohorts

The MRC National Survey of Health and Development (1946 birth cohort) started as a maternity survey of 16,965 children born in one week in March in England, Scotland and Wales, then selected from this a social class-stratified sample of 5,362 individuals to follow over the life-course[15]. The National Child Development Study followed 17,415 children born in England, Scotland and Wales in a single week in March 1958[16]. The 1970 British Cohort Study started following 17,198 children born in England, Scotland and Wales in a single week in April[17].

The analysis sample for the three birth cohorts was everyone for whom mental health data was available for at least one survey sweep in adulthood (18+ years).

Main outcomes and measures

Our main outcome was psychological distress. We collected all measures of psychological distress that were ever used in at least one of the birth cohorts (GHQ-12 and –28[18] Malaise-9[19], Present State Examination [PSE][20], Psychiatric Symptom Frequency Scale [PSF][21]; Supplemental Table 1) and administered them to the calibration sample. Following a counterbalanced design approach[22], order of questionnaire administration was randomised to prevent order effects.

We used the GHQ-12 and Malaise-9 as our index measures for both test-equating methods. Using two index measures has the advantage of being able to assess robustness of the modelling procedure (each

serve as the other's sensitivity analysis). We collected and harmonised data on age in years, sex and highest level of education (none, GSCE or equivalent, A-levels or equivalent, degree or higher) to inform our multiple imputation models.

Statistical analysis

We first composed descriptive statistics of all mental health measures collected in the calibration sample, and estimated correlations between these measures. For our equipercentile linking approach to test-equating, we then cross-tabulated percentile rankings on the GHQ-12 or Malaise-9 with percentile rankings on remaining measures to determine equivalent scores[5,23]. Based on this, we first identified a threshold score on the remaining measures most closely corresponding to that of the Malaise-9 (≥4) and the GHQ-12 (≥12). We applied this calibrated score back to the existing measures in the birth cohorts at each sweep to estimate the prevalence of mental distress. Secondly, using our equipercentile ranking, we converted scores on other measures in all sweeps of the birth cohorts to GHQ-12 or Malaise-9 and estimated means, variance and above-threshold prevalence of mental distress. Details on how we meet the various assumptions associated with equipercentile linking[23] are found in the Supplemental Methods.

Separately, we used a multiple imputation approach to test-equating. Multiple imputation is a more robust extension of linear transformation-based test-equating approaches [7,24], able to take account of stochastic error and uncertainty around single imputation transformed estimates. We coded data on covariates and all psychological distress measures identically across all four datasets, with psychological distress measures coded at the scale-level per age-group. We appended the calibration sample to each of the cohort samples, separately for GHQ-12 and Malaise-9. Each dataset thus consisted of two samples, and (at least) two measures per age group. In at least one of these samples (the calibration sample) both measures were complete and data from this were used to impute values into the cohort sample (Supplemental Table 4). We used multiple imputation by fully-conditional specification using chained equations[25]. Analyses were conducted post-imputation, combining estimates across 50 imputed data sets using Rubin's rule[26]. We estimated means, standard deviations and above-threshold prevalence of mental distress across adulthood in the cohorts.

Sensitivity analysis

In both the 1958 (at age 42) and 1970 (at age 30) cohorts the GHQ-12 and Malaise-9 were jointly administered. We used this opportunity to assess comparability of prevalence yielding from the three methods described above to original estimates as an additional sensitivity analysis. For example, for the GHQ-12 calibration at age 30 in the 1970 cohort, we compared the prevalence yielding from the

equipercentile linking, calibrated cut-off and multiple imputation approaches to the prevalence derived from the original GHQ-12 measure.

Throughout this paper we present results for the GHQ-12 in the main manuscript, with results for the Malaise-9 in the Supplemental Materials.

Our analysis plan was pre-registered on the Open Science Framework and can be accessed here: https://osf.io/7uc4j. We used Stata 16 for all our analyses [27].

Ethics and consent

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. All procedures involving human subjects were approved by the UCL Institute of Education Research Ethics Committee (REC1210). Informed consent was obtained from all participants.

Results

We recruited 5,800 participants into the calibration sample, distributed across five age groups (Supplemental Table 5). The analysis sample for the 1946 cohort consisted of 3,689 participants (68.7% of full cohort); for the 1958 cohort this was 14,814 (85.0% of full cohort); and for the 1970 cohort this was 13,739 (79.9% of full cohort). Missing data in the calibration sample was low (highest: 5.1% on GHQ-28, Supplemental Table 6). Means and standard deviations per questionnaire can be found in Supplemental Table 5. Correlations between measures varied between 0.68 (between Malaise-9 and GHQ-12) and 0.91 (between GHQ-12 and GHQ-28, Supplemental Table 7).

Equipercentile linking

Scores calibrated to the GHQ-12 are detailed in Table 2, and calibrated means and standard deviations across the life-course in each of the three birth cohorts are detailed in Table 3 and Figure 1A. Corresponding calibrated threshold scores were 1 for Malaise-9 and GHQ-28, 3 for Malaise-24 and PSE, and 11 for the PSF (Table 1). Psychological distress peaked in the 1946 birth cohort at age 43 (mean 8.54, SD: 4.45) before declining to 5.20 (SD: 6.65) at age 69. In the 1958 birth cohort, psychological distress peaked at age 42 (mean: 7.26, SD: 6.23). In the 1970 birth cohort it peaked at age 26 (mean: 8.76, SD: 5.68). Across the life-course and across cohorts, distributions tended to be positively skewed (Supplemental Figure 1).

Distribution of above-threshold prevalence is detailed in Figure 2. For the calibrated cut-off scores, prevalence followed an inverse U-shape in the 1946 birth cohort, peaking at 35.9% at age 63, before declining to 27.3% at age 69. In the 1958 birth cohort, the shape was similar to the 1946 cohort and

the peak was observed at age 42 (39.5%). Prevalence of psychological distress was relatively stable across the 1970 birth cohort, peaking at age 26 at 46.5%. Using the equipercentile linking method (where total scores were calibrated before the cut-off was applied), prevalence of psychological distress peaked in the 1946 birth cohort at age 63 at 35.9%, before declining to 27.3% at age 69. In the 1958 birth cohort, prevalence peaked at age 42 at 23.1%, and in the 1970 birth cohort the peak at age 42 was 28.7%.

Multiple Imputation

Means and standard deviations of psychological distress across the life-course based on multiple imputation are detailed in Table 2 and Figure 1A. In the 1946 cohort, mean scores peaked at age 43 (mean 11.41, SD: 0.35) before declining to 9.66 (SD:0.19) at age 69. In the 1958 and 1970 birth cohorts, means were approximately similar across the life-course, varying from 11.13 (SD: 0.35) at age 33 to 11.44 (SD:0.20) at age 50 in the 1958 birth cohort and from 10.74 (SD: 0.04) at age 30 to 12.57 (SD:0.42) at age 26 in the 1970 birth cohort.

Corresponding above-threshold prevalence estimates of psychological distress are detailed in Figure 2. The peak in the 1946 birth cohort occurred at age 43 at 44.6%, before declining to 36.8% at age 69. In the 1958 birth cohort, prevalence declined from a peak at age 23 (45.1%) until age 42 (38.5%) before increasing again at age 50 (44.0%). In the 1970 birth cohort, prevalence was highest at age 26 (52.9%) and lowest at age 30 (34.5%).

Life-course mental health and cross-cohort comparisons

Although broad patterns in distribution and above threshold prevalence over adulthood appear similar across both index measures and the two test-equating methods used in this paper, point estimates differ substantially. For instance, when examining the mean scores and standard deviations at age 36 in the 1946 birth cohort, a three-fold variance is observed for the GHQ calibration (2.91 [SD: 4.25] for the equipercentile linking method, and 9.98 [SD: 0.38] for the imputation approach). Prevalence scores for this sweep varied from 2.7% in the equipercentile linking and calibrated threshold approaches using the GHQ-12 to 35.2% using a multiple imputation approach.

Regarding cross-cohort comparisons of the prevalence of psychological distress calibrated against the GHQ-12, all approaches suggest that prevalence is highest in the 1970 birth cohort, though there is considerable variation in point estimates. Calibration against the Malaise-9 using a multiple imputation approach suggests that prevalence in the 1946 birth cohort is higher than in the other two cohorts, whereas using both a calibrated cut-off and equipercentile linking approach suggest that prevalence is comparable across the three cohorts.

Comparison of the two index measures

Full results for the various methods of calibration against the Malaise-9 can be found in the Supplemental Results. Figure 1 details the mean scores across the adulthood sweeps in all three cohorts for both the equipercentile linking and multiple imputation methods, for both GHQ-12 (Figure 1A) and Malaise-9 (Figure 1B). There appear to be larger differences in means between both methods for the GHQ-12 compared with the Malaise-9, though this was not formally tested. Calibration against the GHQ-12 yielded higher prevalences than calibration against the Malaise-9 (Figure 2, Supplemental Figure 6).

Though broad patterns of psychological distress across the life-course were similar for GHQ-12 and Malaise-9 with prevalence highest in mid-life across the three cohorts, the curve was much flatter for the Malaise-9 across all methods (Supplemental Figure 6). Whereas using the GHQ-12 as an index measures seems to suggest slightly higher psychological distress in the younger cohort, using Malaise-9 suggests that psychological distress is lowest in the 1958 cohort.

Sensitivity analysis

Additional sensitivity analyses resulting from comparing both calibration methods to original prevalence estimates at age 42 in the 1958 cohort and age 30 in the 1970 cohort suggested more accurate estimates using the multiple imputation method with both measures (Supplemental Table 10). When calibrating against the GHQ-12, the equipercentile linking method underestimated mean scores and prevalence in both cohorts. The multiple imputation method was close to original estimates in both cohorts (1958 original mean: 11.06 [SD: 4.79], multiple imputation 11.19 [SD:0.05] and 1970 original mean: 10.67 [SD:4.52], multiple imputation: 10.74 [SD:0.04]). The calibrated cut-off methods yielded relatively similar prevalence to the original estimate in the 1958 cohort (original prevalence= 36.1%, calibrated cut-off: 38.5%), but overestimated prevalence in the 1970 cohort (original prevalence= 32.9%, calibrated cut-off: 46.5%).

Discussion

Whilst broad patterns of psychological distress remained similar across the adulthood, the equipercentile linking test-equating method yielded lower means and standard deviations across the life-course compared the multiple imputation approach. Whilst this held true across both index measures, differences appeared to be larger for the GHQ-12 compared with the Malaise-9. However, cross-cohort comparisons were more susceptible to methodological effects. In general, using the GHQ-12 as an index measure yielded higher prevalence estimates than using Malaise-9. Sensitivity analyses using study sweeps with both index measures suggested that multiple imputation approach lead to

more accurate mean and prevalence estimates, whereas the equipercentile approaches yielded under- or over-estimates.

Comparison with existing literature

The previously reported [8,9] inverse U-shape, across adulthood was observed in the 1946 birth cohort for calibration against the GHQ-12 and Malaise-9 for the calibrated-cut-off and equipercentile linking methods, though for both index measures the multiple imputation method showed a gradual decline in prevalence of psychological distress across the life-course. This pattern was less clearly observed in the 1958 and 1970 cohorts across both index measures and both calibration methods, likely due to the fact that these cohorts are still in middle age and prevalence of psychological distress has only started to decline marginally.

As per previous research comparing age 42 years sweep across the 1958 and 1970 cohorts[13], we find that based on most methods and measures the 1970 cohort has higher prevalence of psychological distress at most ages compared to the 1958. However, this is the first paper comparing these cohorts to the earlier 1946 born cohort and we can draw no clear conclusions about any trends also including this cohort. With GHQ-12 as index measure we see lower prevalence and with the Malaisee-9 see higher prevalence of psychological distress in the 1946 cohort. Previous studies in North America have found that some older and more recent cohorts have higher distress, demonstrating a U-shape in cohort effects[14], and other UK-based studies have observed lower prevalence in more recent cohorts [11]. It important to note that the birth cohorts included in the present study were born across just 24 years in mid-20th century Britain, and hence we cannot extrapolate findings to recent cohorts where higher distress is increasingly reported. Our sensitivity analysis using both sweeps where both GHQ-12 and Malaise-9 were administered could not provide insight into why clear conclusions about cross-cohort trends could not be drawn, although it seems to suggest that multiple imputation might be more reliable than equipercentile linking. However, even just focussing on the multiple imputation findings we see prevalence of psychological distress increasing in more recent cohorts using the GHQ-12, yet lowest distress in the 1958 cohort using the Malaise-9.

Strengths and limitations

These results should be interpreted in the light of the strengths and limitations inherent to this study. We used three nationally representative birth cohorts, and our calibration sample had sufficient coverage across the whole distribution of mental health and was broadly representative of the current general population of the United Kingdom in terms of age, sex, level of education and country of residence. Our study was methodologically robust and was designed to allow for assessing reliability across methods (in contrast to previous literature using one method and measure): we used two

different index measures and test-equating methods, enabling us to describe differences on the basis of these. This is in contrast to previous literature applying these methods which typically only use one method and measure[3,4,6,7] resulting in limited evaluation of the reliability of any findings based on these approaches.

However, there are also some limitations inherent to our study design. As we only used data from the United Kingdom, we are uncertain about the generalisability of our results to an international context. Mode of questionnaire administration differed between our calibration sample (all self-reported online) and the cohort samples (either self-report via a paper questionnaire or interviewer-administered), and this might have led to higher reporting of mental health symptoms in the calibration sample[28]. Finally, we are utilising responses today to equate responses given at a previous point in time, as far back as 1982. However, there appears to be no evidence that within-individuals and cross cohort interpretation of the Malaise-9 changes over time[29], and for the measures in the calibration measurement invariance analyses (Supplemental Table 3) indicate that younger and older respondents today answer these measures similarly, hence increasing confidence in the longitudinal comparisons made.

Interpretation

Whilst life-course patterns of psychological distress were similar across both index measures and test-equating methods, point estimates were not. When comparing methodologies, the imputation method yielded higher means and standard deviations than the equipercentile linking method, and sensitivity analyses indicate the former might be the less biased approach in this scenario. Whilst means are not directly comparable across index measures due to different scale ranges, prevalence estimates were higher using the GHQ-12.

These differences have little bearing on the longitudinal symptom profile (we confirmed an inverted U-shape over the life-course). There are two hypothetical explanations for this inverted U-shape: it might be artefactual because the instruments we use are poor at capturing important aspects of mental health in later life, or it might be a reflection of genuine better mental health in later life (either through a reduced perception of pressure through socioemotional selectivity or through eudaemonic processes) after a period of greater stress in mid-life (potentially reflecting the multiple stressors faced by many of childcare, career pressures and caring for elderly parents[30]) needs to be further understood. However, these differences do have substantial implications for cross-cohort comparisons. For instance, when examining means derived through the equipercentile method calibrated against the Malaise-9 (Figure 1), means are higher in the 1958 and 1970 cohort, and any mid-life peak is earlier in these cohorts. However, when using the same index measure and applying our multiple imputation-

based approach, the mean is highest in the 1946 birth cohort, and there is no discernible mid-life peak in the other two cohorts. This method- and measure- dependency leaves us unable to make strong conclusions about whether more recent generations experience poorer mental health.

It important to note that the Malaise 9 has less variance given the range from 0-9, compared to the GHQ-12 (range 0-36) and most of the other measures that we calibrated. We speculate that some of the discrepancies in the results we observe between these different measures might be due to differences in their scales (for instance a score of between 8 and 10 the GHQ-12 gives a score of 1 on the Malaise-9, Supplemental Table 7). If this is indeed an important part of the consideration, then test-equating between measures with similar ranges and variance is more likely to yield reliable estimates compared with measures with vastly different variances. This might also explain why the multiple imputation approach appears to be more reliable than the other two approaches as it does not try to superimpose substantially larger or smaller variance. An important implication of our findings is the need for formal statistical simulation studies to investigate the conditions where different calibration methods return unbiased results, especially if, as we have shown bias occurs when measures with large discrepancy with respect to their range and standard deviation are calibrated.

This methodological and measure-based heterogeneity in point estimates across the life-course and between cohorts calls into question the robustness of papers using one index measure only to estimate mental health outcomes and we would recommend where possible for studies using test-equating approaches to use two index measures. On the basis of our analyses, it would seem that multiple imputation-based test-equating methods are better suited to scenarios like this where an external calibration sample is used.

Conclusion

We used two different test-equating methods and two different index measures to calibrate psychological distress measures used in three British birth cohorts against an independently recruited calibration sample. Our subsequently derived mean scores and above cut-off prevalences showed heterogeneity across both measure and method. Whilst this had some implications for estimations of psychological distress across the life-course, we mainly observed an inverse-U shaped trajectory across adulthood. However, the method and measure had the most severe consequences for cross-cohort comparisons, where although there are indications that distress is higher is the 1970 compared to 1958 birth cohort, consistent trends across all three cohorts are not observed. We therefore should be cautious in interpreting studies that have relied on one method and one index measure only and we recommend that future studies using test-equating approaches to compare mental health across time-points or datasets use more than one measure to increase reliability of any findings.

Funding

The authors would like to acknowledge the Medical Research Council for ongoing funding of the 1946 birth cohort, and the Economic and Social Research Council for ongoing funding of the 1958 and 1970 cohorts. This particular project was funded by a grant from the Economic and Social Research Council (Grant number: ES/T00116X/1).

Conflicts of interest

All authors have no conflicts of interest to declare.

Author contributions

PP, GP and VGM formulated the research question, all authors were involved in design of the study. EG, PP and HEJ carried out the data collection. HEJ analysed the data and HEJ and PP prepared the manuscript. MR, EG, VGM and GP provided important critical revisions to the manuscript.

Data availability

Data from the 1946 birth cohort is available to bona fide researchers upon request to the NSHD Data Sharing Committee via a standard application procedure. Further details can be found at: http://www.nshd.mrc.ac.uk/data. doi: 10.5522/NSHD/Q101; doi: 10.5522/NSHD/Q102; doi: 10.5522/NSHD/Q103. Data from the 1958 birth cohort is publicly available via the UK Data Service via https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=2000032#!/abstract. Data from the 1970 birth cohort is publicly available from the UK Data Service via https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=200001. Data from the calibration sample will be made publicly available via the UK Data Service as soon as possible.

Bibliography

1. Whiteford HA, Degenhardt L, Rehm J, et al. Global burden of disease attributable to mental and substance use disorders: Findings from the Global Burden of Disease Study 2010. *Lancet*. 2013;382(9904):1575-1586. doi:10.1016/S0140-6736(13)61611-6

2. McManus S, Bebbington P, Jenkins R, Brugha T, eds. *Mental Health and Wellbeing in England: Adult Psychiatric Morbidity Survey 2014*. Leeds: NHS Digital; 2016.

3. Fischer HF, Tritt K, Klapp BF, Fliege H. How to compare scores from different depression scales: Equating the Patient Health Questionnaire (PHQ) and the ICD-10-Symptom Rating (ISR) using Item Response Theory. *Int J Methods Psychiatr Res*. 2011;20(4):203-214. doi:10.1002/mpr.350

4. Fischer HF, Wahl I, Fliege H, Klapp BF, Rose M. Impact of cross-calibration methods on the interpretation of a treatment comparison study using 2 depression scales. *Med Care*. 2012;50(4):320-326. doi:10.1097/MLR.0b013e31822945b4

5. Furukawa TA, Reijnders M, Kishimoto S, et al. Translating the BDI and BDI-II into the HAMD and vice versa with equipercentile linking. *Epidemiol Psychiatr Sci*. 2019:1-13.

6. Choi SW, Schalet B, Cook KF, Cella D. Establishing a common metric for depressive symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS Depression. *Psychol Assess*. 2014;26(2):513-527. doi:10.1037/a0035768

7. Sellers R, Warne N, Pickles A, Maughan B, Thapar A, Collishaw S. Cross-cohort change in adolescent outcomes for children with mental health problems. *J Child Psychol Psychiatry Allied Discip*. 2019;60(7):813-821. doi:10.1111/jcpp.13029

8. Spiers N, Brugha T, Bebbington P, McManus S, Jenkins R, Meltzer H. Age and birth cohort differences in depression in repeated cross-sectional surveys in England: The National Psychiatry Morbidity Surveys, 1993 to 2007. *Psychol Med*. 2012;42(10):2047-2055.

9. Bell A. Life-course and cohort trajectories of mental health in the UK, 1991-2008: a multilevel age-period-cohort analysis. *Soc Sci Med*. 2014;(120):21-30.

10. Jokela M, Batty G, Kivimäki M. Ageing and the prevalence and treatement of mental health problems. *Psychol Med*. 2013;43:2037-2045.

11. Thomson RM, Katikireddi SV. Mental health and the jilted generation: Using age-period-cohort analysis to assess differential trends in young people's mental health following the Great Recession and austerity in England. *Soc Sci Med*. 2018;214:133-143. doi:10.1016/j.socscimed.2018.08.034

12. Blanchflower D, Oswald A. Well-Being Over Time in Britain and the USA. *Natl Bur Econ Res*. 2000. doi:10.3386/w7487

13. Ploubidis GB, Sullivan A, Brown M, Goodman A. Psychological distress in mid-life: evidence from the 1958 and 1970 British birth cohorts. *Psychol Med*. 2017;47(2):291-303. doi:10.1017/S0033291716002464

14. Keyes KM, Nicholson R, Kinley J, et al. Age, Period, and Cohort Effects in Psychological Distress in the United States and Canada. *Am J Epidemiol*. 2014;179(10):1216-1227.

15. Wadsworth M, Kuh D, Richards M, Hardy R. Cohort profile: The 1946 National Birth Cohort (MRC National Survey of Health and Development). *Int J Epidemiol*. 2006;35(1):49-54. doi:10.1093/ije/dyi201

16. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study).

Int J Epidemiol. 2006;35(1):34-41. doi:10.1093/ije/dyi183

17. Elliott J, Shepherd P. Cohort profile: 1970 British Birth Cohort (BCS70). *Int J Epidemiol*. 2006;35(4):836-843. doi:10.1093/ije/dyl174

18. Goldberg D, Williams P. *A User's Guide to the General Health Questionnaire*. Windsor: NFER-Nelson; 1988.

19. Rutter M, Tizard J, Whitmore K. *Education, Health and Behaviour*. London: Longmans; 1970.

20. Wing J, Birley J, Cooper J, Graham P, AD I. Reliability of a Procedure for Measuring and Classifying "Present Psychiatric State." *Br J Psychiatry*. 1967;113(498):499-515.

21. Lindelow M, Hardy R, Rodgers B. Development of a scale to measure symptoms of anxiety and depression in the general UK population: the psychiatric symptom frequency scale. *J Epidemiol Community Heal*. 1997;51(5):549-557.

22. Chen F, Huang X, MacGregor D. Equating or linking: basic concepts and a case study. https://fliphtml5.com/xrgx/bfuj. Published 2009.

23. Kolen MJ, Brennan RL. *Test Equating, Scaling and Linking*. Third Edit. New York: Springer; 2014.

24. Lamprianou I. *An Investigation into the Test Equating Methods Used during 2006, and the Potential for Strengthening Their Validity and Reliability*.; 2007.

25. Little R, Rubin D. *Statistical Analyses with Missing Data*. 2nd Editio. New York: Wiley; 2002.

26. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*. 2011;30(4):377-399. doi:10.1002/sim.4067

27. StataCorp. Stata Statistical Software: Release 16. 2019.

28. Epstein JF, Barker PR, Kroutil LA. Mode Effects in Self-Reported Mental Health Data. *Public Opin Q*. 2001;65(4):529-549. doi:10.1086/323577

29. Ploubidis GB, McElroy E, Moreira HC. A longitudinal examination of the measurement equivalence of mental health assessments in two british birth cohorts. *Longit Life Course Stud*. 2019;10(4):471-489. doi:10.1332/175795919X15683588979486

30. Willis SL, Martin M, Rocke C. Longitudinal perspectives on midlife development: stability and change. *Eur J Ageing*. 2010;7:131-134. doi:10.1007/s10433-010-0162-4

Table 1: Calibrated scores and cut-offs against the GHQ-12

	Calibrated equivalent scores per questionnaire				
GHQ-12 score	**Malaise-9**	**Malaise-24**	**GHQ-28**	**PSE**	**PSF**
0-3	0	0	0	0	0
4					
5					
6					1,2
7		1		1	3-6
8		2		2	7-9
9	1	3		3	10,11,12
10		4		4	13-16
11	**2**	5	1	5	17,18
12		**6**	**2**,3	**6**	19-24*
13	3	7,8	4,5	7	25-30
14	4	9	6	8	31-34
15		10	7,8	9	35-38
16	5	11	9,10		39-42
17		12	11	10	43,44
18	6	13	12,13		45-47
19			14	11	48-50
20	7	14	15		51
21		15	16,17	12	52-54
22		16	18	13	55-58
23	8	17	19		59-61
24		18	20,21	14	62-64
25					65,66
26		19	22,23	15	67-70
27		20			71,72
28	9		24	16	73-75
29		21	25		76,77
30		22,23	26		78-80
31,32		24	27		81-85
33-35			28		86-90

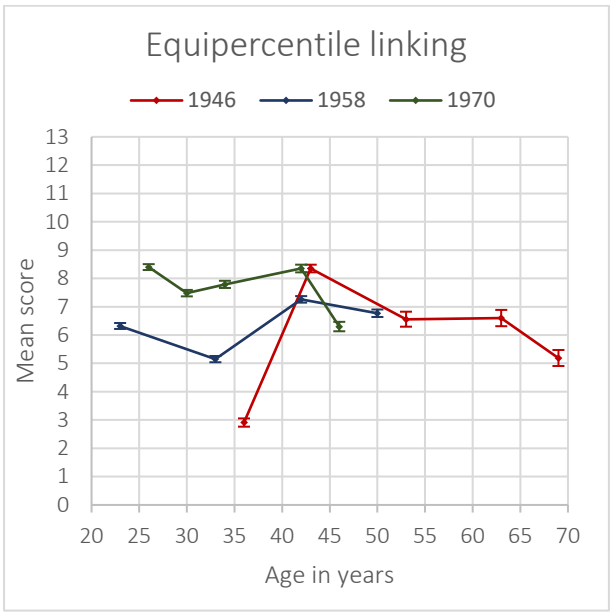
Equivalent cut-off scores are in **bold**

* Equivalent cut-off score for PSF is 20

Table 2: Means and standard deviations of measures calibrated against the GHQ-12 across adulthood

		Multiple Imputation	Equipercentile linking	Original measure (where available)
Cohort	Age	Mean GHQ-12 (SD)	Mean GHQ-12 (SD)	Mean GHQ-12 (SD)
1946	36	9.98 (0.38)	2.91 (4.25)	
	43	11.41 (0.35)	5.35 (3.89)	
	53	11.37 (0.22)	6.56 (7.28)	
	63	10.88 (0.19)	6.60 (6.87)	
	69	9.66 (0.19)	5.19 (6.66)	
1958	23	11.37 (0.35)	6.32 (5.97)	
	33	11.13 (0.35)	5.15 (5.97)	
	42	11.19 (0.05)	7.26 (6.23)	11.06 (4.79)
	50	11.44 (0.20)	6.77 (6.59)	
1970	26	12.57 (0.42)	8.40 (4.93)	
	30	10.74 (0.04)	7.48 (6.11)	10.67(4.52)
	34	11.45 (0.61)	7.79 (6.33)	
	42	12.23 (0.42)	8.35 (6.37)	
	46	12.30 (0.36)	7.63 (6.86)	

Figure 1A: Mean scores derived for calibration against the GHQ-12



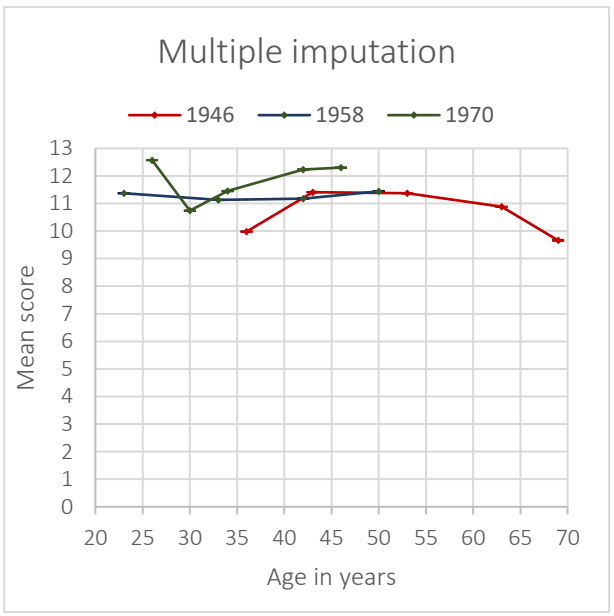
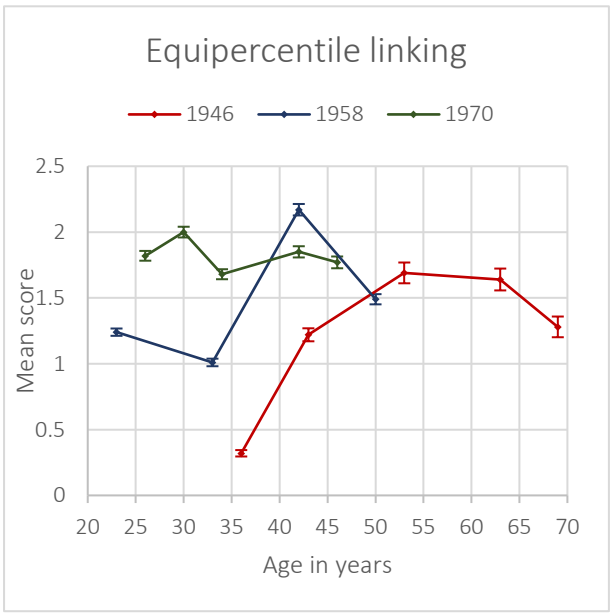


Figure 1B: Mean scores derived for calibration against the Malaise-9



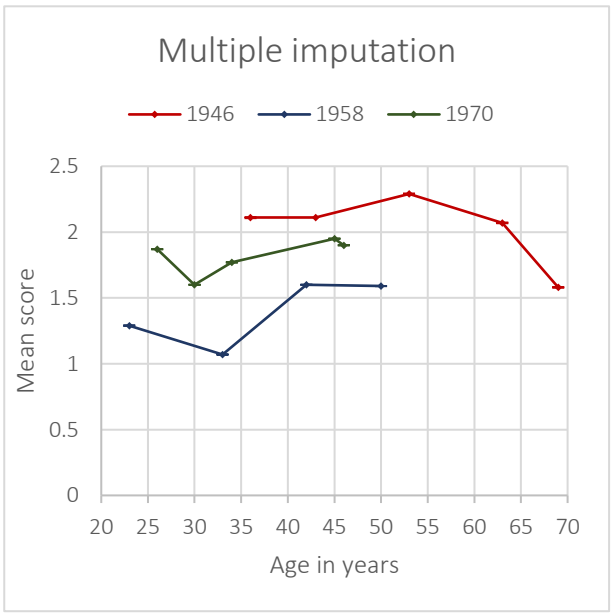


Figure 2:Prevalence of psychological distress across adulthood, calibrated against the GHQ-12

Figure 2A: The 1946 birth cohort

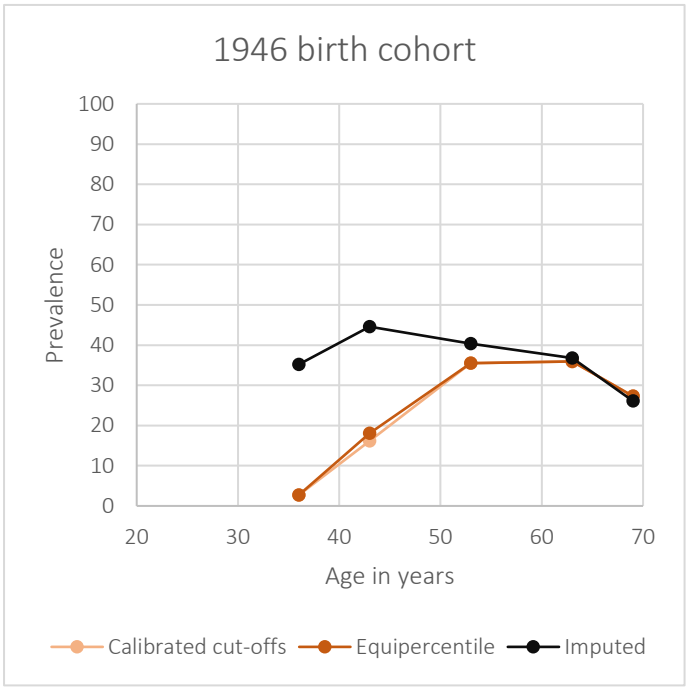


Figure 2B: the 1958 birth cohort

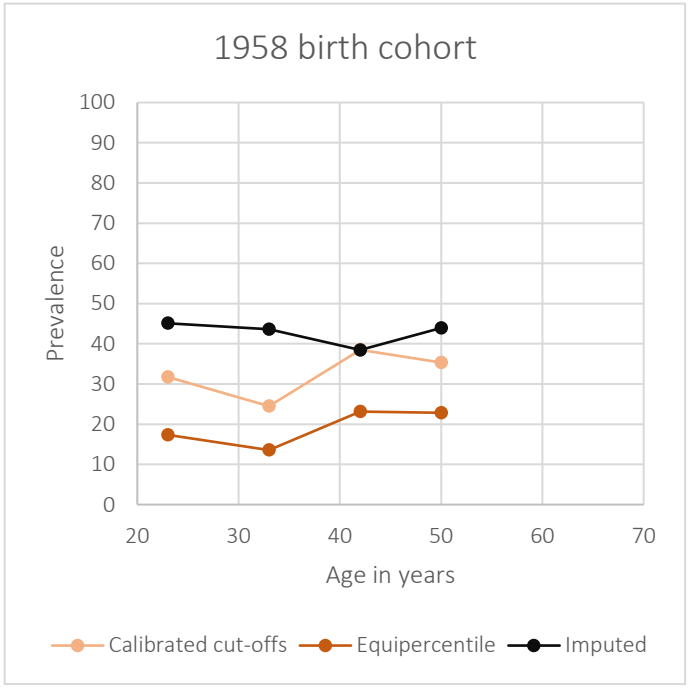


Figure 2C: The 1970 birth cohort