Title: A longitudinal examination of the measurement equivalence of mental health assessments in two British birth cohorts

***George B. Ploubidis** g.ploubidis@ucl.ac.uk, UCL Centre for Longitudinal Studies, Department of Social Science, University College London

Eoin McElroy UCL Centre for Longitudinal Studies, Department of Social Science, University College London

Hugo Cogo Moreira Federal University of Sao Paulo, UNIFESP

***Corresponding author**

Keywords: Mental health; Psychological distress; Measurement invariance; Longitudinal surveys; Measurement error; Malaise Inventory;

Abstract

Valid inference from the investigation of mental health relies - among others - on the assumption of no measurement error. However, it is well known that data from self-reported measures are likely to be biased by some process that is driven by the respondent's personality and/or circumstances. We capitalised on data available in two nationally representative birth cohorts, the National Child Development Study (NCDS – 1958 Birth cohort) and the 1970 British Cohort Study (BCS70 – 1970 Birth cohort) to formally test the longitudinal measurement equivalence of the nine item version of the Malaise Inventory, a measure of psychological distress. The inclusion of identical assessments of mental health in adulthood in both cohorts allowed us to evaluate their measurement properties and investigate whether the passage of time has differentially affected the interpretation of mental health assessments. To do so, we employed methods within the generalised latent variable measurement modeling framework and related extensions for formally testing measurement invariance. We found that the passage of 27 years in the 1958 birth cohort and 20 years in the 1970 birth cohort have not influenced how participants respond to the nine items that comprise the short version of the Malaise Inventory. The observed scalar invariance of the short version of the Malaise Inventory implies that potential sources of bias such as age effects, survey design, period effects, or cohort specific effects did not influence the way participants in the two cohorts respond to the symptoms described

in nine item version of the Malaise Inventory. Our results offer some reassurance for the extent to which self-reported mental health survey questions are affected by systematic sources of error.

Introduction

Major depression and anxiety disorders appear in the top 10 causes of global burden of disease (Vigo, Thornicroft, & Atun, 2016), with major depression also being the second leading cause of disability and a major contributor to the burden of suicide and ischemic heart disease with these impacts projected to increase as a result of population ageing (Murray et al., 2015). Monitoring the evolution of mental health over time within and across groups as well as generations and understanding its antecedents and consequences is therefore of major population health significance, also considering the undisputed huge costs to society, and to the economy, of poor mental health (Layard, 2013). Furthermore, inequalities due to both social causation and selection are well documented (A. Goodman, Joyce, & Smith, 2011; Power, Stansfeld, Matthews, Manor, & Hope, 2002; Stansfeld, Clark, Rodgers, Caldwell, & Power, 2011), but the idea that experiences of mental distress in adulthood are increasing across generations has not been much discussed, and yet if true as recent evidence suggests (Collishaw, Maughan, Natarajan, & Pickles, 2010; Ploubidis, Sullivan, Brown, & Goodman, 2017), is of major societal significance, especially considering the effects of population ageing.

Psychological assessment of populations in the form of self-report questionnaires play a key role in psychiatric research epidemiology and public health (Böhnke & Croudace, 2016). The abundance of longitudinal studies in the UK that collect repeated measures of mental health outcomes from the same individuals over time have resulted in a growing literature investigating mental health longitudinally (Clark, Rodgers, Caldwell, Power, & Stansfeld, 2007; Colman, Ploubidis, Wadsworth, Jones, & Croudace, 2007; Furnham & Cheng, 2015; Sacker & Wiggins, 2002). However, understanding the development of mental health symptoms over the life course, their antecedents and consequences and the investigation of secular mental health trends require comparable measures within and across cohorts, but with

some exceptions (Ploubidis et al., 2017) the measurement conditions that allow reliable comparisons have not been investigated.

Valid inference from the investigation of mental health relies - among others - on the assumption of no measurement error. In instances where the error mechanism is known, it can be modelled accordingly (Freedman, Fainberg, Kipnis, Midthune, & Carroll, 2004), but for the majority of substantive applications the error mechanism is not known. For example, it is well known that data from self-reported measures which constitute the main assessment method of mental health symptoms in longitudinal surveys are likely to be biased by some process that is driven by the respondent's personality and/or circumstances (Adams et al., 2005; Jurges, 2007; Singh-Manoux et al., 2006). Additional sources of error may arise from differences in the comprehension of items and in response tendencies such as response style and social desirability (Adams et al., 2005) which may vary over time, but also between groups, as well as generations (George B Ploubidis & Emily Grundy, 2009; Ploubidis et al., 2017). Within the context of longitudinal surveys, these sources of measurement error can also be thought of as Age, Period and Cohort effects (Keyes et al., 2014; Keyes, Utz, Robinson, & Li, 2010) on the way participants respond to mental health survey questions.

For example, as participants age, they may become more likely to endorse a mental health symptom (age effect), younger generations may have increased awareness of mental health symptomatology (cohort effect) or particular circumstances at the time of the interview may lead survey participants to underreport mental health symptoms, which could be thought of as a manifestation of social desirability (period effect). Furthermore, error can be introduced by differences in the measurement mode between sweeps of longitudinal surveys, or within the same mode by the preceding survey questions before the mental health items appear (Bowling, 2005). It's therefore possible that differences and/or trends observed within and across cohorts may reflect, at least partly, any of the above sources of bias and not true differences/trends in levels of psychological distress.

In order to obtain a meaningful mental health comparison between and within groups, as well as over time, the equivalence of mental health measures has to be

established. Within the generalized latent variable modelling framework, measurement equivalence is analogous to measurement invariance, a set of hypotheses stating that measurement model parameters should function without bias across groups or occasions (Meredith, 1993). Failing to ensure measurement equivalence in the groups of interest is analogous to differential measurement error(Armstrong, 1998), as group membership directly influences measurement error in the outcome. In this paper, we capitalise on the data available in two nationally representative birth cohorts, the National Child Development Study (NCDS) and the 1970 British Cohort Study (BCS70), to the best of our knowledge conducting a prospective investigation of the measurement properties of mental health assessments with the longest follow up to date. The inclusion of identical assessments of mental health in adulthood in both cohorts allowed us to evaluate their measurement properties and investigate whether the passage of time has differentially affected the interpretation of mental health assessments.

INSERT TABLES 1a AND 1b ABOUT HERE

Methods

Measures

Psychological distress was measured in both cohorts with the nine item version of the Malaise Inventory (Rodgers, Pickles, Power, Collishaw, & Maughan, 1999; Rutter, Tizard, & Whitmore, 1970). This version was developed at the Centre for Longitudinal Studies, by John Bynner, for use in multi-purpose questionnaires where it was difficult to find space for all 24 items of the original version of the Malaise Inventory. The nine item short form was constructed using the items with the highest loadings on the first principal factor in analyses of both NCDS and BCS70 to identify the sets of items that when aggregated best reflected the Malaise 24 item score (Johnson, Atkinson, & Rosenberg, 2015). In both surveys the Malaise items were assessed via written self-completion, either on paper or via computer. Descriptive statistics and the full wording of the items are presented in Tables 1a and 1b. The Malaise Inventory has been shown to have good psychometric properties (McGee,

Williams, & Silva, 1986) and has been used in general population studies as well as investigations of high risk groups (Furnham & Cheng, 2015). In both cohorts the nine-item version correlates highly with the 24-item version (r_{NCDS} = 0.91 at age 42 years and r_{BCS70} = 0.92 at age 30) (Ploubidis et al., 2017).

Data

The National Child Development Study (NCDS) follows the lives of 17,416 people born in England, Scotland and Wales in a single week of 1958 (Power & Elliott, 2006). It collects information on physical and educational development, economic circumstances, employment, family life, health behaviour, wellbeing, social participation and attitudes. Since the birth survey in 1958, there have been ten further 'sweeps' of all cohort members at ages 7, 11, 16, 23, 33, 42, 44, 46, 50 and 55. For this paper we use data from the age 23 survey (1981, n = 12357), the age 33 survey (1991, n = 11469), the age 42 survey 2000, n = 11419) and the age 50 survey (2008, n = 9790). The 1970 British Cohort Study (BCS70) follows the lives of 17,915 people born in England, Scotland and Wales in a single week of 1970 (Elliott & Shepherd, 2006). Over the course of cohort members' lives, the BCS70 has collected information on health, physical, educational and social development, and economic circumstances among other factors. Since the birth survey in 1970, there have been nine surveys (or 'waves') at ages 5, 10, 16, 26, 30, 34, 38, 42 and 46. In this paper we use data from five waves, at age 26 (1996, n = 9003), age 34 (2000, n = 11261), age 38 (2004, n = 9665), age 42 (2012, n = 9839) and age 46 (2016, n = 8581).

INSERT GRAPH 1 ABOUT HERE

Statistical modelling

We modelled the probability of response to the binary Malaise inventory items with a latent variable specification of a two parameter probit model (B. Muthén, 1984; Rabe-Hesketh & Skrondal, 2008). The model is presented in Graph 1, where $\boldsymbol{\theta}$

represents latent (unobserved) psychological distress, which is assumed to have a normal distribution N ~ (0, 1), λ is the factor loading that captures the strength of the association between the latent variable $\boldsymbol{\theta}$ and the observed items and τ is the threshold or "difficulty" parameter which quantifies the level of the latent continuum that underlies each item that needs to be reached for a response to an observed item to switch from 0 to 1.

Categorical/ binary observed indicators (y_{ij}) are related to continuous latent variables (θ_j) via a normal ogive response model, such that:

$$y_{ij} = \begin{cases} 1 & \text{if } y_{ij}^* > \tau_i \\ 0 & \text{otherwise} \end{cases}$$

where $y_{ij}^* = \beta_i + \lambda_i \theta_j + \varepsilon_{ij}$

for i=1,…, I_j (I_j being the number of observed indicators for latent variable j). We also assume that

$$\theta_j \sim N(0, \Psi),\ \varepsilon_{ij} \sim N(0,1),\ COV\left(\theta_j, \varepsilon_{ij}\right) = 0$$

where Ψ is a diagonal matrix and COV stands for covariance.

Model (1) can be equivalently expressed as:

$$\Pr\left(y_{ij} = 1 \mid \theta_j\right) = \Pr\left(y_{ij}^* > \tau_{\ i} \mid \theta_j\right) = \Phi\left(\beta_i + \lambda_i \theta_j\right)$$
$$\Phi^{-1}\Pr\left(y_{ij} = 1 \mid \theta\eta_j\right) = \beta_i + \lambda_i \eta \theta_j$$

Where $\Phi(\cdot)$ is the cumulative standard normal distribution and Φ^{-1} is the probit link

With this approach measurement error in the observed Malaise inventory items is controlled since the latent dimension $\boldsymbol{\theta}_j$ captures only the common variation in these and leaves out unique to each item variance (measurement error - $\boldsymbol{\varepsilon}_{ij}$) that is not due to latent psychological distress $\boldsymbol{\theta}$. However, additional sources of error may arise in between cohort comparisons from differences in the comprehension of items and in response tendencies which may vary by cohort so their distribution as sources of error cannot be assumed to be uniform between cohorts (Meredith, 1993). In order to obtain a meaningful comparison of psychological distress ($\boldsymbol{\theta}_j$) levels between NCDS and BCS70, or estimate trends within the two cohorts, the measurement parameters (τ and λ) of the model need to function equivalently within and between

the two cohorts. To empirically test this assumption we estimated a series of multigroup, two parameter probit models, where measurement model parameters (τ and λ) were not allowed to vary, either longitudinally (within cohort), as well as by sex and between cohorts in further specifications. Based on this unidimensional model we report the Scale Information Function (SIF) of the nine item Malaise inventory. The information function the inverse of standard error at each estimated latent score value and provides a graphical evaluation of the precision to which the nine items assess the unobserved (latent) psychological distress (Böhnke & Croudace, 2016). More information (higher values of the y axis) indicates higher measurement precision for a given latent score level (x axis).

The following criteria were used to determine model fit between the configural model where the unidimensional structure is identical between groups but the measurement parameters (τ and λ) are allowed to vary and the scalar model where structure and parameters are identical in all groups: A non-significant chi-square ($p > 0.05$), RMSEA < 0.06, and CFI and TLI > 0.95 (Hu & Bentler, 1999). To evaluate the measurement invariance of the Malaise Inventory across gender, time and cohort and considering that for large samples the chi-square (χ^2) difference between configural and scalar models almost always rejects the null, we report the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), and the Root Mean Square Error Approximation (RMSEA) as well as the difference (Δ) in model fit criteria for not rejecting the null hypothesis of invariance (ΔCFI < 0.01, ΔRMSEA < 0.015 and overlapping RMSEA CIs) (Cheung & Rensvold, 2002; Sass, 2011). All models were estimated with Mplus 8.0 with the Weighted Least Squares Mean and Variance (WLSMV) adjusted estimator, using the Delta parameterization (Bengt Muthén & Asparouhov, 2002) taking into account the dependency in the data due to the clustering of sweeps within cohort members with the Huber - White estimator as implemented in Mplus (Asparouhov, 2005).

INSERT TABLE 2 ABOUT HERE

Results

Longitudinal Within Cohorts Measurement Invariance

Preliminary analysis revealed that the established unidimensional structure of the Malaise Inventory fits the data in all waves of both cohorts. In Table 2 we present the fit criteria for all models and in Graphs 2,3,4 and 5 the measurement model parameters for the configural model where only the factorial structure is restricted to not vary across sweeps as well as the more restrictive scalar model, where model parameters were fixed across time and gender. Within NCDS we fitted configural and scalar invariance eight groups multigroup models, (4 waves * gender). In the configural model, the standardized factor loadings (λi) were all satisfactory and ranged between 0.544 and 0.898, whereas the item thresholds (τi) were mostly located as expected towards the high end of the latent psychological distress continuum (−0.107 to 1.561). As can be seen in Graphs 2 and 3, loadings and thresholds are very similar across waves and genders, indicating that the correlations between items and the level of the psychological distress needed to endorse an item remained relatively stable from age 23 to 50. The measurement model parameters similarity across waves of NCDS and gender was confirmed by the good fit of the model representing longitudinal scalar invariance between the eight groups (4 waves, 2 genders), CFI = 0.979, Tucker Lewis index (TLI) = 0.977, Root Mean Square Error of Approximation (RMSEA) = 0.039, 95% CI = 0.037 to 0.040, indicating the measurement equivalence of the Malaise Inventory in the four waves (ages 23 to 50) and both genders. The less restrictive model representing configural invariance (factor loadings and thresholds freely estimated) had only minimally better fit (CFI = 0.988, TLI = 0.984, RMSEA= 0.033, 95% CI 0.031 to 0.034). The difference (Δ) in model fit was within the criteria for not rejecting the null hypothesis of invariance (ΔCFI < 0.01, ΔRMSEA< 0.015).

INSERT GRAPHS 2 AND 3 ABOUT HERE

BCS70

Within BCS70 we estimated 10 group (5 waves * gender) multigroup configural and scalar models. Similarly to NCDS, in BCS70 the standardized factor loadings (λi) were all satisfactory and ranged between 0.555 and 0.865, whereas the item thresholds (τi) were mostly located as expected towards the high end of the latent psychological distress continuum (−0.393 to 1.858). As can be seen in Graphs 4 and 5, loadings and thresholds from the configural model were of similar magnitude across waves in both genders, indicating that the correlations between items and the level of the psychological distress needed to endorse an item remain relatively stable from age 26 to 42. The measurement model parameters similarity across waves of BCS70 and genders was confirmed by the good fit of the model representing longitudinal scalar invariance across both genders, CFI = 0.981, TLI = 0.979, RMSEA = 0.038, 95%CI = 0.037 to 0.040, indicating the measurement equivalence of the Malaise Inventory in the five waves and both genders. The less restrictive model representing configural invariance (factor loadings and thresholds freely estimated) had as expected better fit (CFI = 0.986, TLI = 0.981, RMSEA= 0.035, 95% CI 0.033 to 0.036. The difference (Δ) in model fit was within the criteria for not rejecting the null hypothesis of invariance for ΔCFI < 0.01 and ΔRMSEA< 0.015, but marginally not overlapping RMSEA Cis.

INSERT GRAPHS 4 AND 5 ABOUT HERE

Between cohorts and genders longitudinal invariance

The restrictive 12 group multigroup model representing scalar invariance between the two cohorts, over time (three overlapping waves per cohort at ages 23/26, 33/34 and 42) and gender had good fit, CFI = 0.976, TLI = 0.974, RMSEA = 0.040, 95%CI = 0.39 to 0.040. The less restrictive model representing configural invariance had better fit as expected (CFI = 0.986, TLI = 0.981, RMSEA= 0.035, 95%CI = 0.033 to 0.036). The difference (Δ) in model fit was within the criteria for not rejecting the null hypothesis of scalar invariance (ΔCFI < 0.01, ΔRMSEA< 0.015, with the exception of overlapping RMSEA CIs). In Table 3 we present factor loadings and thresholds

from the scalar 12 group scalar multigroup model where parameters were fixed to be identical longitudinally at ages 23/26 33/34 and 42, as well as between genders and cohorts. The standardized factor loadings (λi) were all satisfactory and ranged between 0.600 and 0.894, whereas the item thresholds (τi) were mostly located as expected towards the high end of the latent psychological distress continuum (0.558 to 2.137). In Graph 6 we present the SIF of the Malaise Inventory for both cohorts and genders. We see that highest precision (less measurement error) is achieved towards the high end of the latent trait, with the information function of the nine item version of the Malaise Inventory peaking well above the mean (0), indicating that its effective scoring range lies in moderate and high levels of psychological distress. Precision is higher in women and those born in 1970, with the highest precision of the Malaise Inventory being achieved in women at the age 26 sweep of BCS70.

INSERT TABLE 3 ABOUT HERE

Discussion

We investigated the longitudinal measurement equivalence of psychological distress related survey questions in two British birth cohorts. We found that the passage of 27 years in the 1958 birth cohort and 22 years in the 1970 birth cohort have not influenced how participants respond to the nine items that comprise the Malaise Inventory. The observed scalar invariance of the Malaise Inventory implies that potential sources of bias such as age effects, survey design, period effects, did not influence the way participants in the two cohorts respond to the symptoms described in the items of the Malaise Inventory. Furthermore we have shown that the Malaise Inventory functions equivalently between the two cohorts born 12 years apart in all ages across both genders, indicating the absence of cohort effects in the interpretation of the Malaise Inventory items. Further analysis (not presented here) revealed that scalar invariance holds between subgroups defined by parental social class at birth and education, indicating the longitudinal equivalence of the Malaise Inventory between these groups too. However, we note that despite the fact that latent variable measurement models control for measurement error and that scalar

invariance provides evidence against the occurrence of bias due to various sources of measurement error, bias due to unknown forms measurement error is still possible. This could take the form of a common bias within cohort sweep, gender or between cohorts, but due to the properties of the scalar model, this could bias the estimation of the actual value of latent psychological distress means, but not their differences between groups. This implies that the ranking of individuals on latent distress in all subgroups tested, and therefore regression coefficients between or within these groups in models that employ the Malaise Inventory as an exposure or outcome will not be biased from psychological distress related measurement error.

The finding that the passage of time does not differentially affect the interpretation of mental health related survey questions has implications for the validity of self-reported mental health assessment longitudinally and across generations. Indeed the passage of time had such a negligible impact that a very restrictive Rasch type model where the factor loadings (λ) are fixed to the same value across all items did fit the data well, implying that the simple sum of the nine Malaise Inventory items is a sufficient statistic (results not presented here, available from corresponding author). We have also shown that in both cohorts the region on which the Malaise Inventory provides most of its information is towards moderate and high levels of psychological distress. From a population mental health assessment perspective, this is a desirable feature as the interest when psychological distress is the focus of substantive research is for a measure to be able to effectively assess participants with moderate or high symptomatology. Interestingly, it has been shown that the intended measurement range is not always achieved in other measures of well-being and psychological distress (Abbott, Ploubidis, Huppert, Kuh, & Croudace, 2010; Böhnke & Croudace, 2016).

Considering that the public health burden of depression and anxiety – the major components of psychological distress - is estimated to continue to increase (Whiteford et al., 2013) future research to identify modifiable factors to shift the distribution of risk is needed. As such, our findings have the potential to inform the application of dynamic longitudinal models of mental health assessments in longitudinal surveys, studies that seek to identify high-risk life periods and facilitate

prevention and early detection, as well as cross-cohort comparisons of life-course profiles that can help us elucidate whether risk periods are stable or vary according to changing social and economic circumstances. Our findings have implications for survey design, as the longitudinal equivalence and measurement properties of the Malaise Inventory make its inclusion in future sweeps of the two cohorts, but also in other longitudinal surveys, desirable.

Strengths of this study include the availability of prospectively collected data with identically worded mental health survey questions in two population based and representative birth cohorts and the use of methods within the generalised latent variable measurement modeling framework (Rabe-Hesketh & Skrondal, 2008) for formally testing measurment invariance (Bengt Muthén & Asparouhov, 2017; G. B. Ploubidis & E. Grundy, 2009). To the best of our knowledge our study presents results from the longest follow up to date on the formal investigation of longitudinal measurement equivalence of mental health assessments. Limitations include potential bias due to selective attrition and the lack of generalisability of our findings in childhood and adolescence mental health assessments as well as to other longitudinal surveys that do not employ the Malaise Inventory. It's plausible to assume that participants that remain in the studies may interpret mental health items differently compared to those that attrit. However, extensive sensitivity analyses with Multiple Imputation using early life characteristics as auxiliary variables and Full Information Maximum Likelihood using only the variables employed in the multigroup models returned identical results to the ones presented here. Both methods return valid results under the Missing At Random assumption which is largely untestable (Carpenter & Kenward, 2012; Enders, 2010), but various specifications capitalizing on the richness of both birth cohorts that also relaxed the assumption of identical missing data generating mechanism in the two studies, returned similar results to the ones presented here, further reinforcing our interpretation (results available from corresponding author). With respect to generalisability of our findings to other – especially more recently born - cohorts that do not employ the Malaise Inventory, more work is needed to investigate age, cohort, perios and survey design effects on the interpretation of mental health items. However, futher sensitivity analysis using

maternal reports of externalising and internalising symptoms from the Strengths and Difficulties Questionnaire (Anna Goodman, Lamping, & Ploubidis, 2010) in the Millennium Cohort Study and the Avon Longitudinal Study of Parents And Children showed that the passage of time within and between these two cohorts hasn't affected how mothers report the mental health status of their children (results available from corresponding author).

We conclude that despite the presence of various potential sources of bias that may have affected the interpretation of mental health survey questions over time, we did not find evidence that this bias has occurred. On the contrary, measurement model parameters were very similar to the extent that a very restrictive scalar model indicating equivalence within and between cohorts, as well as across gender did fit the data. Our results offer some reassurance for the extent to which self-reported survey questions are affected by systematic sources of error, since despite the effects of age and secular changes that resulted in important differences between the two cohorts the Malaise Inventory was shown to function equivalently across and within cohorts. In future work we will investigate the extent to which non identically worded items that tap into the same symptom function equivalently over time and extend this work to other health phenotypes.

INSERT GRAPH 6 ABOUT HERE

Acknowledgements

George B. Ploubidis and Eoin McElroy were supported by the Economic and Social Research Council (ES/M008584/1 & ES/S000011/1)

References

Abbott, R. A., Ploubidis, G. B., Huppert, F. A., Kuh, D., & Croudace, T. J. (2010). An Evaluation of the Precision of Measurement of Ryff's Psychological Well-Being Scales in a Population Sample. *Social Indicators Research, 97*(3), 357-373. doi:DOI 10.1007/s11205-009-9506-x

Adams, S. A., Matthews, C. E., Ebbeling, C. B., Moore, C. G., Cunningham, J. E., Fulton, J., & Hebert, J. R. (2005). The effect of social desirability and social

approval on self-reports of physical activity. *American Journal of Epidemiology, 161*(4), 389-398. Retrieved from <Go to ISI>://000227145500011

Armstrong, B. G. (1998). Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occup Environ Med, 55*(10), 651-656. Retrieved from https://oem.bmj.com/content/oemed/55/10/651.full.pdf

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling, 12*(3), 411-434.

Böhnke, J. R., & Croudace, T. J. (2016). Calibrating well-being, quality of life and common mental disorder items: psychometric epidemiology in public mental health research. *The British Journal of Psychiatry, 209*(2), 162-168. doi:10.1192/bjp.bp.115.165530

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health, 27*(3), 281-291.

Carpenter, J., & Kenward, M. (2012). *Multiple imputation and its application*: John Wiley & Sons.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255.

Clark, C., Rodgers, B., Caldwell, T., Power, C., & Stansfeld, S. (2007). Childhood and adulthood psychological ill health as predictors of midlife affective and anxiety disorders: The 1958 british birth cohort. *Arch Gen Psychiatry, 64*(6), 668-678. doi:10.1001/archpsyc.64.6.668

Collishaw, S., Maughan, B., Natarajan, L., & Pickles, A. (2010). Trends in adolescent emotional problems in England: a comparison of two national cohorts twenty years apart. *Journal of Child Psychology and Psychiatry, 51*(8), 885-894.

Colman, I., Ploubidis, G. B., Wadsworth, M. E. J., Jones, P. B., & Croudace, T. J. (2007). A Longitudinal Typology of Symptoms of Depression and Anxiety Over the Life Course. *Biological Psychiatry, 62*(11), 1265-1271. doi:http://dx.doi.org/10.1016/j.biopsych.2007.05.012

Elliott, J., & Shepherd, P. (2006). Cohort Profile: 1970 British Birth Cohort (BCS70). *International Journal of Epidemiology, 35*(4), 836-843. doi:10.1093/ije/dyl174

Enders, C. E. (2010). *Applied missing data analysis*. New York: Guilford.

Freedman, L. S., Fainberg, V., Kipnis, V., Midthune, D., & Carroll, R. J. J. B. (2004). A new method for dealing with measurement error in explanatory variables of regression models. *60*(1), 172-181.

Furnham, A., & Cheng, H. (2015). The stability and change of malaise scores over 27 years: Findings from a nationally representative sample. *Personality and Individual Differences, 79*, 30-34. doi:http://dx.doi.org/10.1016/j.paid.2015.01.027

Goodman, A., Joyce, R., & Smith, J. P. (2011). The long shadow cast by childhood physical and mental problems on adult life. *Proc Natl Acad Sci U S A, 108*(15), 6032-6037. doi:10.1073/pnas.1016970108

Goodman, A., Lamping, D. L., & Ploubidis, G. B. J. J. o. a. c. p. (2010). When to use broader internalising and externalising subscales instead of the hypothesised five subscales on the Strengths and Difficulties Questionnaire (SDQ): data from British parents, teachers and children. *38*(8), 1179-1191.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55.

Johnson, J., Atkinson, M., & Rosenberg, R. (2015). *Millennium Cohort Study: Psychological, Developmental and Health Inventories*. Retrieved from Centre for Longitudinal Studies:

Jurges, H. (2007). True health vs response styles: Exploring cross-country differences in self-reported health. *Health Economics, 16*(2), 163-178. Retrieved from <Go to ISI>://000244233500005

Keyes, K. M., Nicholson, R., Kinley, J., Raposo, S., Stein, M. B., Goldner, E. M., & Sareen, J. (2014). Age, period, and cohort effects in psychological distress in the United States and Canada. *Am J Epidemiol, 179*(10), 1216-1227. doi:10.1093/aje/kwu029

Keyes, K. M., Utz, R. L., Robinson, W., & Li, G. (2010). What is a cohort effect? Comparison of three statistical methods for modeling cohort effects in obesity prevalence in the United States, 1971–2006. *Social Science & Medicine, 70*(7), 1100-1108. doi:doi:10.1016/j.socscimed.2009.12.018

Layard, R. (2013). Mental health: the new frontier for labour economics. *IZA Journal of Labor Policy, 2*(1), 1-16. Retrieved from http://www.izajolp.com/content/pdf/2193-9004-2-1.pdf

McGee, R., Williams, S., & Silva, P. A. (1986). An evaluation of the Malaise inventory. *Journal of Psychosomatic Research, 30*(2), 147-152. doi:http://dx.doi.org/10.1016/0022-3999(86)90044-9

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543. doi:10.1007/bf02294825

Murray, C. J. L., Barber, R. M., Foreman, K. J., Ozgoren, A. A., Abd-Allah, F., Abera, S. F., . . . Vos, T. (2015). Global, regional, and national disability-adjusted life years (DALYs) for 306 diseases and injuries and healthy life expectancy (HALE) for 188 countries, 1990–2013: quantifying the epidemiological transition. *The Lancet, 386*(10009), 2145-2191. doi:10.1016/S0140-6736(15)61340-X

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*(1), 115-132. doi:doi:10.1007/BF02294210

Muthén, B., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus web notes, 4*(5), 1-22.

Muthén, B., & Asparouhov, T. (2017). Recent Methods for the Study of Measurement Invariance With Many Groups: Alignment and Random Effects. *Sociological Methods & Research*, 0049124117701488.

Ploubidis, G. B., & Grundy, E. (2009). Later-life mental health in Europe: A country-level comparison. *Journals of Gerontology Series B: Psychological Sciences Social Sciences, 64*(5), 666-676.

Ploubidis, G. B., & Grundy, E. (2009). Later-Life Mental Health in Europe: A Country-Level Comparison. *Journals of Gerontology Series B-Psychological Sciences and Social Sciences, 64*(5), 666-676. doi:10.1093/geronb/gbp026

Ploubidis, G. B., Sullivan, A., Brown, M., & Goodman, A. (2017). Psychological distress in mid-life: evidence from the 1958 and 1970 British birth cohorts. *Psychol Med, 47*(2), 291-303. doi:10.1017/s0033291716002464

Power, C., & Elliott, J. (2006). Cohort profile: 1958 British Birth Cohort (National Child Development Study). *International Journal of Epidemiology, 35*(1), 34-41. doi:10.1093/ije/dyi183

Power, C., Stansfeld, S. A., Matthews, S., Manor, O., & Hope, S. (2002). Childhood and adulthood risk factors for socio-economic differentials in psychological distress: evidence from the 1958 British birth cohort. *Social Science & Medicine, 55*(11), 1989-2004. doi:10.1016/s0277-9536(01)00325-2

Rabe-Hesketh, S., & Skrondal, A. (2008). Classical latent variable models for medical research. *Statistical Methods in Medical Research, 17*(1), 5-32. doi:10.1177/0962280207081236|issn 0962-2802

Rodgers, B., Pickles, A., Power, C., Collishaw, S., & Maughan, B. (1999). Validity of the Malaise Inventory in general population samples. *Social psychiatry and psychiatric epidemiology, 34*(6), 333-341.

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

Sacker, A., & Wiggins, R. (2002). Age–period–cohort effects on inequalities in psychological distress, 1981–2000. *Psychological Medicine, 32*(06), 977-990. doi:doi:10.1017/S0033291702006013

Sass, D. A. (2011). Testing Measurement Invariance and Comparing Latent Factor Means Within a Confirmatory Factor Analysis Framework. *Journal of Psychoeducational Assessment, 29*(4), 347-363. doi:10.1177/0734282911406661

Singh-Manoux, A., Martikainen, P., Ferrie, J., Zins, M., Marmot, M., & Goldberg, M. (2006). What does self rated health measure? Results from the British Whitehall II and French Gazel cohort studies. *Journal of Epidemiology and Community Health, 60*(4), 364-372. Retrieved from <Go to ISI>://000235971500015

Stansfeld, S. A., Clark, C., Rodgers, B., Caldwell, T., & Power, C. (2011). Repeated exposure to socioeconomic disadvantage and health selection as life course pathways to mid-life depressive and anxiety disorders. *Soc Psychiatry Psychiatr Epidemiol, 46*(7), 549-558. doi:10.1007/s00127-010-0221-3

Vigo, D., Thornicroft, G., & Atun, R. (2016). Estimating the true global burden of mental illness. *The Lancet Psychiatry, 3*(2), 171-178. doi:10.1016/S2215-0366(15)00505-2

Whiteford, H. A., Degenhardt, L., Rehm, J., Baxter, A. J., Ferrari, A. J., Erskine, H. E., . . . Vos, T. (2013). Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet, 382*(9904), 1575-1586. doi:10.1016/s0140-6736(13)61611-6

Table 1a. Prevalence of psychological distress symptoms indicated by positive responses (item endorsement) in the nine items of the Malaise Inventory - NCDS

		Men 23	Men 33	Men 42	Men 50	Women 23	Women 33	Women 42	Women 50
Do you feel tired most of the time?	*ƒ*	742	1412	778	1363	1557	2268	1007	1634
	%	11.90%	22.60%	14.00%	23.70%	28.10%	39.50%	21.30%	33.30%
Do you often feel miserable or depressed?	*ƒ*	633	1166	479	823	918	1356	740	1099
	%	10.20%	18.70%	8.60%	14.30%	16.60%	23.60%	15.70%	22.40%
Do you often get worried about things?	*ƒ*	1897	3389	1356	2350	2103	3153	1665	2413
	%	30.40%	54.30%	24.40%	40.80%	38.00%	54.90%	35.20%	49.10%
Do you often get into a violent rage?	*ƒ*	284	446	186	339	231	316	138	127
	%	4.60%	7.10%	3.30%	5.90%	4.20%	5.50%	2.90%	2.60%
Do you often suddenly become scared for no reason?	*ƒ*	234	817	166	386	261	564	257	551
	%	3.80%	13.10%	3.00%	6.70%	4.70%	9.80%	5.40%	11.20%
Are you easily upset or irritated?	*ƒ*	903	1990	654	1193	901	1408	967	1467
	%	14.50%	31.90%	11.70%	20.70%	16.30%	24.50%	20.50%	29.90%
Are you constantly keyed up and jittery?	*ƒ*	205	276	197	254	297	360	306	437
	%	3.30%	4.40%	3.50%	4.40%	5.40%	6.30%	6.50%	8.90%
Does every little thing get on your nerves?	*ƒ*	89	209	96	216	196	325	274	485
	%	1.40%	3.30%	1.70%	3.70%	3.50%	5.70%	5.80%	9.90%
Does your heart often race like mad?	*ƒ*	373	514	243	408	328	604	287	534
	%	6.00%	8.20%	4.40%	7.10%	5.90%	10.50%	6.10%	10.90%

Table 1b Prevalence of psychological distress symptoms indicated by positive responses (item endorsement) in the nine items of the Malaise Inventory – BCS70

		Men 26	Men 34	Men 38	Men 42	Men 46	Women 26	Women 34	Women 38	Women 42	Women 46
Do you feel tired most of the time?	*f*	1173	2006	1560	2318	1167	1562	2167	1403	2058	1711
	%	28.90%	41.20%	29.00%	40.50%	30.92%	34.00%	43.30%	34.60%	45.20%	41.83%
Do you often feel miserable or depressed?	*f*	707	1254	842	1237	741	705	985	810	1009	923
	%	17.50%	25.90%	15.60%	21.60%	19.69%	15.30%	19.70%	20.00%	22.20%	23.56%
Do you often get worried about things?	*f*	1789	3178	2067	3187	1381	1810	2757	1860	2767	2176
	%	44.10%	65.30%	38.40%	55.70%	36.66%	39.40%	55.10%	45.90%	60.90%	53.20%
Do you often get into a violent rage?	*f*	314	402	334	335	167	225	171	157	131	139
	%	7.70%	8.30%	6.20%	5.90%	4.44%	4.90%	3.40%	3.90%	2.90%	3.39%
Do you often suddenly become scared for no reason?	*f*	231	569	342	560	241	261	517	275	490	497
	%	5.70%	11.70%	6.30%	9.80%	6.40%	5.70%	10.30%	6.80%	10.80%	12.15%
Are you easily upset or irritated?	*f*	834	1852	1023	1590	1021	1037	1648	1099	1557	1354
	%	20.60%	38.10%	19.00%	27.80%	27.18%	22.60%	32.90%	27.20%	34.20%	33.10%
Are you constantly keyed up and jittery?	*f*	202	241	282	270	304	317	385	315	371	356
	%	5.00%	5.00%	5.20%	4.70%	8.09%	6.90%	7.70%	7.80%	8.20%	8.70%
Does every little thing get on your nerves?	*f*	148	289	189	275	402	282	467	431	528	478
	%	3.60%	5.90%	3.50%	4.80%	10.68%	6.10%	9.30%	10.70%	11.60%	11.70%
Does your heart often race like mad?	*f*	342	478	417	475	317	339	463	345	507	500
	%	8.40%	9.80%	7.70%	8.30%	8.42%	7.40%	9.30%	8.50%	11.20%	12.23%

Table 2: Goodness of fit criteria

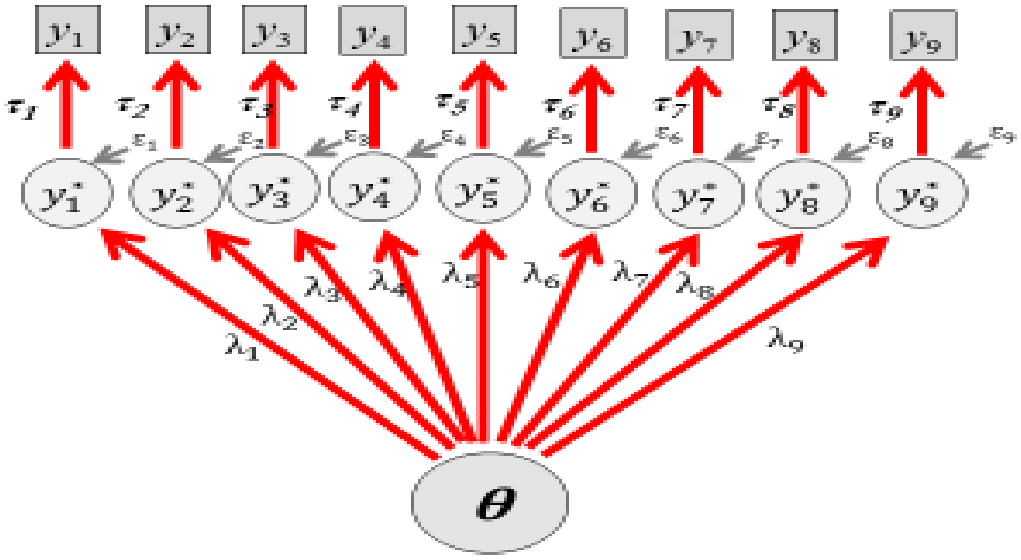
		Chi-square (d.f.)	RMSEA	CFI	TLI	ΔRMSEA	ΔCFI	ΔTLI
***NCDS 23, 33, 42, 50**	Configural	1500.343 (216)	0.033 (0.031 to 0.034	0.988	0.984			
	Scalar	2482.153 (265)	0.039 (0.037 to 0.040)	0.979	0.977	0.006	0.009	0.007
****BCS70 26, 30, 34, 42, 46**	Configural	2169.417 (270)	0.039 (0.038 to 0.041)	0.986	0.982			
	Scalar	2815.072 (333)	0.040 (0.039 to 0.042)	0.982	0.981	0.001	0.004	0.001
*****NCDS & BCS70 23/26, 33/34, 42**	Configural	2354.059 (324)	0.035 (0.033 to 0.036)	0.986	0.981			
	Scalar	3774.957 (401)	0.040 (0.039 to 0.041)	0.976	0.974	0.005	0.010	0.007

***Eight independent groups multigroup models (4 waves, gender)
****Ten independent groups multigroup models (5 waves, gender)
***** Twelve independent groups multigroup models (3 waves, 2 cohorts, gender)

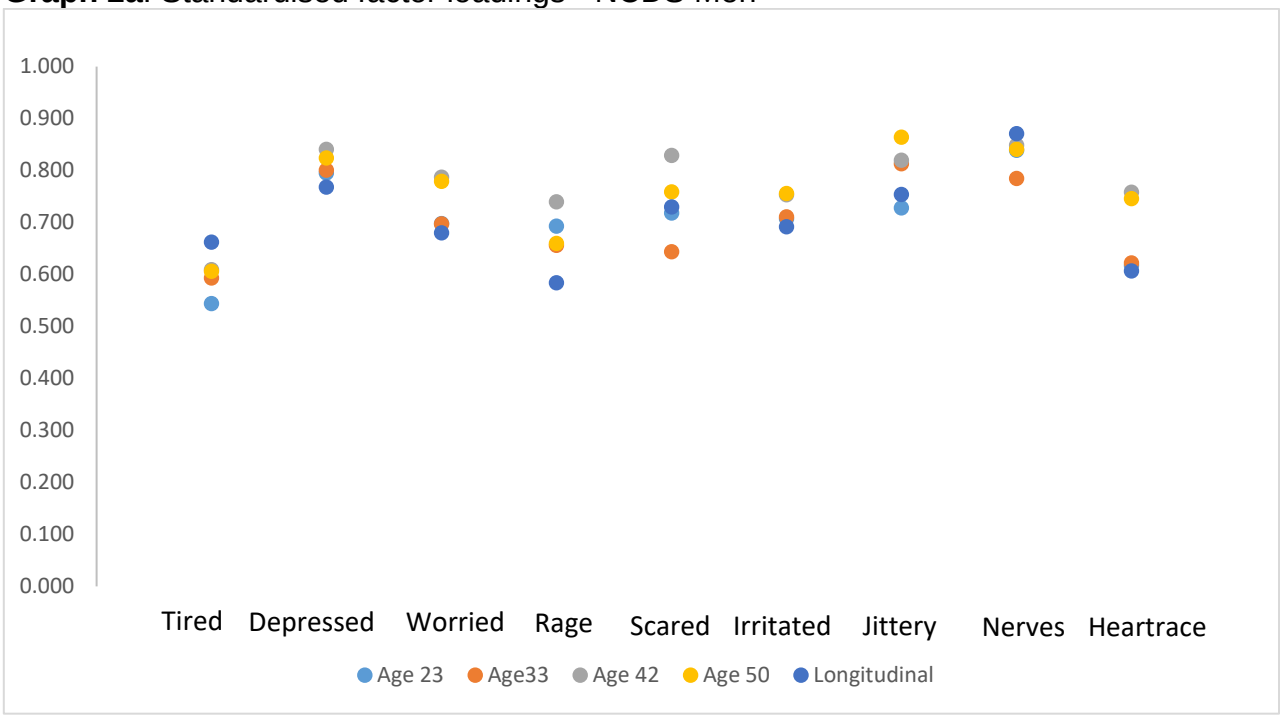
Table 3. Within (age), between (generations) cohorts and genders longitudinal invariance: standardized factor loadings, thresholds and 95% confidence intervals from a twelve independent groups (3 waves, 2 cohorts, gender) multigroup scalar invariance model

	Loading		Threshold	
Do you feel tired most of the time?	0.690	0.662 to 0.718	1.031	0.992 to 1.069
Do you often feel miserable or depressed?	0.749	0.726 to 0.772	1.321	1.282 to 1.360
Do you often get worried about things?	0.671	0.641 to 0.701	0.558	0.526 to 0.591
Do you often get into a violent rage?	0.602	0.574 to 0.630	1.762	1.711 to 1.814
Do you often suddenly become scared for no reason?	0.722	0.691 to 0.753	1.776	1.722 to 1.830
Are you easily upset or irritated?	0.689	0.665 to 0.713	1.077	1.042 to 1.111
Are you constantly keyed up and jittery?	0.740	0.708 to 0.771	1.828	1.772 to 1.885
Does every little thing get on your nerves?	0.882	0.847 to 0.917	2.137	2.062 to 2.212
Does your heart often race like mad?	0.615	0.588 to 0.643	1.558	1.510 to 1.605

Graph 1. Unidimensional measurement model of the Malaise Inventory

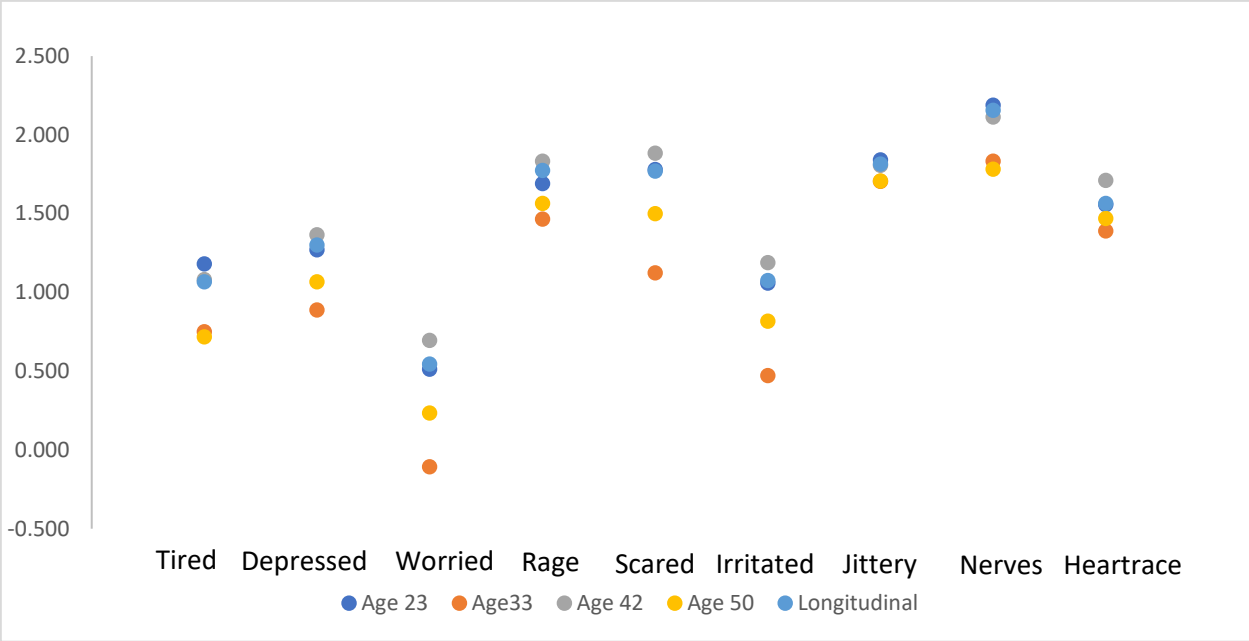


Graph 2a. Standardised factor loadings - NCDS Men

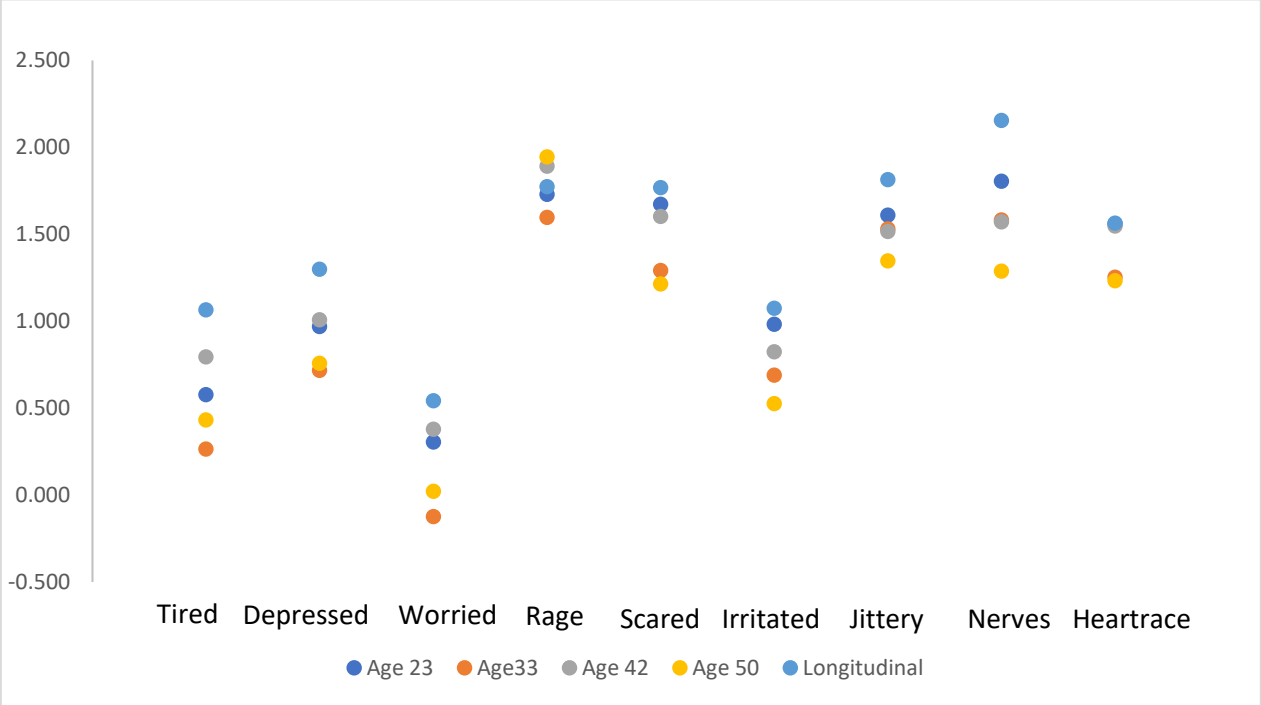


Graph 2b. Standardised factor loadings - NCDS Women

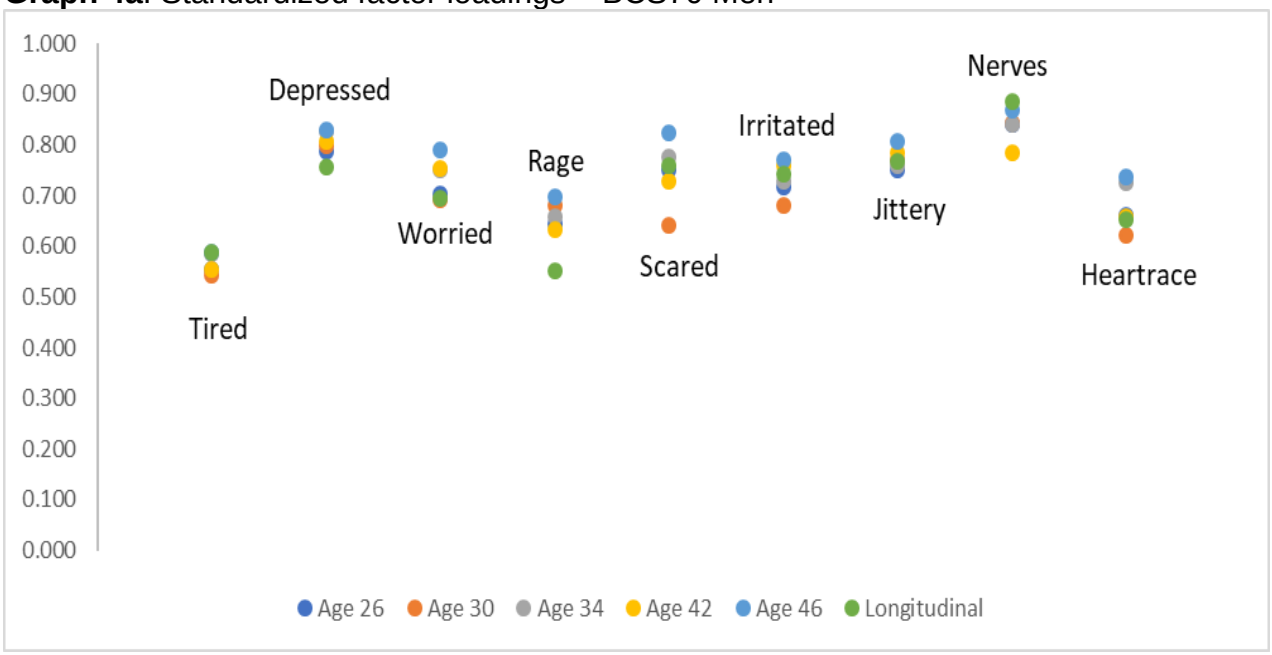
Graph 3a. Standardized thresholds – NCDS Men



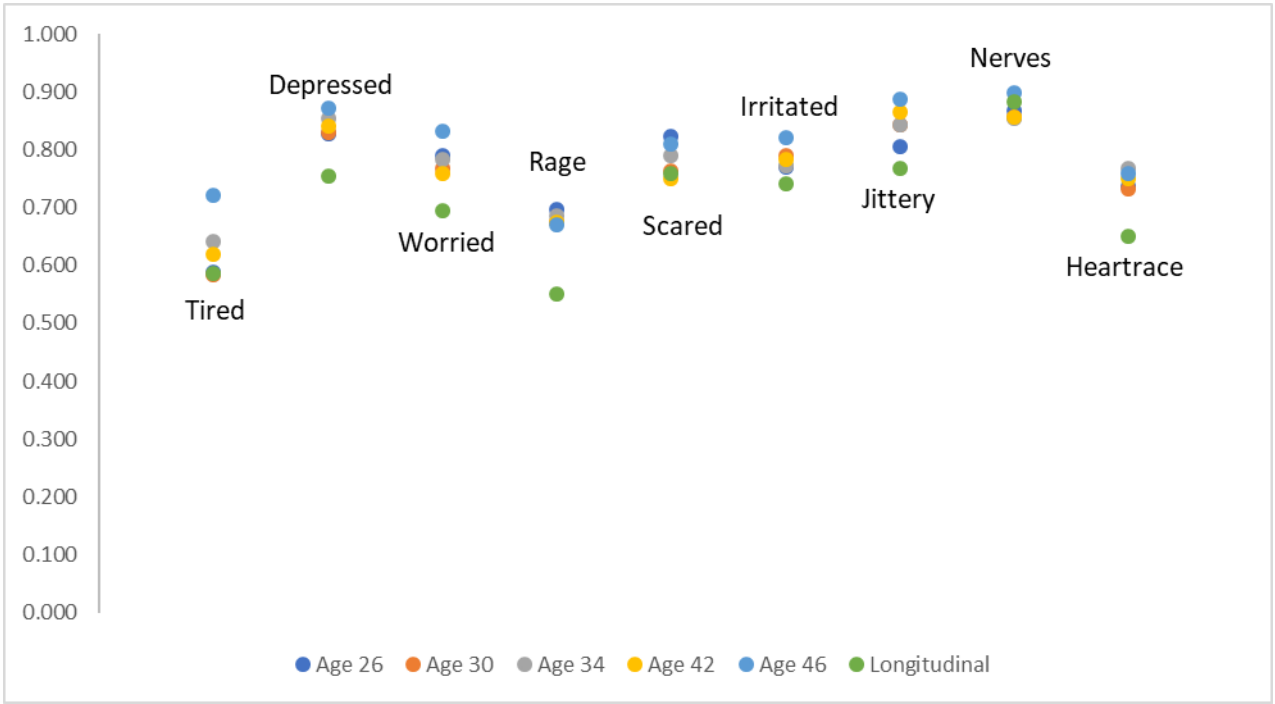
Graph 3b. Standardized thresholds – NCDS Women



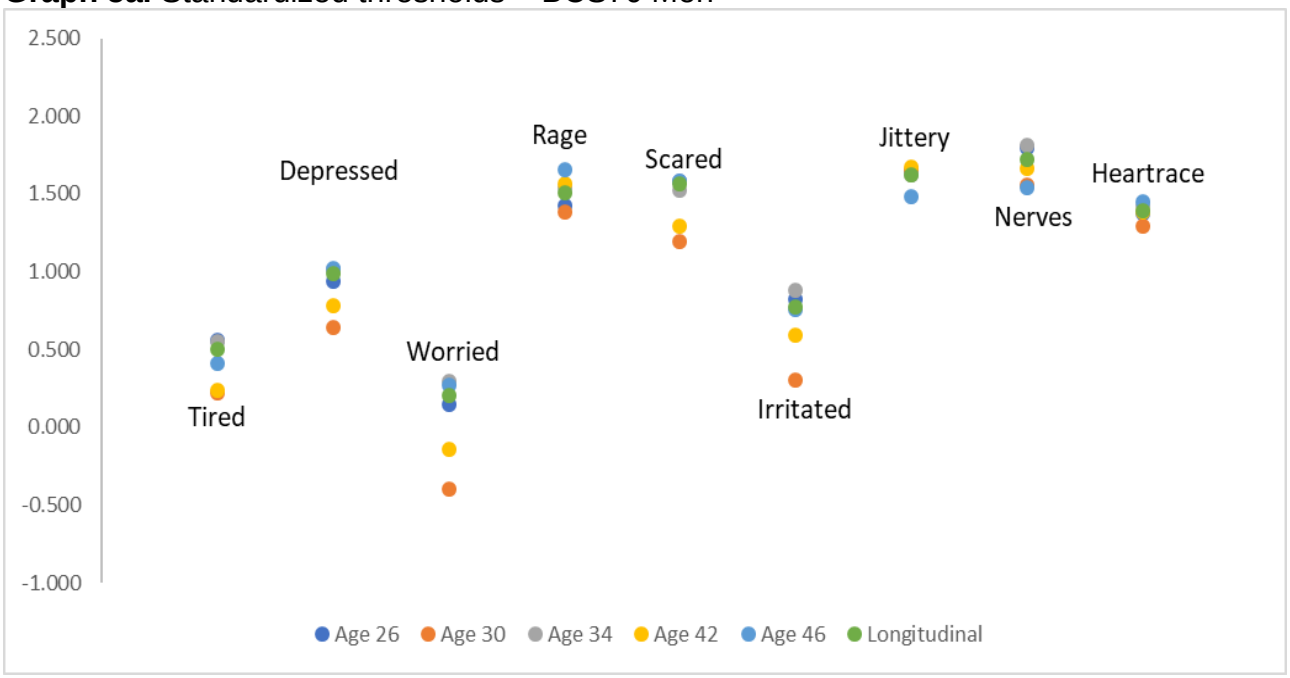
Graph 4a. Standardized factor loadings – BCS70 Men



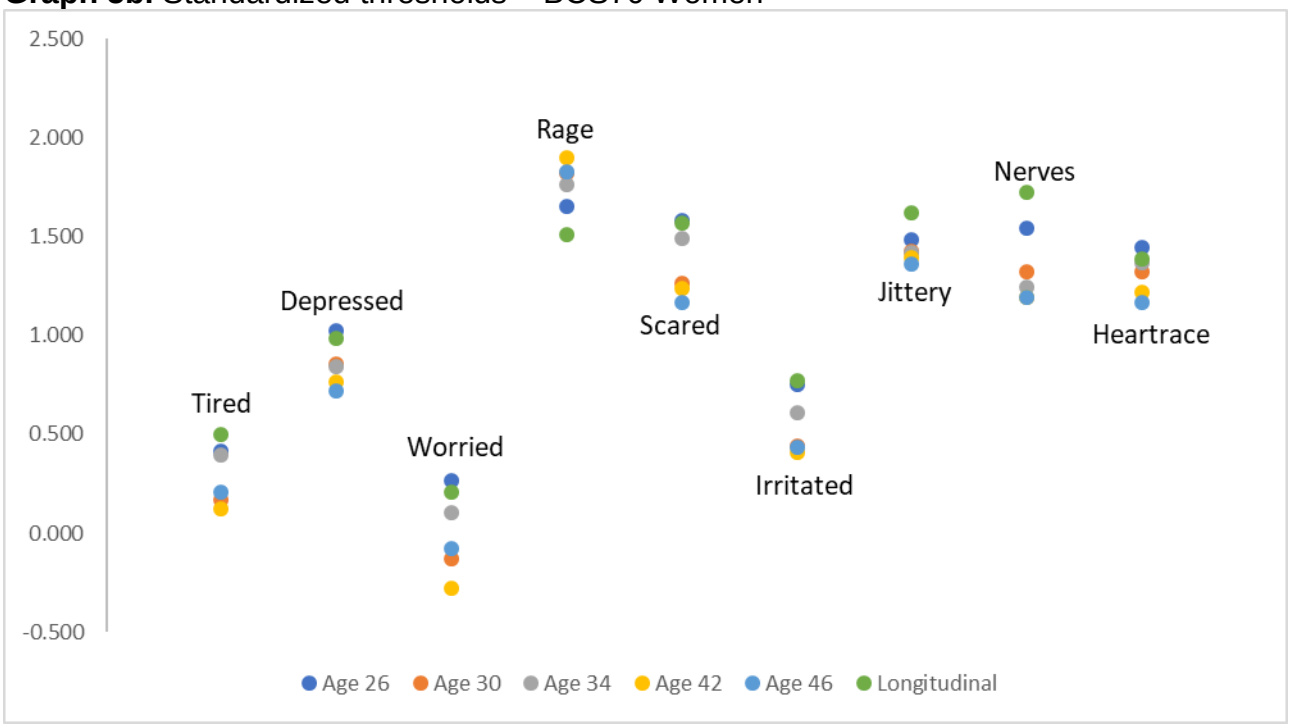
Graph 4b. Standardized factor loadings – BCS70 Women



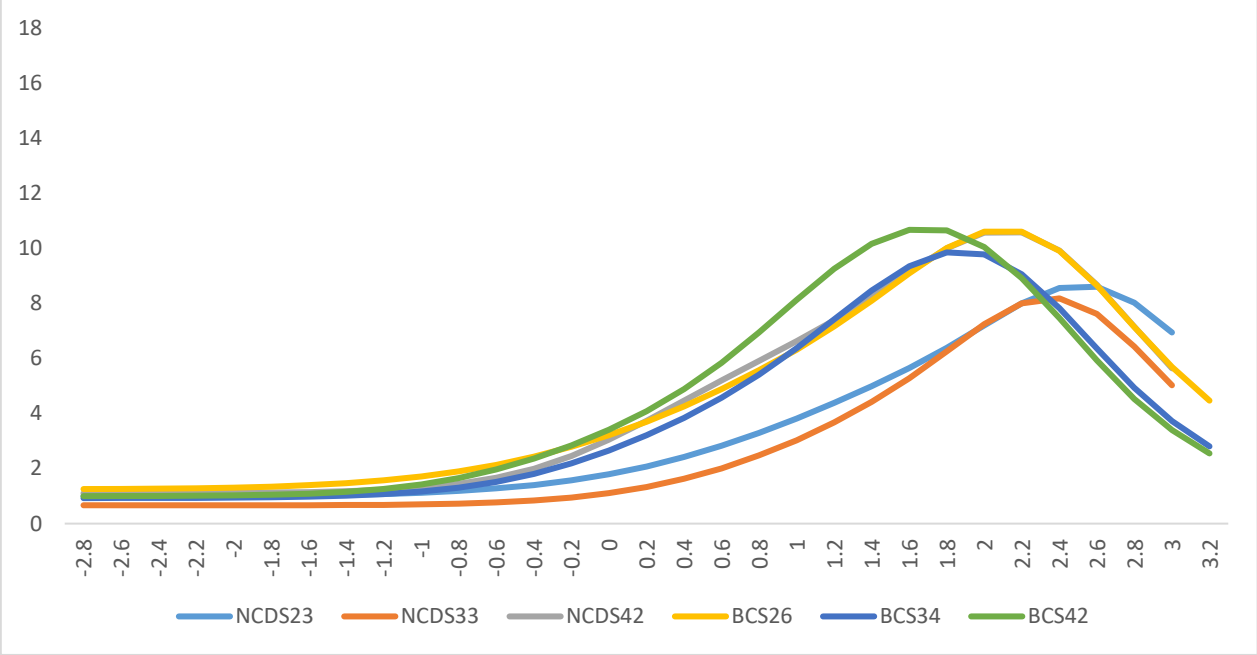
Graph 5a. Standardized thresholds – BCS70 Men



Graph 5b. Standardized thresholds – BCS70 Women



Graph 6a. Scale Information Functions for both cohorts - Men



Graph 6b. Scale Information Functions for both cohorts - Women