Resource report

closer
The home of longitudinal research

Harmonisation and measurement properties of mental health measures in six British cohorts

Eoin McElroy[1], Aase Villadsen[1], Praveetha Patalay[1, 2], Alissa Goodman[1], Marcus Richards[2], Kate Northstone[3], Pasco Fearon[4], Marc Tibber[4], Dawid Gondek[1], George B. Ploubidis[1]

[1] Centre for Longitudinal Studies, University College London

[2] MRC Unit for Lifelong Health and Ageing, University College London

[3] MRC Integrative Epidemiology Unit, University of Bristol

[4] Faculty of Brain Sciences, University College London

UKRI Economic and Social Research Council

Copyright

How to Cite

McElroy, E., Villadsen, A., Patalay, P., Goodman, A., Richards, M., Northstone, K., Fearon, P., Tibber, M., Gondek, D., & Ploubidis, G.B. (2020). Harmonisation and Measurement Properties of Mental Health Measures in Six British Cohorts. London, UK: CLOSER.

Table of contents

List of Tables

List of Figures

Acknowledgements

This project is funded by CLOSER, whose mission is to maximise the use, value and impact of longitudinal studies. CLOSER was funded by the Economic and Social Research Council (ESRC) and the Medical Research Council (MRC) between 2012 and 2017. Its initial five-year grant has since been extended to March 2022 by the ESRC (grant reference: ES/K000357/1). The funders took no role in the design, execution, analysis or interpretation of the data or in the writing up of the findings. www.closer.ac.uk

This work package is supported by CLOSER's Innovation Fund, which supports research that seeks to harmonise existing data from different longitudinal studies in the UK. Data harmonisation is the process of making data from different studies more comparable. By harmonising data from different UK longitudinal studies, researchers will be able to pool data from multiple studies, an exercise that has many benefits; e.g. increased sample sizes, increased heterogeneity of samples. Moreover, data harmonisation provides us with the opportunity to examine factors that may account for between-study differences, thereby providing insight into societal changes over time.

This project brings together data from six British cohorts: i) the MRC National Survey of Health of Development (NSHD), ii) the 1958 National Child Development Study (NCDS), iii) the 1970 British Cohort Study (BCS70), iv) the Avon Longitudinal Study of Parents and Children (ALSPAC), v) Next Steps, and vi) the Millennium Cohort Study (MCS). The NSHD is funded by the Medical Research Council and hosted by the MRC Unit for Lifelong Health and Ageing at University College London (UCL). The NCDS, BCS70, Next Steps and MCS receive core funding from the ESRC, and are hosted by the Centre for Longitudinal Studies, UCL. The NCDS sweep at age 61 is co-funded by the MRC, the US National Institutes of Health and the Department for Work and Pensions. The most recent sweep (at the time of writing) of the BCS70, at age 46, received additional funding from the MRC and the British Heart Foundation. The ALSPAC receives core funding from the MRC, the Wellcome Trust, and the University of Bristol, and is hosted by the University of Bristol. Next Steps was previously managed and funded by the Department for Education (2004-12).

The authors would like to thank the custodians of the six studies included in this report, and the cohort members and their families who have given their time to take part in these studies. We would also like to acknowledge the UK Data Service for providing access to the NCDS, BCS70, Next Steps, and MCS. With regards to the NSHD and ALSPAC, we would like to thank the following individuals for their help in gaining access to the relevant data:

NSHD: Dr Philip Curran and Mr Adam Moore, MRC Unit for Lifelong Health and Ageing, UCL

ALSPAC: Dr Sian Crosweller, Population Health Sciences, Bristol Medical School, University of Bristol

Glossary of terms

Throughout this report we use a variety of terms commonly used in the latent variable modelling literature. Examples include:

Factor analysis	=	A statistical procedure used to infer a latent variable based on a set of observed variables.
Factor loading	=	Parameter that captures the strength of association between a latent and an observed variable.
Item	=	An individual question within a questionnaire.
Latent variable	=	A variable that cannot be measured directly (e.g. psychological distress). Rather its presence can be inferred by assessing various observable variables (self-reported low mood, loss of interest, fatigue) that are purported to be driven by this underlying latent variable.
Measurement precision	=	Ability of a scale to reliably measure a latent construct at *different levels* of the latent construct.
Measurement properties	=	Validity, reliability and responsiveness of a scale/instrument.
Measure/scale/ instrument	=	A questionnaire used to assess mental health problems.
Multi-Group Confirmatory Factor Analysis (MGCFA)	=	Multi-Group Confirmatory Factor Analysis. Simultaneous CFA in various groups, allows for equivalence/invariance testing.
Observed variable	=	A variable that can be observed directly (e.g. self-reported low mood, guilt, fatigue).
Threshold	=	Parameter that captures the level of the latent variable that needs to be reached in order for an individual to transition from one category of an observed variable to another.

List of Acronyms

ALSPAC	=	Avon Longitudinal Study of Parents and Children
BCS70	=	1970 British Cohort Study
CFA	=	Confirmatory Factor Analysis
CFI	=	Comparative Fit Index
EFA	=	Exploratory Factor Analysis
IRT	=	Item Response Theory
MCS	=	Millennium Cohort Study
MGCFA	=	Multi-Group Confirmatory Factor Analysis
NCDS	=	1958 National Child Development Study
NSHD	=	MRC National Survey of Health of Development
RMSEA	=	Root Mean Square Error of Approximation
TIF	=	Total Information Function
TLI	=	Tucker–Lewis Index
WLSMV	=	Weighted Least Squares

Key Findings

1. We tested the respective measurement equivalence of two mental ill-health (i.e. psychological distress) questionnaires that were administered at different time points in six British cohorts (the Malaise Inventory and Strengths and Difficulties Questionnaire). We found full measurement invariance for both questionnaires, meaning there were no systematic differences in measurement error due to cohort membership or time of assessment. As such, findings based on these measures can reliably be compared across sweeps and studies.

2. In cases where different measures were administered across cohorts or assessment waves, we used a content validation approach to identify questions that assessed the same underlying symptom/indicator of psychological distress. We produced a searchable tool that allows researchers to identify these matching questions across different permutations of cohorts and sweeps.

3. We used this tool to identify harmonisable subsets of items for both within-cohort and cross-cohort research, and demonstrated the favourable psychometric properties of these harmonised scales (e.g. good reliability, high correlations with full measures). We also tested the measurement equivalence of these harmonised items sets, and found evidence that the harmonised measures were capturing the same constructs both within and across cohorts.

4. We have also clarified the different levels of measurement invariance, and the level of invariance needed for the specific research questions for which the cohorts are frequently used.

1. Introduction

1.1 Background

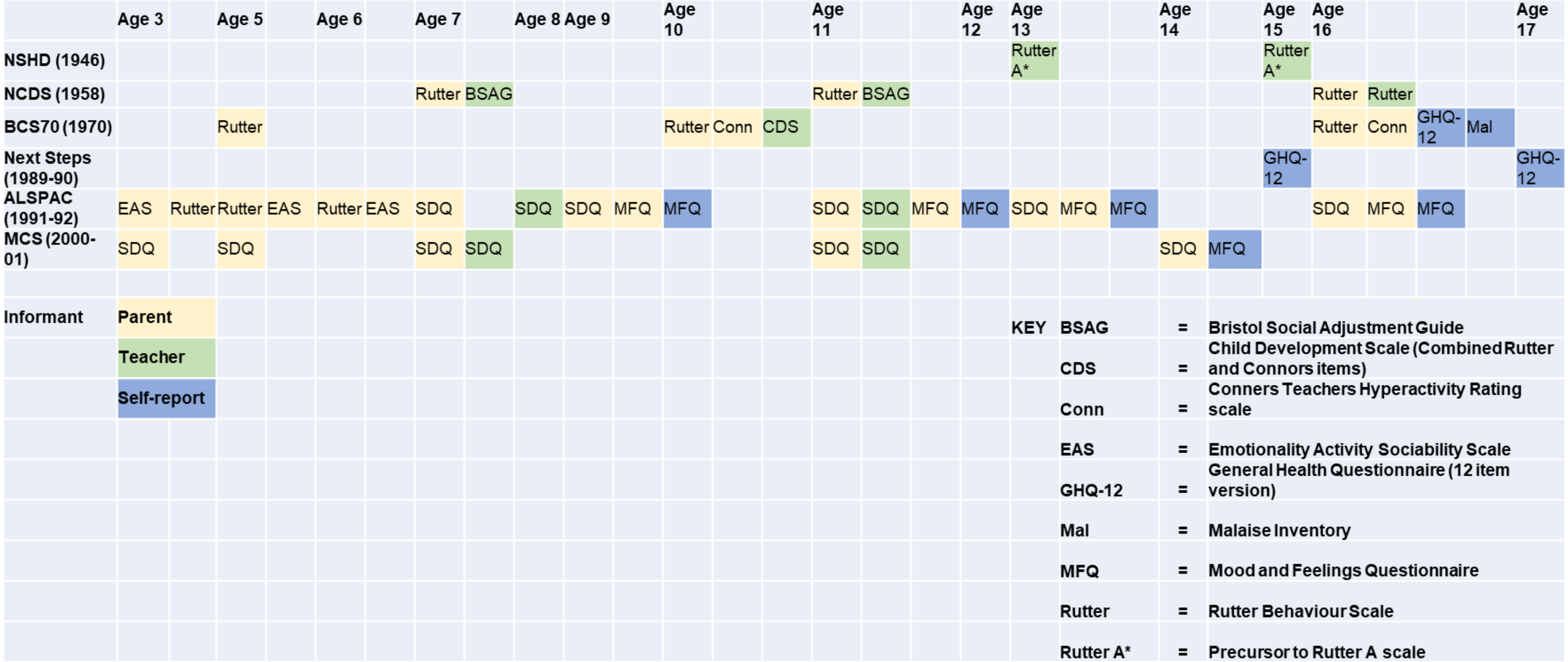
Common mental health problems such as anxiety and depression make a substantial contribution to the global burden of disease (Whiteford et al., 2013). Such difficulties often emerge early in childhood and demonstrate considerable continuity across the life-course (Ormel et al., 2015). Worryingly, recent evidence has suggested that mental health problems are increasing at the population-level (Collishaw, 2015; Patalay & Gage, 2019; Ploubidis, Sullivan, Brown, & Goodman, 2017).

In order to address this considerable public health concern, it is important to understand trends and risk factors that are universal across development (i.e. age effects) and those that are specific to individuals who were born at particular points in history (i.e. cohort effects).

High quality life-course research is required to disentangle age, period and cohort effects, and the British cohorts represent a particularly powerful data resource in this regard. The cohorts contain a wealth of information on the mental health of the UK population across multiple generations (Figure 1 and Figure 2). For a full overview of the mental health data available in the British cohorts, we refer readers to the CLOSER work package, “Maximising the take-up of mental health measures from UK cohorts and longitudinal studies” (https://www.closer.ac.uk/research-fund-2/data-harmonisation/maximising-takeup-mental-health-measures-uk-cohorts-longitudinal/).

Although the depth and breadth of information is a considerable strength of these studies, the specific instruments used vary substantially both within and across cohorts. Upon inspection of the measures available, it becomes clear that they can differ on up to four key features: i) content (number of symptoms assessed and/or wording of questions), ii) response scale (e.g. Likert ratings vs visual analogues), iii) time-frame of reference (i.e. symptoms assessed over past week, month, year etc.), and iv) reporter (e.g. teacher or parent proxies, trained interviewer, self-report).

Given the ultimately subjective nature of these measures, such differences may influence the respondent’s interpretation of the specific questions/items. This in turn may introduce elements of bias into responses, which raises issues about the comparability of these measures within and across the cohorts.

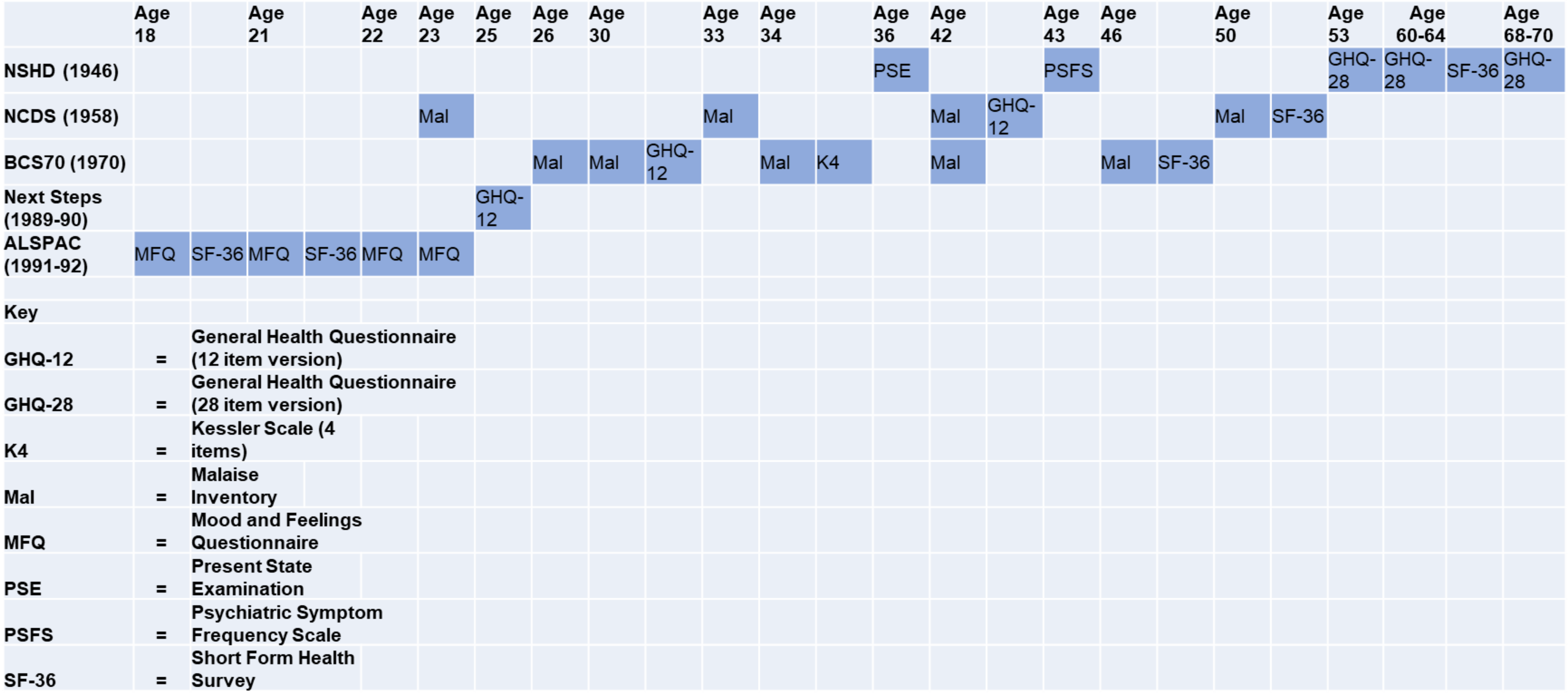


	Age 3		Age 5		Age 6		Age 7		Age 8	Age 9		Age 10			Age 11			Age 12	Age 13			Age 14		Age 15	Age 16				Age 17
NSHD (1946)																			Rutter A*					Rutter A*					
NCDS (1958)							Rutter	BSAG							Rutter	BSAG									Rutter	Rutter			
BCS70 (1970)			Rutter									Rutter	Conn	CDS											Rutter	Conn	GHQ-12	Mal	
Next Steps (1989-90)																								GHQ-12					GHQ-12
ALSPAC (1991-92)	EAS	Rutter	Rutter	EAS	Rutter	EAS	SDQ		SDQ	SDQ	MFQ	MFQ			SDQ	SDQ	MFQ	MFQ	SDQ	MFQ	MFQ				SDQ	MFQ	MFQ		
MCS (2000-01)	SDQ		SDQ				SDQ	SDQ							SDQ	SDQ						SDQ	MFQ						

Informant	
	Parent
	Teacher
	Self-report

KEY			
	BSAG	=	Bristol Social Adjustment Guide
	CDS	=	Child Development Scale (Combined Rutter and Connors items)
	Conn	=	Conners Teachers Hyperactivity Rating scale
	EAS	=	Emotionality Activity Sociability Scale
	GHQ-12	=	General Health Questionnaire (12 item version)
	Mal	=	Malaise Inventory
	MFQ	=	Mood and Feelings Questionnaire
	Rutter	=	Rutter Behaviour Scale
	Rutter A*	=	Precursor to Rutter A scale

Figure 1. Overview of mental health measures administered throughout childhood in six British Cohort Studies



	Age 18		Age 21		Age 22	Age 23	Age 25	Age 26	Age 30		Age 33	Age 34		Age 36	Age 42		Age 43	Age 46		Age 50		Age 53	Age 60-64		Age 68-70
NSHD (1946)														PSE			PSFS					GHQ-28	GHQ-28	SF-36	GHQ-28
NCDS (1958)						Mal					Mal				Mal	GHQ-12				Mal	SF-36				
BCS70 (1970)								Mal	Mal	GHQ-12		Mal	K4		Mal			Mal	SF-36						
Next Steps (1989-90)							GHQ-12																		
ALSPAC (1991-92)	MFQ	SF-36	MFQ	SF-36	MFQ	MFQ																			
Key																									
GHQ-12	=	**General Health Questionnaire (12 item version)**																							
GHQ-28	=	**General Health Questionnaire (28 item version)**																							
K4	=	**Kessler Scale (4 items)**																							
Mal	=	**Malaise Inventory**																							
MFQ	=	**Mood and Feelings Questionnaire**																							
PSE	=	**Present State Examination**																							
PSFS	=	**Psychiatric Symptom Frequency Scale**																							
SF-36	=	**Short Form Health Survey**																							

Figure 2. Overview of mental health measures administered throughout adulthood in six British cohort studies (all measures are self-reports)

1.2 Aims

To conduct meaningful comparisons across sweeps and/or cohorts, it is vital to establish the equivalence of the measures/instruments that were administered. This work aims to conduct a comprehensive investigation into the measurement properties and psychometric equivalence of the mental health measures that are available in six British cohort studies. Specifically, we have the following aims:

Aim 1: Conduct the first systematic investigation into the measurement properties of the existing mental health measures in the cohorts. We will present our findings in the form of a catalogue which will provide researchers with information and guidelines on how best to utilise the available measures.

Aim 2: In cases where *the same measures* have been administered across multiple cohorts or sweeps (e.g. Malaise Inventory, Strengths and Difficulties Questionnaire), we will assess the psychometric equivalence of these measures. This will allow us to provide guidance on how/where we can reliably compare such measures both within and across the cohorts.

Aim 3: In order to facilitate broader comparisons within and across the cohorts, we will conduct *retrospective harmonisation*. This process involves the manipulation of available data in order to make it more comparable across studies (Fortier et al., 2017). We will achieve this by matching specific items from different measures based on content (i.e. matching items that tap the same underlying symptom), recoding these items to a common metric (where necessary), and assessing the psychometric equivalence of these harmonised measures. Again, this will allow us to provide guidance on how and where we can reliably compare measures within and across the cohorts.

Aim 4: A comparison of mother and teacher reports of child mental health will be carried out using harmonised items. We will explore the extent to which parent and teacher questionnaires capture the same underlying dimensions of child mental health; the precisions of measures by these respective reporters; and the extent to which there is agreement between informants on child mental health. These examinations will provide guidance for users on the value of using information from both parents and teachers where this is available in cohort studies.

1.3 Report structure and additional outputs

This report includes a comprehensive catalogue of the measurement properties of the mental health measures available in six British cohorts (Sections 3 and 4). Sections 5 and 6 provide guidance on the comparability of the SDQ and Malaise measures, which were administered in a consistent format within and across cohorts. In sections 7, 8, and 9 we report the results from our retrospective harmonisation procedures in which we assessed the psychometric equivalence of items from different measures (matched based on content). Section 10 details the comparability of mother v teacher proxy reports of child mental health. Finally, section 11 provides a summary of findings and recommended guidelines for researchers looking to conduct comparisons within and across cohorts.

A data deposit of rescaled variables generated as part of this work will be made available in due course on the UK Data Service website (see https://www.closer.ac.uk/research-fund-2/data-harmonisation/harmonisation-mental-health-measures-british-birth-cohorts/). Also we provide a searchable tool which highlights different permutations of comparable items based on our item matching procedure (see section 2.3.1 for further details).

2. Methods

2.1 Studies included

This report documents the measurement properties and explores the feasibility of harmonising the mental health measures in six cohort studies:

MRC National Survey of Health of Development: The MRC NSHD is Britain's oldest birth cohort study. It originally consisted of a socially stratified sample (N=5,362) of men and women born to married parents in England, Scotland and Wales in March 1946. The sample was selected from an initial maternity survey of 13,687 pregnancies, and consisted of all births to non-manual and agricultural families, and a random 1-in-4 sample from manual families. To date, the participants have been followed 24 times between ages 2 and 68-69 years. At age 69, the most recent home visit as of the time of writing, 2,149 cohort members participated.

The 1958 National Child Development Study: The NCDS follows the lives of 17,415 people that were born in England, Scotland or Wales in a single week in March 1958. The NCDS started in 1958 as the Perinatal Mortality Survey and captured 98% of the total births in Great Britain in the target week. The cohort has been followed up a total of 10 times between ages 7 and most recently (as of the time of writing) at 55, when 9,137 cohort members took part.

1970 British Cohort Study: The BCS70 follows the lives of 17,198 people born in England, Scotland and Wales in a single week in March 1970. The BCS70 began as the British Births Survey and participants have since been followed up nine times between ages 5 and 46. A total of 8,581 cohort members took part in the most recent assessment covered by this report, at age 46.

The Avon Longitudinal Study of Parents and Children: The ALSPAC is a prospective cohort study of children born in the English county of Avon between April 1st 1991 and December 31st 1992 (N = 14,062). Data is collected on both parents and children, and more recently ALSPAC has started to recruit and collect data on the children of the original cohort members.

Next Steps (formerly the Longitudinal Study of Young People in England; LSYPE): Next steps, follows the lives of around 16,000 people in England born in 1989-90. Although not a birth cohort (the study began in 2004 when the cohort members were aged 14), respondents were selected to be representative of young people in England. Cohort members were surveyed annually until 2010, and the next sweep after this was when they were aged 25, in 2015-16.

The Millennium Cohort Study: The MCS a UK-wide birth cohort study of individuals born in England, Scotland, Wales and Northern Ireland at the start of the millennium (Sept. 2000 – Jan. 2002). The initial sample consisted of 19,517 children. Since the initial birth survey at 9 months, the cohort has been followed up five times at ages 3, 5, 7, 11 and at the latest sweep covered by this report, age 14, when 11,872 cohort members took part.

More details on each of the cohorts, along with links to cohort profiles can be found at https://www.closer.ac.uk/closer/explore-the-studies/.

2.2 Measurement properties

We investigated the measurement properties of the mental health scales using a latent variable modelling approach. Given the measures were questionnaires answered using Yes/No or Likert ratings, we employed the appropriate model for binary and ordered categorical data; i.e. the multivariate probit model estimated using the robust Mean and Variance Adjusted Weighted Least Squares (WLSMV) estimator. All analyses of measurement properties were conducted using Mplus version 8.1 (Muthén & Muthén, 2018).

2.2.1 Structural properties

We used factor analysis to examine the latent structure of each instrument. For measures with well-established factor structures, confirmatory factor analysis (CFA) was used to evaluate the fit of said structures in the cohort data. Model fit was assessed using the following indices; 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). For both the CFI and TLI, values of greater than 0.90 and 0.95 were judged to reflect adequate and good model fit respectively (Barrett, 2007). For the RMSEA, values of less than 0.05 were taken to reflect good fit, and values up to 0.08 acceptable fit (Hu & Bentler, 1998). In cases where models approached but did not reach acceptable fit, or demonstrated acceptable fit on some indices but not others, we inspected Mplus' modification indices, and allowed correlations between the unique/residual variances of certain item pairs within the same factor. This strategy can improve model fit by increasing the proportion of variance explained, without changing the substantive conclusions regarding the adequacy of a given factor structure in describing a set of data (Bollen, 1989).

For measures without an established factor structure, exploratory factor analysis (EFA) with oblique (geomin) rotation was used to examine the underlying dimensionality of the measures, and CFA was used to evaluate the fit of the optimal EFA model. For the EFA, we decided on the number of factors to extract using the Kaiser-Guttman rule (eigenvalues above 1), and confirmed this decision by inspecting the scree plot and factor loadings (Yong & Pearce, 2013).

2.2.2 Precision of measurement

The precision of measurement of each scale was evaluated by plotting total information functions (TIFs). A TIF plot presents Fisher information, which is inversely related to the standard error of measurement, and therefore illustrates the precision (or reliability) of a measure at different levels of the underlying latent variable ($\boldsymbol{\theta}$) (Betz & Turner, 2011). As such, TIF plots are useful for defining a range of scores over which a measure may be considered precise/reliable.

2.3 Harmonisation

In certain cases, the exact same measure was administered across multiple sweeps and/or cohorts. For example, 9 items from the Malaise Inventory of general psychological distress were administered at various assessment waves between ages 16 – 50 in both the NCDS and BCS70. As such, it was relatively straightforward to test the measurement equivalence of this instrument across the two cohorts, and within each cohort over time (see Section 6). In order to conduct broader comparisons (i.e. include a broader range of sweeps and/or cohorts), it was necessary to conduct *retrospective harmonisation.* This term is used to describe the broad process of modifying existing data to make it more directly comparable across studies. For a review of general principles and methods, see (Fortier et al., 2017). Our retrospective harmonisation strategy consisted of three stages. First, we identified items from different measures that captured the same symptom. Second, in situations where the identified items were administered on different scales, we recoded or transformed these items to a comparable metric. For specific examples, see sections 3.3.1 (for Rutter scale visual analogue conversion) and 8.1 for rescaling in adulthood. Third, we assessed the measurement equivalence of these harmonised items within and across cohorts (see Section 2.4 for details).

2.3.1 Item matching process

In order to identify items from different scales that could be considered candidates for harmonisation, we adopted the following two-step process:

i. Two raters (a research associate specialising in psychiatric epidemiology and psychometrics; an experienced clinical psychologist) independently and systematically inspected all of the mental health measures for items that overlapped in content. First, the raters worked through each measure item-by-item and assigned a code to each individual item that summarised the content of that item at the most basic descriptive level. For example, the item *"Have you been in low spirits or felt miserable"* (question # 2 from the Psychiatric Symptom Frequency Scale in the NSHD), was coded as 'low mood' by both raters independently of one another.

ii. The coded items were then compiled in a single spreadsheet, and inter-rater agreement (coded agree/disagree) was recorded for each item. This information was then used to calculate an overall inter-rater agreement score (calculated as the number of items which the raters agreed upon divided by the total number of items they inspected). In instances where the two raters disagreed on the coding of an item, a third independent rater (an experienced clinical psychologist) made the decision on which item code (if either) was appropriate.

Heatmaps were produced to illustrate interrater agreement across the different measures. Summary tables were also produced highlighting the content overlap of items from different measures. A searchable item-mapping tool, based on this content analysis, is available on the CLOSER website (https://www.closer.ac.uk/research-fund-2/data-harmonisation/harmonisation-mental-health-measures-british-birth-cohorts/).

2.4 Measurement equivalence

Even when psychometric measures are ostensibly similar within or across studies, various factors (e.g. respondent age, study design, mode effects, period effects, cohort effects) can impact the manner in which participants interpret and ultimately respond to questions. A failure to account for such measurement error can bias any comparisons made either within or across studies.

In order to accurately compare scores on a latent variable across cohorts or sweeps, it is important that the underlying measurement model is equivalent (Van De Schoot et al., 2013). In other words, the relationship between the latent variable (in this case, psychological distress) and its measured indicators (in this case, the specific items/questions asked) should be consistent across cohorts/assessments. We assessed the psychometric equivalence of the mental health questionnaires by testing for measurement invariance. Failing to ensure measurement invariance in the groups of interest is analogous to differential measurement error (Armstrong, 1998), as group membership directly influences measurement error in the outcome. Although it is beyond the scope of this report to provide an in-depth technical account of measurement invariance, we provide a short conceptual overview of this process. For further in-depth discussions, see (Little, 2013; Putnick & Bornstein, 2016; Van De Schoot, Schmidt, De Beuckelaer, Lek, & Zondervan-Zwijnenburg, 2015; Wicherts & Dolan, 2010).

In summary, we tested for measurement invariance using multiple group confirmatory factor analysis (MGCFA). This involves the fitting of a series of nested confirmatory factor models (CFAs), in which increasingly strict equality constraints are placed on specific measurement parameters across different cohorts/assessment waves. The two relevant measurement parameters are the factor loadings (λ) and thresholds (τ). A factor loading reflects the strength of association between the unobserved latent variable (e.g. psychological distress) and a measured indicator (e.g. "Q1: Do you often feel low, miserable or depressed"). A high factor loading indicates that a particular question/item can be considered a good indicator of the underlying latent variable. For ordered categorical data (e.g. Likert responses that are typically used in mental health questionnaires), the threshold parameter reflects the level of the latent trait that must be exceeded for an individual to be in a particular category (see Figure 3).

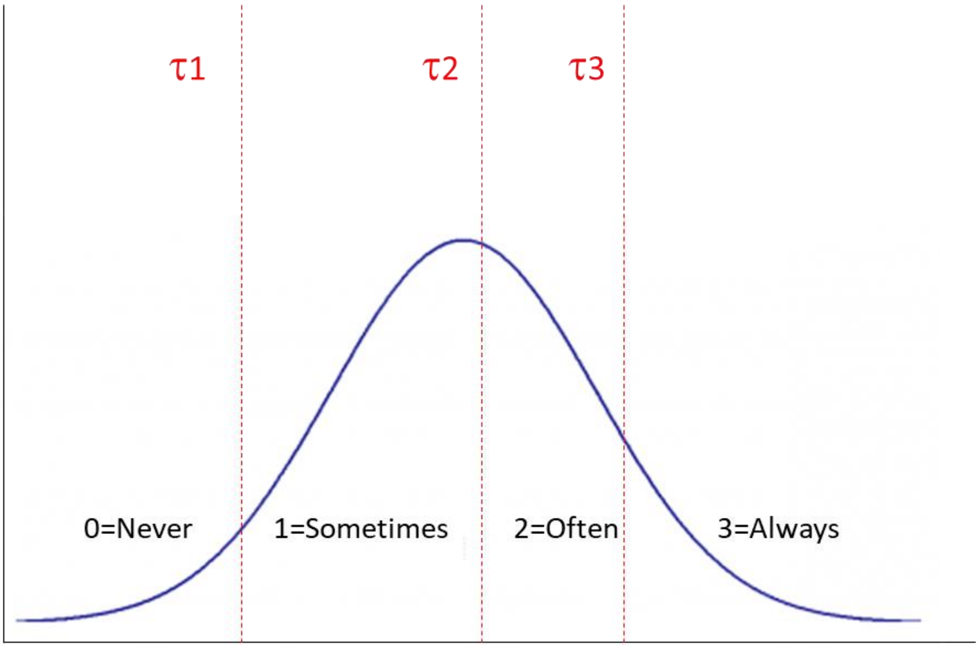


Figure 3. Illustration of threshold parameters (τ) for a hypothetical question asked on a 4-point Likert scale, assuming a responses reflect an underlying normal distribution

If, after fitting equality constraints across cohorts/sweeps, we do not observe a worsening of overall model fit, then said level of measurement invariance is judged to hold, and the parameters in question can be considered equivalent (i.e. group membership is *not* directly influencing measurement error). Importantly, different conclusions regarding the comparability of the measures can be drawn depending on the level of measurement invariance that is supported. Four levels of invariance are typically discussed in the literature:

i. **Configural invariance:** This is the least restrictive model. The same measurement model is specified in each cohort/sweep; however no equality constraints are placed on the parameters (i.e. factor loadings), and thresholds are allowed to differ across cohorts/sweeps. This tests whether the same measurement model is appropriate in each cohort/sweep (i.e. whether the data is adequately described by the same number of factors and pattern of indicators), and it serves as a baseline by which to compare more restrictive models.

ii. **Metric invariance:** This is tested by holding the factor loadings equal across cohorts/sweeps (Figure 4). If metric invariance holds, we can conclude that the associations between the underlying latent variable and its measured indicators are consistent across cohorts/sweeps. In other words, metric invariance ensures that the same construct is being measured across sweeps/cohorts. At this level of invariance, we can be confident that we can compare variances and covariances at the latent level. In the case of the discussed British cohorts, this level of invariance is important for researchers looking to examine whether particular associations between mental health variables and predictor/outcome variables are consistent across cohorts/sweeps (i.e. regression coefficients will not be biased due to group membership).

iii. **Scalar invariance:** This is tested by holding both the factor loadings and thresholds equal across different cohorts/sweeps (Figure 4). If scalar invariance holds, this indicates that participants from different cohorts or sweeps are interpreting the response scales of questions in a consistent manner. To illustrate, if scalar invariance holds across two cohorts, we can conclude that individuals from both cohorts had a similar interpretation of the differences in severity implied by the response options “0=Never”, “1=Sometimes”, and “2=Always”. For researchers interested in using the discussed British cohorts, scalar invariance is particularly important for those who wish to compare mean scores on mental health measures across time/cohorts (e.g. studies of change over time).

iv. **Strict invariance:** Strict invariance is tested by holding the factor loadings, thresholds and residuals (ε) equal across cohorts/sweeps. If strict invariance holds, then any difference observed between cohorts/sweeps can be attributed solely to a difference in the underlying latent variable. Methodologists note, however, that the conditions for strict invariance are rarely satisfied in practice (Van De Schoot et al., 2013). Moreover, others question whether it is even appropriate to test for strict invariance. For instance, Little (Little, 2013) notes that the residual of each indicator/test is comprised of both random and item-specific error. While it is plausible that the item-specific error could be consistent across cohorts/sweeps, random error, by definition, should be considered unique in each instance. Strict invariance conflates both random and item-specific error, and therefore introduces an element of bias into the solution. As such, we do not test for strict invariance.

In practice, it can often be challenging to obtain full scalar invariance (Van De Schoot et al., 2015). In this situation, many researchers opt to test for partial measurement invariance (PMI) by releasing equality constraints (intercepts, loadings, or both) to the point where acceptable levels of fit are achieved. This PMI solution can then be used to explore differences in latent means or associations, with the obvious caveat that there will be some unquantifiable element of bias in the estimates that can be attributed to the freed parameters. Research in this this area is still rather limited (see (Putnick & Bornstein, 2016) for an overview), and there is no consensus as to how many parameters can be released whilst maintaining meaningful comparisons. Chen (Chen, 2007) demonstrated that the bias in mean estimates across groups increased in proportion to the number of non-invariant factor indicators, therefore it is clearly desirable to have as many invariant indicators as possible. Most guidelines suggest that at least half of the indicators should be invariant across groups/time in order to conduct meaningful comparisons (Little, 2013; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). In the present report, we approach this issue on a case-by-case basis; in instances where only PMI is supported, we comment on the number of noninvariant items/questions, and what this means when comparing means and regression coefficients within and across cohorts. There are numerous methods for selecting the parameters that are to be freed when testing for PMI. In this project, we followed the guidelines of Yoon and Kim (Yoon & Kim, 2014), who proposed a 'backwards method' of releasing parameters one at a time based on the size of their relevant modification index.

2.5 Analysing harmonised scales

The ultimate goal of harmonising mental health measures is to create a set of scores that can validly be compared within or across studies. As such, it can be viewed as an attempt to reduce measurement error. After measurement invariance has been established, the next step is to implement these harmonised scores in subsequent analysis in order to answer substantive research questions. There are several options open to the researcher at this stage, and below we discuss these in descending order of recommendation.

i. The preferred method for incorporating latent variables into any analysis is to include them in your model directly. In other words, estimate both the measurement model (with equality constraints placed on loadings and threshold) and any additional parameters (e.g. path coefficients) jointly within a SEM framework. This approach is not always possible however, as practical issues such as sample size, model complexity and data type may cause issues with convergence (Devlieger & Rosseel, 2017; Hoshino & Bentler, 2011). There may also be other practical issues, for example software availability, as for many statistical analyses the specification of latent variables is not currently possible with existing software.

ii. A practical approach to dealing with these issues is to employ a two-step approach in which measurement models (with equality constraints placed on loadings and thresholds) are estimated and used to produce factor scores. Factor scores are numerical values that represent estimates of an individual's relative standing on a latent variable. By placing equality constraints on the measurement parameters used to derive these factor scores across cohorts/assessment waves, the estimated scores are placed on a comparable metric, which allows for valid comparisons between cohorts or within cohorts over time (Curran et al., 2014). These factor scores can then be used in subsequent models in place of summed total scores (Bauer & Hussong, 2009; Curran et al., 2014). Before using these scores in further analyses, we recommend researchers assess the quality of factor score estimates; see Ferrando & Lorenzo-Seva (2018) for an overview of this topic.

iii. In instances where full scalar invariance has been supported, the estimation and interpretation of factor scores is relatively straightforward. However, as long as at least one item is invariant, it is possible to produce factor scores within and across groups that are anchored to a consistent metric. This practice remains debated however, and the number of invariant indicators required to make valid comparisons is an area that requires further research (Curran et al., 2014). At present, it is recommended that the majority of indicators are invariant within and across cohorts (Curran et al., 2014; Little, 2013). One limitation of this approach is that the factor scores are treated as observed variables and not as estimates, as they really are. Not taking into account the uncertainty in the estimation of factor scores may lead to underestimation of standard errors of regression coefficients in subsequent analysis. We recommend when the two-step approach has to be employed, that - where possible - standard errors are estimated by a resampling technique such as bootstrapping.

iv. Having investigated and established satisfactory measurement equivalence within and/or across cohorts, a third strategy could be to simply construct summed total scores based on the invariant items. This approach brings in an additional assumption, namely that the items have equal factor loadings, which would imply a Rasch type model. This assumption is testable, and if supported by the data, as for example has been shown for the Malaise Inventory on NCDS and BCS70 (Ploubidis, McElroy, & Moreira, 2019), the sum score can be used instead of factor scores. As the sum score is directly observed and not estimated, there is no need to correct for estimation uncertainly in subsequent analysis. However, Rasch type models are very restrictive and might not fit the data. In these instances, although the establishment of measurement invariance rules out differential measurement error due to group membership, using a summed score might introduce bias by conflating true score and measurement error. A detailed discussion of the disadvantages of using summed totals is beyond the scope of this report. We refer those unfamiliar with these issues to introductory SEM resources (Bollen, 2014; Kline, 2015).

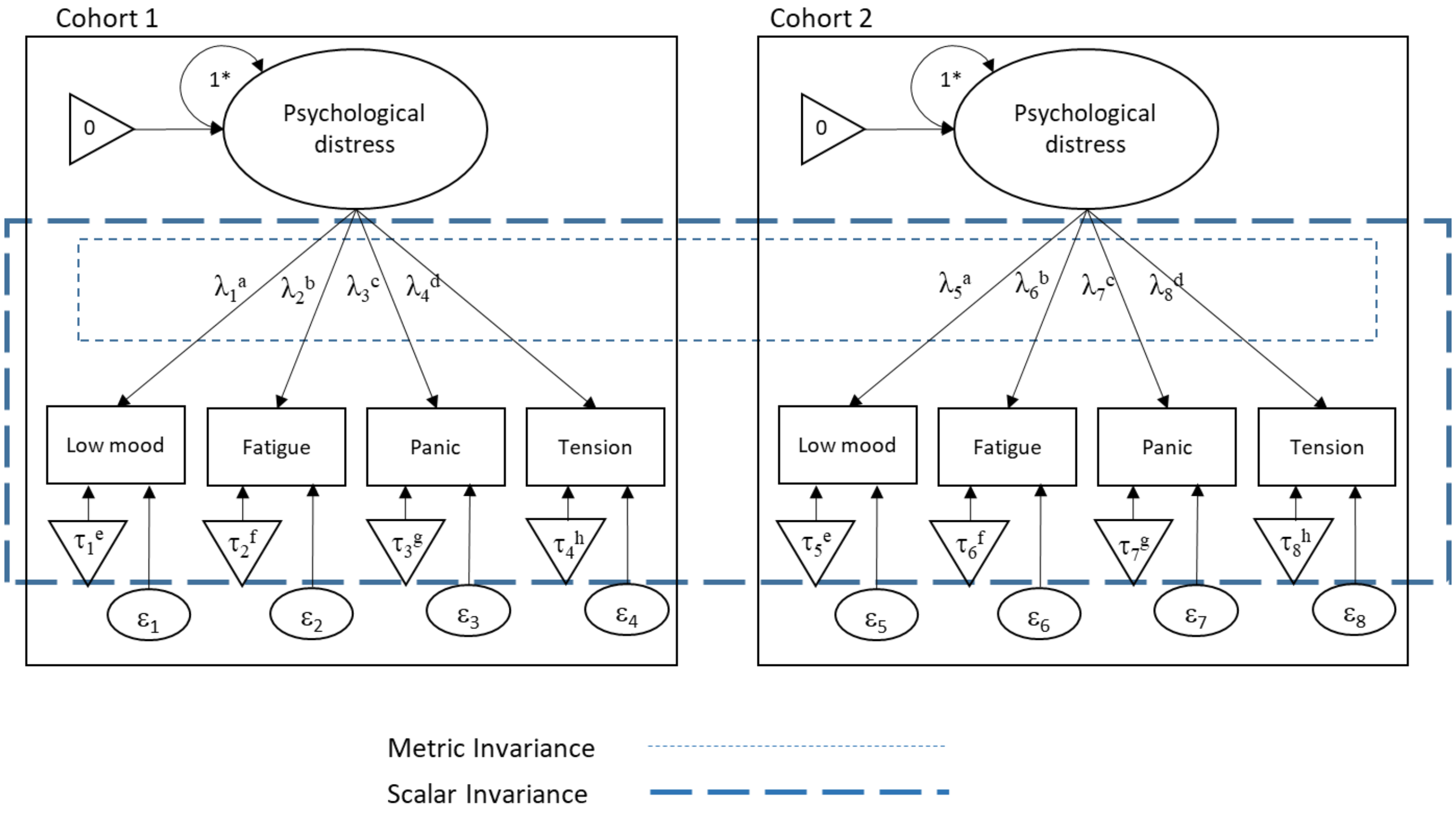


Figure 4. Graphical illustration of multiple group confirmatory factor analysis, with four measured indicators of a general psychological distress factor, assessed across two cohorts

λ = Factor loadings; τ = Thresholds; ⍰ = residuals (theta parameterisation); a-d = loadings held equal across cohorts in test for metric invariance; e-h = thresholds held equal across cohorts in test for scalar invariance

3. Measurement Properties of Mental Health Scales in Childhood

3.1 NSHD

The first comprehensive assessment of mental health in the NSHD was conducted using teacher-reports at ages 13 and 15 years. The measure administered was a precursor to the Rutter A scale (Elander & Rutter, 1996). This scale assessed emotional and behavioural problems, with responses indicated on a 3-point scale that roughly corresponded to *absent*, *normal* and *severe*.

To date, the only examination of the latent structure of this measure was conducted by Xu and colleagues (Xu et al., 2013). Xu et al. (Xu et al., 2013) noted that several of the items in the measure did not align with an *absent-normal-severe* scoring system, but rather assessed qualitatively distinct symptoms of mental health. Take for example the following item; *"Which statement best describes your child: A dare devil (1); As cautious as the average child (2); Extremely fearful (3)"*. This item appears to capture both risk taking/reckless behaviour, and fear/anxiety. As such, Xu et al. (2013) separated the above item into two separate binary items; "*As cautious as the average child (0); A dare devil (1)*" and "*As cautious as the average child (0); Extremely fearful (1)"*. Using this recoded pool of items (see Table 1), they conducted EFA and found a three factor solution fit the data best at both ages. The three factors corresponded to emotional problems (e.g. low mood, anxiety), conduct problems (e.g. aggression, disobedience) and a self-organisation factor (e.g. concentration, neatness, daydreaming).

Table 1. Items from teacher reported Rutter precursor scale, as coded by Xu et al. (2013)

Factor	Split item	Which statement in each group best describes this child?
Self-organisation		A very hard worker (0); average–works moderately well (1); a poor worker or lazy (2)
Self-organisation		One with high power of concentration (0); Average–concentrates moderately well (1); little or no power of sustained concentration (2)
Self-organisation		Extremely neat and tidy in class work (0); average–moderately neat and tidy (1); very untidy in class work (2)
Self-organisation		Seldom or never daydreams in class (0); sometimes daydreams in class (1); frequently daydreams in class (2)
Conduct		Has this child been punctual in attending school during the past year? Never late unless with good reason (0); Sometimes late (1); Persistently late (2)
Conduct		Has this child played truant during the last year? Yes, frequently (2); yes, occasionally (1); Never (0)
Conduct		Seldom or never disobedient (0); Sometimes disobedient (1); Frequently disobedient (2)
Conduct		Seldom or never difficult to discipline (0); sometimes difficult to discipline (1); frequently difficult to discipline (2)
Conduct		Seldom or never restless in class (0); Sometimes restless in class (1); Frequently restless in class (2)
Conduct		Seldom or never cribs (0); Sometimes cribs (1); Frequently cribs (2)
Conduct		Seldom or never evades the truth to keep out of trouble (0); Sometimes evades the truth to keep out of trouble (1); Frequently evades the truth to keep out of trouble (2)
Conduct	a	Takes a normal part in rough games (0); Liable to get unduly rough during playtime (1)
Conduct	b	Does not unduly avoid or seek attention (0); Shows off; seeks attention (1)
Conduct	c	As cautious as the average child (0); A dare devil (1)
Conduct	d	Average–not particularly quarrelsome (0); A quarrelsome and aggressive child (1)
Conduct	e	Normally competitive (0); Overcompetitive with other children (1)
Conduct	f	How does this child react to criticism or punishment? Normal attitude to criticism and punishment (0); Tends to become unduly resentful (1);

Table 1 continued

Factor	Split item	Which statement in each group best describes this child?
Emotional		Unusually happy and contented child (0); Generally cheerful and in good humor (1); Usually gloomy and sad (2)
Emotional		Makes friends extremely easily (0); Takes usual amount of time to make friends (1); Does not seem able to make friends (2)
Emotional		Would you describe this child as an anxious child (i.e., apprehensive, worrying, and fearful)? Not at all anxious (0); Somewhat anxious (1); Very anxious (2)
Emotional		Do you regard this child as Extremely energetic, never tired (0); Normally energetic (1); Always tired and “washed out” (2)
Emotional	a	Takes a normal part in rough games (0); Rather frightened of rough games (1)
Emotional	b	Does not unduly avoid or seek attention (0); Avoids attention, hates being in the limelight (1)
Emotional	c	As cautious as the average child (0); Extremely fearful (1)
Emotional	d	Average—not particularly quarrelsome (0); A timid child (1)
Emotional	e	Normally competitive (0); Diffident about competing with other children (1)
Emotional	f	How does this child react to criticism or punishment? Normal attitude to criticism and punishment (0); Tends to become unduly miserable or worried (1)

a-f denotes split items

Using the item pool and recoding procedure adopted by Xu et al. (Xu et al., 2013), we fitted CFA models to the data at ages 13 and 15 years. Unidimensional factor models (reflecting general psychological distress) and three factor models (self-organisation, emotional, and conduct) were fitted to the data, and the results are presented in Table 2. The single factor model provided poor overall fit at both ages. The three factor model demonstrated adequate fit on all three indices at age 13. At age 15, the three factor model had acceptable fit based on the RMSEA, but (marginally) missed the criteria for acceptable fit on to the CFI and TLI.

Table 2. Fit statistics for mental health measures administered in childhood in NSHD

Age	Measure	Model	N	X^2	DF	RMSEA	CFI	TLI
13	Precursor to Rutter	1-factor	4109	14794.608*	324	0.104	0.657	0.628
		3-factor	4109	4385.190*	321	0.056	0.904	0.895
15	Precursor to Rutter	1-factor	4050	15642.641*	324	0.108	0.677	0.650
		3-factor	4050	5446.548*	321	0.063	0.892	0.882

Standardised factor loadings from the 3-factor model are presented in Figure 5. All items demonstrated moderate-to-high loadings, and the patterns of loadings were highly consistent across ages, suggesting measurement equivalence.

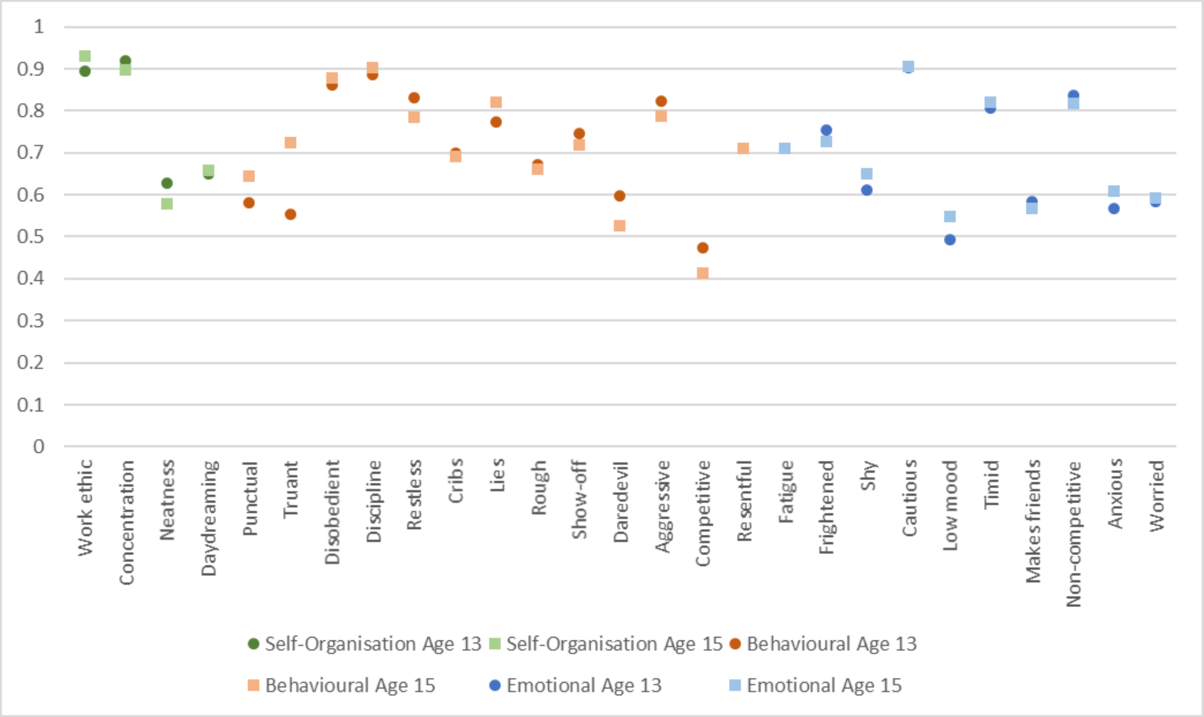


Figure 5. Standardised factor loadings of the Rutter precursor items in NSHD at ages 13 and 15

TIFs for the emotional, conduct and self-organisation factors are presented in Figure 6. At both age 13 and 15, the TIF for the self-organisation scale resembled a bimodal distribution, indicating highest precision at moderately high and low levels of the latent trait, but low precision around the mean level. For the emotional problems sub-scale, the TIF demonstrate the highest level of precision at moderately high levels of the trait (approximately 1.2 – 2.4 SDs above the mean). The conduct scale had highest precision from approximately 0.8 to 3.0 SDs above the mean.

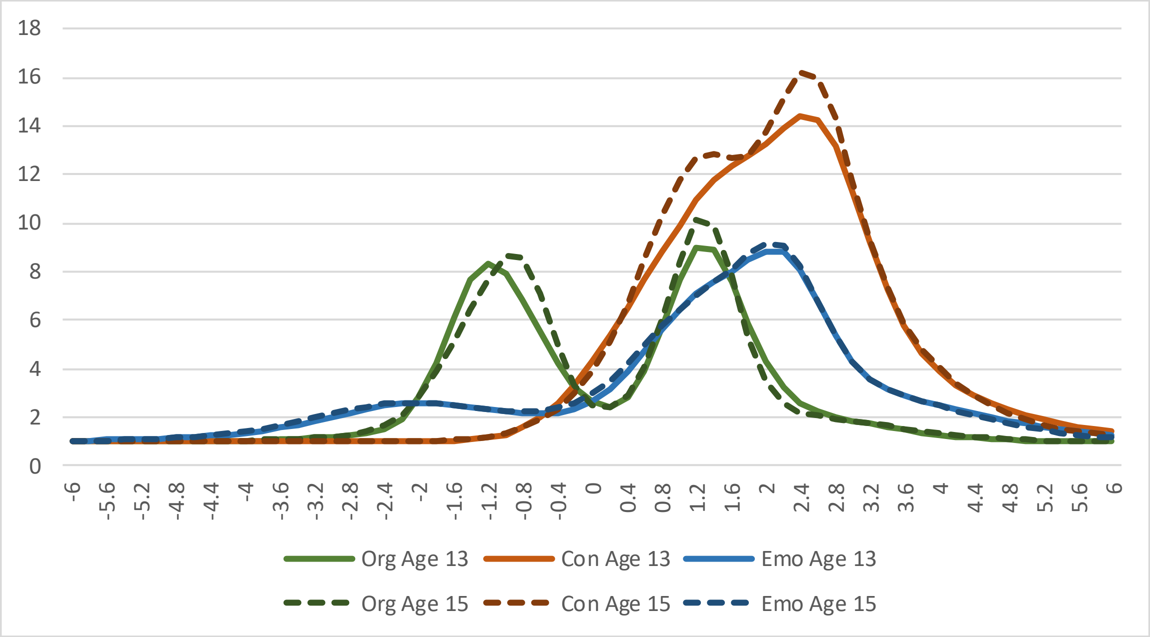


Figure 6. TIFs for the emotional, conduct and self-organisation subscales of the Rutter A precursor in NSHD

The level of the latent factor is presented on the X-axis (expressed in standard deviations from a mean of 0). Fisher information (inverse of standard error of measurement) is presented on the Y-axis, with higher values reflecting greater precision of measurement.

3.2 NCDS

3.2.1 Rutter Behaviour Scales

The main measures of child and adolescent mental health in the NCDS were modified versions of the parent and teacher Rutter behaviour scales (Rutter, Tizard, & Whitmore, 1970). These scales were completed by study parents when children were aged 7, 11 and 16, and by teachers when children were aged 16. The exact items administered are presented in Table 3. Items covered a range of emotional and behavioural difficulties, with responses indicated on a 3-point Likert response scale (0 = 'Does not apply', 1 = 'Applies somewhat', 2 = 'Certainly applies').

Table 3. Alternative versions of the Rutter behaviour scale administered in NCDS

14-item parent-report version at ages 7 and 11	18-item parent-report version age 16	26-item teacher-report version age 16
1. Has difficulty in settling to anything for more than a few moments	1. Very restless. Has difficulty staying seated for long.	1. Very restless. Has difficulty staying seated for long.
2. Prefers to do things on his/ her own rather than with others	2. Squirmy, fidgety child.	2. Truants from school
3. Is bullied by other children	3. Often destroys own or others' property.	3. Squirmy, fidgety child.
4. Destroys own or others' belongings (e.g. tears or breaks)	4. Frequently fights or is extremely quarrelsome with other children	4. Often destroys own or others' property.
5. Is miserable or tearful	5. Not much liked by other children	5. Frequently fights or is extremely quarrelsome with other children
6. Is squirmy or fidgety	6. Often worried, worries about many things	6. Not much liked by other children
7. Worries about many things	7. Tends to do things on his/ her own - rather solitary	7. Often worried, worries about many things
8. Is irritable, quick to fly off the handle	8. Irritable, is quick to fly off the handle	8. Tends be on own - rather solitary
9. Sucks thumb or finger during the day	9. Often appears miserable, unhappy, tearful or distressed	9. Irritable, touchy, is quick to fly off the handle
10. Is upset by new situation, by things happening for the first time	10. Has twitches, mannerisms or tics of the face or body	10. Often appears miserable, unhappy, tearful or distressed
11. Has twitches or mannerisms of the face, eyes or body	11. Frequently sucks thumb or finger	11. Has twitches, mannerisms or tics of the face or body
12. Fights with other children	12. Frequently bites nails or fingers	12. Frequently sucks thumb or finger
13. Bites nails	13. Is often disobedient	13. Frequently bites nails or fingers
14. Is disobedient at home	14. Cannot settle anything for more than a few moments	14. Tends to be absent from school for trivial reasons
	15. Tends to be fearful or afraid of new things or new situations	15. Is often disobedient
	16. Fussy or over particular	16. Cannot settle anything for more than a few moments
	17. Often tells lies	17. Tends to be fearful or afraid of new things or new situations
	18. Bullies other children	18. Fussy or over particular
		19. Often tells lies

Table 3 continued

14-item parent-report version at ages 7 and 11	18-item parent-report version age 16	26-item teacher-report version age 16
		20. Has stolen things on one or more occasions in the past 12 months
		21. Unresponsive, inert or apathetic
		22. Often complains of aches or pains
		23. Has had tears on arrival at school or has refused to come into the building in the past 12 months
		24. Has a stutter or stammer
		25. Resentful or aggressive when corrected
		26. Bullies other children

Over the years, various different scoring conventions have been applied to the Rutter scales in the NCDS. Different combinations of items have been summed to create subscales that most often taking the form of distinct emotional/internalizing, behavioural/externalizing, and psychomotor agitation scales. However, the exact items used to create these sum-scores have varied slightly across different studies (Anderson, 2018; Collishaw, Maughan, Goodman, & Pickles, 2004; Schoon, Sacker, & Bartley, 2003). Although many studies have explored the psychometric properties of the Rutter scales, the majority of these studies have focussed on the unmodified versions which contain several additional items (Elander & Rutter, 1996; Rutter et al., 1970). To our knowledge, no studies have as of the time of writing explored the latent structure of the modified versions of the Rutter scales that were administered in the NCDS. Therefore, we conducted an EFA using all available items in order to uncover the underlying structure of the data.

In the parent report versions of the scales, a three-factor solution was judged the best fitting model at ages 7, 11, and 16. The first two factors corresponded to the previously discussed dimensions of emotional/internalizing, and behavioural/externalizing problems. The third factor encompassed symptoms of hyperactivity (e.g. difficulty settling, squirmy/fidgeting), but also included several items capturing psychomotor agitation/habits e.g. (biting nails, sucking thumb, twitches/mannerisms of the face). For convenience, this factor will henceforth be referred to as 'psychomotor agitation'. In the teacher report version at age 16, a 4-factor solution was extracted. Three of the factors corresponded to the internalizing, externalizing and psychomotor agitation factors discussed above. The fourth factor encompassed three items that reflected interest/engagement with school (truancy, absent due to trivial reasons, apathetic/unresponsive). This factor was labelled 'truancy'.

Having identified the best fitting models, we fitted CFAs to the data (Table 4). We also estimated unidimensional models, and 3-factor models consisting of the 13 items that were common across all versions of the scales. The 3-factor models (4-factor in the case of the teacher report at 16) provided acceptable fit at all ages. The modified 3-factor models (common items only) also provided adequate fit.

Table 4. Fit statistics for Rutter scales administered in childhood in NCDS

Age	Measure	Reporter	Model	N	X^2	DF	RMSEA	CFI	TLI
7	Rutter (14 items)	Parent	1-factor	14,608	5267.567	77	0.068	0.817	0.783
			3-factor	14,608	2683.544	74	0.049	0.908	0.887
			3-factor[c]	14,608	2528.921	62	0.052	0.903	0.878
11	Rutter (14 items)	Parent	1-factor	13,805	4765.998	77	0.066	0.820	0.787
			3-factor	13,805	2591.752	74	0.050	0.903	0.881
			3-factor[c]	13,805	2459.401	62	0.053	0.898	0.872
16	Rutter (18 items	Parent	1-factor	11,653	7808.033	135	0.070	0.800	0.773
			3-factor	11,653	2920.718	132	0.043	0.927	0.916
			3-factor[c]	11,652	1259.226	62	0.041	0.936	0.919
16	Rutter (26 items)	Teacher	1-factor	12,551	26803.096	299	0.084	0.898	0.889
			4-factor	12,551	13492.173	293	0.060	0.949	0.944
			3-factor[c]	12,533	2719.879	62	0.058	0.963	0.954

[c] 13-item model containing all common items across ages and reporters

Standardised factor loadings from the best fitting models are presented in Figure 7, and loadings for the common item models are presented in Figure 8. The rank ordering of factor loadings (lowest to highest) was broadly consistent across ages and reporters; however loadings were generally highest for the teacher-report administered at age 16 years. A similar pattern of loadings emerged when only the items common to all waves/reporters were modelled (Figure 8).

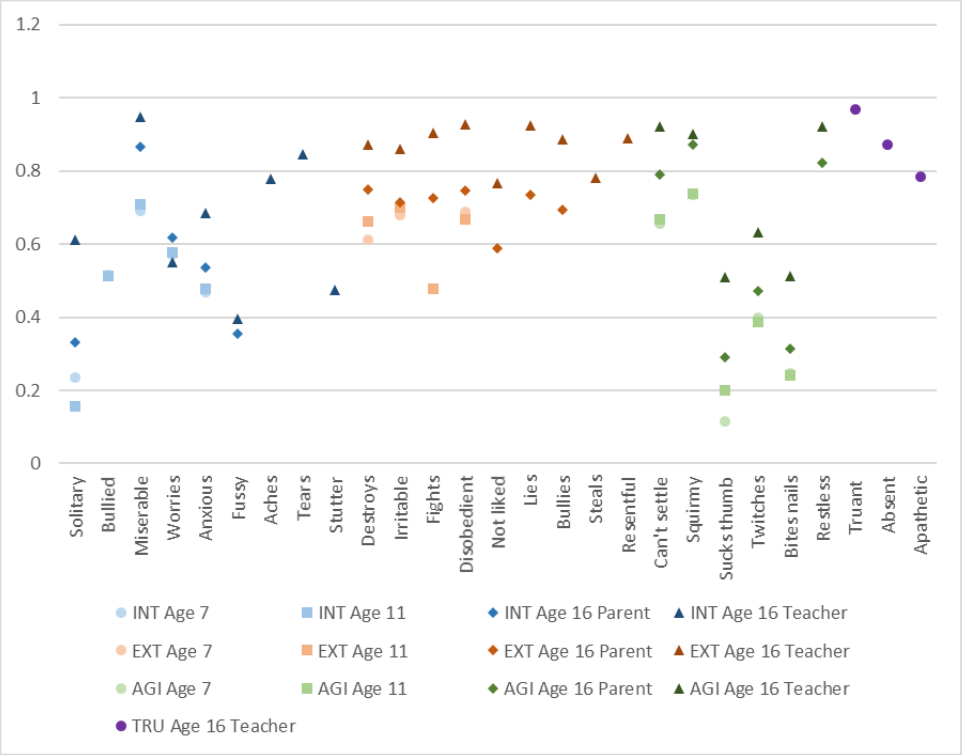


Figure 7. Standardised factor loadings for Rutter behaviour scales (all items) in NCDS

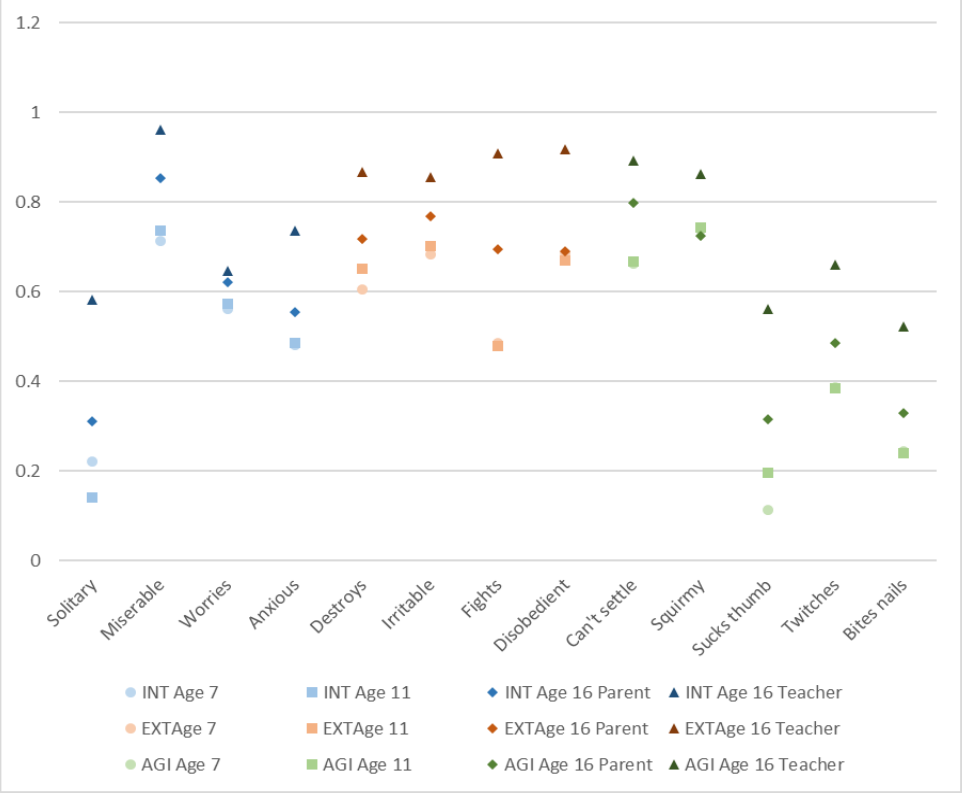


Figure 8. Standardised factor loadings for Rutter behaviour scales (common items only) in NCDS

Several items loaded relatively poorly on certain factors. Across all assessment waves and reporters, the items that measured twitches, thumb sucking and nail-biting did not load strongly on the psychomotor agitation factor. Similarly, items capturing solitary behaviour and being bullied/disliked by other children had relatively low loadings on the emotional factor. These low loadings in the CFAs, coupled with high cross-loadings in the EFAs, suggest that these items can be considered poor indicators of their relevant factors. We leave it to the discretion of the researcher as to whether these variables should be included in future analyses.

The TIFs for the various Rutter subscales are presented in Figure 9. Overall the teacher reports at age 16 had the highest precision over the broadest range of the latent trait, followed by parent reports at age 16. Parent reports at ages 7 and 11 demonstrated low precision. TIFs are influenced by the number of items in a given measure; therefore models were also estimated using only the items that were common to all assessment waves/reporters (13 items; see Table 3 for further details). The TIFs for the Rutter subscales, as measured using only the common items, are also presented in Figure 9. A similar pattern emerged, with teacher reports demonstrating good precision compared with parent reports.

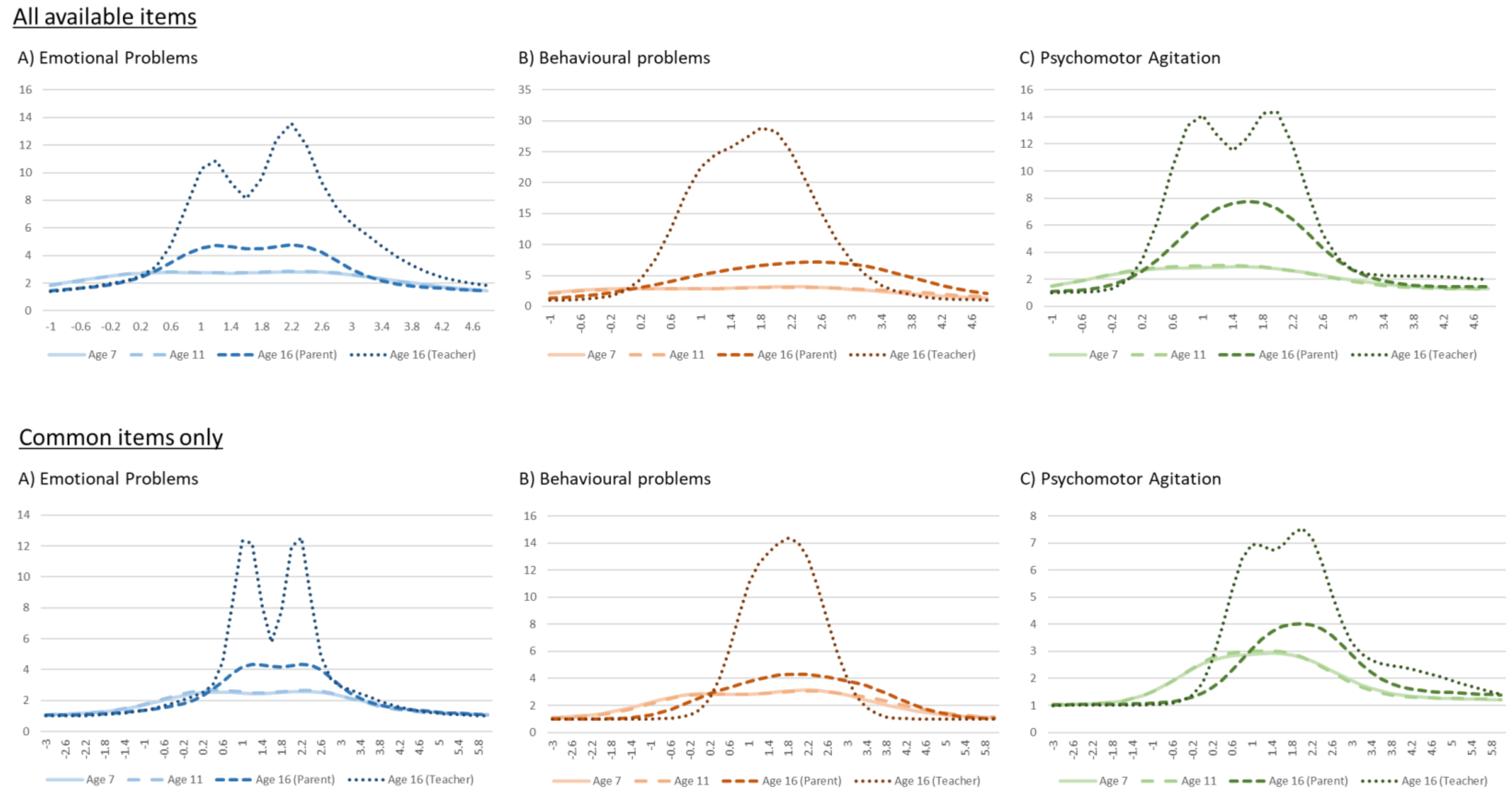


Figure 9. TIFs for the emotional, conduct and psychomotor agitation subscales of the Rutter behaviour questionnaires as administered in NCDS

The level of the latent factor is presented on the X-axis (expressed in standard deviations from a mean of 0). Fisher information (inverse of standard error of measurement) is presented on the Y-axis, with higher values reflecting greater precision of measurement.

3.2.2 Bristol Social Adjustment Guide (BSAG)

At ages 7 and 11, teachers also completed the Bristol Social Adjustment Guide (BSAG) (Stott, 1974). The BSAG is designed to measure behaviour in a school setting. It consists of a 4 page booklet that contains over 250 phrases/descriptions of behaviour. Teachers are required to underline the descriptions that best fit the child, and these can be transformed via a coding system into a smaller number of quantitative 'syndromes'. Here we analyse these derived summary scales (i.e. item parcels), not the raw items (not available at time of writing). For a more detailed description of the BSAG and the relevant variables available in the NCDS, we refer readers to a resource report conducted by Shepherd (Shepherd, 2013).

A previous factor analytic study found that the various syndromes could be grouped under a stable two-factor solution broadly reflecting internalizing and externalizing problems (Ghodsian, 1977). We examined this model in the NCDS at ages 7 and 11 using CFA with robust maximum likelihood estimation (MLR) due to the continuous and skewed nature of the 'syndrome' data. Fit statistics are presented in Table 5 and standardised factor loadings in Figure 10.

Table 5. Fit statistics for BSAG administered in childhood in NCDS

Age	Model	N	X^2	DF	RMSEA	CFI	TLI
7	1 factor	14,930	8515.757	35	0.127	0.626	0.520
	2 factor	14,930	5222.963	34	0.101	0.771	0.697
11	1 factor	14,158	6885.311	35	0.118	0.677	0.584
	2 factor	14,158	4828.763	34	0.100	0.774	0.700

Overall, both 1-factor (general psychological distress) and 2-factor (internalizing/externalizing) models failed to meet the criteria for acceptable model fit, indicating poor psychometric properties. Standardised factor loadings from the 2-factor model are presented in Figure 10. Loadings were highly consistent at ages 7 and 11 years.

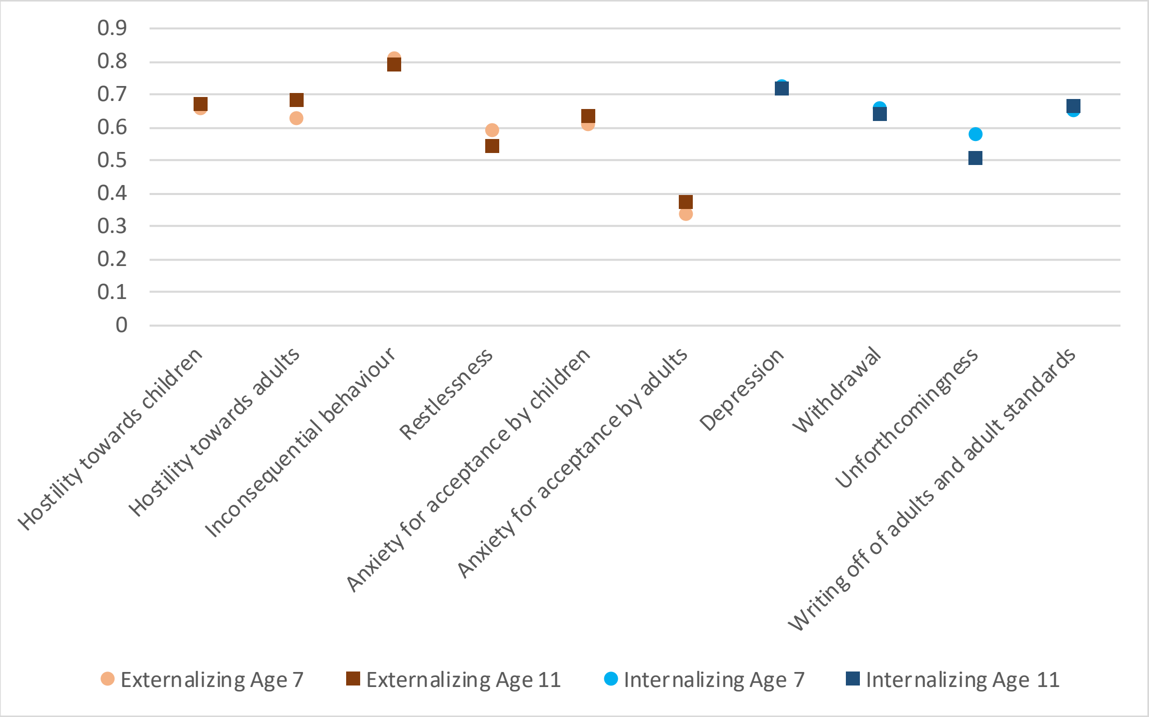


Figure 10. Standardised factor loadings of BSAG syndromes in NCDS

3.3 BCS70

The primary measures of mental health in childhood were the Rutter Behaviour Scales (Rutter et al., 1970) and the Conners Teachers Hyperactivity Rating Scale (Conners, 1969). Both scales are parent or teacher proxy reports that assess emotional and behavioural difficulties in children. The normal response scale for the Rutter measure is a 3-point Likert rating (1='Certainly applies', 2='Applies Somewhat', 3='Doesn't Apply'), and for the Conners scale a 4-point rating is used (1='Not at all', 2='Just a Little', 3='Pretty Much', 4='Very Much'). However at the age 10 assessment, responses were indicated on a visual analogue (see 3.3.1 below). The Rutter scale was administered to study mothers when the children were aged 5, and both the Rutter and Conners scales were administered at ages 10 and 16 years. A composite measure, dubbed the 'Child Development Scale', was administered to teachers at the age 10 assessment. This measure consisted of select items from the Rutter and Conners scales, along with additional items from the Swansea Assessment Battery (Butler & Bynner, 1997). The specific items that were administered in the Rutter, Conners and Child Development Scales are presented in Table 6.

Self-report measures of mental distress (the GHQ-12 and Malaise Inventory) were administered to study children when they were aged 16 years. The GHQ-12 (Goldberg, 1988) assesses general psychological distress. It asks respondents to rate the degree to which they experience a symptom 'generally' on a 4-point scale ('Less than usual', 'No more than usual', 'Rather more than usual', or 'Much more than usual'). The Malaise Inventory (Rutter et al., 1970), hereafter referred to in shorthand as the Malaise, is another self-report measure of general distress, in which emotional and somatic symptoms are endorsed in a simple 'yes/no' format. The full version of the Malaise consists of 24 items; however the version administered at this assessment consisted of 22 items (questions regarding rheumatism or fibrosis, and having had a nervous breakdown were removed).

Table 6. Overview of Rutter, Connors, and Child Development scale items completed by mothers and/or teachers at age 10, with corresponding BCS70 variable names

Variable	*Rutter Parent Scale (19 items)*	*Variable*	*Connors Hyperactivity Scale (19 items)*	*Variable*	*Child Developmental Scale (53 items)*
m43	Very restless	m63	Is noticeably clumsy [1*]	j122	Child's popularity with peers (reverse code) [d¥]
m44	Squirmy or fidgety [a]	m64	Trips or falls easily or bumps into objects or other children [2]	j123	Friends
m45	Destroys belongings [b]	m65	Inattentive, easily distracted [3]	j124	Boldness
m46	Fights with other children [c]	m66	Hums or makes other odd noises at inappropriate times [4]	j125	Cooperative
m47	Not much liked by other children [d]	m67	Has difficulty picking up small objects [5]	j126	Negotiate child's behaviour
m48	Worried [e]	m68	Drops things which are being carried [6]	j127	Child is daydreaming
m49	Does things on own-rather solitary [f]	m69	Becomes obsessional about unimportant things [7]	j128	Afraid of new things/situations [i]
m50	Irritable	m71	Requests must be met immediately, easily frustrated [8]	j129	Cannot concentrate on particular task [19]
m51	Appears miserable or distressed [g]	m72	Shows restless or overactive behaviour [9]	j130	Wetting pants during class
m52	Takes others' belongings	m73	Is impulsive, excitable [10]	j131	Complains about things
m53	Has twitches, mannerisms or ticks [h]	m74	Interferes with the activity of other children [11]	j132	Trips falls bumps [2]
m54	Sucks thumb or finger	m75	Is sullen or sulky [12]	j133	Works deftly with hands
m55	Bites nails or fingers	m76	Fails to finish things he/ she starts, short attention span [13]	j134	Displays outbursts of temper [17]
m56	Often disobedient	m77	Given to rhythmic tapping or kicking [14]	j135	Teases other children
m57	Cannot settle to do anything	m78	Cries for little cause [15]	j136	Clumsy at games [1*]
m58	Afraid of new things/situations [i]	m79	Changes mood quickly and drastically [16]	j137	Cries for little cause [15]

Table 7 continued

Variable	*Rutter Parent Scale (19 items)*	*Variable*	*Connors Hyperactivity Scale (19 items)*	*Variable*	*Child Developmental Scale (53 items)*
m59	Fussy or over particular [j]	m80	Displays outbursts of temper, explosive or unpredictable behaviour [17]	j138	Becomes bored during class
m60	Often tells lies	m81	Has difficulty using scissors [18]	j139	Shows perseverance
m61	Bullies other children [k]	m82	Has difficulty concentrating on any particular task though may return to it frequently [19]	j140	Difficulty kicking ball
				j141	Dresses/undresses competently
				j142	Interferes with others [11]
				j143	Confused or hesitant
				j144	Difficulty picking up small objects [5]
				j145	Behaves 'nervously'
				j146	Fussy or over-particular [j]
				j147	Changes mood quickly [16]
				j148	Excitable impulsive [10]
				j149	Worried and anxious [e]
				j150	Shows restless or over-active behaviour[9]
				j151	Squirmy and fidgety [a]
				j152	Easily distracted [3]
				j153	Manipulates small objects with hands
				j154	Drops things being carried [6]
				j155	Pays attention in class
				j156	Relations with others unhappy/tearful [g*]

Table 8 continued

Variable	*Rutter Parent Scale (19 items)*	*Variable*	*Connors Hyperactivity Scale (19 items)*	*Variable*	*Child Developmental Scale (53 items)*
				j157	Obsessional about unimportant tasks [7]
				j158	Forgetful on complex task
				j159	Rather solitary [f]
				j160	Quarrels with other kids [c]
				j161	Can use manipulative equipment (e.g. scissors) [18*¥]
				j162	Shows lethargic/listless behaviour
				j163	Destroys belongings [b]
				j164	Hums or makes odd vocals [4]
				j165	Rhythmic tapping in class [14]
				j166	Inadequate control of pencil/paint brush
				j167	Soils pants during class
				j168	Accident prone
				j169	Bullies other children [k]
				j170	Sullen or sulky [12]
				j171	Has twitches, mannerisms/tics [h]
				j172	Truants from school
				j173	Fearful in movements
				j174	Completes tasks
				j175	Is easily frustrated [8]
				j176	Holds instruments appropriately

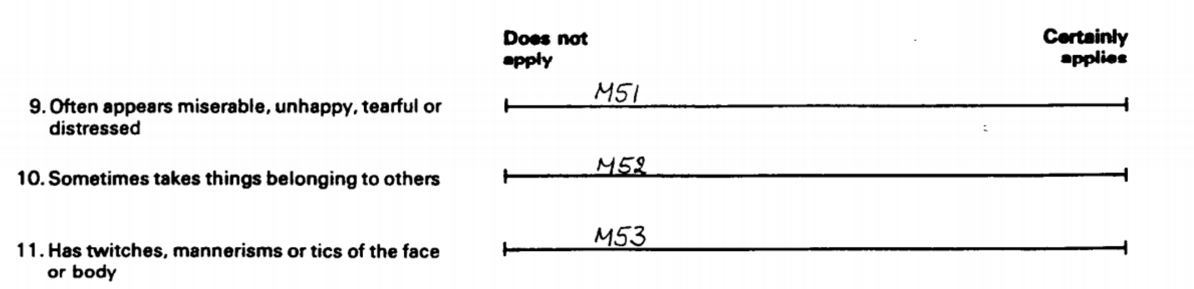
Table 9 continued

Variable	*Rutter Parent Scale (19 items)*	*Variable*	*Connors Hyperactivity Scale (19 items)*	*Variable*	*Child Developmental Scale (53 items)*
				j177	Fails to finish tasks [13]
				j178a	Extrovert-introvert
				j178b	Scale anxious-unworried

[a-i] Rutter scale items completed by mothers and teachers. [1-18] Connors scale items completed by mothers and teachers. *Closest matching item. [¥] Note direction of scoring.

3.3.1 Scale issues at age 10 assessment

At the age 10 assessment, both the Rutter and Connors scales were administered to mothers and teachers in the form of a visual analogue, in which they were asked to mark on a horizontal line how much a symptom statement applied to their child:



Does not apply

Certainly applies

9. Often appears miserable, unhappy, tearful or distressed — M51

10. Sometimes takes things belonging to others — M52

11. Has twitches, mannerisms or tics of the face or body — M53

Figure 11. Example of visual analogue responses from BCS70 age 10 maternal questionnaire

Each maternal-report item was then converted into a 100-point continuous variable, whereas each teacher-response was converted to 47-point continuous scale (an inspection of the original documentation failed to uncover a reason for this discrepancy in scales). This scoring system differed markedly from every other scale that was administered in childhood, both within the BCS70 and across other cohorts. In order to make the mental health variables more comparable within and across cohorts, the visual analogue data were recoded to 3- and 4-point Likert scales in line with the original scoring formats for the Rutter and Conners scales. Rather than use arbitrary cut-offs to force a Likert structure on the data (e.g. 0-33 = 'Doesn't apply', 34-66 = 'Applies somewhat', 67-100 = 'Definitely applies'), we employed finite mixture modelling, specifically latent profile analysis (LPA), to derive cut-offs empirically, thus minimising the loss of information. LPA is a form of latent variable model that can be used to uncover an unobserved discrete variable (e.g. ordinal categories) from a continuous distribution (Oberski, 2016). This allowed us to transform scores from the 0-100 and 0-47 scales into 3- and 4-point Likert scales as appropriate (Figure 12).

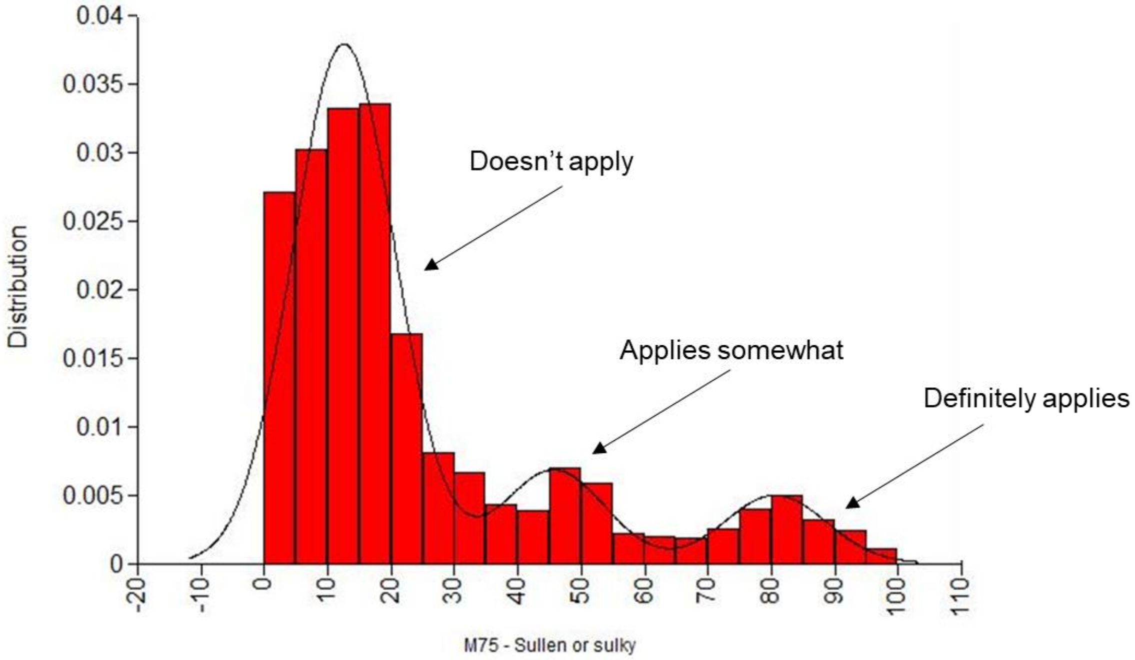


Figure 12. Graphical illustration of ordered categories recovered from continuous responses to the question 'Does your child often appear sullen or sulky' using latent profile analysis

We checked the validity of these derived ordinal variables by correlating each one with its continuous counterpart, with correlations ranging from 0.80 – 0.97. As an additional validity check, we created another set of variables by recoding the data manually as: 0-33 = 'Doesn't apply', 34-66 = 'Applies somewhat','67-100 = 'Definitely applies'[1]. We then correlated these variables with the original continuous variables. In each case, the categorical variables derived via LPA demonstrated a higher correlation with the continuous variables compared with the manually derived equivalent (full list of correlations available upon request). This was particularly evident in data that were highly skewed. For instance, for the highly positively skewed teacher-report item "Truants from school", the correlation between the original continuous variable and the LPA derived ordinal variable was 0.89. When we correlated the original continuous variable with the manually recoded variable, this correlation was 0.87. This demonstrates that the LPA method was superior at preserving the rank ordering of individuals compared with manual recoding. The LPA-derived variables were used for all of the age 10 analyses

[1] A similar manual recoding process was adopted for the 0-47 scale teacher-report variables

discussed in the present report. These derived variables will be made available in due course from the UK Data Service website (see https://www.closer.ac.uk/research-fund-2/data-harmonisation/harmonisation-mental-health-measures-british-birth-cohorts/).

3.3.2 Measurement properties

An EFA of the Rutter scale (parent-report) at age 5 revealed a 3-factor structure identical to that found in the NCDS, with factors corresponding to emotional and behavioural problems, and a third factor encompassing psychomotor agitation (e.g. "Can't settle", "Twitches", "Bites nails"). This three factor structure was therefore fitted to the mother–reported Rutter data at ages 5, 10, and 16. Only a subset of items (19 from 39) were administered to mothers from the Conners Hyperactivity Rating Scale, therefore we used EFA to examine the underlying structure of these items. At the age 10 and 16 assessments, a 3-factor solution provided the best description of the data. These factors reflected motor/coordination difficulties (e.g. "Noticeably clumsy", "Difficulty picking up small objects"), attention problems (e.g. "Inattentive, easily distracted") and a behavioural factor (e.g. "Is sullen or sulky", "Displays outbursts of temper"). For the self-report measures (GHQ-12 and Malaise Inventory), established 1- and 2-factor models were fitted to the data (Gnambs & Staufenbiel, 2018; Rodgers, Pickles, Power, Collishaw, & Maughan, 1999).

For the only teacher-report measure, the composite 'Child Development Scale', an EFA was conducted to explore its underlying dimensionality. A 6 factor solution was judged to fit best; however the patterns of loadings made these factors difficult to interpret. Therefore we fit separate models using the available Rutter and Conners scale items, applying the same factor structures to these scales as discussed above. The fit statistics for these models are presented in Table 7.

Table 10. Fit statistics for mental health measures administered in childhood in BCS70

Age	Measure	Reporter	Model	N	X2	DF	RMSEA	CFI	TLI
5	Rutter (19 items)	Mother	1-factor	13053	10296.795	152	0.072	0.817	0.795
			3-factor	13053	4864.409	149	0.049	0.915	0.903
10	Rutter (19 items)	Mother	1-factor	13548	10201.564	152	0.070	0.825	0.803
			3-factor	13548	4631.287	149	0.047	0.922	0.910
10	Conners (19 items)	Mother	1-factor	13524	43679.579	152	0.146	0.728	0.694
			3-factor	13524	22321.897	149	0.105	0.861	0.841
10	Rutter (11 items)	Teacher	1-factor	12702	10882.509	44	0.139	0.803	0.754
		Teacher	3-factor	12702	4228.298	41	0.090	0.924	0.898
10	Connors (19 items)	Teacher	1-factor	12702	25260.608	152	0.114	0.872	0.856
		Teacher	3-factor	12702	13114.282	149	0.083	0.934	0.924
16	Rutter (19-items)	Mother	1-factor	8931	6391.627	152	0.068	0.852	0.834
			3-factor	8931	3184.737	149	0.048	0.928	0.918
16	Conners (19 items)	Mother	1-factor	8921	13158.445	152	0.098	0.832	0.811
			3-factor	8921	3940.915	149	0.053	0.951	0.944
16	GHQ-12	Self	1-factor	5631	5550.375	54	0.134	0.864	0.833
			2-factor	5631	5434.199	53	0.134	0.867	0.834
16	Malaise (22 items)	Self	1 factor	5539	3818.619	209	0.056	0.894	0.883
16	Malaise (9 items)	Self	1 factor	5522	589.355	27	0.061	0.960	0.946

The previously identified 3-factor models for both the Conners and Rutter scales provided acceptable levels of fit across sweeps and reporter, with the exception of the maternal-report at age 10, which narrowly missed the cut-offs. In contrast, 1-factor models fit these measures poorly, suggesting that subscale scores should be used instead of overall total scores. For the Malaise Inventory, a 1-factor model using all 22 items had acceptable fit on the RMSEA, and approached acceptable fit on the CFI and TLI. An alternative model was specified in which we used only the 9 items from the abbreviated version of the scale (this version was administered frequently in the adult sweeps). A 1-factor model, reflecting general psychological distress, demonstrated excellent model fit.

Two alternative models for the GHQ-12 were fitted based on recent meta-analytic findings regarding the structure of this instrument: a 1-factor model reflecting general psychological distress, and a correlated 2-factor model that account for positively and negatively worded questions (Gnambs & Staufenbiel, 2018). Both models failed to demonstrate acceptable fit. For the 2-factor model, the model covariance matrix was not positive definite, due to a correlation of greater than 1 between the two factors. Based on this result and findings from the meta-analysis, a bifactor model was also fit to the data. This model consisted of a general psychological distress factor, and two orthogonal method factors (reflecting positively and negatively worded items). Each item was loaded onto the general factor and also on one of either the positively or negatively worded factors, depending on item content. This model demonstrated slightly better fit, although was still below acceptable cut-offs on the RMSEA and TLI. For this model, we also computed the 'explained common variance' (ECV). This statistic is calculated as the proportion of variance explained by the general factor divided by the overall variance, and can be used to determine the unidimensionality of a measure (Rodriguez, Reise, & Haviland, 2016). This value, which ranges from 0 to 1, was 0.803, indicating the general factor was by far the dominant source of shared variance. Overall these findings indicate that the GHQ-12 in the BCS70 should be treated as a unidimensional scale.

Factor loadings from the Rutter and Conners scales are presented in Figures 13 and 14 respectively. With reference to the Rutter scales, the items relating to habits had low factor loadings, as was the case in the NCDS. All of the Conners items had moderate-to-high loadings, with the exception of the item assessing.

TIFs for the Rutter, Conners and scales are presented in Figure 15. Across age and reporter, both scales demonstrated the highest level of precision at moderate-to-high levels of the latent trait (approx. 1.2-2.0 SDs from the mean). Across all three subscales of the maternal-report Rutter scales, higher and wider precision was observed as children aged. The teacher-report version of the Rutter (age 10) demonstrated comparable levels of measurement precision with the maternal reports, despite consisting of fewer items (11 vs. 19). The TIFs for the Conners scale were relatively consistent across age and reporter, although teachers had more precision when it came to behavioural problems, and parents had higher precision at higher ends of the trait for motor and hyperactivity problems when children were aged 16.

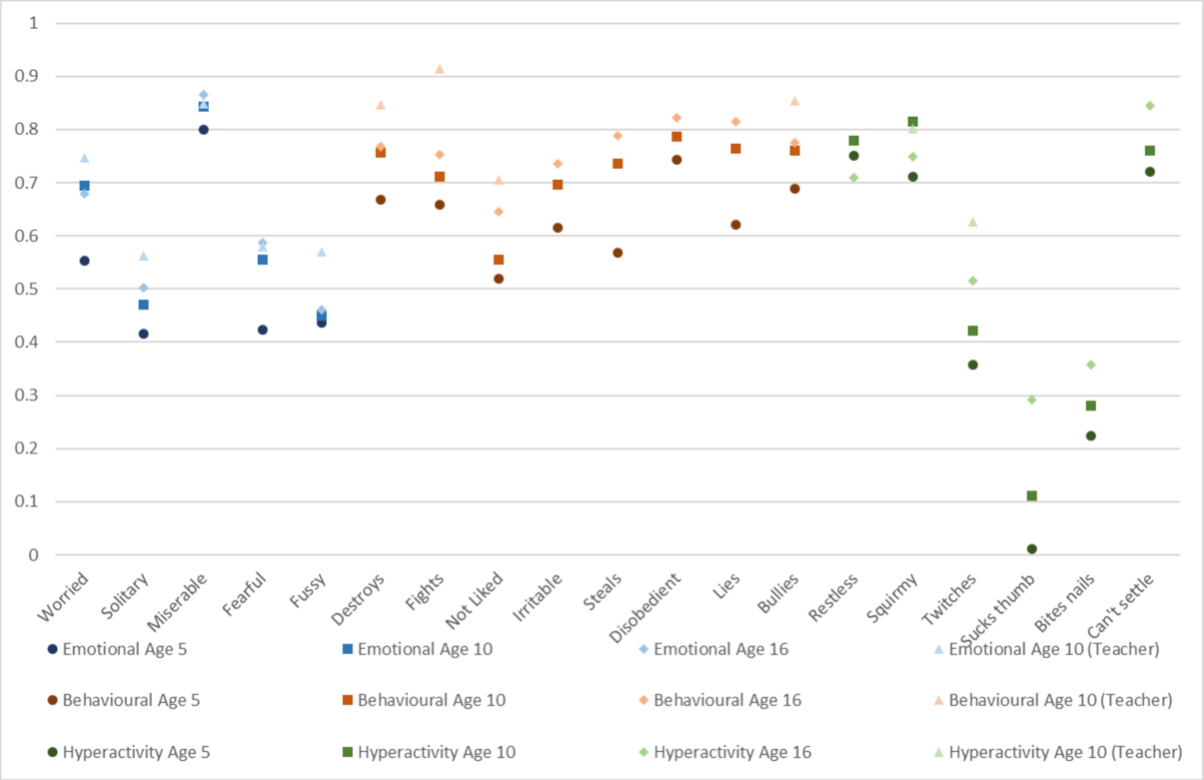


Figure 13. Factor loadings of Rutter items in BCS70

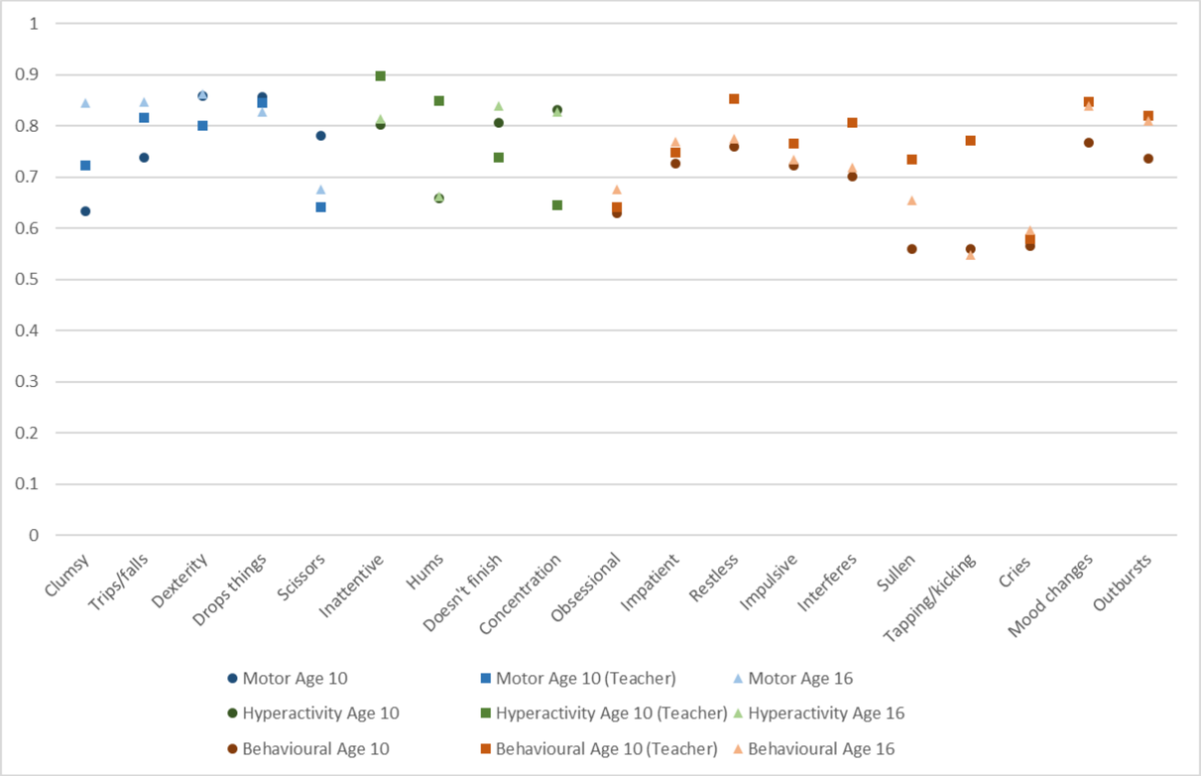


Figure 14. Factor loadings of Conners items in BCS70

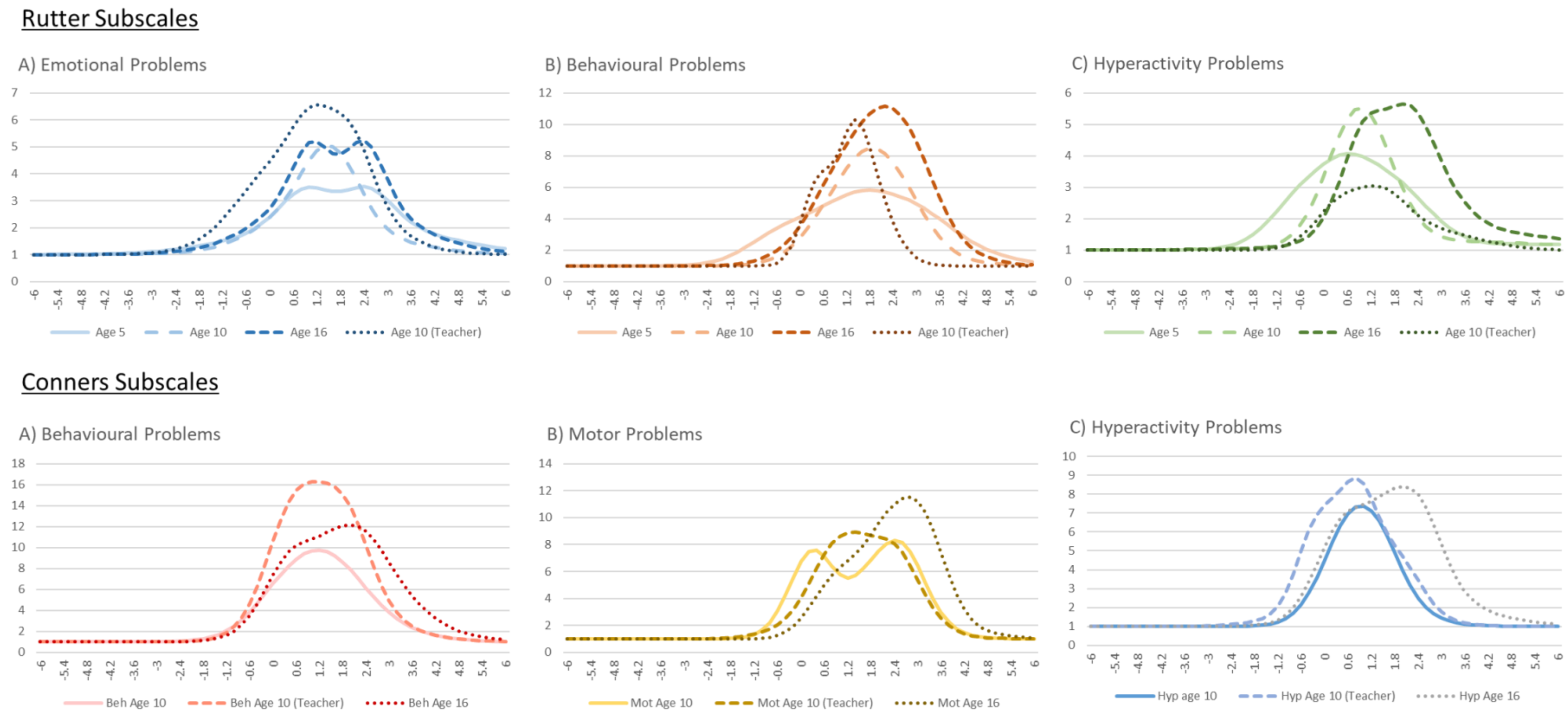


Figure 15. TIFs for the Rutter and Conners subscales in BCS70

The level of the latent factor is presented on the X-axis (expressed in standard deviations from a mean of 0). Fisher information (inverse of standard error of measurement) is presented on the Y-axis, with higher values reflecting greater precision of measurement.

TIFs for the GHQ-12 and Malaise scales (1-factor; psychological distress) are presented in Figure 16. Again, precision was highest at moderate-to-high levels of the latent trait, although the GHQ-12 demonstrated better precision at average and lower ends of the trait, suggesting it may be particularly useful for capturing distress in general populations. Although the effective measurement range was similar for the 22-item and 9-item versions of the Malaise Inventory, the 22-item version had higher precision, which can be attributed to the increased number of items.

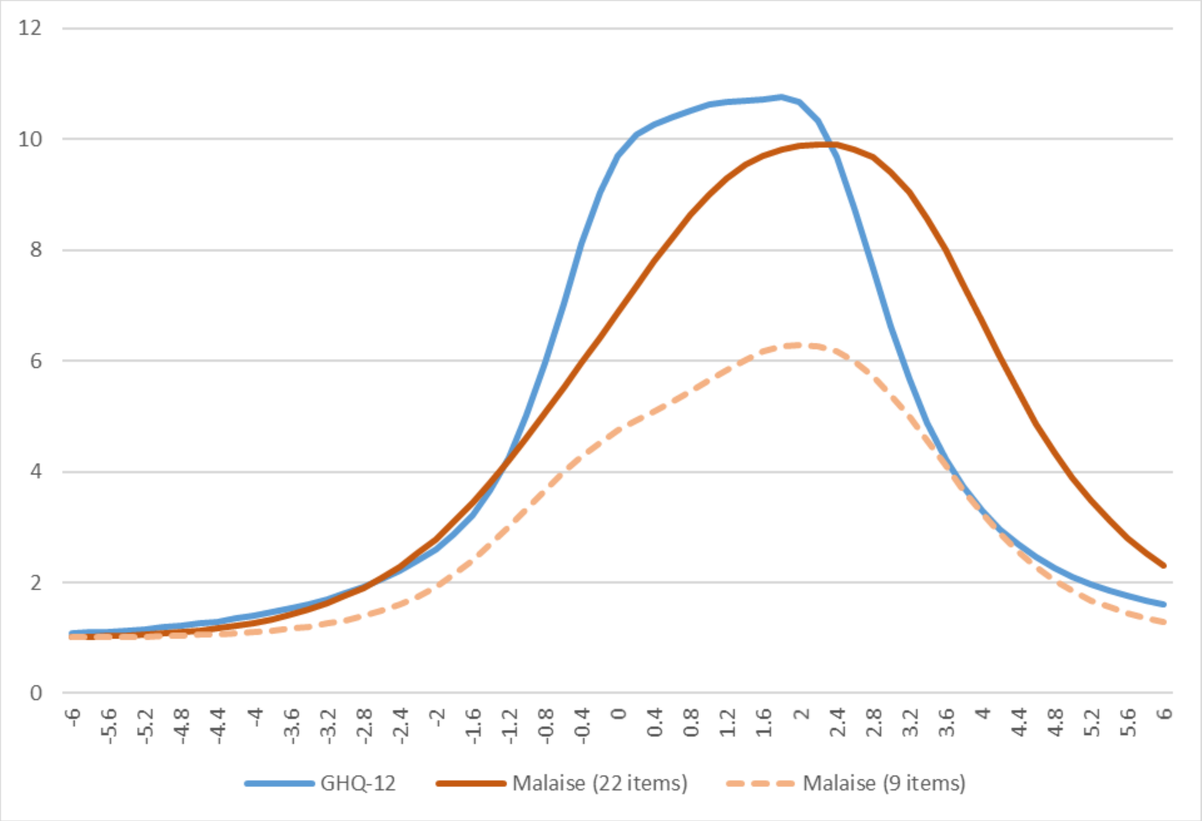


Figure 16. TIFs for the GHQ-12 and Malaise Inventory in BCS70

The level of the latent factor is presented on the X-axis (expressed in standard deviations from a mean of 0). Fisher information (inverse of standard error of measurement) is presented on the Y-axis, with higher values reflecting greater precision of measurement.

3.4 Next Steps

The main measure of psychological distress in Next Steps was the GHQ-12 (Goldberg, 1988). It asks respondents to rate the degree to which they experience a symptom ‘generally’ on a 4-point scale (1=‘Less than usual’, 2=‘No more than usual’, 3=‘Rather more than usual’, or 4=‘Much more than usual’). This was administered at 3 assessment waves: wave 2 (age 15), wave 4 (age 17) and wave 8 (age 25 years). As this was the only measure of general psychological distress administered in this cohort, the measurement properties in both childhood and early adulthood will be discussed in this section. As the GHQ was designed to capture general psychological distress, a unidimensional model was fit to the data at each wave. As in section 3.3.2, two alternative models for the GHQ-12 were fitted based on recent meta-analytic findings regarding the structure of this instrument: a 1-factor model reflecting general psychological distress, and a correlated 2-factor model that account for positively and negatively worded questions (Gnambs & Staufenbiel, 2018). The results are presented in Table 8. As the GHQ was administered only once in adulthood, results from this assessment wave are also presented here. At each wave, only the 2-factor models met the criteria for acceptable fit.

Table 11. Fit statistics for GHQ-12 in Next Steps

Age	Model	N	X^2	DF	RMSEA	CFI	TLI
15	1-factor	13,134	14598.696	54	0.143	0.882	0.856
	2-factor	13,134	5249.405	53	0.086	0.958	0.948
17	1-factor	11,476	10196.964	54	0.128	0.899	0.876
	2-factor	11,476	4790.649*	53	0.088	0.953	0.941
25	1-factor	7,436	8012.058	54	0.141	0.921	0.903
	2-factor	7,436	3449.182	53	0.093	0.966	0.958

ECV statistics (calculated using confirmatory bifactor modelling) were all above 0.7, suggesting a single factor is the dominant source of shared variance, and therefore summing all 12 items of the GHQ to form an overall total scale score can be justified (Rodriguez et al., 2016).

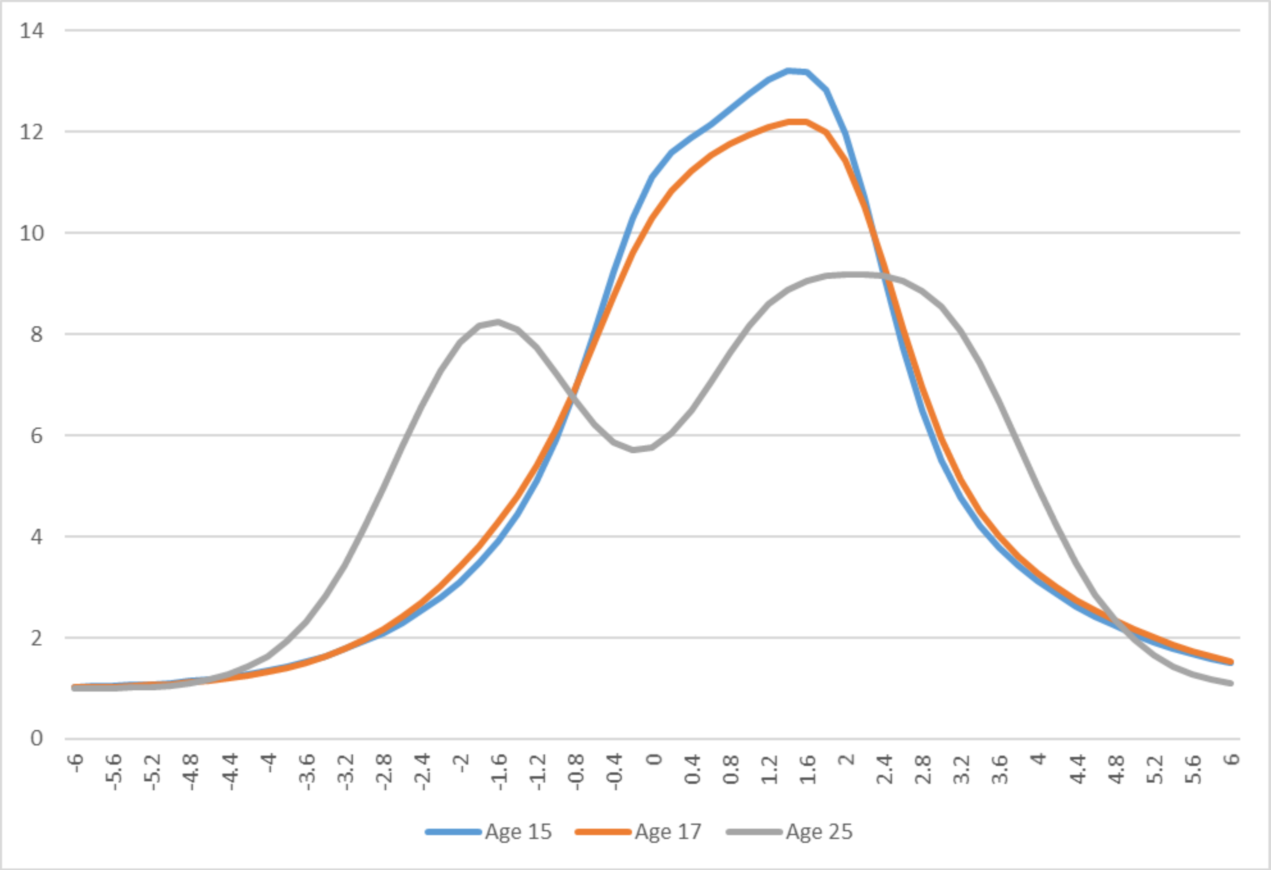


Figure 17. TIFs for the GHQ-12 over time in Next Steps

The level of the latent factor is presented on the X-axis (expressed in standard deviations from a mean of 0). Fisher information (inverse of standard error of measurement) is presented on the Y-axis, with higher values reflecting greater precision of measurement.

Measurement precision (Figure 17) was very similar at ages 15 and 17; information was highest at approximately 1.5 SDs from the mean, suggesting the GHQ is appropriate as a screener for psychological distress in the general population. At age 25, greater precision over a wider range of the latent trait was observed, with more information at the extremes, and less information towards the mean.

3.5 ALSPAC

The earliest measures of child mental health administered in the ALSPAC were the maternal report versions of the Emotionality Activity Sociability Temperament Survey (EAS) (Buss & Plomin, 1984) and the Revised Rutter Parent Scale for Preschool Children (Elander & Rutter, 1996). The EAS is a multi-dimensional measure of temperament; 20 items assess four dimensions (emotionality, activity, shyness and sociability) using 5-point Likert response (1= ‘Not at all’, 5 = ‘Exactly’) (Bould, Joinson, Sterne, & Araya, 2013). The version of the Rutter scales administered in the ALSPAC consisted of 42 items (3-point Likert) that can form 4 subscales (emotional, conduct, hyperactivity and prosocial) or a total behavioural difficulties score (Elander & Rutter, 1996).

Beginning at age 7, the Strengths and Difficulties Questionnaire (SDQ) (Goodman, 1997) became the primary assessment of mental health. This scale consists of 25 items (3-point Likert; 0 = ‘Not true’, 1 = ‘Somewhat true’, 2 = ‘Certainly true’) that assess 5 domains: emotional problems, peer problems, behavioural problems, hyperactivity and prosocial behaviour (Goodman, 1997). The emotional, peer, behavioural and hyperactivity subscales can be summed to form a total difficulties score (Goodman, 1997). Both maternal and teacher versions of the SDQ were administered at different time points.

As children approached adolescence, maternal and self-report versions of the Moods and Feelings Questionnaire (MFQ) were also administered (in conjunction with the SDQ in the case of study mothers). Depending on the assessment wave, either 13-, 16 or 17-item versions of the scale were administered. The MFQ is a unidimensional measure of depressive symptoms (Sharp, Goodyer, & Croudace, 2006), with responses indicated on a 3-point Likert scale (1=‘Not true’, 2=‘Sometimes true’, 3=‘True’).

All of the measures administered in the ALSPAC had established factor structures; therefore these models (along with unidimensional models for comparison) were fitted to the data. Fit statistics are presented separately by measure in Tables 9 - 11.

Table 12. Fit statistics for early childhood mental health measures in ALSPAC

Age	Measure	Reporter	Model	N	X^2	DF	RMSEA	CFI	TLI
3	Emotionality Activity Sociability (EAS)	Mother	1-factor	10082	64592.419	170	0.194	0.623	0.578
			4-factor	10082	18925.372	164	0.107	0.890	0.873
			4-factor*	10082	14250.610	161	0.093	0.917	0.903
3.5	Revised Rutter Parent Scale for Preschool Children (42 items)	Mother	1-factor	10017	14893.689	350	0.064	0.755	0.735
			4-factor	10017	11801.030	371	0.055	0.886	0.875
			4-factor*	10017	10326.742	370	0.052	0.901	0.891
5	Emotionality Activity Sociability (EAS)	Mother	1-factor	9474	58620.891	170	0.191	0.606	0.559
			4-factor	9474	16389.282	164	0.102	0.891	0.873
			4-factor*	9474	11647.801	161	0.087	0.922	0.909

* = correlated residuals.

In order to achieve acceptable fit for the established 4-factor model of the EAS (Bould et al., 2013), we had to allow the residuals from two item-pairs to correlate: i) ‘Friendly to strangers – Takes a long time to warm to strangers’, and ii) ‘Something of a loner – Prefers quiet games’. Similarly for the Rutter scale at age 3.5 years, the residual correlation between ‘Blames others’ and ‘Tells lies’ was estimated in order to obtain acceptable fit.

Table 13. Fit statistics for SDQ administered in childhood in ALSPAC

Age	Measure	Reporter	Model	N	X^2	DF	RMSEA	CFI	TLI
7	SDQ	Mother	1-factor	8459	14076.887	170	0.098	0.719	0.686
			5-factor	8461	7837.653	265	0.058	0.888	0.873
			5-factor*	8461	6813.801	264	0.054	0.903	0.890
8	SDQ	Teacher	1-factor	6364	16574.273	104	0.158	0.797	0.766
			5-factor	6364	7251.757	179	0.079	0.948	0.938
9	SDQ	Mother	1-factor	8111	13428.956	170	0.098	0.730	0.699
			5-factor	8113	6863.042	265	0.055	0.891	0.877
			5-factor*	8113	6005.749	264	0.052	0.906	0.893
11	SDQ	Teacher	1-factor	7656	18975.590	104	0.154	0.805	0.775
			5-factor	7656	8350.166	179	0.077	0.948	0.939
11	SDQ	Mother	1-factor	7393	12542.058	170	0.099	0.742	0.711
			5-factor	7397	6334.369	265	0.056	0.900	0.887
13	SDQ	Mother	1-factor	7087	12664.439	170	0.102	0.732	0.700
			5-factor	7089	7021.749	265	0.060	0.891	0.876
			5-factor*	7089	5631.711	264	0.054	0.913	0.901
16	SDQ	Mother	1-factor	5693	9504.782	170	0.098	0.742	0.711
			5-factor	5693	5712.060	265	0.060	0.891	0.877
			5-factor*	5693	5115.032	263	0.057	0.903	0.889

* = correlated residuals.

The 5-factor model of the SDQ was above the acceptable thresholds on all three fit indices, provided the residuals between items relating to ‘fidgeting/squirming’ and ‘restless/overactive’ were correlated. Unidimensional had poor fit, suggesting that the use of subscales should be preferred over total/overall difficulty scores.

Table 14. Fit statistics for MFQ administered in childhood in ALSPAC

Age	Measure	Reporter	M	N	X^2	DF	RMSEA	CFI	TLI
9	MFQ	Mother	1-factor	8074	1709.146	65	0.056	0.967	0.960
10	MFQ	Child	1-factor*	7409	996.498	113	0.032	0.976	0.971
11	MFQ	Mother	1-factor	7363	1300.802	65	0.051	0.975	0.970
12	MFQ	Child	1-factor*	6773	1134.260	101	0.039	0.979	0.975
13	MFQ	Mother	1-factor	7104	1334.442	65	0.052	0.977	0.972
	MFQ	Child	1-factor	6076	1839.992	101	0.053	0.971	0.965
16	MFQ	Mother	1-factor	5683	1267.589	65	0.057	0.976	0.971
	MFQ	Child	1-factor	5094	2976.932	113	0.071	0.967	0.960

* = correlated residuals.

For the maternal-report versions of the MFQ, only 13 items were administered, all of which were negatively worded (e.g. “Teenager felt miserable or unhappy”). For the self-report versions, between 16 and 17 questions were asked, with the addition of 3-4 positively worded items (e.g. “Teenager has felt happy”). The 13-item maternal-report version of the MFQ demonstrated excellent fit across assessments. For the self-report measures, in order to achieve excellent levels of fit (loading all 16/17 items onto a single psychological distress factor), we had to allow the residuals between the positively worded items to correlate.

TIFs for the EAS, SDQ and MFQ are presented in Figures 18, 19, and 20, respectively. The EAS emotionality scale demonstrated good precision over a wide range of the latent trait (approximately ± 2 SDs from the mean); however the activity, shyness and sociability scales had higher precision at relatively lower levels of the trait.

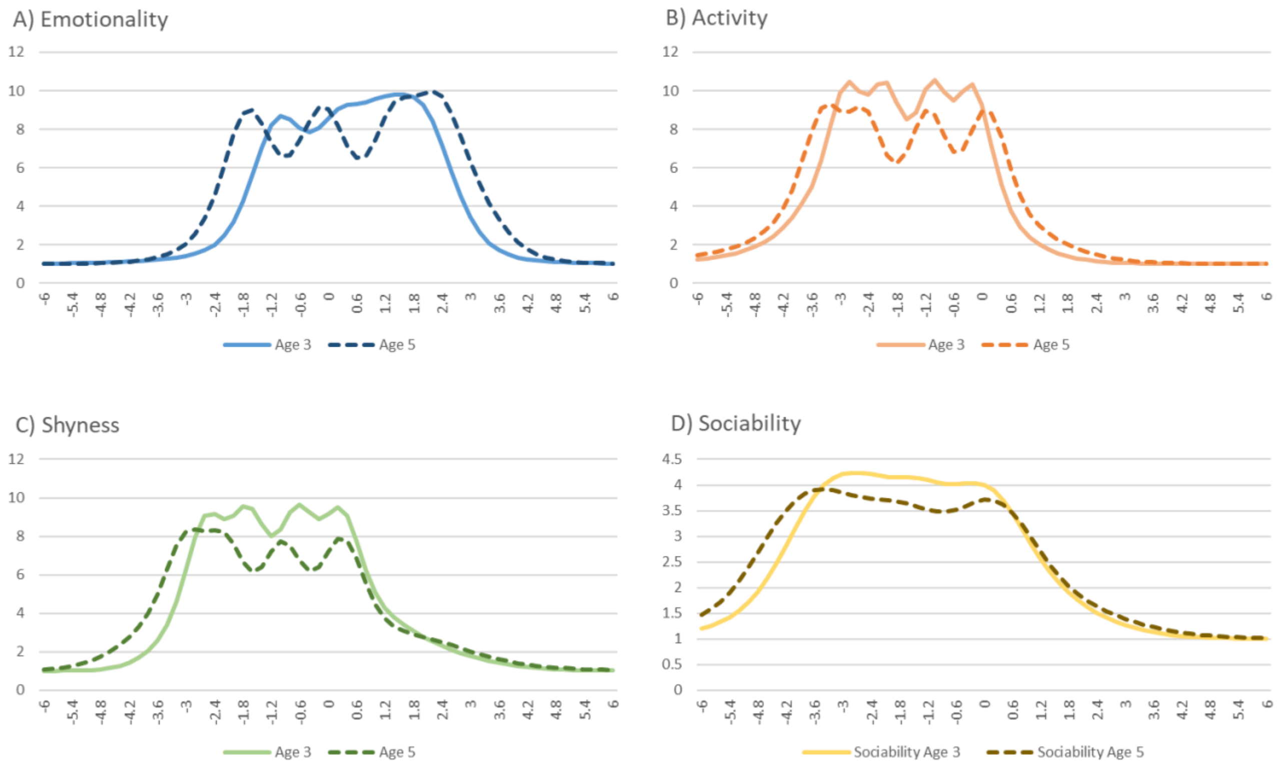


Figure 18. TIFs for the EAS at ages 3 and 5 years in ALSPAC

The level of the latent factor is presented on the X-axis (expressed in standard deviations from a mean of 0). Fisher information (inverse of standard error of measurement) is presented on the Y-axis, with higher values reflecting greater precision of measurement.

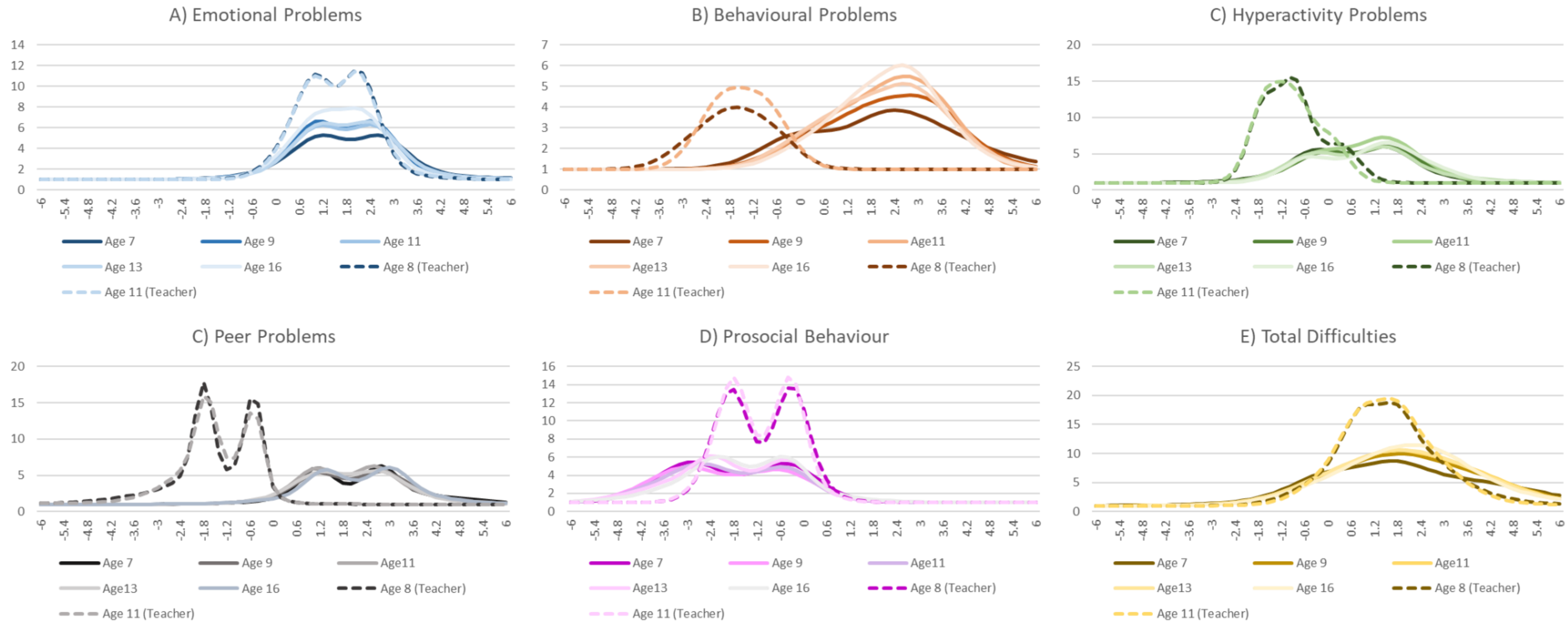


Figure 19. TIFs for the SDQ subscales administered in childhood in ALSPAC

The level of the latent factor is presented on the X-axis (expressed in standard deviations from a mean of 0). Fisher information (inverse of standard error of measurement) is presented on the Y-axis, with higher values reflecting greater precision of measurement.

The various SDQ subscales demonstrated similar levels of precision across assessment waves, however there were notable differences by rater. Parents had higher precision at higher levels of the latent trait for behavioural, hyperactivity and peer problems. Information was high over a similar range of the latent trait for the emotional and prosocial scales, with teachers providing more information.

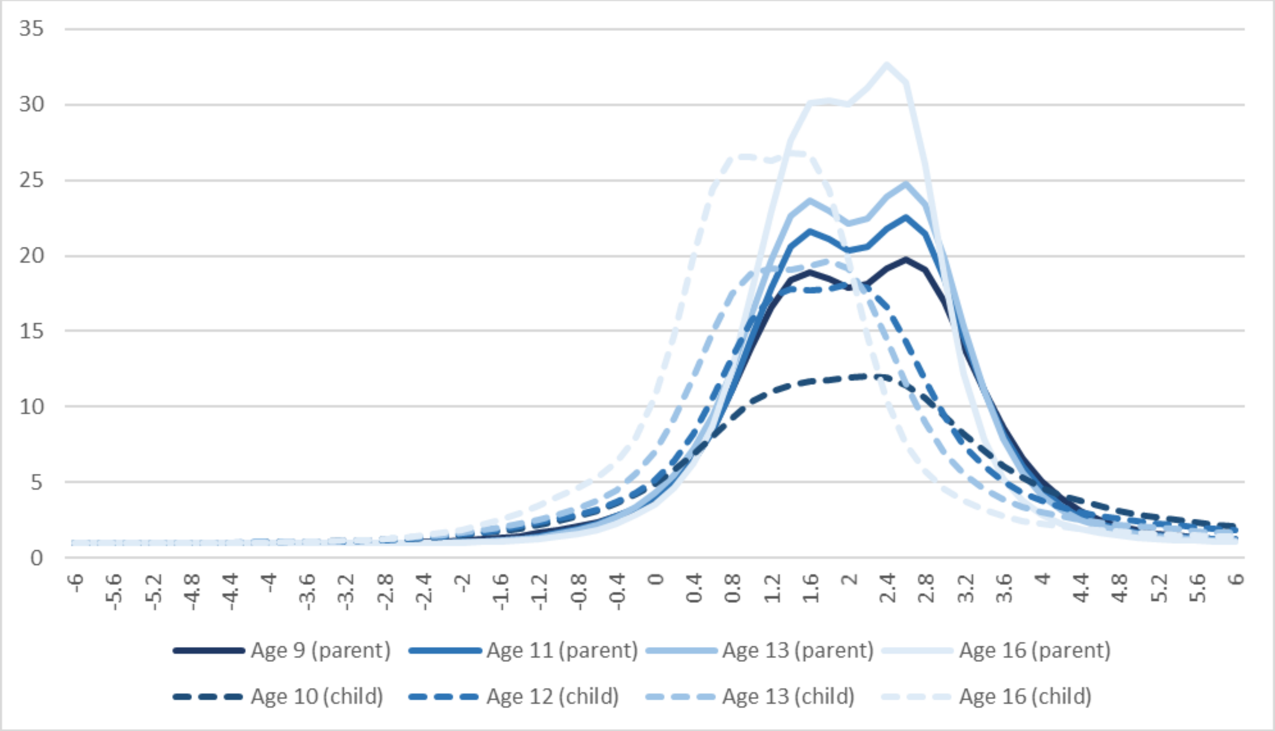


Figure 20. TIFs for the MFQ administered in childhood in ALSPAC

The level of the latent factor is presented on the X-axis (expressed in standard deviations from a mean of 0). Fisher information (inverse of standard error of measurement) is presented on the Y-axis, with higher values reflecting greater precision of measurement.

The MFQ (Figure 20) had highest precision (i.e. lowest standard errors) at moderate-to-high levels of the latent trait (0.8-3.5 SDs above the mean). Precision increased as children aged, with similar levels and ranges of precision observed across both parent and child reporters.

3.6 MCS

The primary measure of child/adolescent mental health in the MCS was the parental report version of the Strengths and Difficulties Questionnaire (SDQ) (Goodman, 1997). This scale consists of 25 items (3-point Likert; 0 = ‘Not true’, 1 = ‘Somewhat true’, 2 = ‘Certainly true’) that assess 5 domains: emotional problems, peer problems, behavioural problems, hyperactivity and prosocial behaviour (Goodman, 1997). The emotional, peer, behavioural and hyperactivity subscales can be summed to form a total difficulties score (Goodman, 1997). Two versions of this questionnaire were administered in the MCS (SDQ 2-4 and SDQ 4-17). At age 3, the SDQ 2-4 year version of the questionnaire was administered. This version is almost identical to the 4-17 version of the scale, with the exception of two different items: i) 'Often argumentative with adults' replaces 'Often lies or cheats', and ii) 'Can be spiteful to others' replaces 'Steals from home, school or elsewhere'.

Five-factor and 1-factor (total difficulty) models were fitted to the data at each wave. The MCS uses a complex survey design. Appropriate weights were used to account for the complex survey design of the MCS (i.e. stratified, clustered random sample design, and oversampling from areas that were disadvantaged or had high ethnic minority populations).

Table 15. Fit statistics for mental health measures administered in childhood in MCS

Age	Measure	Model	N	X2	DF	RMSEA	CFI	TLI
3	SDQ	1-factor (total difficulties)	14836	9635.322	170	0.061	0.797	0.774
		5-factor	14836	8142.602	265	0.045	0.860	0.841
5	SDQ	1-factor (total difficulties)	14773	9261.052	170	0.060	0.820	0.798
		5-factor	14773	5696.481	265	0.037	0.914	0.902
7	SDQ	1-factor (total difficulties)	13489	9911.518	170	0.065	0.834	0.815
		5-factor	13489	5162.176	265	0.037	0.929	0.920
11	SDQ	1-factor (total difficulties)	12821	9704.616	170	0.066	0.817	0.795
		5-factor	12821	4454.660	265	0.035	0.925	0.915
14	SDQ	1-factor (total difficulties)	11267	22095.521	275	0.084	0.755	0.733
		5-factor	11267	7971.206	265	0.051	0.913	0.902

At each wave, the 5-factor models provided good fit the data well. One-factor models (reflecting total difficulties and consisting only of items from the emotional, peer, behavioural and hyperactivity scales) demonstrated acceptable fit on the RMSEA, but not the CFI and TLI. These findings suggest that subscales should be preferred over the total difficulties score.

TIFs for the SDQ subscale and total difficulties scales are presented in Figure 21. Relatively similar patterns of precision were observed across ages, with greater measurement precision observed as children grew older. The emotional, behavioural, peer and hyperactivity subscales were most precise (i.e. had lowest standard error) at moderate-to-high levels of the latent factor (approximately two SDs above the mean), whereas the prosocial scale demonstrated the reverse pattern (higher levels of information at the lower end of the trait).

Teacher report versions of the SDQ were administered in the MCS at ages 7 and 11 (for further details see section 10.2)

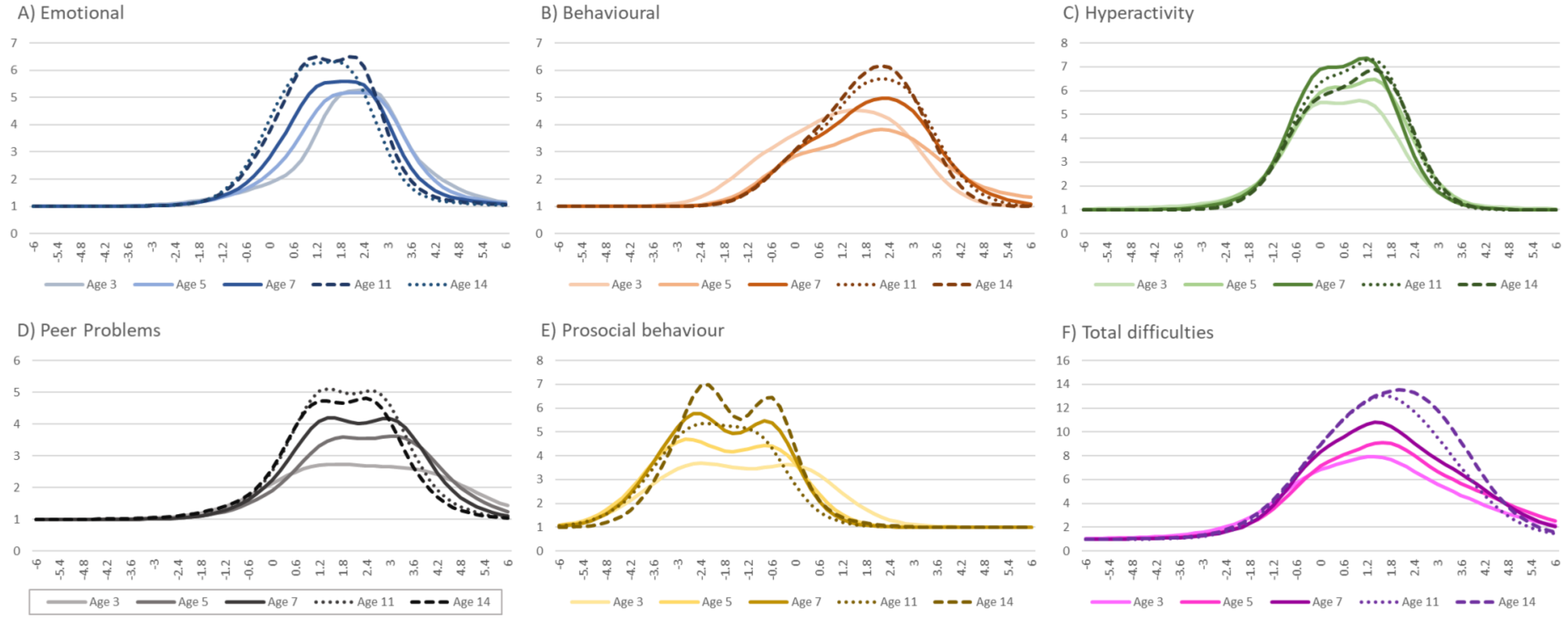


Figure 21. TIFs for the SDQ subscales administered in childhood in MCS

The level of the latent factor is presented on the X-axis (expressed in standard deviations from a mean of 0). Fisher information (inverse of standard error of measurement) is presented on the Y-axis, with higher values reflecting greater precision of measuremen

4. Measurement Properties in Adulthood

4.1 NSHD

At the age 36 assessment of the NSHD, study participants completed a shortened version of the Present State Examination (PSE) (Wing, Cooper, & Sartorius, 2011). This semi-structured interview was administered by trained interviewers. It assessed 48 symptoms, 41 of which were rated on a 3-point Likert scale ('0 = Not present', '2 = Clinically intense') and can be summed to form a total PSE score (Rodgers & Mann, 1986). At age 43, the Psychiatric Symptom Frequency Scale (PSF) (Lindelow, Hardy, & Rodgers, 1997) was delivered by trained nurses. This interview was developed to assess affective symptoms in the general population, and measures 18 symptoms using a 6-point Likert response format ('0 = Never', '5 = Always'). It is worth noting that this questionnaire asks participants to indicate responses based on how they have been feeling over the *past year,* a time frame of reference considerably longer than most of the other mental health measures in adulthood. At ages 53, 60-64 and 68-70 years, participants completed the General Health Questionnaire 28-item version (GHQ-28) (D. Goldberg, 1978), a self-report measure of emotional distress. Each item is rated on a 4-point Likert scale ('0 = Not at all, 3 = Much more than usual'). Furthermore, the Short form 36 (SF36) (Jenkinson, Coulter, & Wright, 1993) health survey questionnaire was also administered at age 50. The SF36 is a measure of health-related quality of life, and contains a mental health subsection consisting of 10 questions that assess general psychological wellbeing. Each response is indicated on a 6-point ordinal scale.

In line with the recommended scoring systems of the PSE and PSF (Lindelow et al., 1997; Rodgers & Mann, 1986; Wing et al., 2011), simple 1-factor models were estimated in which all symptoms were treated as measured indicators of a general psychological distress factor. For the GHQ, a 4-factor (somatic; anxiety; social dysfunction, and depression) structure was fitted to the data based on the extant factor analytic literature (Werneke, Goldberg, Yalcin, & Üstün, 2000). Fit statistics are presented in Table 13.

Table 16. Fit statistics for adulthood mental health measures administered in adulthood in NSHD

Age	Measure	Model	N	X^2	DF	RMSEA	CFI	TLI
36	PSE	1-factor	3317	4971.87	779	0.040	0.811	0.801
43	PSF	1-factor	3247	3274.79	135	0.085	0.891	0.877
		1-factor*	3247	2260.15	134	0.070	0.926	0.916
53	GHQ-28	1-factor	2964	16213.70	350	0.124	0.830	0.817
		4-factor	2964	6541.09	344	0.078	0.934	0.927
60-64	GHQ-28	1-factor	2227	11074.10	350	0.117	0.797	0.781
		4-factor	2227	4749.610	344	0.076	0.917	0.908
68-70	GHQ-28	1-factor	2144	10361.30	350	0.116	0.811	0.796
		4-factor	2144	4328.68	344	0.074	0.925	0.917

*Correlated residuals between items 9 (‘Have you had trouble getting off to sleep’) and 10 (‘have you had trouble with waking up and not being able to get back to sleep?’)

Overall, the established factor structures were well supported in the NSHD data. The 4-factor model of the GHQ was above the cut-off for acceptable fit on all three indices. The 1-factor model of the PSF also demonstrated acceptable fit, after the residual correlation between items 9 (‘Have you had trouble getting off to sleep’) and 10 (‘have you had trouble with waking up and not being able to get back to sleep?’) was estimated. The 1-factor model of the PSE fit well based on the RMSEA; however the CFI and TLI were below the recommended values for acceptable fit. An inspection of the modification indices, however, revealed no indices above the minimum value, which suggests that there were no meaningful correlations between the residual variances of items.

TIFs for the various adult measures of mental health in the NSHD are presented in Figure 22. The PSE had higher precision (i.e. lower standard error) at higher levels of the latent trait, and relatively low precision at mean/lower levels. This is perhaps unsurprising given it was a semi-structured clinical interview. The GHQ and PSF had highest precision at moderate levels of the latent trait (0-2 SDs above the mean), which suggests they are reliable measures for assessing emotional distress in the general population. However, the GHQ demonstrated higher precision over a similar range when compared with the PSF.

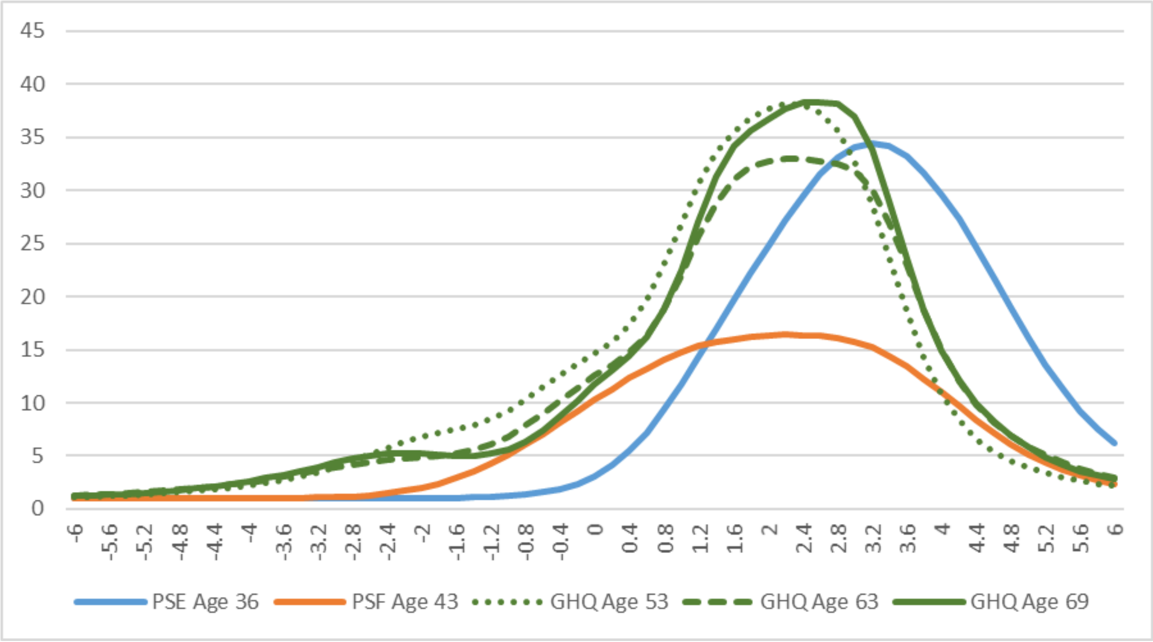


Figure 22. TIFs for mental health measures administered in adulthood in NSHD

*The level of the latent factor **θ** (expressed in standard deviations from a mean of 0) is presented on the X-axis. Fisher information (inverse of standard error of measurement) is presented on the Y-axis, with higher values reflecting greater reliability.*

4.2 NCDS

The most frequently administered measure of mental health in the NCDS was the Malaise Inventory (Rutter et al., 1970). This 24-item measure is designed to assess general psychological distress, with items scored using a simple 'Yes/No' response. An abbreviated version of the scale (consisting of 9 of the original 24 items) was administered at age 50. At age 42, the GHQ-12 was administered alongside the Malaise Inventory (see section 3.3 for a description). Furthermore, the Short form 36 (SF36) (Jenkinson et al., 1993) health survey questionnaire was also administered at age 50. The SF36 is a measure of health-related quality of life, and contains a mental health subsection consisting of 10 questions that assess general psychological wellbeing. Each response is indicated on a 6-point ordinal scale.

As the Malaise Inventory is established as a unidimensional measure of psychological distress, 1-factor models were fitted to both the 24- and 9-item versions of the scale. In line with the existing literature (Gnambs & Staufenbiel, 2018), unidimensional and a correlated 2-factor models (accounting for positively and negatively worded questions) were fitted to the GHQ-12 data. A unidimensional factor (psychological wellbeing) was fit to the wellbeing subscale of the SF36 as per its scoring manual (Jenkinson et al., 1993). Results from the CFAs are presented in Table 14.

Table 17. Fit statistics for mental health measures administered in adulthood in NCDS

Age	Measure	Model	N	X2	DF	RMSEA	CFI	TLI
23	Malaise (24 item)	1-factor	12,490	4484.405	252	0.037	0.913	0.905
	Malaise (9 item)	1-factor	12,489	280.740	27	0.027	0.987	0.983
33	Malaise (24 item)	1-factor	11,343	6257.600	252	0.046	0.892	0.881
	Malaise (9 item)	1-factor	11,343	303.595	27	0.030	0.989	0.985
42	Malaise (24 item)	1-factor	11,277	8240.915	252	0.053	0.882	0.870
	Malaise (9 item)	1-factor	11,276	412.392	27	0.036	0.986	0.982
42	GHQ-12	1-factor	11,279	12327.463	54	0.142	0.899	0.877
		2-factor	11,279	5485.809	53	0.095	0.955	0.944
50	Malaise 9-item	1-factor	9,634	354.294	27	0.035	0.992	0.989
50	SF36	1-factor	8,762	15590.326	35	0.225	0.919	0.896

Unidimensional models fit the Malaise data well. Indeed, the fit statistics were all in the ‘excellent’ range for the 9-item version of the scale across time. The 2-factor model of the GHQ-12 fit the data well. To further explore the dimensionality of the GHQ-12, the explained common variance (ECV) was calculated by fitting a bifactor model (1 general factor, 1 positively worded factor, 1 negatively worded factor) to the data. The ECV was 0.68, indicating a general factor accounted for the majority of shared variance, therefore summing all twelve items to form a total score is justified.

TIFs for the various measures are presented in Figure 23. The Malaise Inventory demonstrated highest levels of precision at moderate-to-high levels of the latent trait (approximately 2 SDs above the mean). The GHQ-12 demonstrated similar levels of precision at this level of the trait, but greater levels of precision than the Malaise at lower ends of psychological distress. The SF36 had particularly high levels of precision at low ends of the trait, suggesting that for this age group it is perhaps more reliable at capturing psychological wellbeing than distress.

TIFs for the 9-item version of the Malaise Inventory (ages 23 - 50) are presented in Figure 24. Again, the 9-item version had the highest measurement precision at moderate-to-high levels of the latent trait (approx. 2 SDs from the mean), which indicates that the Malaise is particularly reliable when measuring moderately high levels of psychological distress, and therefore can be considered a reliable measure for use in general populations. Measurement precision also appeared to increase as study members aged.

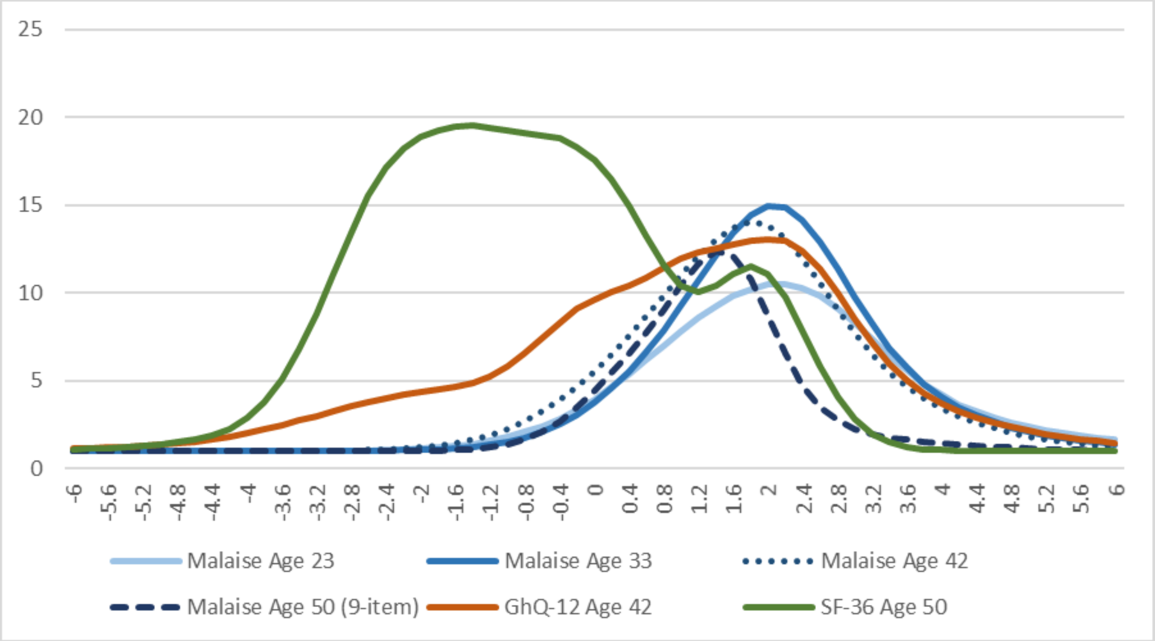


Figure 23. TIFs for mental health measures administered in adulthood in NCDS

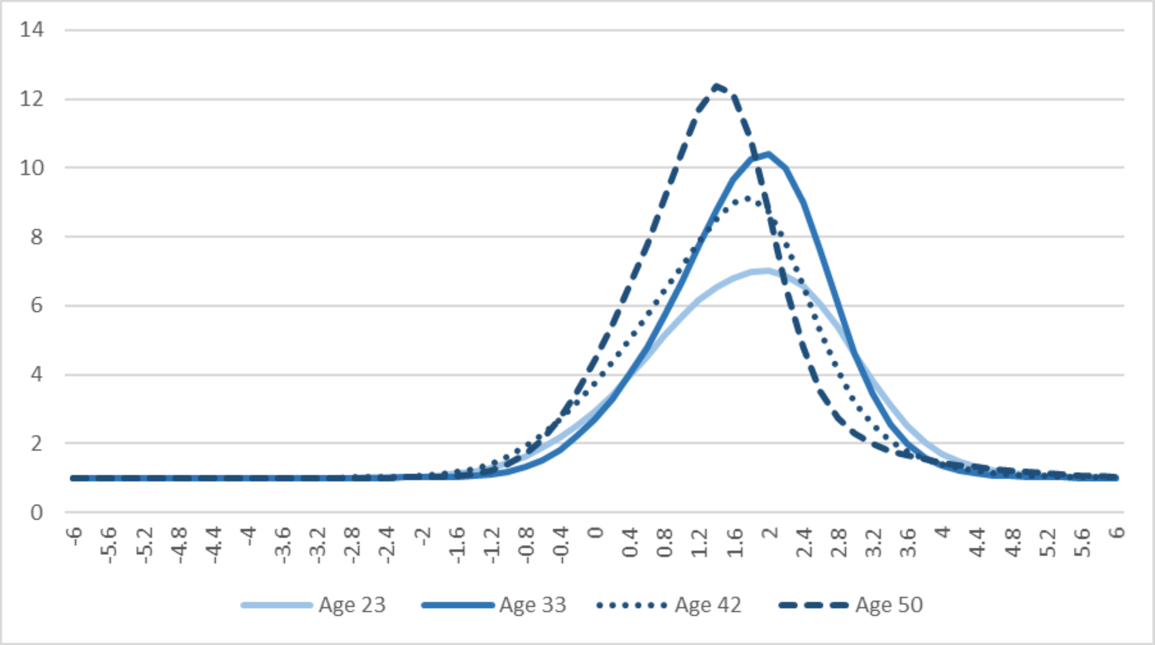


Figure 24. TIFs for Malaise Inventory (9-item version) in NCDS

4.3 BCS70

As with NCDS, the most frequently administered measure of mental health in the BCS70 was the Malaise Inventory (Rutter et al., 1970). This 24-item measure is designed to assess general psychological distress, with items scored as 'Yes/No'. The 24-item version of the scale was administered at ages 26 and 30, with the 9-item version used thereafter at ages 34 and 42. At age 30, the GHQ-12 was administered alongside the Malaise Inventory (see section 3.3 for a description). At age 34, 4 items from the Kessler Psychological Distress Scale (Kessler et al., 2002) were administered along with the Malaise. The Kessler scale is a measure of general psychological distress for use in large population-based surveys population.

Each measure in administered in adulthood in the BCS70 was designed to assess general psychological distress; therefore 1-factor models were fitted. A 2-factor model was also fitted to the GHQ in order to capture positively and negatively worded items (Gnambs & Staufenbiel, 2018). Fit statistics are presented in Table 15.

Table 18. Fit statistics for mental health measures administered in adulthood in BCS70

Age	Measure	Model	N	X^2	DF	RMSEA	CFI	TLI
26	Malaise (24 item)	1-factor	8,968	4690.517	252	0.044	0.892	0.882
	Malaise (9 item)	1-factor	8,962	362.629	27	0.037	0.980	0.973
30	Malaise (24 item)	1-factor	11,112	5484.803	252	0.043	0.908	0.899
	Malaise (9 item)	1-factor	11,111	440.976	27	0.037	0.982	0.976
	GHQ-12	1-factor	11,115	10933.917	54	0.135	0.879	0.852
		2-factor	11,115	4560.733	53	0.087	0.950	0.938
34	Malaise (9 items)	1-factor	9,598	396.676	27	0.038	0.987	0.983
34	Kessler scale (4 items)	1-factor	9,596	530.794	2	0.166	0.986	0.959
42	Malaise (9 items)	1-factor	8,636	395.056	27	0.040	0.988	0.984

For each measure, 1-factor models reflecting general psychological distress had acceptable or near-acceptable levels of fit. The fit of the 1-factor model was particularly good for the 9-item version of the Malaise. A 2-factor model provided acceptable fit for the GHQ. To further explore the dimensionality of the GHQ-12, the explained common variance (ECV) was calculated by fitting a bifactor model (1 general factor, 1 positively worded factor, 1 negatively worded factor) to the data. The ECV was 0.60, indicating a general factor accounted for the majority of shared variance, therefore summing all twelve items to form a total score is justified.

TIFs for the measures are presented in Figure 25. Again, the Malaise Inventory had the highest measurement precision at moderate-to-high levels of the latent trait (approx. 2 SDs from the mean), and therefore can be considered a reliable measure for use in general populations. As with NCDS, the GHQ-12 had similarly high levels of precision to the Malaise at moderate-to-high levels of the trait, but with higher precision at average and lower levels of the trait. The four-item version of the Kessler scale had good precision at lower levels of the trait, but poor precisions at higher levels. This appears to contradict its intended use as a screener for severe psychological disorders (Kessler et al., 2002).

TIFs for the Malaise 9-item versions are presented in Figure 26. As was the case in the NCDS, the 9-item version had the highest measurement precision at moderate-to-high levels of the latent trait (approx. 2 SDs from the mean), which indicates that the Malaise (9-item) is particularly reliable when measuring moderately high levels of psychological distress, and therefore can be considered a reliable measure for use in general populations. Again, measurement precision also appeared to increase as study members aged.

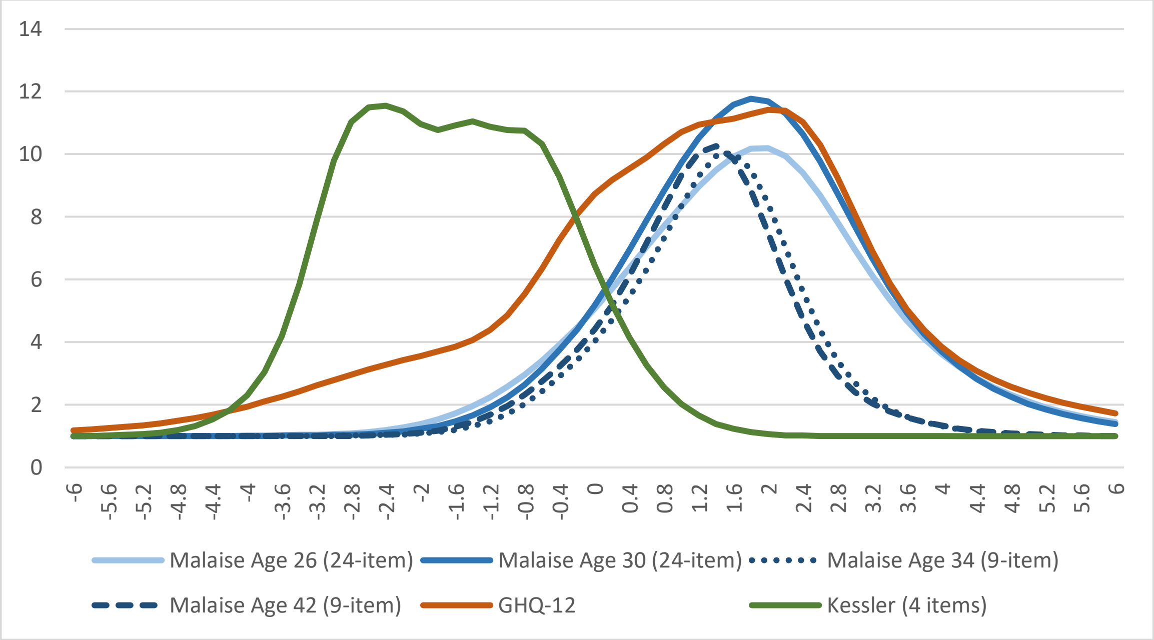


Figure 25. TIFs for mental health measures administered in adulthood in BCS70

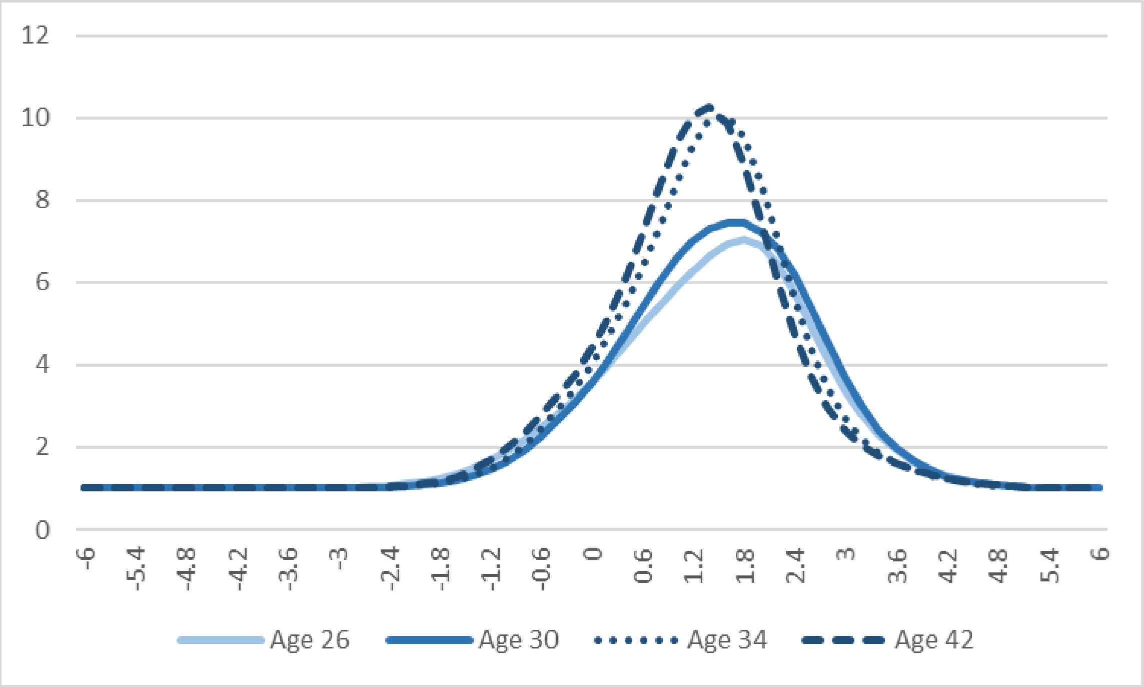


Figure 26. TIFs for Malaise Inventory (9-item version) in BCS70

4.4 ALSPAC

The main measures of psychological distress in adulthood that are available in the ALSPAC are the self-report versions of the Moods and Feelings Questionnaire (MFQ) and the 10-item mental health subscale of the Short form 36 (SF36) (Jenkinson et al., 1993). The MFQ is a unidimensional measure of depressive symptoms (Sharp et al., 2006), with responses indicated on a 3-point Likert scale ('Not true', 'Sometimes true', 'True'). The SF36 mental health subsection assess general psychological wellbeing. Each response is indicated on a 6-point ordinal scale. Given both measures capture general psychological distress, unidimensional factor models were tested.

Table 19. Fit statistics for mental health measures administered in early adulthood in ALSPAC

Age	Measure	Model	N	X^2	DF	RMSEA	CFI	TLI
18	MFQ	1-factor	3354	1700.229	65	0.087	0.969	0.963
	SF36	1-factor	3356	2963.476	29	0.174	0.921	0.877
21	MFQ	1-factor	3405	1236.534	65	0.073	0.980	0.976
	SF36	1-factor	3297	2687.632	29	0.167	0.950	0.923
22	MFQ	1-factor*	3978	6412.770	128	0.111	0.914	0.898
23	MFQ	1-factor*	4092	5940.501	129	0.105	0.937	0.925

*Correlated residuals for positively worded items.

A one-factor model of the MFQ (Table 16) demonstrated excellent fit at ages 18 and 21. At these assessments, only 13 negatively worded items were administered. At the 22 and 23 year assessments, additional positively worded items were included. In order to achieve acceptable fit, residual correlations between these positively worded items were included in the model.

The mental health subscale of the SF36 demonstrated acceptable fit according to the CFI and TLI, however not according to the RMSEA.

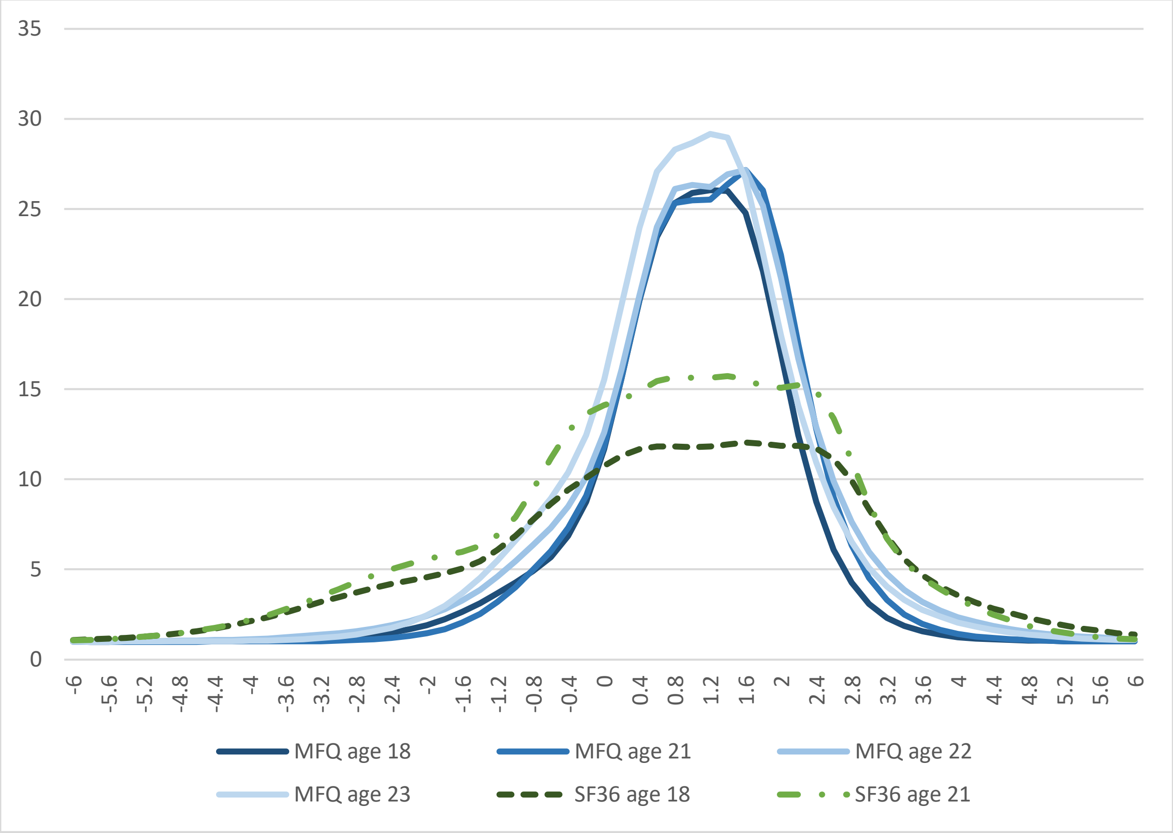


Figure 27. TIFs for MFQ and SF36 (mental health subscale) administered in early adulthood in ALSPAC

TIFs for the MFQ (Figure 27) were highly consistent across adulthood in the ALSPAC. Measurement precision was highest at moderate levels of the latent trait (approximately 1.2 SDs from the mean). The SF36 demonstrated a similar pattern of precision.

5. The Strengths and Difficulties Questionnaire: A longitudinal examination of the measurement equivalence in the ALSPAC (1990-1992) and MCS (2000-2002) cohorts

The second aim of this report was to assess the psychometric equivalence of mental health measures that were administered within and across cohorts. Here we present an example in childhood, focussing on the Strengths and Difficulties Questionnaire (SDQ) (Goodman, 1997). Parental-report versions of the SDQ were administered in both the ALSPAC (at ages 7, 9, 11, 13 and 16) and MCS (3, 5, 7, 11 and 14). There were three instances in which these measurement occasions overlapped across the two cohorts: age 7, 11, and 13/14 years. We therefore tested the measurement invariance of the SDQ across both age and cohort using MGCFA. First, we merged data from the MCS and ALSPAC into a single file, and created a grouping variable that denoted every age x cohort permutation. This resulted in 6 groups; $\text{ALSPAC}_{\text{age7}}$, $\text{ALSPAC}_{\text{age11}}$, $\text{ALSPAC}_{\text{age13}}$, MCS_{age7}, $\text{MCS}_{\text{age11}}$, $\text{MCS}_{\text{age14.}}$ We then tested for configural, metric and scalar invariance based on these groupings, assuming a 5-factor model of the SDQ (emotional problems, peer problems, behavioural problems, hyperactivity and prosocial behaviour). The results are presented in Table 17.

Table 20. Fit statistics for MGCFA models of SDQ in ALSPAC and MCS

Model	N	X^2 (df)	RMSEA	CFI	TLI	∆RMSEA	∆CFI	∆TLI
Configural	60,290	52582.334 (1590)	0.056	0.907	0.895			
Metric		43167.537 (1690)	0.049	0.925	0.920	0.007	0.018	0.025
Scalar		56879.747 (1790)	0.055	0.900	0.899	0.001	0.007	0.004

Previous analyses (sections 3.5 and 3.6) indicated that the 5-factor model of the SDQ fit the data well at all ages in the MCS and ALSPAC. The full longitudinal scalar model of the

SDQ did not result in a worsening of overall fit compared to the configural/baseline model (Table 17). Therefore measurement invariance for the SDQ was supported across both age and cohort. This indicates that the interpretation of items by participants across cohorts or age groups did *not* influence the observed scores on this measure; therefore the SDQ can reliably be compared in terms of both covariances and means across age groups and cohorts. An inspection of the factor loadings and thresholds from the configural model (Figures 28 and 29 respectively) corroborates this; the measurement parameters for each individual question in the SDQ were highly similar across age and cohorts.

Overall these findings demonstrate that, in the MCS and ALSPAC cohorts, the scores on the SDQ measure were not biased due to such factors as age, survey design, period effects, or cohort effects, and therefore can reliably be compared across cohorts.

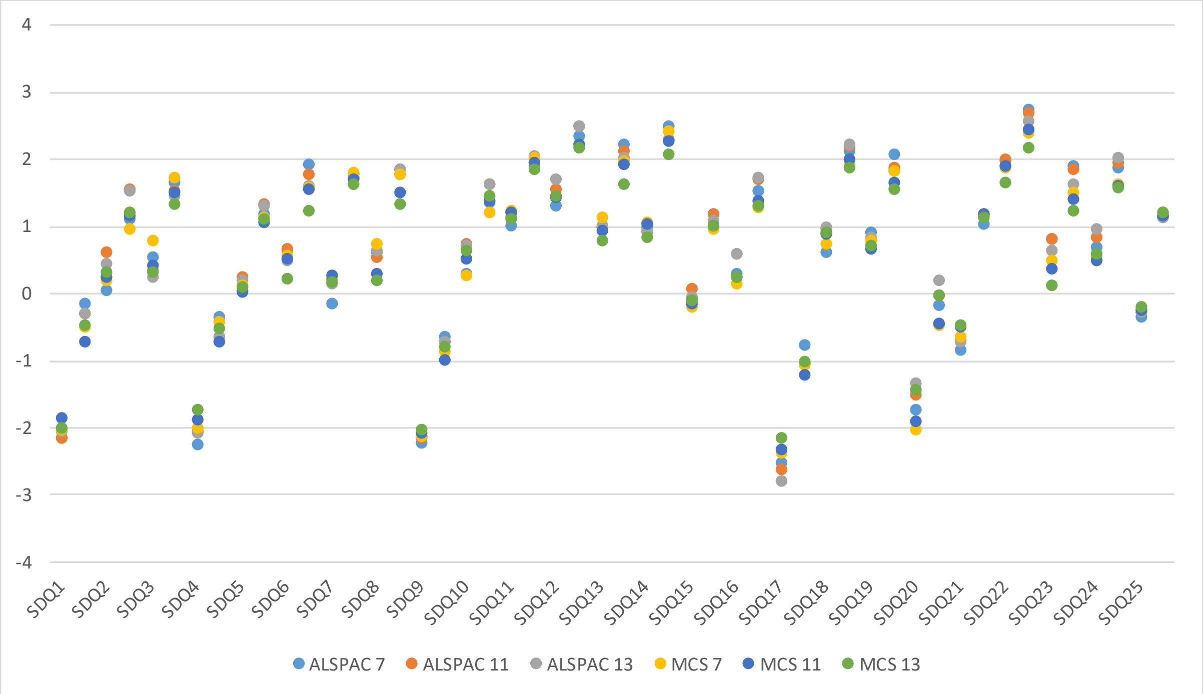


Figure 28. Standardised threshold parameters of SDQ items in MCS and ALPSAC (configural model)

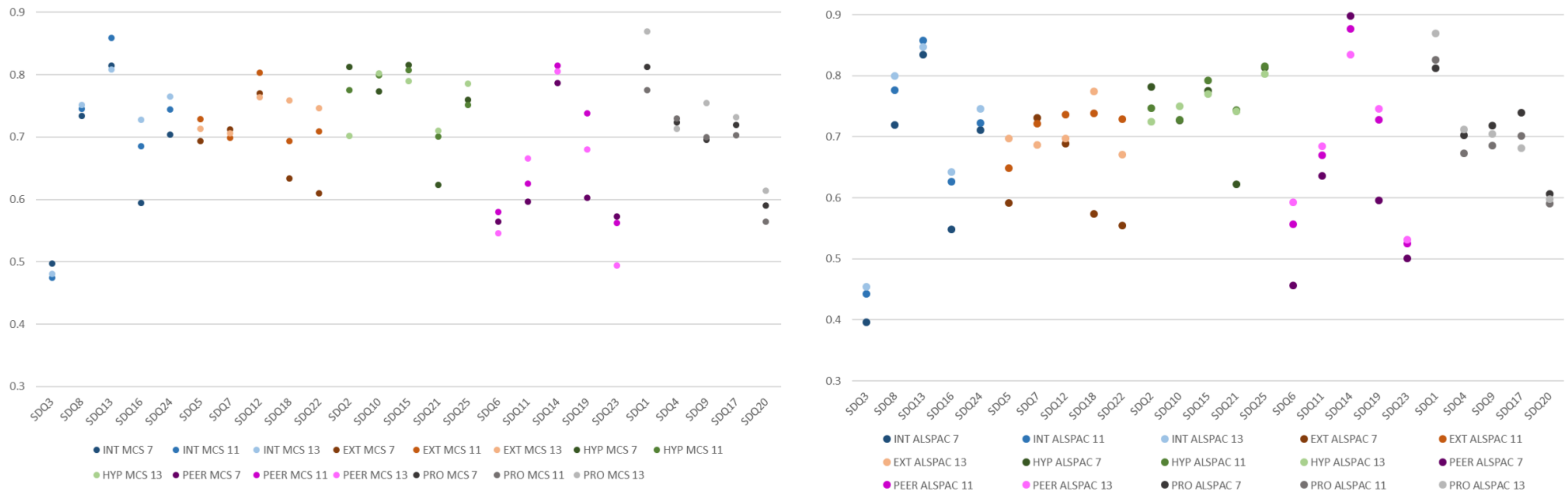


Figure 29. Standardised factor loadings of SDQ items in MCS and ALPSAC (configural model)

6. The Malaise Inventory: A longitudinal examination of measurement equivalence in the NCSD (1958) and BCS70 (1970) cohorts

The most consistent measure of general psychological distress administered in adulthood in the cohorts was the Malaise Inventory (Rutter et al., 1970). This self-report measure was administered in both the NCDS (at ages 23, 33, 42, and 50 years) and BCS70 (at ages 26, 30, 34, 42 and 46 years).

There were four instances in which these measurement occasions overlapped across the two cohorts, roughly corresponding to different decades of life: 20's (age 23 and 26), 30's (age 33 and 34) years, 40's (age 42) and 50's (age 50 and 46). We therefore tested the measurement invariance of the Malaise across both cohort and assessment waves using MGCFA. First, we merged data from the NCDS and BCS70 into a single file, and created a grouping variable that denoted every wave x cohort permutation. This resulted in 8 groups. We then tested for configural, metric and scalar invariance using the 9-item version of the scale, assuming a unidimensional model reflecting general psychological distress. The results are presented in Table 18.

Table 21. Fit statistics for MGCFA models of Malaise Inventory in NCDS and BCS70

Model	N	X^2 (df)	RMSEA	CFI	TLI	ΔRMSEA	ΔCFI	ΔTLI
Configural	79,812	2987.754 (216)	0.036	0.988	0.984			
Metric		2426.987 (272)	0.028	0.991	0.990	0.008	0.003	0.006
Scalar		4814.203 (265)	0.041	0.980	0.978	0.005	0.008	0.006
Rasch		7023.577 (280)	0.049	0.971	0.970	0.013	0.017	0.014

The full longitudinal scalar model of the Malaise did not result in a worsening of overall fit compared to the configural/baseline model (Table 18). Therefore measurement invariance for the Malaise was supported across both age and cohorts. This indicates that the interpretation of items by participants across cohorts or age groups did not influence the observed scores on this measure, therefore the Malaise can reliably be compared in terms of both covariances and means across age groups and cohorts. An inspection of the factor loadings and thresholds from the configural model (Figure 30) corroborates this; the measurement parameters for each individual item in the Malaise Inventory were highly similar across age and cohorts. Subsequent tests of invariance by gender are reported elsewhere (Ploubidis et al., 2019).

Overall these findings demonstrate that, in the NCDS and BCS70 cohorts, the scores on the Malaise inventory were not biased due to such factors as age, survey design, period effects, or cohort effects, and therefore can reliably be compared across cohorts.

In addition, a Rasch model was specified, in which the factor loadings of all 9 indicators were held equal within cohorts. This model provided excellent fit (not worse than configural based on RMSEA), supporting the use of sum scores.

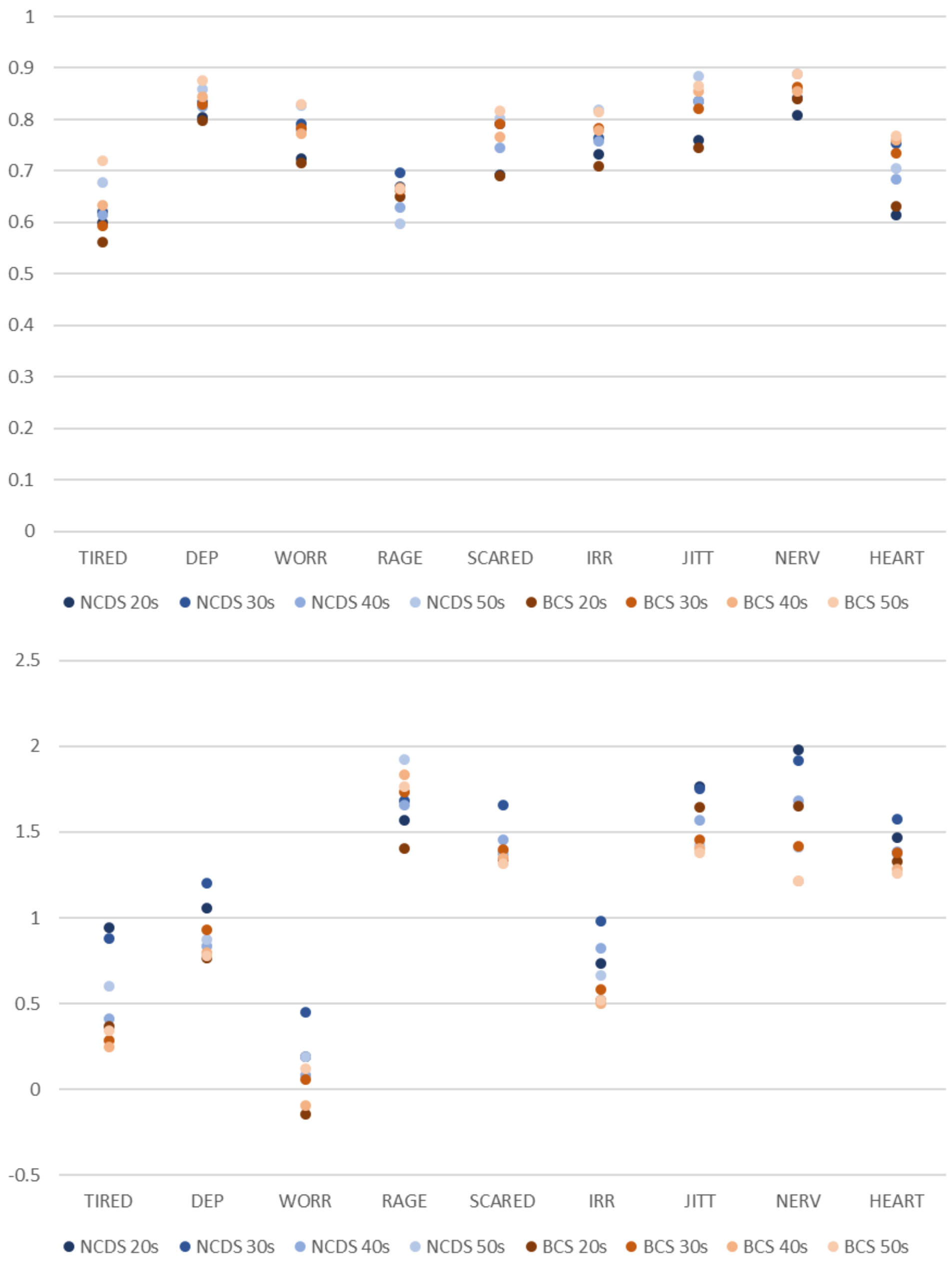


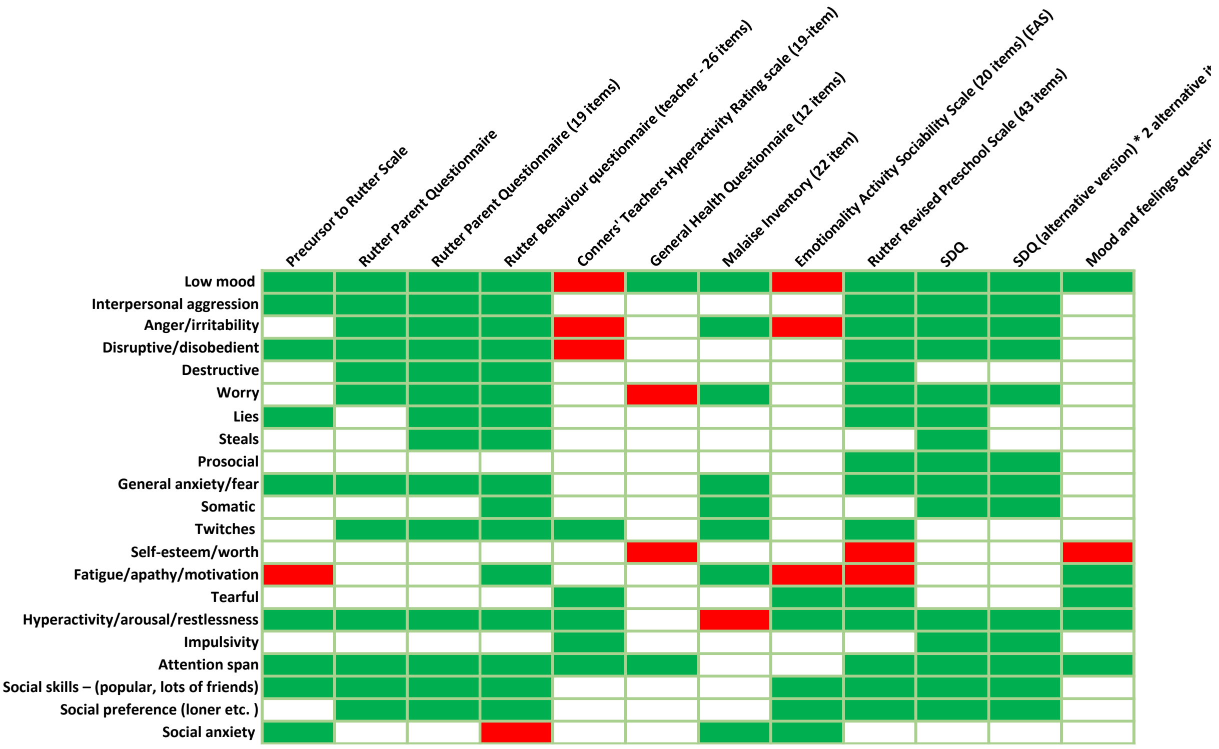
Figure 30. Standardised factor loadings (top) and thresholds (bottom) from configural model of Malaise Inventory in NCDS and BCS70

7. Harmonisation of Mental Health Measures in Childhood

The final aim of this work package was to facilitate broader comparisons within and across the cohorts, particularly when different measures were administered. We therefore conducted retrospective harmonisation. This process involved the manipulation of available data in order to make it more comparable across studies (see section 2.3 for details). We present the results from this harmonisation process below.

7.1 Item selection and inter-rater agreement

A content validation approach was adopted in order to identify conceptually similar items from different measures. Candidate items for harmonisation were identified by two independent raters. Both raters scrutinised every available item within each measure administered in the six cohorts, and assigned each item a code reflecting its core content. Figure 31 presents a heatmap of inter-rater agreement regarding the codes that were assigned to the items (childhood measures only). Overall, agreement was high (88%). The greatest source of disagreement was with items that assessed self-esteem/worth, due in part to the fact that several measures contained multiple items assessing this construct.



	Precursor to Rutter Scale	Rutter Parent Questionnaire	Rutter Parent Questionnaire (19 items)	Rutter Behaviour questionnaire (teacher - 26 items)	Conners' Teachers Hyperactivity Rating scale (19-item)	General Health Questionnaire (12 items)	Malaise Inventory (22 item)	Emotionality Activity Sociability Scale (20 items) (EAS)	Rutter Revised Preschool Scale (43 items)	SDQ	SDQ (alternative version) * 2 alternative it	Mood and feelings questio
Low mood	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟥	🟩	🟩	🟩	🟩
Interpersonal aggression	🟩	🟩	🟩	🟩					🟩	🟩	🟩	
Anger/irritability		🟩	🟩	🟩	🟥		🟩	🟥	🟩	🟩	🟩	
Disruptive/disobedient	🟩	🟩	🟩	🟩	🟥				🟩	🟩	🟩	
Destructive		🟩	🟩	🟩					🟩			
Worry		🟩	🟩	🟩		🟥	🟩		🟩	🟩	🟩	
Lies	🟩		🟩	🟩					🟩	🟩		
Steals			🟩	🟩						🟩		
Prosocial									🟩	🟩	🟩	
General anxiety/fear	🟩	🟩	🟩	🟩			🟩		🟩	🟩	🟩	
Somatic				🟩			🟩			🟩	🟩	
Twitches		🟩	🟩	🟩	🟩		🟩		🟩			
Self-esteem/worth						🟥			🟥			🟥
Fatigue/apathy/motivation	🟥			🟩			🟩	🟥	🟥			🟩
Tearful					🟩			🟩	🟩			🟩
Hyperactivity/arousal/restlessness	🟩	🟩	🟩	🟩	🟩		🟥	🟩	🟩	🟩	🟩	🟩
Impulsivity					🟩					🟩	🟩	
Attention span	🟩	🟩	🟩	🟩	🟩	🟩			🟩	🟩	🟩	🟩
Social skills – (popular, lots of friends)	🟩	🟩	🟩	🟩				🟩	🟩	🟩	🟩	
Social preference (loner etc.)		🟩	🟩	🟩				🟩	🟩	🟩	🟩	
Social anxiety	🟩			🟥			🟩	🟩				

Figure 31. Heat map representing inter-rater agreement on item content in mental health measures available in childhood

Green blocks reflect agreement; red blocks equal disagreement. Empty blocks indicate that neither researcher identified a corresponding item. SDQ = Strengths and Difficulties Questionnaire.

In cases where the two raters disagreed, a third independent rater decided between the two assigned codes. This process resulted in a comprehensive list of matched items across the various measures administered in childhood. This is information available as a searchable spreadsheet at https://www.closer.ac.uk/research-fund-2/data-harmonisation/harmonisation-mental-health-measures-british-birth-cohorts/.

7.2 Assessment of quality of harmonised items

Given the amount of symptoms assessed in the British cohorts, the number of permutations of overlapping items is vast, and will vary considerably depending on i) the cohorts the researcher is interested in, and ii) the developmental periods/assessment waves that are relevant to the research question (see above spreadsheet). Having selected a set of overlapping items, it is important to assess the measurement equivalence of these items before using them in substantive research. Below we present an example in which we selected overlapping parent-report items that were administered within and across the NCDS, BCS70, ALSPAC and MCS (NSHD was not included due to the fact that only teacher reports were available; Next Steps was not included due to a lack of overlap in terms of age of assessment).

The most consistent measures of mental health problems administered in these four cohorts were the parent-report versions of the Rutter behaviour scales (Rutter et al., 1970) and the SDQ (Goodman, 1997). These scales were administered at three overlapping developmental periods, which we categorised as: i) mid-childhood (5-7 years), ii) early-adolescence (10-11 years), and adolescence (14-16 years, see Table 19).

Table 22. Overlapping parent-report measures in NCDS, BCS70, ALSPAC and MCS

Age	Period	NCDS	BCS70	ALSPAC	MCS
5	Mid -childhood		Rutter (19-item)		
7	Mid -childhood	Rutter Parent Questionnaire		SDQ	SDQ
10	Early-adolescence		Rutter (19-item)		
11	Early-adolescence	Rutter (14- item)		SDQ	SDQ
14	Adolescence				SDQ
16	Adolescence	Rutter scale (18-item)	Rutter (19-item)	SDQ	

Note: Row shading reflects broadly overlapping age ranges.

Table 20 presents the overlapping items from the 4 cohorts, as identified based on our item-matching process. Both the Rutter scale and SDQ asked parents to rate the presence/severity with which children displayed each symptom on comparable 3-point metric scales, which roughly corresponded to 0 = ‘Does not apply’, 1 = ‘Applies somewhat’ and 2 = ‘Certainly applies'. As such, no processing/manipulation of data were required, with the exception of the age 10 sweep of the BCS70, which was converted from a visual analogue scale to a 3-point scale (see 3.3.1).

The measurement equivalence of 6 items was tested: a 3-item subscale of emotional problems and a 3-item subscale of behavioural problems.

Table 23. Comparable items (i.e. overlapping content) from parent-reported measures across NCDS, BCS70, ALSPAC and MCS

Construct	Rutter Parent Questionnaire (14-item version)	Rutter Parent Questionnaire (19-item version)	SDQ
Low mood	5. Is miserable or tearful	9. Often appears miserable, unhappy, tearful or distressed.	13. Is often unhappy, down hearted or tearful
Worry	7. Worries about many things	6. Often worried, worries about many things	8. Has many worries, often seems worried
Fear/anxiety	10. Is upset by new situation, by things happening for the first time	16. Tends to be fearful or afraid of new things or new situations.	16. Is nervous or clingy in new situations, easily loses confidence
Peer problems	3. Is bullied by other children	5. Not much liked by other children.	11. Has at least one good friend
Solitary	2. Prefers to do things on his/ her own rather than with others	7. Tends to do things on his/her own – rather solitary.	6. Is rather solitary, tends to play alone
Aggression	12. Fights with other children	4. Frequently fights other children.	12. Often fights with other children or bullies them
Disobedience	14. Is disobedient at home	14. Is often disobedient	7. Is generally obedient, usually does what adults request
Irritability/temper	8. Is irritable, quick to fly off the handle	8. Irritable. Is quick to fly off the handle.	5. Often has temper tantrums or hot tempers
Restlessness	6. Is squirmy or fidgety	1. Very restless. Often running about or jumping up and down. Hardly ever still.	2. Is restless, overactive, cannot stay still for long
Concentration problems	1. Has difficulty in settling to anything for more than a few moments	15. Cannot settle to anything for more than a few moments.	15. Is easily distracted, concentration wanders

Blue = emotional problems; Grey = peer problems; Orange = Behavioural problems; Green = attention/hyperactivity problems

A two-factor model was estimated and measurement invariance explored using MGCFA. First, data from all cohorts were merged into a single file, and a variable was derived that denoted all possible age x cohort permutations. This resulted in 12 groups (i.e. 4 cohorts x 3 assessment periods). Configural, metric, and scalar models were fit to the data and the results are presented in Table 21.

Table 24. Measurement invariance of emotional and behavioural subscales in childhood (SDQ and Rutter only)

Model	N	X²	DF	RMSEA	CFI	TLI	∆RMSEA	∆CFI	∆TLI
Configural	131,503	9292.564	96	0.093	0.931	0.871			
Metric		9260.930	140	0.077	0.932	0.912	0.016	0.001	0.041
Scalar		32714.148	184	0.127	0.757	0.762	0.034	0.174	0.109
Partial scalar*		29077.834	118	0.102	0.900	0.848	0.009	0.031	0.023

*Thresholds of worry, anxiety, irritability and aggression items freed.

The configural and metric models demonstrated acceptable levels of model fit. Therefore, variances and covariances can reliably be compared at the latent level. The scalar model however resulted in a considerable worsening of model fit, with indices falling below acceptable levels. As such, the means of these latent emotional and behavioural variables cannot reliably be compared within or across cohorts due to differential measurement error within and/or across groups. Even after inspecting the modification indices and freeing over half of the thresholds across groups, the model did not achieve acceptable fit.

In order to determine whether measurement invariance could be achieved in isolated sections of the above model, we tested for invariance separately within each cohort over the 3 assessment points, and across cohorts at the 3 overlapping ages (5-7, 10-11, and 14-16 years). Out of this set of models, full scalar invariance was observed only once, across the four cohorts at age 15. The results from these models are presented in Table 22.

Table 25. Measurement invariance of emotional and behavioural subscales in childhood (SDQ and Rutter only)

Model	N	X²	DF	RMSEA	CFI	TLI	∆RMSEA	∆CFI	∆TLI
Configural	36,535	1880.407	32	0.080	0.960	0.925			
Metric		1835.650	44	0.067	0.961	0.947	0.013	0.001	0.022
Scalar		3776.941	56	0.085	0.919	0.913	0.005	0.041	0.008

Although the CFI did demonstrate a drop of more than the recommended 0.01, the full scalar model fit the data well. Therefore, it is justifiable to compare means of the latent construct across the four cohorts at this age.

Standardised factor loadings and thresholds from the configural model are presented in Figure 32. The rank ordering and magnitude of these parameters were highly consistent across the four cohorts, further confirming their measurement equivalence.

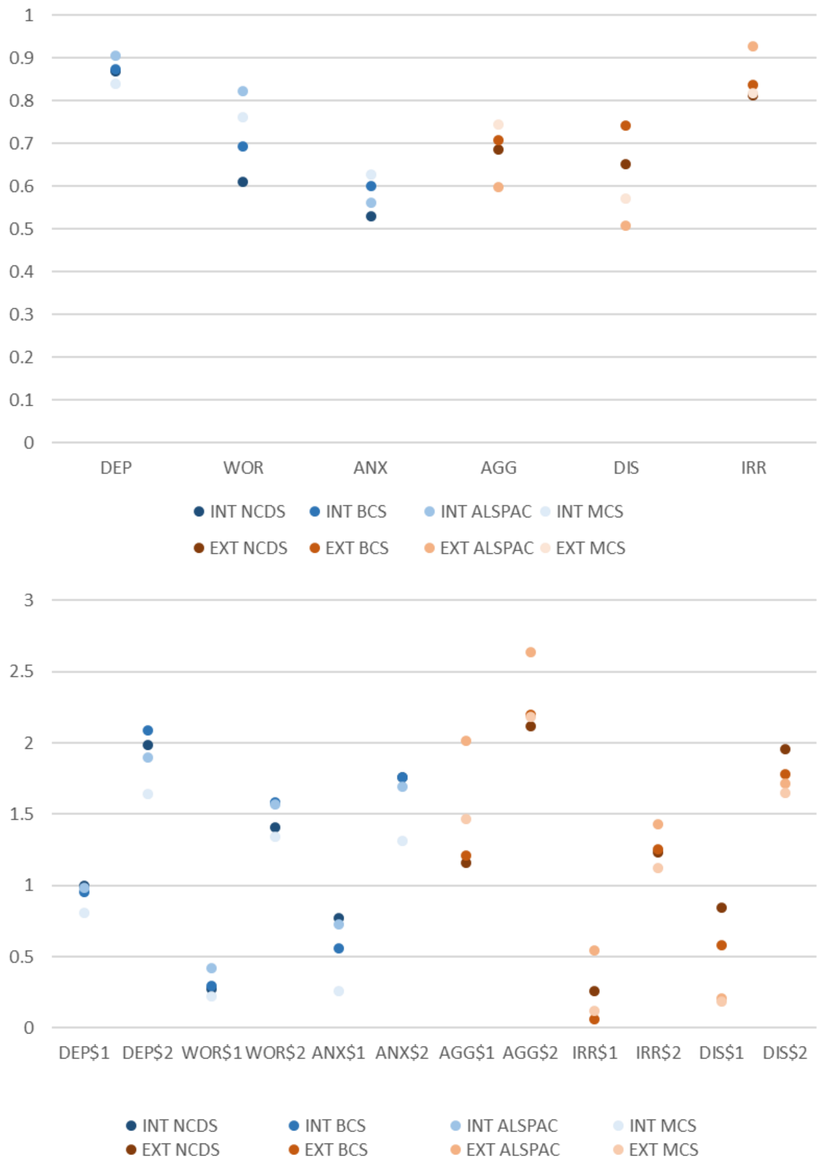


Figure 32. Standardised factor loadings (top) and thresholds (bottom) of harmonised set of items at ages 14-16 across four cohorts

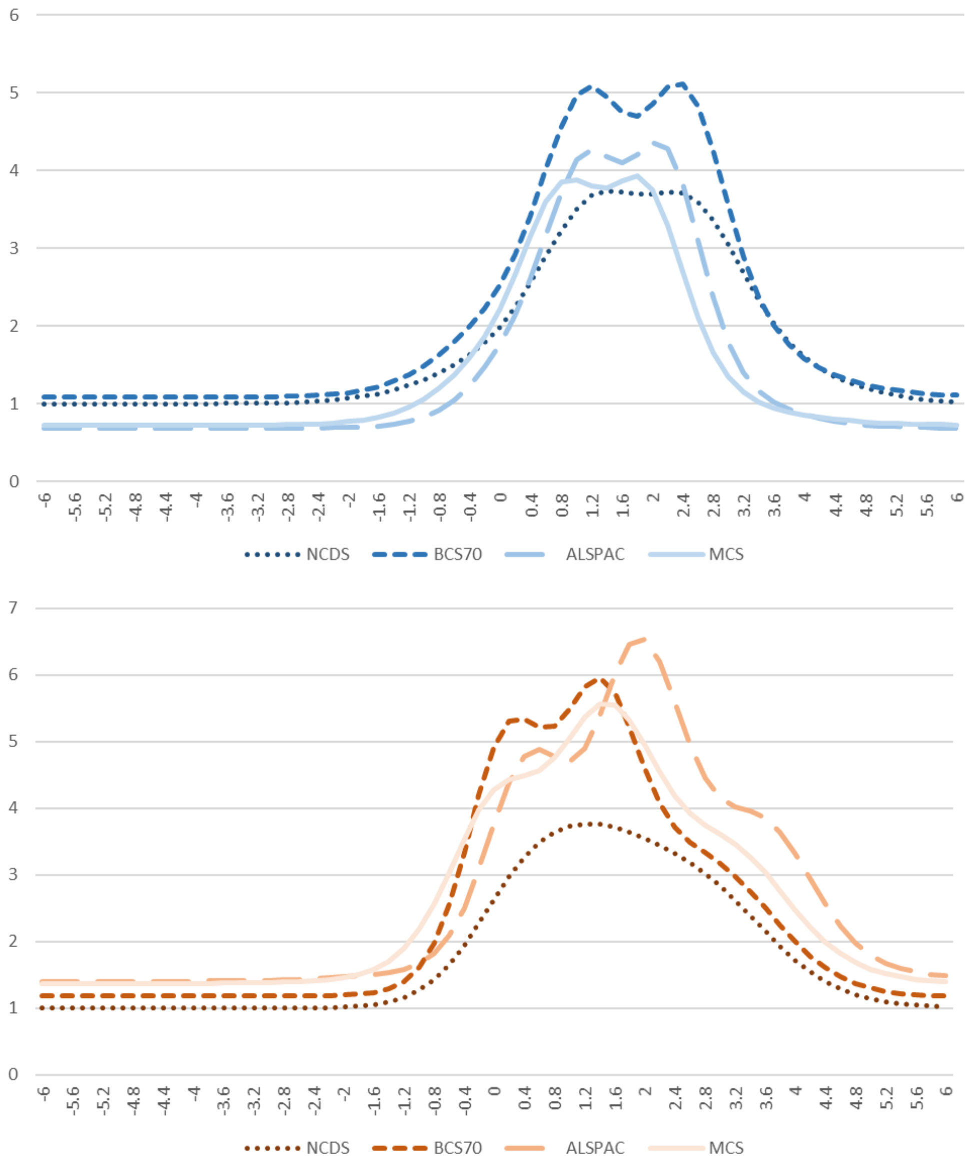


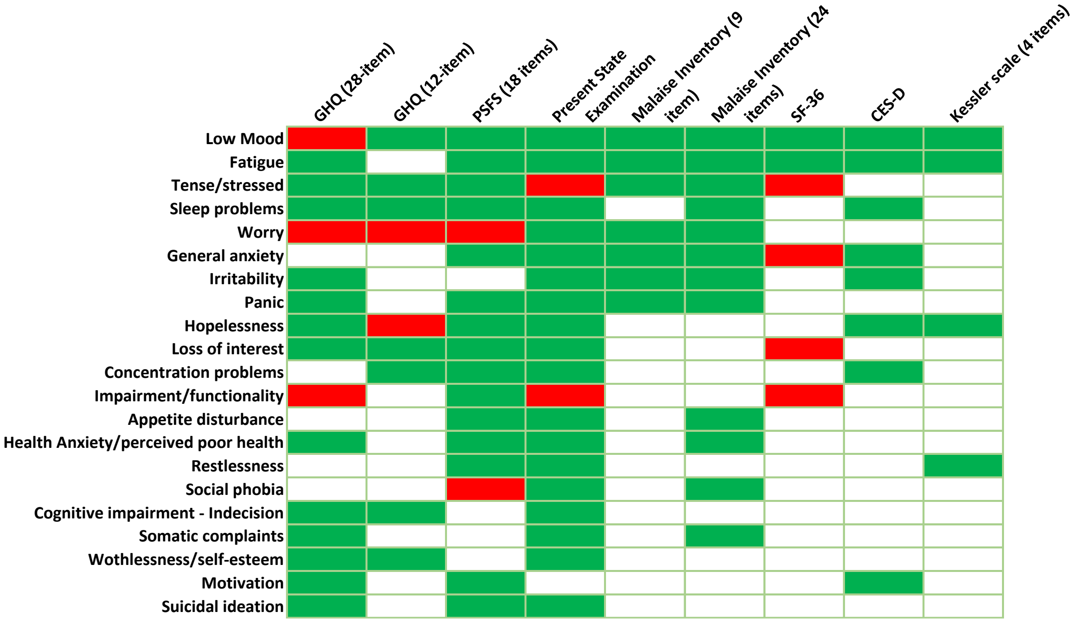
Figure 33. TIFs of harmonised emotional (top) and behavioural (bottom) problems across four cohorts

The measurement precision (Figure 33) of the harmonised item set was comparable across all four cohorts, with the most information at moderate-to-high levels of the latent trait (0.89-2.8 SDs above the mean).

8. Harmonisation of Mental Health Measures in Adulthood

As with the measures administered in childhood, we identified candidate items for harmonisation in the adult measures using a content validation approach (see section 2.3.1). Two raters scrutinised every available item within each measure administered in the 6 cohorts, and assigned each item a code reflecting its core content. Figure 34 presents a heat map of the inter-rater agreement in adulthood (88% agreement rate).

8.1 Item selection and inter-rater agreement



	GHQ (28-item)	GHQ (12-item)	PSFS (18 items)	Present State Examination	Malaise Inventory (9 item)	Malaise Inventory (24 items)	SF-36	CES-D	Kessler scale (4 items)
Low Mood	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Fatigue	🟩		🟩	🟩	🟩	🟩	🟩	🟩	🟩
Tense/stressed	🟩	🟩	🟩	🟥	🟩	🟩	🟥		
Sleep problems	🟩	🟩	🟩	🟩		🟩		🟩	
Worry	🟥	🟥	🟥	🟩	🟩	🟩			
General anxiety			🟩	🟩	🟩	🟩	🟥	🟩	
Irritability	🟩			🟩	🟩	🟩		🟩	
Panic	🟩		🟩	🟩	🟩	🟩			
Hopelessness	🟩	🟥	🟩	🟩				🟩	🟩
Loss of interest	🟩	🟩	🟩	🟩			🟥		
Concentration problems		🟩	🟩	🟩				🟩	
Impairment/functionality	🟥		🟩	🟥			🟥		
Appetite disturbance			🟩	🟩		🟩			
Health Anxiety/perceived poor health	🟩		🟩	🟩		🟩			
Restlessness			🟩	🟩					🟩
Social phobia			🟥	🟩		🟩			
Cognitive impairment - Indecision	🟩	🟩		🟩					
Somatic complaints	🟩			🟩		🟩			
Wothlessness/self-esteem	🟩	🟩		🟩					
Motivation	🟩		🟩					🟩	
Suicidal ideation	🟩		🟩	🟩					

Figure 34. Heat map representing inter-rater agreement on item content in mental health measures available in adulthood

Green blocks reflect complete agreement. Red blocks equal disagreement. Empty blocks indicate neither researcher identified an item corresponding to that symptom. PSF = Psychiatric Symptom Frequency Scale; GHQ = General Health Questionnaire; CES-D = The Centre for Epidemiological Studies-Depression; SF-36 = 36-Item Short Form Survey.

Again, if the two raters disagreed, a third independent rater decided which item code (if either) was most appropriate. A searchable spreadsheet containing all possible permutations of matched items (modifiable by assessment wave and/or cohort) is available at https://www.closer.ac.uk/research-fund-2/data-harmonisation/harmonisation-mental-health-measures-british-birth-cohorts/.

8.2 Assessment of quality of harmonised items

As previously discussed, the number of conceptually-matched items varies depending on the number of cohorts and assessment waves in question. It is therefore important to assess the measurement equivalence of any matched items. Here we provide an example using the adult measures of mental health available in the NSHD, NCDS, and BCS70. In the present analysis, we wished to examine the measurement equivalence of a sub-set of items that would be comparable both within and across the three cohorts. An inspection of the available measures revealed that four instruments would offer this level of coverage: the PSE, PSF, GHQ, and Malaise Inventory (see Table 23). These measures covered all three cohorts, and could be grouped into three broadly overlapping assessment periods: participants in their 30's, 40's and 50's.

Table 26. Overlapping self-report measures administered in adulthood in NSHD, NCDS and BCS70

Age	Period	NSHD	NCDS	BCS70
33	30's		Malaise (24 items)	
34	30's			Malaise (9 items)
36	30's	Present State Examination		
42	40's		Malaise (24 items)	Malaise (9 items)
43	40's	Psychiatric Symptom Frequency Scale		
46	50's			Malaise (9 items)[2]
50	50's		Malaise (9 items)	
53	50's	General Health Questionnaire		

[2] Most recent assessment available.

The sub-set of overlapping items, based on our matching procedure, is presented in Table 24.

Table 27. Comparable items in overlapping self-report measures administered in adulthood across NSHD, NCDS, and BCS70

Symptom	GHQ (28-item)	PSF	PSE	Malaise
Low mood	17. Been able to enjoy your normal day-to-day activities	2. Have you been in low spirits or felt miserable	20. Do you keep reasonably cheerful or have you been very depressed or low-spirited recently? Have you cried at all? (Rate depressed mood)	2. Do you often feel miserable or depressed?
Fatigue	2. Been feeling in need of a good tonic	14. Have there been days when you tired out very easily?	3. Have you been exhausted and worn out during the day or evening even when you haven't been working very hard? (rate tiredness/exhaustion) (slightly doubtful about this one)	1. Do you feel tired most of the time?
Tension	16. Felt constantly under strain	1. Have you felt on edge, keyed up or mentally tense	7. Do you often feel on edge, or keyed up, or mentally tense or strained? (rate nervous tension)	7. Are you constantly keyed up and jittery?
Panic	19. Been getting scared or panicky for no good reason	8. Have you been in situations when you felt shaky or sweaty, or your heart pounded or you could not get your breath?	11. Have you had times when you felt shaky or you heart pounded or you felt sweaty and you simply had to do something about it? (rate panic attacks)	9. Does your heart often race like mad?

All four measures employed different response scales, ranging from the simple 'yes/no' responses of the Malaise to the 6-point ordinal scale used in the PSF. As transforming the binary response of the Malaise was not possible, a decision was made to recode the other three measures to a binary format.

Although the Malaise uses a 'yes/no' scoring format, this does not simply reflect the presence or absence of a symptom. Rather, questions are phrased in such a way that endorsing the 'yes' option requires the symptom in question to be both present and frequent/recurrent/severe; e.g. 'Do you feel tired *most of the time*?', 'Are you *constantly* keyed up and jittery?'. This element of the Malaise Inventory informed our decisions regarding where to place the binary split when recoding items from the GHQ, PSFS, and PSE, all of which used different Likert responses. A breakdown of our recoding scheme is presented in Table 25.

Table 28. Recoding of variables from PSE, PSFS and GHQ

PSE		PSFS		GHQ	
Value	**Label**	**Value**	**Label**	**Value**	**Label**
0	Not present	0	Never	1	Not at all
1	Symptom definitely present during past month, but of moderate clinical intensity	1	Occasionally	2	No more than usual
2	Intense form of symptom present for more than 50% of past month	2	Sometimes	3	Rather more than usual
		3	Quite often	4	Much more than usual
		4	Very often		
		5	Always		

Note: Dotted line indicates placement of binary split. Above the line coded as 0 (no symptom), below the line coded as 1 (symptom present).

As a validity check, we summed the new binarized items, and correlated these harmonised sub-scales with the full, original scale scores. These correlations ranged from 0.78 (PSF) to 0.98 (Malaise), which demonstrated that the recoding process did not interfere unduly with the rank ordering of participants and that the harmonised 4 - item subscales measures the same construct as the original scales.

We tested the measurement equivalence of these 4 items across cohorts and assessment waves using MGCFA. A 1-factor model (reflecting general psychological distress) was tested. The grouping variable was every possible permutation of cohort and assessment periods (3 cohorts x 3 assessment periods = 9 groups). Unlike in previous analyses, a metric invariance model was not tested as this model is not identified when indicators are binary. Results from the tests of measurement invariance are presented in Table 26.

Table 29. Measurement invariance for full longitudinal and cross-cohort model

Model	N	Chi-square (DF)	RMSEA	CFI	TLI	∆RMSEA	∆CFI	∆TLI
Configural	65,997	269.102 (18)	0.044	0.994	0.983			
Scalar		4087.048 (66)	0.091	0.914	0.929	0.047	0.08	0.054
Partial Scalar*		1129.715 (58)	0.050	0.976	0.977	0.006	0.018	0.006
Partial Scalar 2**		444.496 (50)	0.033	0.991	0.990	0.011	0.003	0.007

*Threshold for ‘tense’ freed

The 1-factor model (capturing general psychological distress) provided excellent fit, and the metric model led to virtually no decrease in model fit. The scalar model however, resulted in a worsening of overall model fit by conventional model comparison guidelines. However, the fit statistics were still in the acceptable range in terms of absolute fit, and therefore an argument could be made that using this model is justifiable. As such, it may be possible to compare reliably the means of this latent psychological construct both within and across the cohorts using this harmonised subset of items.

An inspection of the modification indices revealed several areas for model improvement. We also fitted two partial invariance models, each of which demonstrated increasingly good model fit. The first partial invariance model (freeing the threshold for the item reflecting tension) resulted in a trivial decrease in RMSEA (<0.015), but a decrease in CFI of greater than the recommended cut-off (>0.01).

The second partial model (thresholds for tension and fatigue freed) did not lead to a deterioration of model fit according to any conventional guidelines (Barrett, 2007). Thus we can conclude that this 4-item subset is highly comparable across cohorts and age ranges. It may therefore be possible to compare mean-level scores; however the researcher is faced with a choice of whether to use full or partial scalar models to do so. The former approach ensures no differential measurement error due to group membership, however may introduce an element of misspecification to the model. The latter approach ensures a well specified model; however the freely estimated thresholds will introduce an element of bias when comparing means, the degree of which is difficult to quantify.

9. Harmonisation within the NSHD (1946 cohort)

In this section, we demonstrate how to derive a harmonised measure of mental health within a particular cohort (the NSHD) across adulthood. The NSHD, which is the oldest of the CLOSER cohorts, has assessed mental health at the following ages across adulthood: 36 (PSE), 43 (PSFS), 53 (GHQ-28), 60-64 (GHQ-28), and 68-70 years (GHQ-28). Using our item-mapping tool (available at https://www.closer.ac.uk/research-fund-2/data-harmonisation/harmonisation-mental-health-measures-british-birth-cohorts/) we identified 7 conceptually similar items that captured general psychological distress across mid-to-later life (Table 28).

As each of the three measures used a different Likert rating scale, we collapsed each question into a binary item, using the strategy outlined in section 8.2. Once again, we conducted a validity check by summing these binarized variables, and correlating these scores with the full, original scale scores. These correlations ranged from 0.81 (GHQ) to 0.85 (PSFS), which demonstrated that the rank ordering of participants remained relatively consistent after the recoding process and that the 7 – item subscale captures psychological distress similarly to the original scales.

We tested the measurement equivalence of the 7 harmonised items across the 5 assessment waves using MGCFA. A 1-factor model (reflecting general psychological distress) was tested. Results from the tests of measurement invariance are presented in Table 27.

Table 30. Measurement invariance for full longitudinal and cross-cohort model

Model	N	Chi-square (DF)	RMSEA	CFI	TLI	∆RMSEA	∆CFI	∆TLI
Configural	13,886	544.328 (70)	0.049	0.979	0.968			
Scalar*		1175.735 (94)	0.064	0.952	0.946	0.015	0.027	0.02
Partial Scalar*¥		1173.975 (95)	0.064	0.952	0.947			

*Latent variances fixed to 1
¥Residual for ‘tense’ fixed to 0

The configural model fit the data well; all 7 indicators tapped a general psychological distress factor across time, regardless of the measure used. Although the scalar model was just within the acceptable range of ∆RMSEA, problems with the estimation of the variances of certain measured indicators (mood items after age 50) led to implausible estimates of the latent variances at later time points. To address, this the variances were fixed to 1 across all assessment waves. This in turn resulted in a small negative residual value for the indicator ‘tense’ at age 60. Fixing this parameter at 0, which is an appropriate strategy when negative residuals are small (http://www.statmodel.com/discussion/messages/11/555.html?1358188287), resulted in model convergence, and a plausible set of latent parameter estimates. This model fit the data well, and is therefore recommended for any analysis in which the latent means are to be compared.

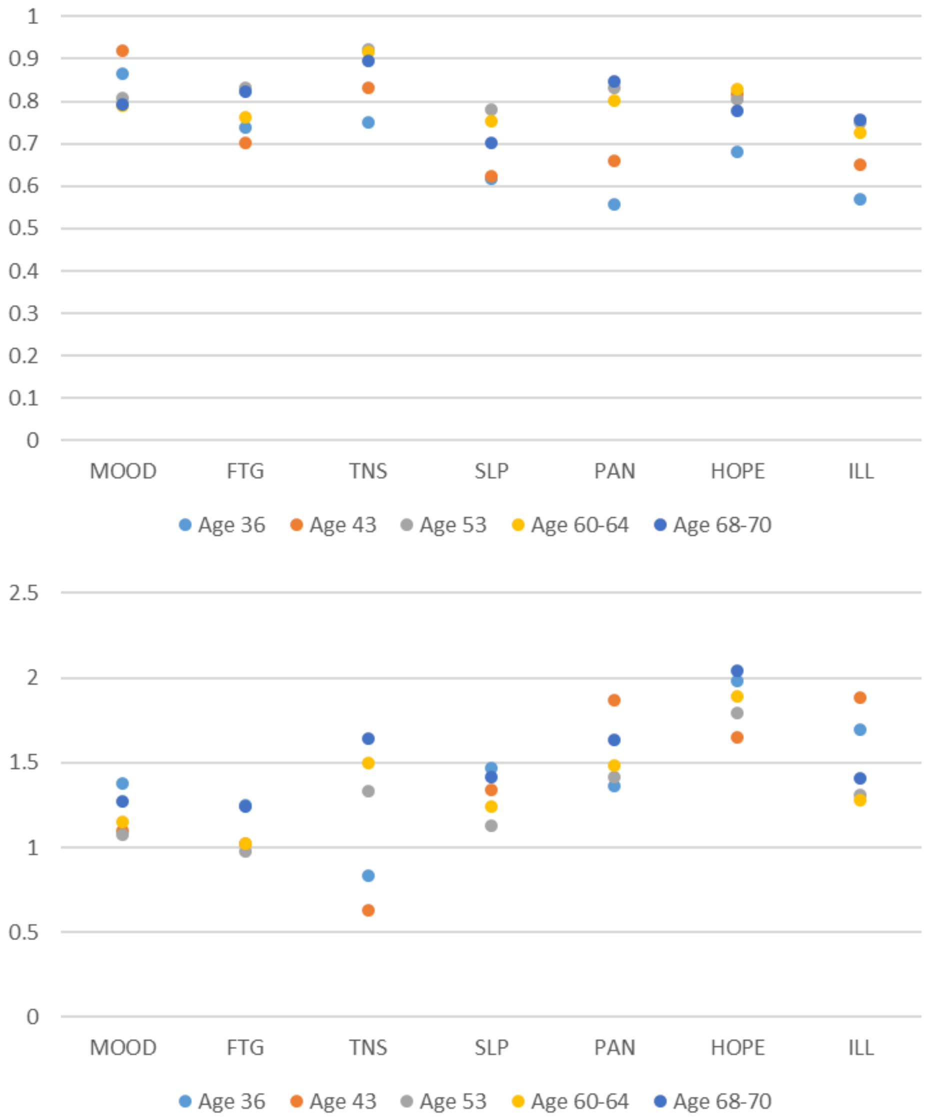


Figure 35. Standardised factor loadings and thresholds from configural model within NSHD

Patterns of loadings and thresholds (Figure 35) were relatively consistent across measures and assessment waves, with the (expected) exception of the ‘tension’ indicator, which had notably lower thresholds at ages 36 (PSE) and 43 (PSFS).

Table 31. Overlapping self-report measures within NSHD

	Present State Examination (selected items)	Psychiatric Symptom Frequency Scale (18 items)	General Health Questionnaire (28-item version) (GHQ-28)
Low Mood	20. Do you keep reasonably cheerful or have you been very depressed or low spirited recently? Have you cried at all? (Rate depressed mood)	2. Have you been in low spirits or felt miserable	17. Been able to enjoy your normal day-to-day activities
Fatigue	3. Have you been exhausted and worn out during the day or evening even when you haven't been working very hard? (rate tiredness/exhaustion)	14. Have there been days when you tired out very easily?	2. Been feeling in need of a good tonic
Tense/stressed	7. Do you often feel on edge, or keyed up, or mentally tense or strained? (rate nervous tension)	1. Have you felt on edge or keyed up or mentally tense	23. Been feeling nervous and strung-up all the time
Sleep problems	30. Have you had any trouble getting off to sleep in the last month? (rate delayed sleep)	9. Have you had trouble getting off to sleep	8. Lost much sleep over worry
Panic	11. Have you had times when you felt shaky or you heart pounded or you felt sweaty and you simply had to do something about it? (rate panic attacks)	8. Have you been in situations when you felt shaky or sweaty or your heart pounded or you could not get your breath	19. Been getting scared or panicky for no good reason
Hopelessness	21. How do you see the future? (rate hopelessness)	16. Have you had the feeling that the future does not hold much for you?	22. Felt that life is entirely hopeless
Health Anxiety	6. Do you tend to worry over your physical health? (rate hypochondriasis)	11. Have you been frightened or worried about becoming ill or about dying?	4. Felt that you are ill

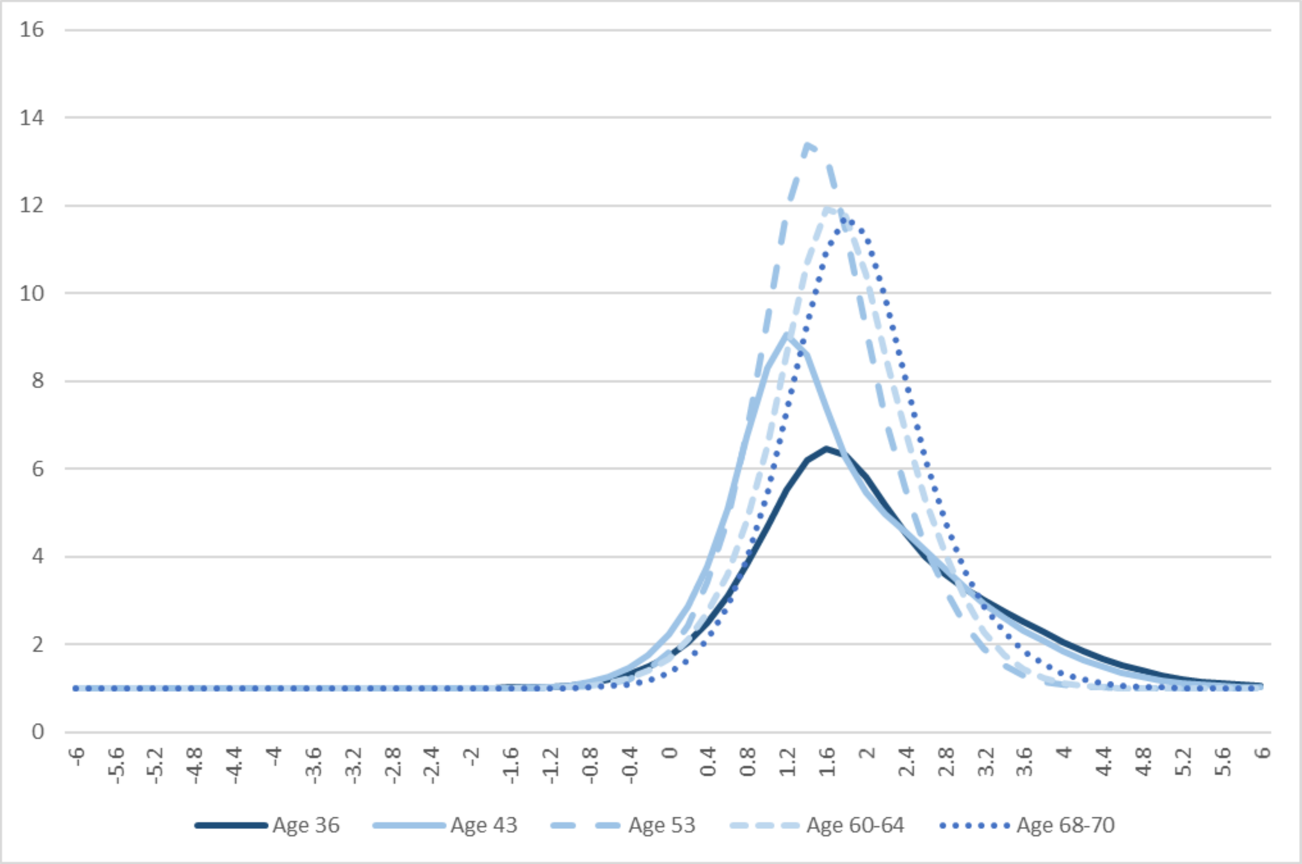


Figure 36. TIFs of harmonised item set (7 items) in NSHD

The measurement precision (Figure 36) of the harmonised item set peaked at similar levels of the latent trait, with the most information at moderate-to-high levels of the latent trait (0.89-2.8 SDs above the mean). The item set had more precision at later assessments.

10. Comparison of Mother and Teacher Reports of Mental Health in Childhood

This section focuses on the comparison of mother and teacher reports of mental health in childhood. Analyses were limited to the BCS70 and MCS, both of which have available information on parent and teacher ratings of cohort members mental health at similar ages in childhood; BCS70 at age 10, and MCS at age 11. The NCDS also had parent and teacher reports of child mental measured in the same sweep, although only much later in childhood at age 16 and was not included in the current examination. Comparisons of parent and teacher reports were restricted to child conduct and emotional problems, thereby excluding hyperactivity and other dimensions of child mental health. The main objective was to compare parent and teacher reports within each cohort, and therefore within cohort harmonised items are used rather than items harmonised between cohorts. Further details of the harmonisation methods and psychometric modelling used are provided above in Section 2 of this report.

10.1 Parent and teacher reports compared in the British Cohort Study

10.1.1 Items on conduct and emotional problems in the BCS70

In the BCS70 at age 10 parents and teachers reported child mental health using the Rutter and the Connor scale, as outlined previously in section 3.3. A reduced version of the Rutter scale (14 items) were administered to teachers whilst parents were administered the full 19 items scale. Both reporters completed the Connor hyperactivity scale, consisting of 19 items. The harmonised items between parents and teachers that relate to child conduct and emotional problems are shown in Table 29, with five items measuring each of these dimensions of mental health. All items are drawn from the Rutter scale, except one item from the Connor scale that substitutes one of the Rutter items omitted on the teacher form. These items were deemed to be comparable in the two-rater item harmonisation process (section 2).

Question wording varies slightly between parents and teachers. This is especially evident in item 3 relating to disobedience on the conduct problems scale; items are conceptually similar, but the phrasing is quite different, and the parent item is negatively worded whilst the teacher item is positively worded. Parents provided responses on a visual analogue scale from 0 (‘Does not apply’) to 100 (‘Certainly applies’), whilst teachers used an analogue scale of 1 (‘Not at all’) to 47 (‘A great deal’). As described in section 3.3.1, these response scales were harmonised by way of latent profile analysis, reducing each analogue scale to a three-level response scale.

A total of 10,644 cohort members in the BCS70 had valid responses from both parents and teachers on all items measuring the two dimensions of child mental health.

Table 32. Parent and teacher items measuring child conduct and emotional

	Parent items		**Teachers items**	
Conduct				
Item 1	m45	Often destroys own or	j163	Destroys own or other
Item 2	m46	Frequently fights with	j160	Quarrels with other children
Item 3	m56	Is often disobedient	j126	To what extent can you
Item 4	m61	Bullies other children	j135	Teases other children to excess
Item 5	m80	Displays outbursts of	j134	Displays outbursts of temper,
Emotional				
Item 1	m48	Often worried, worries	j149	Is worried about many things
Item 2	m49	Tends to do things on	j159	Tends to do thing on his or her
Item 3	m51	Often appears miserable,	j156	In relations with others appear
Item 4	m58	Tends to be fearful or	j128	Is fearful or afraid of new
Item 5	m59	Is fussy or over particular	j146	Is fussy or over-particular

Parent response scale: 0=Does not apply – 100=Certainly applies

10.1.2 Confirmatory factor analysis and measurement invariance

In Table 30 we present fit statistics for various model configurations of parent and teacher reported conduct and emotional problems at age 10. The unidimensional model, with all 20 items (10 parent, 10 teacher items) loading on one factor, did not provide a very good fit to the data. The fit was much improved in the 4-factor configural model (parent conduct, teacher conduct, parent emotional, teacher emotional), which demonstrated a good overall fit. The metric model where factor loadings were fixed to be equal across raters had good fit. Finally, the scalar model that imposes invariant thresholds and loadings also had acceptable fit. The increase in RMSEA in these restricted models compared to the configural is well within the acceptable level of <.015, although the change in CFI for the scalar model exceeds the cut off (<.01), as specified by Little (2013).

Table 33. Fit statistics for parent and teacher reported child conduct and emotional problems at age 10 in BCS70

Measure	Model	N	χ^2	DF	RMSEA	CFI	TLI
Rutter[a] (20 items)	1-factor	10,644	22930.763	170	0.112	0.627	0.583
	Configural: 4-factor[b]	10,644	4175.175	164	0.048	0.934	0.924
	Metric: 4-factor[b]	10,644	4653.584	174	0.049	0.926	0.920
	Scalar: 4-factor[b]	10,644	5387.132	184	0.052	0.915	0.912

[a] Two Connor items replace two Rutter items (one parent and one teacher item).

10.1.3 Factor loadings

Factor loadings for parent and teacher reported measures are shown below for conduct problems (Figure 37) and for emotional problems (Figure 38). Loadings were generally higher for conduct than for emotional problems. Factor loadings tended to be higher for teachers than for parents. However, there were exceptions; one being item 3 (disobedience) on the conduct scale, which is the items that varied somewhat in terms of wording and which is positively worded on the teacher form. The other exception is item 4 (fearful or afraid) of the emotional problems measure which had an almost identical factor loading for parents and teachers.

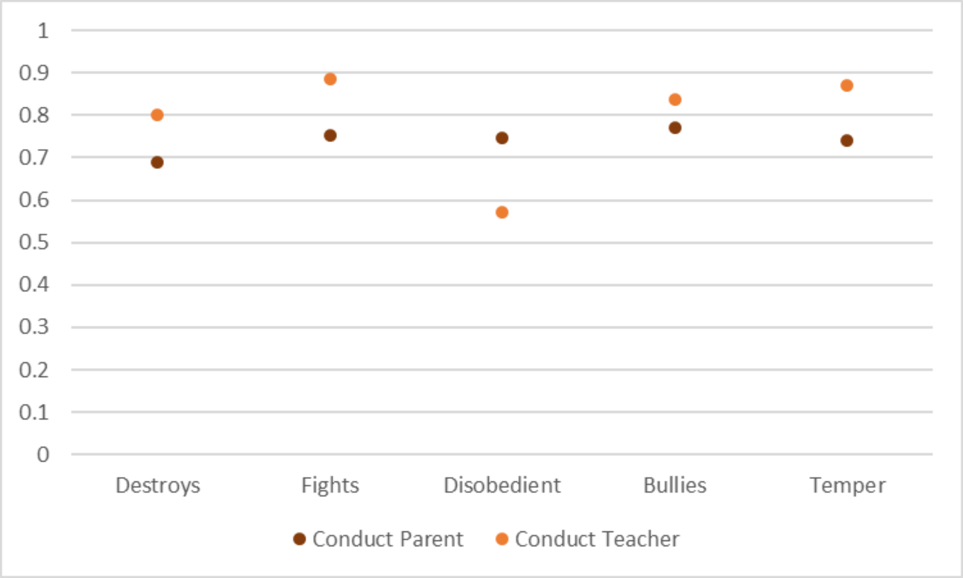


Figure 37. Factor loadings for conduct problems (5 items, Rutter), parents and teachers in BCS70

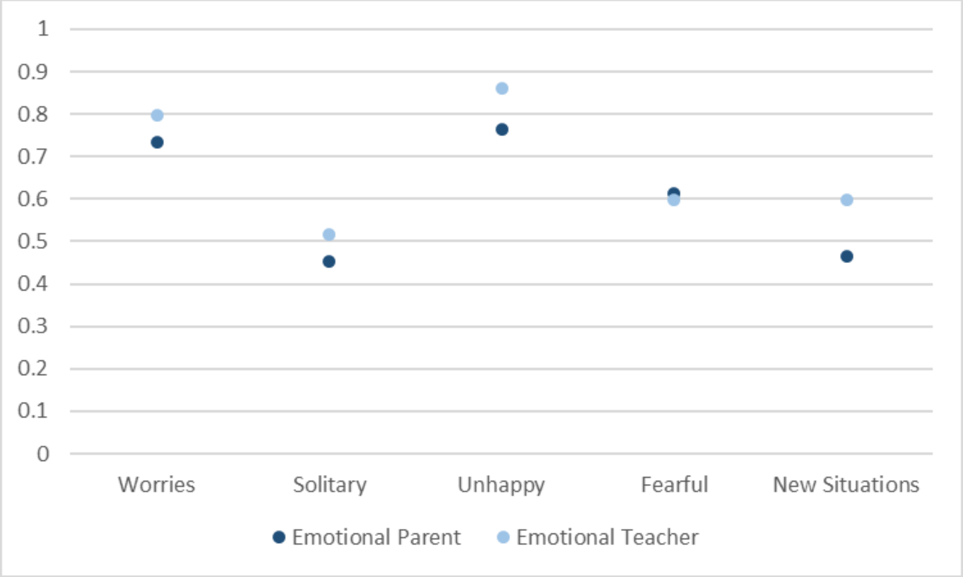


Figure 38. Factor loadings for emotional problems (5 items, Rutter), parents and teachers in BCS70

The TIFs for conduct and emotional problems, which are shown in Figure 39 for both parents and teachers. Teachers provided more information on both conduct and emotional scales, compared to parents where information curves are 'flatter' for both types of symptoms. Information was especially high for the teacher conduct scale. Both parents and teachers assess conduct problems more precisely at the higher end of the distribution (more severe symptoms) than emotional problems which are measured better slightly further down the distribution (less severe symptoms).

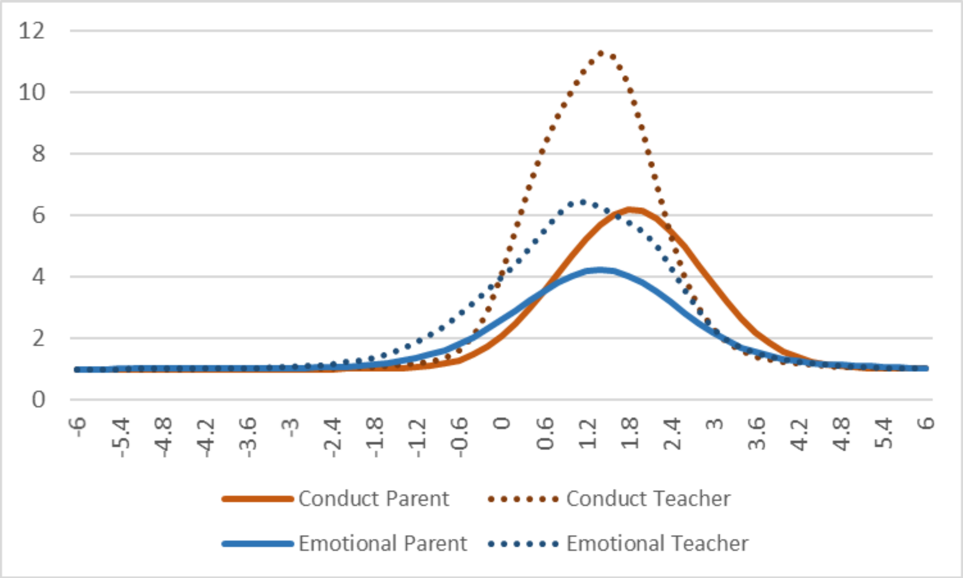


Figure 39. TIFs for child conduct and emotional problems in BCS70, by parent and teacher reporters

10.1.4 Agreement between parent and teacher reports

Agreement between parents and teachers in their reports of child mental health is shown in Table 31 below. Correlations are shown both for measures where raw items are simply summed, and for latent measures based on the 4-factor CFA reported above. Agreement between parents and teachers was low to modest (r=.18 to r=.43), and higher for conduct problems (r=.27 and r=.43) than for emotional problems (r=.18 and r=.29). For both types of symptoms, agreement was as expected higher for the latent measures (r=.29 and r=.43) than for the measures based on summed items (r=.18 and r=.27).

Table 34. Agreement between parent and teacher reports of child mental health (BCS70 aged 10)

Raw items summed	Conduct problems	r = .27
	Emotional problems	r = .18
Latent factors	Conduct problems	r = .43
	Emotional problems	r = .29

10.2 Parent and teacher reports compared in the Millennium Cohort Study

10.2.1 Items on conduct and emotional problems in the MCS

Child mental health was assessed in the MCS at age 11 using the Strengths and Difficulties Questionnaire (SDQ) which was completed in full (25 items) by parents and teachers (see section 3.6 above for further details). For the current comparative analyses between parents and teachers, the conduct problems subscale (5 items) and the emotional problems subscale (5 items) of the SDQ were used. Items and question wording are shown in Table 32 below. As seen, question wordings are identical across the parent and the teacher forms, with the only slight difference being one response option, which for parents is worded ‘certainly true’ and for teachers ‘very true’, whilst the remaining response categories are identical (’not true’, ‘somewhat true’). The sample size for these comparative analyses were based on 5,651 cohort members who have complete data from both parents and teachers on all items making up the conduct and emotional subscales.

Table 35. Parent and teacher items measuring child conduct and emotional problems at age 11 in MCS

	Parent items		Teachers items	
Conduct				
Item 1	EPSDTT00	Often has temper tantrums or hot tempers	EQ5E	Often has temper tantrums or hot tempers
Item 2	EPSDOR00	Is generally obedient, usually does what adults request*	EQ5G	Is generally obedient, usually does what adults request*
Item 3	EPSDOA00	Often lies or cheats	EQ5R	Often lies or cheats
Item 4	EPSDFB00	Often fights with other children or bullies them	EQ5L	Often fights with other children or bullies them
Item 5	EPSDCS00	Steals from home, school or elsewhere	EQ5V	Steals from home, school or elsewhere
Emotional				
Item 1	EPSDHS00	Often complains of headaches, stomach aches or sickness	EQ5C	Often complains of headaches, stomach aches or sickness
Item 2	EPSDMW00	Has many worries, often seems worried	EQ5H	Has many worries, often seems worried
Item 3	EPSDUD00	Is often unhappy, down-hearted or tearful	EQ5M	Is often unhappy, down-hearted or tearful
Item 4	EPSDNC00	Is nervous or clingy in new situations, easily loses confidence	EQ5P	Is nervous or clingy in new situations, easily loses confidence
Item 5	EPSDFE00	Has many fears, is easily scared	EQ5X	Has many fears, is easily scared

*reverse coded
Parent response scale: 1=Not true, 2=Somewhat true, 3=Certainly true
Teacher response scale: 1=Not true, 2=Is somewhat true, 3=Very true

10.2.2 Confirmatory factor analysis and measurement invariance

Several CFA models of parent and teacher measures of child conduct and emotional symptoms were tested, as shown in Table 33. The unidimensional model with all items loading onto a single factor did not provide a satisfactory fit. An improved factor solution was seen in the 4-factor configural model (parent conduct, teacher conduct, parent emotional, teacher emotional), which provided good model fit. Measurement invariance was tested in two further models; first in terms of metric invariance by fixing factor loadings to be equal across raters which provided a good model fit (RMSEA=0.045), and finally scalar invariance was tested by additionally holding thresholds equal across raters, also with a good model fit (RMSEA=0.049). The small decrease in model fit in each step of the invariance test was within the acceptable level of change in the RMSEA (<.015), but exceeded slightly the recommendation cut off for change in CFI between model (<.01). However, considering the good fit of the full scalar model we recommend that this be used when teacher and more ratings are compared.

Table 36. Fit statistics for parent and teacher reported child conduct and emotional problems at age 11 in MCS

Measure	Model	N	X2	DF	RMSEA	CFI	TLI
SDQ (20 items)	1-factor	5651	12008.175	170	0.111	0.674	0.636
	Configural: 4-factor[a]	5651	1745.964	164	0.041	0.956	0.950
	Metric: 4-factor[a] Invariant factor loadings	5651	2171.410	174	0.045	0.945	0.940
	Scalar: 4-factor[a] Invariant factor loadings Invariant thresholds	5651	2702.159	184	0.049	0.931	0.928

[a] Parent conduct, Teacher conduct, Parent emotional, Teacher emotional. Each factor with 5 items.

Factor loadings from the configural model are shown below for each reporter for items making up the conduct problems scale (Figure 40) and for the emotional problems scale (Figure 41). Generally, factor loadings were high, around .70 or higher, for conduct and emotional problems, with the exception of "somatic complaints" on the emotional problems scale which was .50 for parents and .60 for teachers. For both problem scales, factor loadings were generally higher for teachers, the only exception being "often unhappy" on the emotional problems scale for which the factor loading was near identical for parents and teachers. Factor loadings for items 4 and 5 on the conduct problems scale were also close in value for parents and teachers.

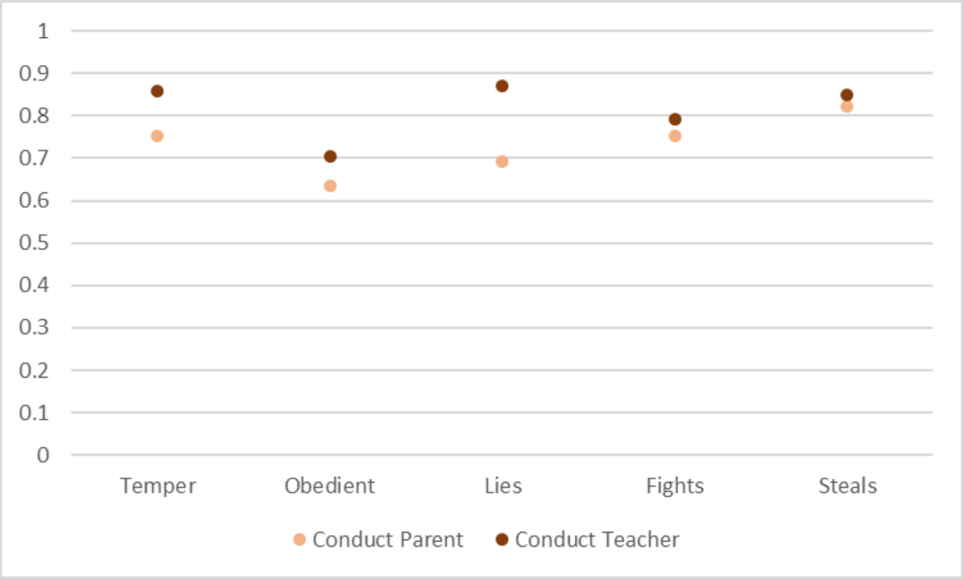


Figure 40. Factor loadings for conduct problems (5 items, SDQ), parents and teachers in MCS (configural model)

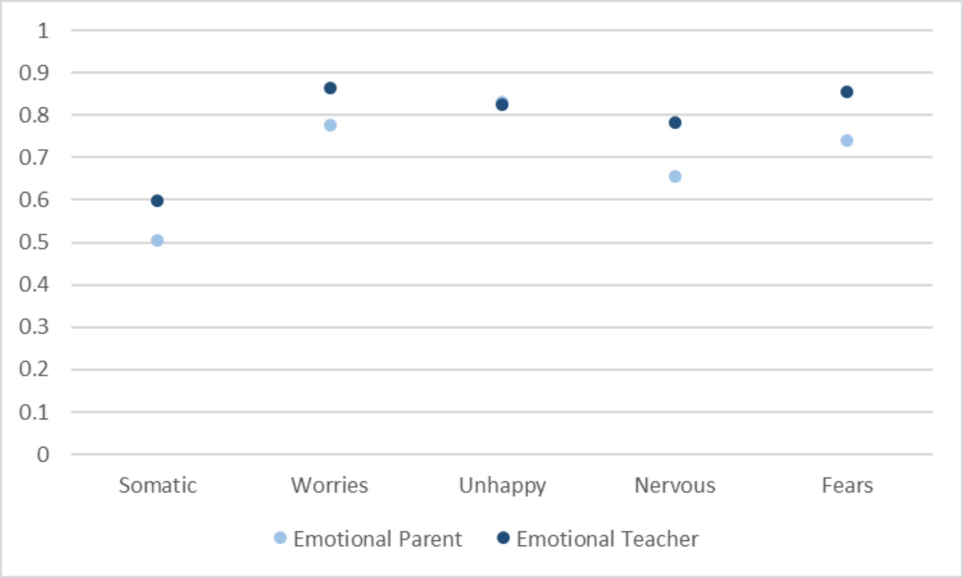


Figure 41. Factor loadings for emotional problems (5 items, SDQ), parents and teachers in MCS (configural model)

In Figure 42 we present TIFs for measures of child conduct and emotional problems. Teachers provided more information than parents on both scales. As in the BCS70, for parents and teachers, the TIF for conduct problems was located towards the higher end of the distribution, which means that the scale has good precision at measuring more severe levels of symptoms. The TIF for the emotional problems scale (for both parents and teachers) was located slightly further down the distribution (less severe symptoms) although still measuring problems with more precision above the mean.

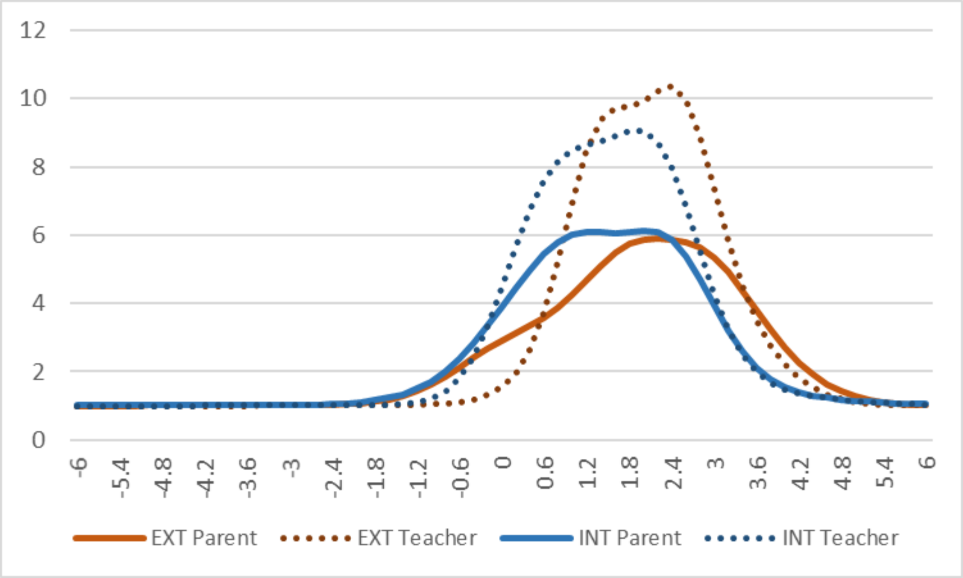


Figure 42. TIFs for child conduct and emotional problems in MCS, by parent and teacher reporters

10.2.3 Agreement between parent and teacher reports

Shown in Table 34 below are correlations between parent and teacher reports of child conduct and emotional problems. Correlations are shown for the latent measures from the 4-factor CFA model reported above, and for measures were items simply were summed. Agreement between reporters was moderate to high (r=.36 – r=.61), with higher agreement for conduct problems (r=.39 and r=.61) than for emotional problems (r=.36 and r=.50). Correlations were as expected higher for the latent measures (r=.61 and r=.50) than for the measures based on summing items (r=.39 and r=.36).

Table 37. Agreement between parent and teacher reports of child mental health at age 11 in MCS

Raw items summed	Conduct problems	r = .39
	Emotional problems	r = .36
Latent factors	Conduct problems	r = .61
	Emotional problems	r = .50

10.3 Summary and conclusion

Parent and teacher reports of child mental health were compared in the MCS and BCS70, cohorts born thirty years apart. Comparable results were largely seen in these cohorts, although there were also some differences.

Within each cohort, parents and teachers completed similar harmonised items that measure two latent dimensions of child mental health – conduct problems and emotional problems. Parent and teacher measures were found to be invariant, suggesting that informants reported on similar underlying aspects of child mental health. In both the BCS70 and the MCS, teachers provided more information (higher precision) both on children's conduct and emotional problems as revealed by the shape of the total information functions and higher factor loadings. One explanation may be that teachers observe children in the school setting in interaction with various other children and that may lead to more precise assessments.

Information curves for emotional problems (for both parents and teachers) were higher in the younger MCS cohort compared to BCS70. Information curves for conduct problems appeared to have a similar shape and were located similarly across the latent trait in both cohorts. An important caveat is that information curves are not directly comparable between cohorts because they are not based on similar (harmonised) items, although the same number of items with a three-level response were used in both the BCS70 and the MCS. The main aim of this exercise was to harmonise between parent and teachers and not between cohorts. Nevertheless, the difference between cohorts for emotional problems is interesting and may indicate that adult reporters have become better at identifying child emotional problems. However, bias in the form systematic measurement error due to a method effect remains a possible explanation for the observed cross-cohort difference in measurement precision

Regarding correlations (agreement) between parents and teachers, these were larger for latent measures than for summed measures. This is an expected finding as latent scores capture all possible patterns of responses, while sum scores combine different response patterns in a single score, which results in measurement error reflected in the attenuation of correlation coefficients.

Reporter agreement on latent measures was moderate in the BCS70 and high in the younger MCS cohort. It is possible that parents and teachers more recently have come to view child behaviour more similarly. However, methods effect is again a plausible explanation, especially considering differences in question wording on the parent and teacher forms in the BCS70. In both cohorts there was a higher level of agreement for conduct problems than for emotional problems. Overall results of the correlational analyses in the BCS70 and the MCS correspond reasonably well with those found in previous studies. The recent meta-analysis by De Los Reyes et al. (2015) reported low-to-moderate parent and teacher correlations (although it is unclear how many of the included studies use latent variables as in the current study), and a higher agreement for conduct than for emotional symptoms was also shown, as in the current study.

Although there is some agreement between parents and teachers, both provide a unique insight on child functioning. If information from both raters cannot be used and the decision is solely based on measurement properties, we recommend the inclusion of teacher reports as they provide more information on child conduct and emotional problems.

11. Summary and Recommendations

This project had 4 primary aims: i) investigate and document the measurement properties of the existing mental health measures in six British cohorts, ii) evaluate the psychometric equivalence of measures that have been administered across multiple sweeps/cohorts (e.g. Malaise Inventory, SDQ), iii) retrospectively harmonise and evaluate the measurement equivalence of items from *different* instruments within and across cohorts, and iv) explore the extent to which parent and teacher questionnaires capture the same underlying dimensions of child mental health.

In this section we briefly summarise our conclusions and provide recommendations for researchers looking to use the cohort studies to investigate mental health.

11.1 Measurement properties of the existing mental health measures

The measurement properties of the various questionnaires are documented in sections 3 and 4 of this report. In summary, the majority of measures demonstrated good psychometric properties. Established factor structures were supported in the cohort data, and the measurement precision of most instruments peaked at mid-to-high levels of the latent trait. This is a desirable feature for population measures of psychological distress, as it is crucial for such measures to be able to effectively assess participants with moderate or high symptomatology.

11.2 Evaluating the psychometric equivalence of the Malaise Inventory and SDQ within and across cohorts

The SDQ, which was administered multiple times in childhood in ALSPAC and MCS, demonstrated excellent psychometric equivalence both within and across the two cohorts. Full scalar invariance was supported, which indicated that factors such as age effects, survey design, period effects, or cohort specific effects did not bias the ways in which participants responded to the questions that were asked. The same results and interpretation were observed for the Malaise Inventory, which was administered multiple times across adulthood in the NCDS and BCS70. As such, both the Malaise and SDQ can be used to make valid comparisons both within and/or across cohorts. These measures are appropriate for research questions related to covariances (i.e. stability/change in associations within or across cohorts) and changes in means (e.g. studies of change within or across cohorts)

Collectively, these findings offer reassurance for the extent to which self-reported mental health items are affected by systematic sources of error, since despite the effects of age and secular changes that resulted in important differences within and between cohorts the SDQ and the Malaise Inventory were shown to function equivalently.

11.3 Retrospective harmonisation of items from different questionnaires

11.3.1 Harmonisation in childhood

Using a content validation approach, we identified six items (three emotional and three behavioural) from the Rutter scales and the SDQ that demonstrated topological/content overlap across four cohorts (NCDS, BCS70, ALSPAC, and MCS).

We evaluated the measurement equivalence of these harmonised items within and across the four cohorts using MGCFA. Metric invariance was supported, therefore these items are suitable for researchers who wish to determine whether associations between emotional/behavioural problems and predictor/outcome variables are consistent across cohorts/sweeps (i.e. regression coefficients will not be biased due to group membership).

Scalar invariance was not supported in the full longitudinal x cohort model, however full scalar invariance was observed across all four cohorts in adolescence (ages 14-16 years). As such, mean levels of emotional and behavioural problems can reliably be compared at the latent level across the four cohorts at this age.

11.3.2 Harmonisation in adulthood

Using the same approach described above, we identified four items that assessed similar symptoms across adulthood in the NSHD, NCDS, and BCS. Each of these items tapped a general psychological distress factor. We tested the measurement equivalence of these items within and across the cohorts. We found the best fitting model was a partially invariant model in which the threshold parameters for two of the four items (tension, fatigue) were freed. A full scalar model also provided acceptable levels of model fit; therefore we conclude that this 4-item subset is highly comparable across cohorts and age ranges, and it is justifiable to compare means and/or regression coefficients related to this item-set.

11.4 Equivalence of parent and teacher reports

Parent and teacher reports of child mental health were compared in the MCS and BCS70 cohorts. A harmonised item set was identified which was completed by both parents and teachers. Items corresponded to two dimensions: conduct and emotional problems. Parent and teacher measures were found to be reasonably invariant, suggesting that a lack of systematic bias due to reporter. In both the BCS70 and the MCS, teachers provided more information both on children's conduct and emotional problems. Although there is some agreement between parents and teachers in terms of predictive associations, both provide a unique insight on child functioning.

11.5 General guidance for retrospective harmonisation

In this final section we provide some general guidance on how to identify and harmonise an item pool based on the mental health data available in the British cohorts.

i. **Establish your research question:** Are you interested in comparing associations? In other words, is your cross-cohort research focussed on whether certain predictors or outcomes are differentially associated with mental health problems at different developmental periods or across generations? Alternatively, is your research question related to mean-levels of mental health within or across cohorts? For example, are you interested in growth/change/trends in mental health over the lifecourse or across generations?

ii. **Identify a harmonisable item pool:** When attempting to retrospectively harmonise different measures, the exact number of harmonisable items will vary depending on the number of cohorts and/or assessment waves that are relevant to your research question. Our searchable tool (https://www.closer.ac.uk/research-fund-2/data-harmonisation/harmonisation-mental-health-measures-british-birth-cohorts/) can be used to identify overlapping items that may be good candidates for within/cross-cohort research.

iii. **Establish measurement equivalence of your item pool:** Once you have identified your item pool, it is important to establish the measurement equivalence of these items (i.e. are they assessing the same underlying construct and to the same degree). *Metric invariance* (i.e. equality constraints placed on loadings) establishes whether the same construct is being assessed by your item set, and is important to establish if you wish to compare regression coefficients within or across cohorts. *Scalar invariance* (equality constraints placed on threshold parameters) is required in order to make valid comparisons of mean-levels of mental health problems at different time points or across cohorts. This form of invariance tests whether respondents from different cohorts or at different assessment waves are interpreting the items similarly and are attributing the same level of severity to responses. It is important to establish measurement equivalence even when even when identical measures are administered within/across cohorts, in order to ensure there are no systematic differences in measurement error due to age/cohort.

iv. **Using your item pool in subsequent analysis:** If you have established the requisite level of measurement invariance, the final step is to use your harmonised item pool to answer your substantive research question. As discussed in section 2.5, there are three methods of doing this: i) simultaneous estimation (i.e. include latent variables in your model using SEM), ii) produce and analyse factor scores, and iii) use observed score based on your harmonised item pool.

12. References

Anderson, L. R. (2018). Adolescent mental health and behavioural problems, and intergenerational social mobility: A decomposition of health selection effects. *Social Science & Medicine, 197*, 153-160.

Angold, A., & Stephen, C. (1995). Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents. *Age (years), 6*(11).

Armstrong, B. G. (1998). Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occupational and environmental medicine, 55*(10), 651-656.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences, 42*(5), 815-824.

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological methods, 14*(2), 101.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin, 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.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research, 17*(3), 303-316.

Bollen, K. A. (2014). *Structural equations with latent variables* (Vol. 210): John Wiley & Sons.

Bould, H., Joinson, C., Sterne, J., & Araya, R. (2013). The Emotionality Activity Sociability Temperament Survey: Factor analysis and temporal stability in a longitudinal cohort. *Personality and Individual Differences, 54*(5), 628-633.

Buss, A., & Plomin, R. (1984). *Temperament: Early developing personality traits.* Hillsdale, NH: Earlbaum.

Butler, N., & Bynner, J. (1997). 1970 British Cohort Study: ten-year follow-up. In: UK Data Archive Colchester, England.

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.

Collishaw, S. (2015). Annual research review: secular trends in child and adolescent mental health. *Journal of child psychology and psychiatry, 56*(3), 370-393.

Collishaw, S., Maughan, B., Goodman, R., & Pickles, A. (2004). Time trends in adolescent mental health. *Journal of Child Psychology and Psychiatry, 45*(8), 1350-1362.

Conners, C. K. (1969). A teacher rating scale for use in drug studies with children. *American journal of Psychiatry, 126*(6), 884-888.

Curran, P. J., McGinley, J. S., Bauer, D. J., Hussong, A. M., Burns, A., Chassin, L., . . . Zucker, R. (2014). A moderated nonlinear factor model for the development of commensurate measures in integrative data analysis. *Multivariate Behavioral Research, 49*(3), 214-231.

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological bulletin, 141*(4), 858.

Devlieger, I., & Rosseel, Y. (2017). Factor score path analysis. *Methodology*.

Elander, J., & Rutter, M. (1996). Use and development of the Rutter parents' and teachers' scales. *International Journal of Methods in Psychiatric Research*.

Ferrando, P. J., & Lorenzo-Seva, U. (2018). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement, 78*(5), 762-780.

Fortier, I., Raina, P., Van den Heuvel, E. R., Griffith, L. E., Craig, C., Saliba, M., . . . Ferretti, V. (2017). Maelstrom Research guidelines for rigorous retrospective data harmonization. *International journal of epidemiology, 46*(1), 103-105.

Ghodsian, M. (1977). Children's behaviour and the BSAG: some theoretical and statistical considerations. *British Journal of Social and Clinical Psychology, 16*(1), 23-28.

Gnambs, T., & Staufenbiel, T. (2018). The structure of the General Health Questionnaire (GHQ-12): two meta-analytic factor analyses. *Health psychology review, 12*(2), 179-194.

Goldberg, D. (1978). *Manual of the general health questionnaire*: Nfer Nelson.

Goldberg, D. P. (1988). User's guide to the General Health Questionnaire. *Windsor*.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Journal of child psychology and psychiatry, 38*(5), 581-586.

Hoshino, T., & Bentler, P. M. (2011). Bias in factor score regression and a simple solution.

Hu, L.-t., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological methods, 3*(4), 424.

Jenkinson, C., Coulter, A., & Wright, L. (1993). Short form 36 (SF36) health survey questionnaire: normative data for adults of working age. *Bmj, 306*(6890), 1437-1440.

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

Kline, R. B. (2015). *Principles and practice of structural equation modeling*: Guilford publications.

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

Little, T. D. (2013). *Longitudinal structural equation modeling*: Guilford press.

Muthén, L., & Muthén, B. (2018). *Mplus User's Guide. Eighth Edition.* Los Angeles, CA: Muthén & Muthén.

Oberski, D. (2016). Mixture models: Latent profile and latent class analysis. In *Modern statistical methods for HCI* (pp. 275-287): Springer.

Ormel, J., Raven, D., Van Oort, F., Hartman, C., Reijneveld, S., Veenstra, R., . . . Oldehinkel, A. (2015). Mental health in Dutch adolescents: a TRAILS report on prevalence, severity, age of onset, continuity and co-morbidity of DSM disorders. *Psychological Medicine, 45*(2), 345-360.

Patalay, P., & Gage, S. H. (2019). Changes in millennial adolescent mental health and health-related behaviours over 10 years: a population cohort comparison study. *International journal of epidemiology, 48*(5), 1650-1664.

Ploubidis, G., 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.

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

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental review, 41*, 71-90.

Rodgers, B., & Mann, S. A. (1986). The reliability and validity of PSE assessments by lay interviewers: a national population survey. *Psychological Medicine, 16*(3), 689-700.

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.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of personality assessment, 98*(3), 223-237.

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

Schoon, I., Sacker, A., & Bartley, M. (2003). Socio-economic adversity and psychosocial adjustment: a developmental-contextual perspective. *Social Science & Medicine, 57*(6), 1001-1015.

Sharp, C., Goodyer, I. M., & Croudace, T. J. (2006). The Short Mood and Feelings Questionnaire (SMFQ): a unidimensional item response theory and categorical data factor analysis of self-report ratings from a community sample of 7-through 11-year-old children. *Journal of abnormal child psychology, 34*(3), 365-377.

Shepherd, P. (2013). Bristol social adjustment guides at 7 and 11 years. *Centre for Longitudinal Studies*.

Steenkamp, J.-B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of consumer research, 25*(1), 78-90.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*(2), 173-180.

Stott, D. (1974). Manual to the Bristol Social Adjustment Guides. *San Diego: Educational and Industrial Testing*.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1-10.

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*, 770.

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Measurement invariance. *Frontiers in psychology, 6*, 1064.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods, 3*(1), 4-70.

Werneke, U., Goldberg, D. P., Yalcin, I., & Üstün, B. (2000). The stability of the factor structure of the General Health Questionnaire. *Psychological Medicine, 30*(4), 823-829.

Whiteford, H. A., Degenhardt, L., Rehm, J., Baxter, A. J., Ferrari, A. J., Erskine, H. E., . . . Johns, N. (2013). Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *The Lancet, 382*(9904), 1575-1586.

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice, 29*(3), 39-47.

Wing, J. K., Cooper, J. E., & Sartorius, N. (2011). *Measurement and classification of psychiatric symptoms: An instruction manual for the PSE and CATEGO program*: Cambridge University Press.

Xu, M. K., Jones, P. B., Barnett, J. H., Gaysina, D., Kuh, D., Croudace, T. J., & Richards, M. (2013). Adolescent self-organization predicts midlife memory in a prospective birth cohort study. *Psychology and aging, 28*(4), 958.

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology, 9*(2), 79-94.

Yoon, M., & Kim, E. S. (2014). A comparison of sequential and nonsequential specification searches in testing factorial invariance. *Behavior research methods, 46*(4), 1199-1206.