Resource report

closer
The home of longitudinal research

Feasibility of retrospectively harmonising cognitive measures in five British birth cohort studies

Eoin McElroy[1], Marcus Richards[2], Emla Fitzsimons[1], Gabriella Conti[1], George B. Ploubidis[1], Alice Sullivan[1], Vanessa Moulton[1]

[1] Centre for Longitudinal Studies, University College London
[2] MRC Unit for Lifelong Health, University College London

March 2021

Economic and Social Research Council

Copyright

How to Cite

McElroy, E., Richards, M., Fitzsimons, E., Conti, G., Ploubidis, G.B., Sullivan, A., Moulton, V. (2021). Feasibility of retrospectively harmonising cognitive measures in five British birth cohort studies. London, UK: CLOSER.

Table of Contents

List of Tables

List of Figures

Acknowledgements

This project is supported 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 ESRC took no role in the design, execution, analysis or interpretation of the data or in the writing up of the findings in this report

This resource report is part of a broader work package (CLOSER Work Package #19 'Assessment and harmonisation of cognitive measures in British birth cohorts'), that is supported by CLOSER's Innovation Fund. This initiative supports research that seeks to enhance and extend the research possibilities of 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 or 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 5 British birth cohorts: i) MRC National Survey of Health and 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); and v) 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 UCL. The NCDS, BCS70, and MCS receive core funding from the ESRC, and are hosted by the Centre for Longitudinal Studies, UCL. The next 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 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.

The authors would like to thank the owners of the five 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 and MCS. We would also like to thank the following contributions throughout this project:

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

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

The authors would also like to thank Dara O'Neill and Jennie Blows, CLOSER, for feedback on the report manuscript and for the creation of the online resource based on selected content from this report.

1 Introduction

The CLOSER British birth cohorts have collected a wealth of information on cognition over the life course of different generations. However, cross-study comparisons of cognitive ability are challenging as a multitude of different tests have been administered over the years (for a comprehensive guide see Moulton et al. (2020)). Such inconsistencies within and across studies represents a significant challenge for researchers who wish to pool and compare individual-level data from multiple studies or assessment waves.

The most common approach to calibrating measures of cognitive ability within and across longitudinal studies is the 'standardise and average approach' (Gross et al., 2014). Using this approach, raw scores on cognitive tests are standardised (typically converted to *Z*-scores), then averaged together. This produces a standardised composite score reflecting general cognitive ability. Although this approach is both intuitive and easy to interpret (i.e. measures now have a common metric), there are both conceptual and methodological weaknesses to this approach.

First, standardising *within* assessments removes the means and standard deviations; therefore these newly constructed composite variables cannot be used to answer research questions pertaining to mean levels of ability, e.g. group differences in ability or individual growth and decline in ability over time (McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009). Even when used to investigate and compare covariances, this approach *assumes* that each test captures the same underlying cognitive ability construct, and does so to the same degree.

Another common practice is to use principal components analysis (PCA) to derive a weighted composite variable using all or some (typically those that are at least conceptually consistent) of the available tests (Schoon, 2010). PCA is a data reduction technique that linearly transforms an original set of variables into a substantially smaller set of uncorrelated variables (i.e. components) that represent the maximum amount of information from the original set of variables (Dunteman, 1989). Although this goes some way to addressing the issue of equal weighting discussed above, the other issues persist.

When using PCA to conduct within- or cross-cohort comparisons, it remains assumed that the measures included capture the same underlying cognitive construct and do so to the same degree across different populations and/or measurement occasions.

To our knowledge, no attempts have been made as of writing to empirically test this assumption using the cognitive measures in the British cohorts. This report aims to explore the measurement equivalence of cognitive measures in the cohorts by taking a latent variable modelling approach (Skrondal & Rabe-Hesketh, 2004) to retrospective harmonisation.

1.1 Aims

This report is part of a broader work package that has two primary aims: i) document all of the measures of cognitive ability that have been administered in five of the CLOSER British birth cohorts using a consistent format, and ii) explore the feasibility of retrospectively harmonising (i.e. recoding/manipulating) these data in order to allow comparisons across cohorts and within cohorts over time. Our first aim was achieved in a companion report (Moulton et al., 2020), in which we documented 180 cognitive tests. We described various aspects of the tests (e.g. method, procedure, scoring). We also categorised each instrument according to the Cattell-Horn-Carroll (CHC) model of cognitive ability (Schneider & McGrew, 2018). The CHC model is the most psychometrically established model of cognitive ability[1]. It conceptualises cognitive ability as both multidimensional and hierarchical in nature, ranging from general ability ('*g*') to broad, narrow, and specific abilities (See Figure 1). By convention, abilities at the broad-stratum level are denoted with an abbreviation that begins with a capital 'G' (standing for 'general'), followed by lowercase letters; e.g. Gc (crystallised intelligence), Gf (fluid intelligence) (Schneider & McGrew, 2012). At the highest level of the hierarchy, a general cognitive ability factor ('*g*') is posited. This model has demonstrated a high degree of generality; a wide range of measures spanning multiple disciplines have been shown to conform to this structure

[1] Note that we categorise tests based on a psychometric model derived from covariances of test scores. Underlying neuropsychological processes may overlap across these domains.

(Jewsbury, Bowden, & Duff, 2017). For a detailed description of the CHC model, we refer readers to Schneider and McGrew (2018).

Our second aim, exploring the feasibility of retrospectively harmonising cognitive variables, will be addressed in the present report. Retrospective harmonisation (described in greater detail in Section 2 below) is a broad term used to describe the process of making existing data more comparable. There is no 'one-size fits all' approach to data harmonisation (Fortier et al., 2017; Griffith et al., 2015), rather, harmonisation strategies can vary widely depending on i) the nature of the available data, and ii) the purpose (i.e. research question) for which the data will be used. Given the conceptual and methodological heterogeneity of the measures (again see Moulton et al. (2020)), and the broad age ranges over which these measures were administered, deriving a universally applicable retrospectively harmonised measure of cognitive ability for use both within and across all five studies is currently highly problematic. This report however, aims to highlight key learnings from documenting the cognitive tests; and explore the measurement equivalence of identical or conceptually similar tests that were administered within cohorts over time, or across cohorts when assessments overlapped by age of the participants. This report is intended to provide guidance for researchers, offering advice on *where* and *to what degree* the various cognitive measures in the British cohort are comparable from a psychometric standpoint. This will be achieved by i) inspecting the available cognitive data for conceptual overlap, ii) retrospectively harmonising candidate variables (i.e. performing transformations in order to place variables on a common metric), and iii) testing for measurement invariance to determine the degree to which variables can be compared.

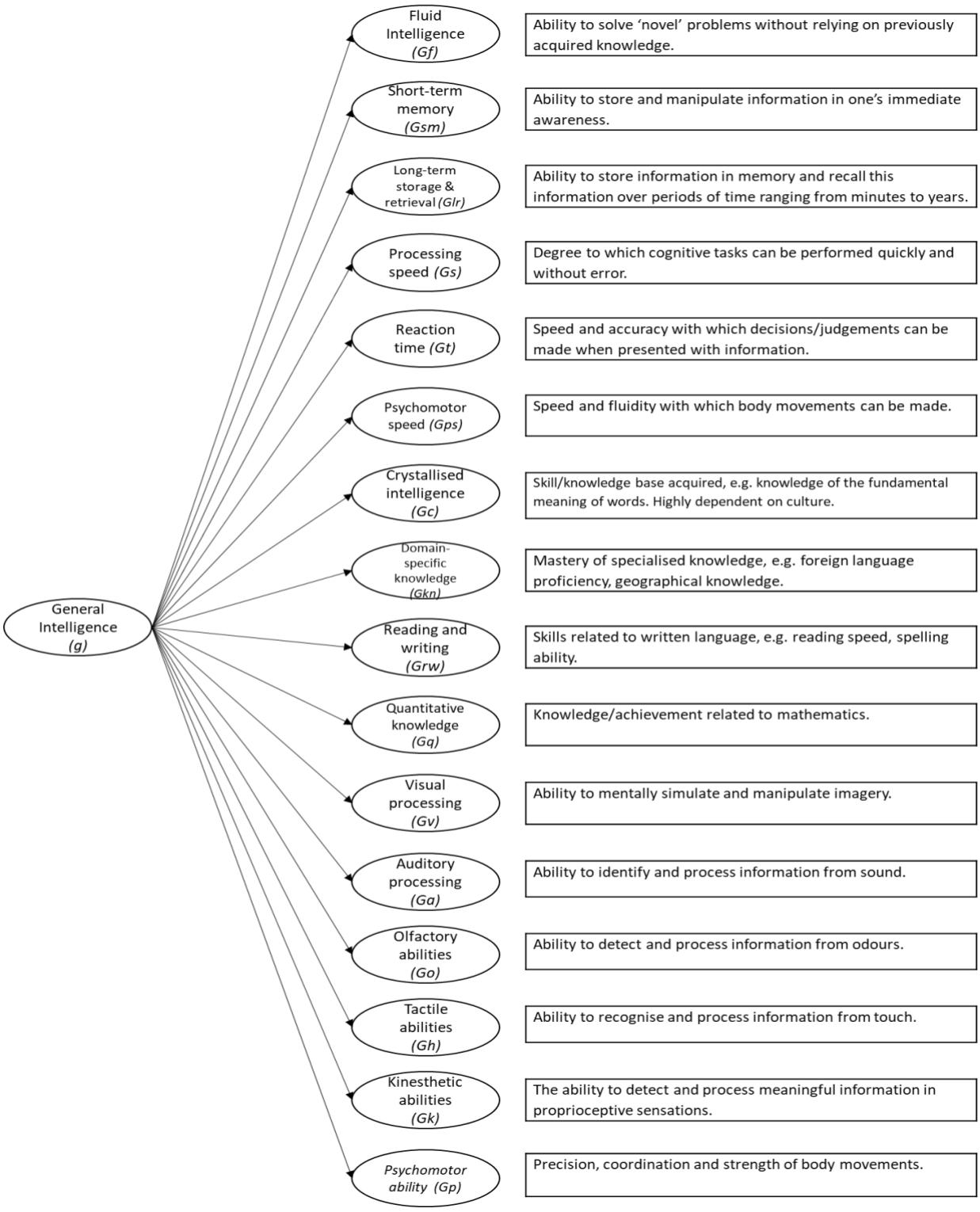

Figure 1. CHC model of cognitive ability. Adapted from Schneider and McGrew (2012).

1.2 Cohorts and tests included

This resource report explores the cognitive measures that have been administered in the following studies: i) MRC National Survey of Health of Development (NSHD), ii) the 1958 National Child Development Study (NCDS), iii) 1970 British Cohort Study (BCS70), iv) the Avon Longitudinal Study of Parents and Children (ALSPAC), and v) the Millennium Cohort Study (MCS). A brief description of each study follows:

MRC National Survey of Health and Development: The MRC NSHD is Britain's longest running 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. In addition to the main BCS70 sweeps, several sub-studies have been conducted, including 'Insight 46' - a neuroscience sub-study (2017), and the women's health survey (1989-1998). http://www.nshd.mrc.ac.uk/

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 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 at 55 (including a biomedical survey in 2002). A total of 9,137 cohort members took part in what is at the time of writing the most recent sweep. https://cls.ucl.ac.uk/cls-studies/1958-national-child-development-study/

1970 British Cohort Study: The BCS70 follows the lives of 17,198 people born in England, Scotland and Wales in a single week in 1970. The BCS70 began as the British Births Survey and participants have since been followed up nine times to date, spanning ages 5 and 46. A total of 8,581 cohort members took part in the most recent assessment at age 46. In addition to the main BCS70 sweeps, the following sub-studies (i.e. focussing on select sub-

samples) have been conducted: 1) Twins study (2008-2009), 2) Age 21 sweep (1992), 3) Age 7 sweep (1977) 4) 22 month and 42 month sweeps (1972-1973). https://cls.ucl.ac.uk/cls-studies/1970-british-cohort-study/

The Avon Longitudinal Study of Parents and Children: ALSPAC charts the lives of 14,541 people born in the former county of Avon between April 1991 and December 1992. Assessments have been administered frequently, with 68 data collection time points between birth and 18 years of age. Data is collected on parents and children, and more recently ALSPAC has started to recruit and collect data on the children of the original cohort members. http://www.bristol.ac.uk/alspac/

The Millennium Cohort Study: The MCS follows the lives of 19,517 children born in England, Scotland, Wales and Northern Ireland in 2000-02. Since the initial birth survey at 9 months, the cohort has been followed up six times at ages 3, 5, 7, 11, 14 and most recently as of the time of writing at age 17, when 10,757 cohort members took part. https://cls.ucl.ac.uk/cls-studies/millennium-cohort-study/

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/.

In the present report, we focus on measures that were administered to entire cohorts only (i.e. cognitive tests administered solely to targeted sub-samples were not considered for harmonisation due to their smaller sample sizes and lack of generalisability). Furthermore, we focus only on measures that were administered to the cohort members; any measures administered to the cohorts' parents, children of the cohort members or other parties were not included. Finally, we focus only on measures designed specifically to assess theoretically defined cognitive abilities (e.g. fluid reasoning, working memory, lexical knowledge, verbal comprehension), tests used to assess basic levels of skills (e.g. basic adult literacy) or subjective reports of cognitive difficulties were not included.

2 Retrospective harmonisation approach

Retrospective harmonisation is a term used to describe the process whereby existing data, either within or across different studies, are manipulated in some way in order to make them more directly comparable. Harmonisation projects are inherently idiosyncratic; each case depends on the nature of the data and the use for which it is intended. However, Fortier and colleagues (2017) offer broad methodological guidelines for the process, consisting of the following steps:

1. Assemble pre-existing knowledge and select studies
2. Select core variables to be harmonised
3. Process the data (i.e. convert data to a common format/scale where necessary)
4. Estimate quality of the harmonised variables generated
5. Disseminate and preserve final harmonisation products

2.1 Step 1: Assembling pre-existing knowledge

This step was completed in our companion report (Moulton et al., 2020) in which we catalogued all of the extant cognitive tests in five British birth cohorts. The additional steps will be discussed below.

2.2 Step 2: Selecting variables to be harmonised

The above catalogue allowed us to inspect the available cognitive data for overlap, by which we mean measures (either identical or conceptually similar) that were administered across multiple sweeps within a cohort, or at similar ages across different cohorts. In subsequent sections, we provide summary tables of the same and conceptually similar cognitive tests that were administered across different ages and cohorts. As a rule of thumb, when investigating the measurement equivalence of overlapping measures within and across cohorts, we prioritised cognitive tests that were identical. If identical tests

were not administered, we focussed on conceptually similar tests, and processed these data to place them on comparable metrics.

2.3 Step 3: Processing the data

In order for data from different measures to be compared in terms of mean levels, it is crucial that data are on the same metric (note this is not a necessary prerequisite if the research question concerns covariances). This is typically done using some form of ad-hoc algorithmic/recoding approach. The complexity of this process depends largely on the data in question. To give an example using physical health data, say a researcher wishes to harmonise the weight of participants across two studies, one of which has used the British Imperial System (lbs) and the other which has measured weight in metric kilogrammes (kg). A simple algorithmic transformation could be performed to place the imperial data on a metric scale (i.e. by multiplying all values by 0.45359).

When we were unable to locate identical tests within/across cohorts, we focussed on tests that assessed the same underlying cognitive construct. These tests often differed in terms of content, methods and scoring metrics. As these scales have not been calibrated with one another (as the lbs – kg conversion above), for testing purposes we primarily used simple monotonic linear transformations to place raw scores on comparable metrics (e.g. 0-50) using the following formula:

$$s_i = s_1 + \left(\frac{s_n - s_1}{p_n - p_1}\right) \cdot (p_i - p_1)$$

Where p denotes the primary scale and S the common secondary numerical scale, and p_i the primary test score. The discrete response options consecutively numbered from P_1 to P_n, (with n the number of response options) are projected onto the common secondary numerical scale, ranging from a lower bound S_1 to an upper bound S_n.

This process simply changes the absolute metric of a test[2]. This approach is different from standardisation, such as converting data to Z-scores. Z-scores place variables on a metric that is relative to the population mean and standard deviation. When using Z-scores, it is not possible to compare mean levels of a given variable across cohorts or assessment waves as information about the scale is lost; i.e. data points are expressed in terms standard deviations from a mean of 0.

2.4 Step 4: Estimating the quality of harmonised variables using a latent variable modelling approach

Once the candidate variables were identified and processed (where necessary), it was vital to estimate the overall quality of this harmonisation process. To do this, we employed a latent variable modelling approach in which we tested the measurement equivalence of these tests across assessment waves and/or cohorts. Under this framework, and in line with the CHC model of cognitive ability, each test was conceptualised as an observable indicator of an unobservable (i.e. latent) general cognitive ability variable (Figure 2).

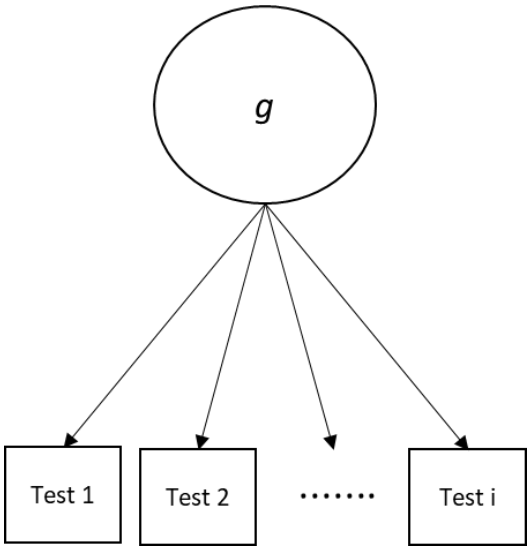

Figure 2. Graphical illustration of latent variable model of general cognitive ability (*g*).

[2] Readers should note that rescaling a variable may alter the variance of the test in question, although such alterations are common and permissible in SEM, e.g. http://www.statmodel.com/discussion/messages/14/11947.html?1362454606

To conduct valid comparisons of '*g*' across time and/or populations, it is important that the underlying measurement model of *g* is equivalent (Van De Schoot et al., 2013). In other words, the relationship between '*g*' and its measured indicators (i.e. the various cognitive tests) should be consistent across assessment sweeps and/or studies. After selecting identical or conceptually similar tests and applying the necessary transformations to place them on a common metric, we assessed the psychometric equivalence of these scales by testing for measurement invariance (MI). A failure to support MI would suggest that '*g*' has a different structure or meaning across cohorts/sweeps, and therefore cannot be meaningfully compared. Although it is beyond the scope of this report to provide an in-depth technical discussion and/or tutorial, we provide a short conceptual overview of this process. For further in-depth discussions of MI, see (Putnick & Bornstein, 2016; Van De Schoot, Schmidt, De Beuckelaer, Lek, & Zondervan-Zwijnenburg, 2015).

In summary, we tested for MI by fitting a series of nested confirmatory factor models (CFAs), in which increasingly strict equality constraints were placed on measurement parameters across different cohorts/assessment waves. The CFA model used to test for MI can be thought of as a simple linear regression model in which the observed score of a cognitive test (Y) is predicted by the unobserved general ability factor (η) (Mellenbergh, 1994). The relevant measurement parameters are the regression slope/weight known as the factor loading (λ), the intercept (τ) which reflects the point at which the slope crosses 0, and an error term (ε) (Wicherts & Dolan, 2010). As such, an individual's score on a particular test (Y1) can be calculated using the formula:

$$Y1 = \tau 1 + \lambda 1 \eta 1 + \varepsilon 1$$

If, after fitting equality constraints across cohorts/sweeps, we do not observe a worsening of absolute model fit, then said level of MI is judged to hold, and the parameters in question can be considered equivalent.

Before conducting any harmonisation using the cognitive data in the British cohorts, we encourage researchers to thoroughly consider their research question, and the intended use for the harmonised data. In particular, researchers should be clear on whether their primary question concerns i) associations between cognitive variables and

predictor/outcome variables within or across cohorts, *or* ii) comparisons of mean levels of cognitive ability (e.g. studies of growth/decline/change in cognitive ability, cross-cohort comparisons of population-level cognitive ability). Depending on the answer to the type of research question, different levels of measurement invariance of harmonised variables need to be satisfied:

I. *Configural invariance:* This is the least restrictive model. The same measurement model is specified in each cohort/wave. However, no equality constraints are placed on the parameters; i.e. factor loadings, intercepts and errors are allowed to differ the across cohorts/sweeps (Figure 3A). 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:* Metric invariance is tested by holding the factor loadings equal across cohorts/sweeps (Figure 3B). If metric invariance holds, we can conclude that the tests are associated with 'g' in a consistent manner across cohorts/sweeps. At this level of MI, we can be confident that we can compare variances and covariances at the latent level. In simple terms, if metric invariance holds, we can compare correlation/regression coefficients across cohorts/sweeps (provided any predictor or outcome variables included in the models are also consistent). *In the case of the British cohorts, this level of invariance is important for researchers looking to examine whether particular associations between cognitive ability and predictor/outcome variables are consistent across cohorts and/assessment waves.*

III. *Scalar invariance:* This is tested by holding both the factor loadings and intercepts equal across different sweeps/cohorts (Figure 3C). If scalar invariance holds for a given test, the underlying level of the test can be considered equivalent across groups. In other words, individuals from two different cohorts who have the same level of 'g' will demonstrate the same score on a scalar invariant cognitive test. Scalar invariance allows us to compare latent means across groups; therefore *it is*

particularly important for researchers who are interested in using the British cohorts to compare mean scores on cognitive tests across cohorts/assessment waves.

IV. *Strict invariance:* Strict invariance is tested by holding the factor loadings, intercepts and residuals (ε) equal across sweeps/cohorts. If strict invariance holds, then any difference observed between cohorts/sweeps can be attributed solely to a difference in the underlying latent variable ('g'). 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 time/groups, random error, by its very 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 area is still rather limited (see Putnik et al., 2016 for an overview), and there is no consensus as to how many parameters can be released while maintaining meaningful comparisons. Chen (2008) demonstrated that the bias in mean estimates across groups increased in proportion to the number of non-invariant factor indicators, therefore it is desirable to have as many invariant indicators as possible. Most guidelines suggest that at least half of the indicators should be invariant across cohorts/sweeps 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 non-invariant cognitive tests, 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 (2014), who proposed a 'backwards method' of releasing parameters one at a time based on the size of their relevant modification index.

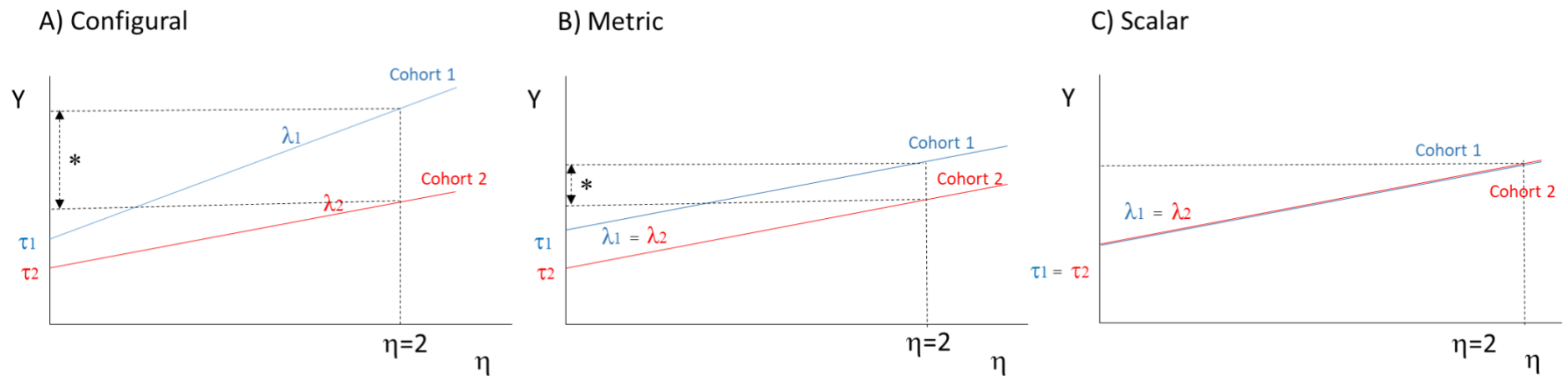

Figure 3. Conceptual illustration of different levels of measurement invariance of a single item across two cohorts. Y = Observed score. η = Unobserved score on latent variable. λ = factor loading of a test (slope of regression line between observed and latent score). τ = intercept of a test (point at which slope/loading crosses 0). Dotted line = predicted score on observed test at η = 2. * = bias in observed score attributable to group/cohort membership. Adapted from Wicherts and Dolan (2010).

In the present work, we were primarily interested in whether measurement parameters were consistent within cohorts over time (i.e. at different ages), and across cohorts at similar ages. Depending on whether we were testing differences within or across cohorts, different configurations of the CFA model were used to test for MI. When comparing measures across cohorts, we use multiple-group confirmatory factor analysis (MGCFA; Figure 4), in which the measurement models and equality constraints were fitted to independent groups (Meredith, 1993). In the case of within-cohort invariance, we used a series of longitudinal confirmatory factor models (Little, 2013; Figure 5). Although the general principle remained the same (i.e. the same measurement model is specified at different cohorts or sweeps, and equality constraints are then placed on parameters), the longitudinal model differs slightly in that it includes correlations between the latent variables over time, and residual correlations between the same indicators over time.

Both the MGCFA and longitudinal factor models were estimated using the robust maximum likelihood (MLR) estimator, which adjusts for violations of non-normality in continuous data. The fit of configural (i.e. baseline) models 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 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).

When comparing the metric and scalar models with the configural model, we used common conventions for determining whether the imposed equality constraints led to an overall worsening of model fit. A change of less than 0.015 in the RMSEA and -0.01 in the CFI and TLI were considered acceptable; i.e. the equality constraints did not lead to a

significant worsening of it (Chen, 2007). However, it is important to stress that, although these guidelines are frequently used, there remains no consensus about the best fit indices or cut-off values when comparing fit indices across all conditions (Putnick & Bornstein, 2016).

All tests of MI were conducted in Mplus Version 8.0 (Muthén & Muthén, 2017).

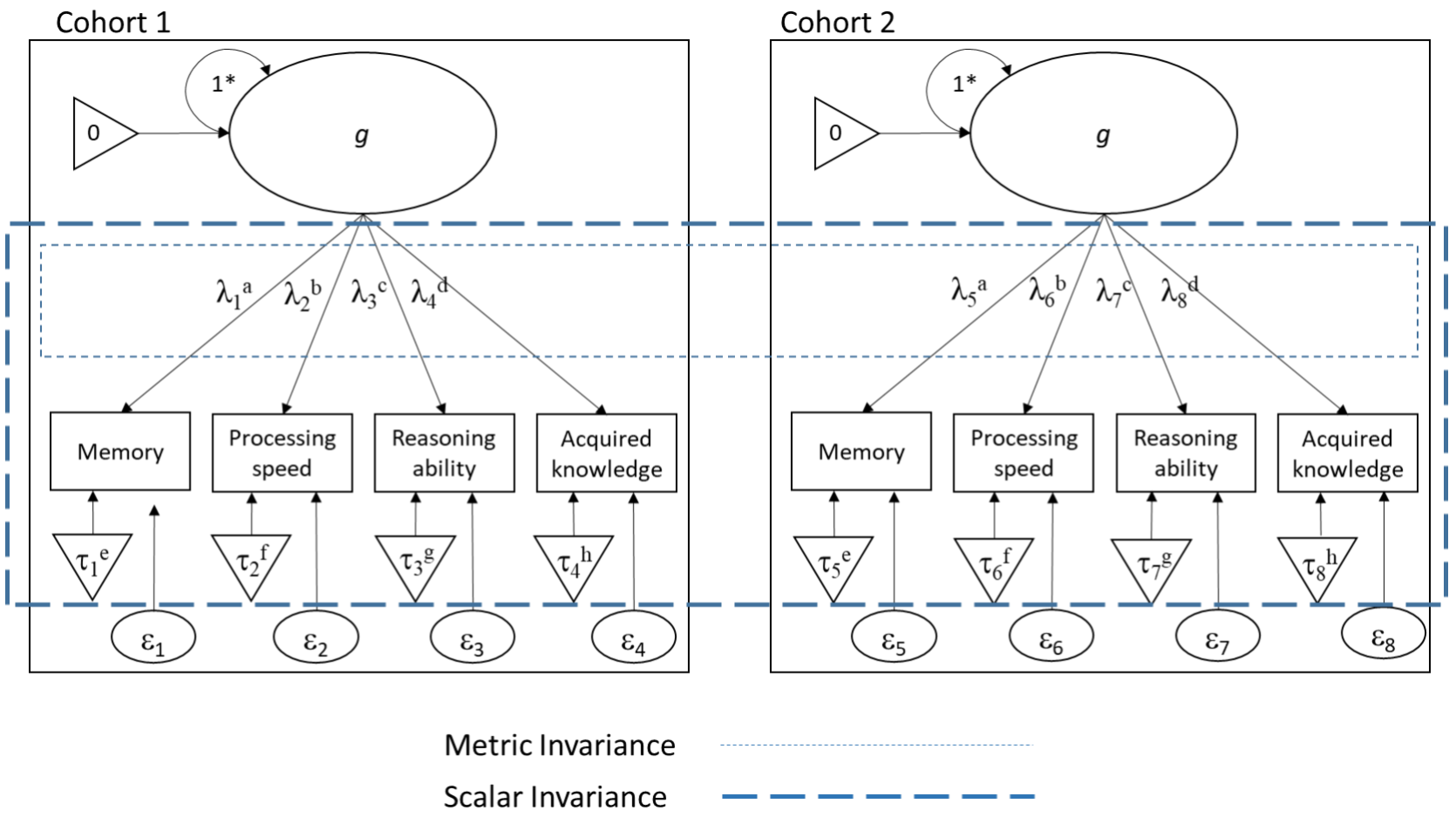

Figure 4. Graphical illustration of a multiple-group confirmatory factor analysis (MGCFA). Example consists of 4 measured indicators (i.e. cognitive tests) loading onto a general cognitive ability factor (G). λ = factor loading. τ = intercept. ε = residuals. a-d = factor loadings held equal. e-h = intercepts held equal.

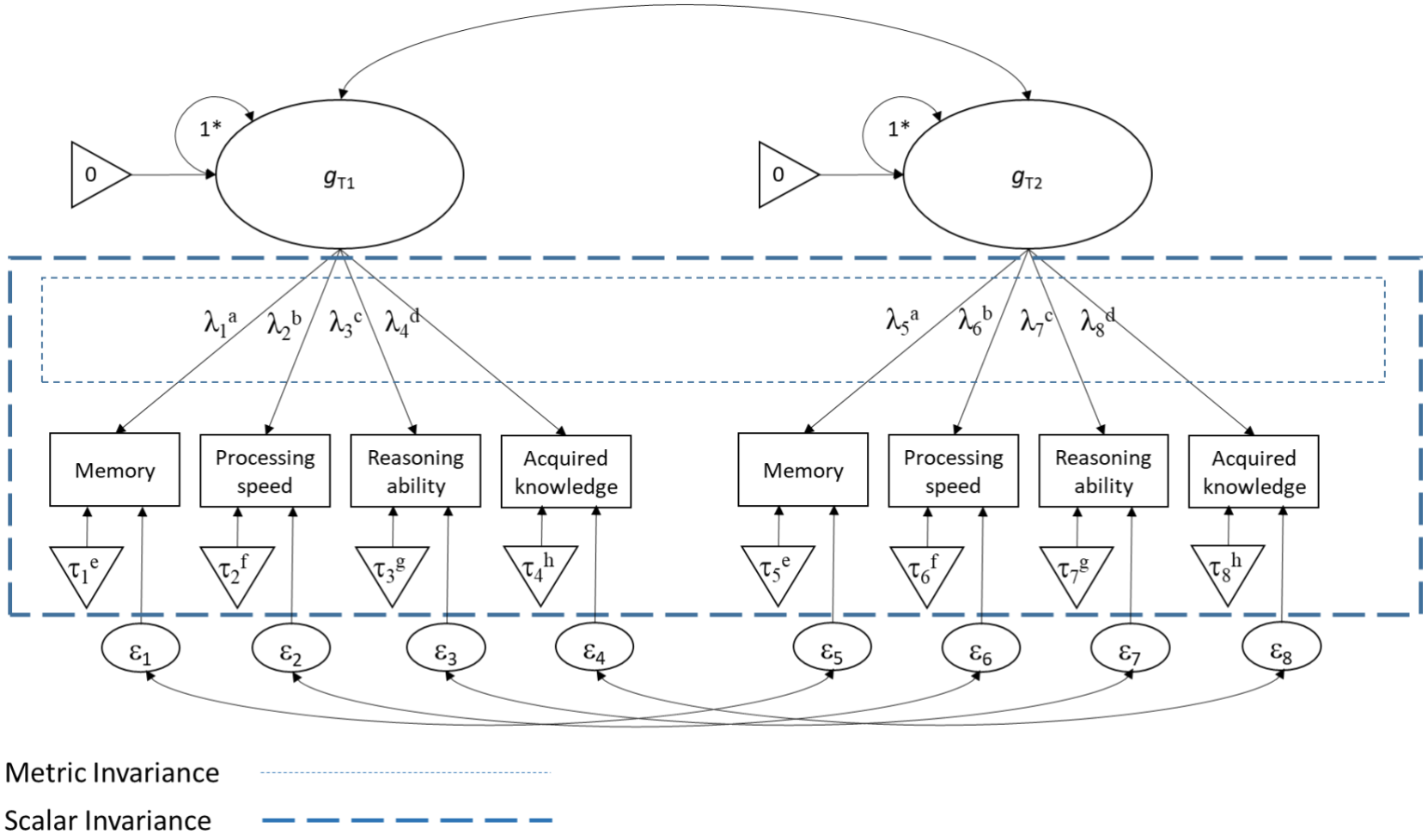

Figure 5. Graphical illustration of a longitudinal confirmatory factor analysis. Example consists of 4 measured indicators (i.e. cognitive tests) loading onto a general cognitive ability factor (G). λ = factor loading. τ = intercept. ε = residuals. a-d = factor loadings held equal. e-h = intercepts held equal.

2.5 Step 5: Disseminating and preserving final harmonisation products

The ultimate goal of harmonising cognitive ability measures is to create a set of scores that can be compared within or across studies. After measurement invariance has been established, the next step is to implement these harmonised scores in subsequent analyses 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.

1. The preferred method for incorporating latent variables into any analysis is to include them in the model directly. In other words, estimate both the measurement model (with equality constraints placed on loadings and intercepts) 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.

2. A practical approach to addressing these issues is to employ a two-step approach in which measurement models (with equality constraints placed on loadings and intercepts) 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.

3. In instances where full scalar invariance has been supported, the estimation and interpretation of factor scores is relatively straightforward. 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.

Dissemination is also important, for transparency and to allow other research to replicate and/or adapt harmonisation approaches. It is important that researchers provide detailed descriptions of their harmonisation strategies and analysis, share their code, and where possible make their harmonised variables available to others (Fortier et al., 2017). See Appendix II for syntax used in this report.

3 Learnings from assessing the cognitive measures

Before describing our attempts at harmonisation, we discuss several learnings from our work documenting the cognitive tests and assessing the feasibility of the tests for harmonisation (Moulton et al., 2020). We hope that by sharing this knowledge researchers will have a better understanding of some of the challenges and issues they may need to take into account of when using the cognitive tests in their work.

3.1 Features of the cognitive tests administered in the cohorts

Mapping all of the cognitive tests by cohort and age at administration resulted in a clear overview of the range of cognitive tests in the cohorts, as well as any gaps (Moulton et al., 2020). As outlined in our companion report (Moulton et al., 2020), there were several conventions in the available tests. By definition, the cognitive tests administered in the cohort studies were period specific. In the earlier cohorts, tests in childhood appeared to i) reflect the curricula of the period, ii) were administered using a traditional pen and paper approach, and iii) were often devised specifically for the cohort study in question. In the more recent cohorts, standardised ability tests (e.g. British Ability Scales (BAS) and Wechsler (WISC)) were employed, as well as more varied modes (e.g. computer-assisted personal interviewing) and specialist cognitive domains (e.g. decision making, speed, short-term memory). These differences in how the cognitive tests were conceptualised and administered across each of the cohorts should be considered.

In addition, cognitive tests tend to be devised for specific populations, with age of the subject population an important criterion, particularly in childhood. Therefore, repeating exactly the same cognitive test to a cohort member is atypical and problematic without some form of age relevant adaptation. For example, in the BCS70, a shortened version of the Edinburgh Reading Test originally devised for the cohort at age 10 was adapted for the cohort when they were 16 years old. In another example in the MCS, the BAS II test Naming Vocabulary administered at ages 3 and 5 was developed for use with children in their 'early years' from the ages of 2 and 6 months to 7 years and 11 month; the starting points on the test were dependent on the child's age, while the finishing point and items

included, on their age and ability (Elliott, Smith, & McCulloch, 1997). If the same test has been administered over different sweeps within cohort or at different ages across cohorts any test adaptations should be noted. In addition, some of the standardised ability tests (e.g. since the 1970's the BAS has been revised on three occasions) have been updated and revised over time.

Similarly, the cognitive domains tested in adulthood were very different from those featured in childhood. In later life, the focus was on capturing cognitive functions purported to be important in everyday contexts (e.g. executive function, memory and processing speed), as opposed to tests of reasoning and acquired knowledge which were typical in childhood. In addition, there were very few cognitive tests administered in any of the cohorts between the ages of 20 and 40; thus, early to mid-adulthood has been an overlooked period for the study of cognitive development in the cohorts. This is not unique to the British cohort studies; it is also a feature of cognitive tests in general, as highlighted by Ackerman (2017, p.987), "little thought has been given on how adult '*intelligence*' may differ from child and adolescent *'intelligence'*." As a result, our approach has been to retain the inherent life stage divide and focus our analyses, particularly within cohort, on the very distinct cognitive measures in childhood (up to age 16/17), and mid to late adulthood.

3.2 Age at which the cognitive tests were administered

In general, birth cohort studies are homogenous samples, wherein there are minimal differences in participant age at study entry. However, depending on when and how the cognitive tests were administered, as the cohort members became older the exact age of testing began to diverge for cohort members. Table 1 below shows the age mean, range and standard deviation (SD) in months for major sweeps in childhood. Depending on the life course stage, the cohort and the type of cognitive test administered, as well as the research question and hypothesis, accounting for age may or may not be an issue. For example, in harmonising cognitive measures within a cohort, the age difference for an individual from one sweep to another may vary considerably. When making comparisons across studies, the age range within sweeps can be wide; however for the majority the age

may be centred within a restricted range. For example, at age 11 in the NSHD, although the age varied by 10 months, 80% of the cohort were aged 130 or 131 months, while in the aged 10 BCS70 sweep although the age ranged by 23 months, 84% were assessed between the ages 120 to 123 months.

Researchers are advised to check the age each test was administered; details, where possible are outlined in (Moulton et al., 2020).

Table 1. Age (in months) at time of test administration

Age range:	NSHD	NCDS	BCS70	ALSPAC	MCS
Age 4/5	-	-	60-77 m=62 SD=1.3	48 – 51 m=49 SD=0.4	53-74 m=63 SD=3
Age 7/8	8 years 6 months[3]	82 – 93 m=85 SD=1.6	-	89 – 127 m=104 SD=3.9	76 – 98 m=87 SD=3
Age 10/11	128 - 137 m=130 SD=1.1	130 – 152 m=134 SD=1.7	117-139 m=122 SD=2.7	125 - 164 m=141 SD=2.9	122 – 148 m=134 SD=4
Age 14/16	172 - 182 m=175 SD=2.1	190 – 201 m=193 SD=1.4	189-212* m=197 SD=4.5	171 - 212 m=186 SD=4.2	157 - 184 m=171 SD=4.1

*School sample only

3.3 Administration of the same cognitive test across the cohorts

There are very few occasions in which the same cognitive test was administered in different cohorts, as shown in Table 2. Indeed, there are only three cognitive tests where the test items, administration and procedure was exactly the same in two or more cohorts; a general ability test (National Foundation for Educational Research (NFER)) administered at age 11 in the NSHD and NCDS, and in mid to later life the verbal fluency (animal naming) and the timed letter search tests in the NSHD, NCDS and BCS70.

[3] Age at time of interview not available. 8 years 6 months is a best estimate based on date of interview.

Table 2. Overview of same test completed by cohort member (CM) and repeated across cohorts by test and cohort

Test	NSHD	NCDS	BCS70	ALSPAC	MCS
General ability test (NFER)	11 (128- 137)	11 (130 - 152)			
BAS similarities (word or verbal)			10 (117-139) [BAS]		11 (122-148) [BAS II]
Copying Designs Test (CDT)		7 (82 - 93) 11 (130-152)	5 (60-77)		
Human Figure Drawing (HFD)		7 (82-93)	5 (60-77)		
APU Vocabulary test			16 (189-212) 42 (500-517)		14 (157-184)
Verbal Learning/Word List Recall	43 (514-533) 53 (636-650) 60-64 (724-780) 68-70 (828-848)	50 (598-614)	46 (542-578)		
Timed Letter Search/Letter Cancellation	43 (514-533) 53 (636-650) 60-64 (724-780) 68-70 (828-848)	50 (598-614)	46 (542-578)		
Verbal Fluency (animal naming)	53 (636-650)	50 (598-614)	46 (542-578)		

Age administered in years and (age range in months) in parentheses

The other tests outlined in Table 2, although putatively the same test, had differences in either the mode, scoring, wording or number of questions asked (discussed below).

BAS Similarities

The BAS Similarities (word) test administered in the BCS70 at age 10 is the predecessor to the BAS II Verbal Similarities measured in the MCS at age 11. In both versions, cohort members were asked to describe how word-pairs were similar, and across both tests there were common items. However, in the MCS version, there were age relevant start and finish points. As such, although there are 12 similar items, only 3 would have been attempted by the MCS majority, the rest dependent on their responses to a set of age determined items. In addition, in the BCS70 version for each groups of items (e.g. apple, orange, banana), children were asked to name another word consistent with the group; this element is not included in the latter versions.

Copying Designs Test (CDT)

The CDT in the NCDS at age 7 and 11, consisted of the same 6 designs to copy. In the BCS70 at age 5 there were 8 designs, 5 of which were also in the NCDS versions.

Human Figure Drawing (HFD)

In both the NCDS age 7 and BCS70 age 5 the child was asked to draw a picture of a man (or a lady), and to draw a whole person, not just the face and head. The scoring scheme differed in the two cohorts - the BCS70 used the Harris point scoring system (0-30), while the NCDS used the Koppitz scoring system (0-100).

Applied Psychological Unit (APU) Vocabulary test

The APU Vocabulary test was asked in the BCS70 at age 16 and age 42, and in the MCS at age 14. The original test included 75 words and was completed in the BCS70 at age 16. The test in the BCS70 at age 42 and the MCS at age 14, was a shortened version including 20 items from the original test.

Verbal Learning/Word List Recall

When participants were aged 53 years in the NSHD, they were shown a list of 15 words at a rate of one word every two seconds. They were then asked to write down as many words recalled as possible. In contrast, participants in the NCDS and BCS70 (age 50) were played a list of 10 words via a recording, one word every two seconds, and were then asked to orally recount as many words as they could remember.

Timed Letter Search/letter Cancellation

At the mid-life (age 46-53) assessments in NSHD, NCDS, and BCS70, cohort members were given a page of random letters arranged in rows and columns and were asked to cross out as many target letters ("Ps" and "Ws") as possible within a one-minute timeframe. The number of rows and columns and placement of target letters differed across cohorts. Furthermore, three trials were administered in the NSHD, whereas only one trial was administered in the other two cohorts.

Other tests

In addition, some tests administered across cohorts measure the same domain, but have been devised by different test developers e.g. BAS Matrices in the BCS70 at age 10 and 16 and the Wechsler Abbreviated Scale of Intelligence (WASI) Matrix reasoning in ALSPAC, age 15; BAS II Pattern construction in the MCS ages 5 and 7 and the Wechsler Intelligence Scale for Children (WISC-III) Block design in ALSPAC ages 4 and 8.

Further details on each particular test are outlined in our companion report (Moulton et al., 2020). In preparation for the harmonisation process and to aid researchers, further details on similar tests repeated and comparable constructs within and across cohorts are outlined in Appendix I.

3.4 Cognitive test scores and scales

As well as the content, domain, administration and age appropriateness of the cognitive tests, the scores and scales employed in each cognitive test should also be considered.

Compared to physical measurements, which can be clearly defined and consistently measured using precise instrumentation, the measurement of unobservable constructs such as cognitive ability and knowledge is complex. As outlined earlier in this section, there is a wide range of tests even within a specific domain such as reading. These tests also have heterogenous scoring systems and scales.

In the main, there are three types of scales operationalised in the cognitive tests in the cohorts. Most cognitive scales in the cohort are deemed interval, whereby the test score is based on the number correct, where each item in a test score is worth the same amount when calculating the total. The assumption is that a difference in score points reflects a consistent difference in the construct no matter where the test taker is on the scale. However, most tests (although treated as interval scales) are not, as each item on the test does not have a similar level of difficulty. Most tests consist of items ranging from easy to more difficult, which get progressively harder throughout the test. It is important to note that tests used in the more recent cohorts, use more complex test designs (e.g. BAS II in the MCS) which take account of the item difficulties in both administering and scoring the cognitive tests.

Several of the cognitive tests in the cohorts employ a ratio scale, with a meaningful absolute zero. For example, a number of tests use total count scores, such as the Verbal Fluency test in the NSHD, NCDS and BCS70 where each named animal contributes the same amount to the total score, ranging from zero to total number of named animals, regardless of whether the animal named is commonplace or unusual. Other tests consider speed taken to complete the test or reaction times, both of which are on ratio scales.

There are various rubrics applied to the item-level scores in the different cognitive tests in the cohorts. The main scoring mechanism employed is dichotomous. (polytomous scoring refers to three or more outcome responses for a given item, and tends to relate to rating scales, but also include categorical responses such as the correctness of a response, for example incorrect, partially correct and fully correct). Traditionally, the raw score has been applied, defined as the number, proportion or percentage of test items a participant answers correctly (Petersen, Kolen, & Hoover, 1989). This implies items are scored right (1)

or wrong (0) and the raw score is based on the sum of the number of items correctly answered. In addition, nonresponse is not an option and is treated as incorrect. Although, some tests involve multiple-choice questions, they are usually scored dichotomously.

Computerisation has led to the use of innovative item formats, which can apply more complicated scores than right/wrong or a simple sum of items (Kolen & Brennan, 2014). For example, the use of item response theory (IRT) has enabled test developers to construct scales incorporating fewer items from all the test items. IRT makes assumptions related to the probability a test taker will produce a particular response to a particular item, given a particular ability level (Yen & Fitzpatrick, 2006). For example, the BAS II tests in the MCS use the Rasch (1960) model of item analysis and test scaling, as well as start and stop rules dependent on age and responses to items on the test. The total raw scores are not equivalent between individuals. In addition, the adjusted total scores, are based on which items are answered correctly and not the number of items completed.

In addition, score scales are often modified to have certain properties, such as size of the score intervals, different midpoints and variability. A common approach is to convert the score to a z-score, with a mean of zero and standard deviation (SD) of 1; this has been applied to some of the total scores available, for example in the BCS70. Other transformations employed include the t-scale, with a mean of 50 and SD of 10, derived in the NSHD. For the standard score, the transformation most allied with IQ scores, a mean of 100 and SD of 15 was applied to total scores on the Weschler IQ scales in ALSPAC and for childhood measures in the NSHD.

More complex transformations to total cognitive scores are also available in the cohorts, by incorporating normative information to test scores. The cognitive test is administered to a norm group, and the scale score distribution is set relative to this norm group. Therefore, the scale score is meaningful to the extent that the norm group is central to score interpretation (Kolen, Tong, & Brennan, 2009). For example, in the MCS, as well as a total raw and ability adjusted scores for the BAS II cognitive tests, there are also age-normed adjusted scores. These are based on the results of the BAS II tests administered to a representative sample of 3 to 17.11 year olds in 1995 (Elliott,1997).

Along with different approaches to the measurement of cognitive abilities/knowledge and modifications, if any, to the total cognitive scores there are also inconsistencies in what and how the data has been deposited in each of the cohorts. In some of the earlier cohorts e.g. NCDS in childhood, there is no item level data provided, only a total raw score. Therefore, no measurement invariance analysis can be conducted to compare these tests with other cognitive tests at an individual test level or to identify discriminating items. Other cohorts have provided modified total scores only (although individual items may be available on request), e.g. CANTAB Cambridge Gambling Task in the MCS, and Weschler IQ scores in ALSPAC with no raw or item level information provided. In contrast, for some cognitive tests the data has been deposited at an item only level; researchers will have to derive their own total scores. In addition, with some of the cognitive tests there is also additional information provided on the test environment and any difficulties the CM might have encountered during the test.

It is also worth noting that some of the cognitive tests have severe floor and ceiling effects. In some cases, the test has been administered as a benchmark and included in a cohort sweep earlier than the age the test was developed for. For example, in the BCS70 most children did not respond to the Schonell Reading test, administered at age 5 as they could not read and thus were given a score of zero resulting in serious floor effects. Ceiling effects are found in the Addenbrooke's Cognitive Examination (ACE-III) administered at age 68-70 in the NSHD; the ACE-III was designed to detect cognitive impairment, which is positively related to age.

One further aspect to consider when equating cognitive tests is the difficultly of the test and the differential ability of the test takers on each test. Differences in the distribution of the resulting scores can be a result of both these factors. With particular reference to cross-cohort comparisons, differential ability of test takers is a confounding factor which needs to be excluded before adjusting for the difficulty of the test (Dorans, 2018). When comparing tests (measuring the same construct), there are two approaches which can be utilised to separate these issues: to use a common population of test takers or to use an anchor measure of the construct being assessed by tests X and Y, ideally measuring the same construct. The anchor approach assumes that the performance on a set of items or

test can quantify the ability differences between two distinct but not necessarily equivalent test takers. As outlined in section 3.3 there are few occasions where the same test (or common items) has been measured across cohorts, and therefore an 'anchor' approach could not be used. Additional external information on how tests (measuring the same construct) relate to each other is needed.

4 Feasibility of retrospective harmonisation: Testing for measurement invariance in the British birth cohorts

This first step in our approach to retrospective harmonisation was to scan the available data for overlap both within and across cohorts. By overlap we mean measures (either identical or conceptually similar) that were administered across multiple sweeps within a cohort, or at a similar age across different cohorts. As noted in section 3.3, there were only a few instances where the exact same cognitive test was administered either within or across different cohorts. Even when we expanded our criteria to measures that were conceptually similar, there were still limited instances of overlap. To compare a simple unidimensional factor model (e.g. '*g*') across cohorts, four measured indicators are required to ensure model identification while allowing for an assessment of model fit. For comparisons across assessment waves (i.e. within-cohort), three measured indicators are required across at least two time points (Little, 2013). Therefore, to identify and assess the measurement equivalence of a latent '*g*' factor, we required a minimum of four cognitive tests to be common across cohorts or three common tests within cohorts. Based on our examination of the available data, we found only two examples in which a sufficient number of tests overlapped across cohorts to explore the measurement equivalence a '*g*': i) middle childhood (age approximately 10/11 years) and ii) midlife (age approximately 46-53 years). We found no examples where three or more tests overlapped within any of the cohorts. As a result, a modified form of longitudinal factor analysis was explored as a means of incorporating overlapping and non-overlapping tests into a common measurement model (Curran et al., 2014), see section 4.3 for further details.

4.1 Testing for measurement invariance at age 10/11 years

The five British birth cohorts measured cognition to varying degrees around the age of 10 to 11, as shown in Table 3 below. For this age-group, the exact same General Ability Test comprising of both Verbal and Non-verbal sub-scales was administered in both the NSHD and NCDS. The Verbal sub-scale was a verbal reasoning test, conceptually akin to the Word/Verbal similarities test in the British Ability Scales administered in the BCS70 and

MCS respectively, albeit using different procedures and response conventions. In addition, the non-verbal reasoning test in the NSHD and NCDS was similar to the BAS Matrices Test, administered in the BCS70. Also, in the NSHD, NCDS and BCS70, a mathematics test, although different across the three cohorts, was administered at this age-group.

The ALSPAC tests at age 10 and 11 captured working memory and decision speed, and reaction time, attention, processing speed, and higher conceptual reasoning, respectively. These cognitive tests were conceptually very different to tests that were administered in the earlier cohorts. In addition, the speed of test completion was also an important factor in these tests, which incorporated another element of difference compared to the earlier cohorts. For both of these reasons, ALSPAC was excluded from the analysis at this age-group.

Although, in the MCS, the BAS II Verbal similarities test was measured at age 11, the two other tests that were administered at this age were the CANTAB Cambridge Gambling Task and Spatial Working Memory Task. Both of these tasks were specialist tests. Although conceptually similar to some tests administered in ALSPAC, they applied different modes and assessed different cognitive processes from tests in the earlier cohorts. Both tests had several key summary outcome variables which were derived by the test developers; for more details, researchers are directed to Atkinson (2015). We conducted a series of CFAs to investigate the relation between the key cognitive MCS measures; however no suitable latent construct could be identified (as model fits were extremely poor). Therefore, the MCS was also excluded from the analysis to test for measurement invariance at age 10/11. Both ALSPAC and MCS contain detailed measures of cognitive ability at younger ages (from age 4 months); however these were not included in the present analyses as participants were judged to be too young at assessment (8.5 years or younger).

Table 3. Comparable constructs assessed at age 10-11

	NSHD	NCDS	BCS70	ALSPAC	MCS
Gc (Crystallised ability)	General ability (NFER) Verbal Test Vocabulary	General ability (NFER) Verbal Test	Edinburgh Reading Test (ERT) (Word) Similarities (BAS) Word Definitions (BAS)		Verbal similarities (BAS II)
Gc/Grw	Word Reading	Reading Comprehension test (NFER)	Pictorial Language Comprehension Test (PLCT) Spelling Dictation Task (SDT)		
Gf (Fluid ability)	General ability (NFER) Non-Verbal Test	General ability (NFER) Non-verbal Test	Matrices (BAS)	Higher Conceptual Reasoning (Bike Drawing)	
Gsm (Working Memory)			Recall of Digits (BAS)	Working Memory (Counting Span Task) (TEACh) – Sky task and Dividing Attention: Dual Task	Spatial working memory (CANTAB)
Gq (Quantitative Knowledge)	Arithmetic Test (NFER)	Mathematics Test	Friendly Maths Test (ERT)		
Gv (Visual Processing)		Copying Designs Test (CDT)			
Gt (Decision Speed)				Inhibition (Stop Signal Task)	Cambridge Gambling Task (CANTAB)
Gs (processing speed)				(TEACh) – Attentional control: Opposite Worlds	

Of the three remaining cohorts (NSHD, NCDS and BCS70), there were comparable measures of Gc (verbal reasoning) and Gf (non-verbal reasoning), as well as similar constructs of Gq (mathematics), and other tests measuring crystallised ability (reading, comprehension and vocabulary). In the NCDS, there was only one measure of Gc: reading comprehension. While in the NSHD, there were two further tests measuring Gc: NFER Word Reading and Vocabulary. Compared to the word reading test, the vocabulary test was more akin to the verbal comprehension measure available in the NCDS and was more normally distributed. In the BCS70, there were a number of additional tests measuring Gc at age 10. However, in the first instance as a measure of language comprehension, the Pictorial Language Comprehension Test (PLCT) was chosen as the most comparable alternative (although both the BAS word similarities and Edinburgh Reading Test (ERT) were reasonable comparisons) to the Gc measures in the earlier cohorts.

A breakdown of the variables used in our final psychometric analyses are presented in Table 4. To compare the items included in the latent ability construct, we tested for MI using both the raw and transformed data. The metric for all the measures were transposed to the same scale (0-50).

Table 4. Variables used in cross-cohort psychometric analyses at age 10/11

CHC	Measure	Cohort	Variable	Harmonisation	N	Mean (SD)	Range
Gc [1]	Verbal Ability (NFER)	NSHD	V1157	Metric transformed to (0-50)	4,032	21.61 (7.50)	0-40
	Verbal Ability (NFER)	NCDS	n914	Metric transformed to (0-50)	14,131	22.06 (9.36)	0-40
	(Word) Similarities	BCS70	i3575-i3616 (example group correct (sum))	Metric transformed to (0-50)	11,482	12.06 (2.61)	0-21
Gf	Non-verbal Ability (NFER)	NSHD	NV1157	Metric transformed to (0-50)	4,032	23.40 (9.17)	0-40
	Non-verbal Ability (NFER)	NCDS	n917	Metric transformed to (0-50)	14,131	20.88 (7.61)	0-40
	Matrices (BAS)	BCS70	I3617-i3644 (sum)	Metric transformed to (0-50)	11,494	15.35 (5.40)	0-28
Gq	Arithmetic Test (NFER)	NSHD	A1157	-	4,025	26.39 (11.74)	0-50
	Mathematics Test	NCDS	n926	Metric transformed to (0-50)	14,126	16.63 (10.35)	0-40
	Friendly Maths Test	BCS70	BD3MATHS	Metric transformed to (0-50)	11,633	43.95 (12.32)	0-72
Gc[2]	Vocabulary	NSHD	VOC1157	-	4,027	29.99 (7.45)	0-50
	Reading Comprehension (NFER)	NCDS	n923	Metric transformed to (0-50)	14,130	15.98 (6.29)	0-35
	Pictorial Language Comprehension Test (PLCT)	BCS70	I8-i62, i66-i110 (sum)	Metric transformed to (0-50)	12,790	61.10 (10.69)	0-100

Gc[1] and Gc[2] are both measures of crystallised ability, albeit Gc[1] are measures of verbal reasoning, whereas Gc[2] are measures more akin to verbal knowledge.

In the three cohorts, four variables loaded onto a general ability factor: i) Gc^1, ii) Gf, iii) Gq and iv) Gc^2. The results from the multiple group CFA using the raw and the transformed data are presented in Table 5.

Table 5. Results from multiple group CFA at age 10/11 for 3 cohorts (NSHD, NCDS and BCS70).

	Model	RMSEA	CFI	TLI	ΔRMSEA	ΔCFI	ΔTLI
Configural	Raw	0.114	0.991	0.959			
	Transformed	0.114	0.991	0.959			
Metric	Raw	0.301	0.841	0.714	0.187	-0.15	-0.245
	Raw*	0.361	0.727	0.590	0.247	-0.264	-0.369
	Transformed	0.159	0.955	0.920	0.045	-0.036	-0.039
Scalar	Transformed	0.269	0.797	0.771	0.155	-0.194	-0.188

* residual variance of verbal reasoning fixed at to the appropriate observed variance for each cohort.

Although the configural model fit the data fairly well, when using the raw data the metric model resulted in a worsening of model fit well outside the suggested conventions. In addition, the BAS Similarities measure was identified as having a negative residual variance; this can be for several reasons including model misspecification. However, sometimes this is owing to the relation between the other items in the model. As there were other cognitive tests measuring Gc in the BCS70 at this age we did test for measurement invariance, replacing PLCT consecutively with two other Gc tests: the ERT and Word Definitions (BAS). This did not improve the model fit, nor did it result in model convergence. We also reran the model, fixing the residual variances for the Gc^1 (verbal reasoning) items only to the appropriate observed variance for each cohort. This resulted in a worsening model fit; we conclude that the latent general ability factor when using the raw cognitive scores cannot be compared across the three cohorts.

When testing for MI with the transformed cognitive measures, the metric model converged, possibly a result of reducing the variance of the indicators (see http://www.statmodel.com/discussion/messages/12/17.html?1438886834). Although the model fit improved and the change in metric model fit compared to the configural model was superior compared to the raw data, the change was still outside the 'acceptable' cut-offs. As there were two identical measures (a rare occurrence across the cohorts) of verbal and non-verbal reasoning in both the NSHD and NCDS, we tested for MI between just these two cohorts. As outlined previously the four variables were loaded onto a general ability factor: i) Gc, ii) Gf, and iii) Gq and iv) Gc (Figure 6). The results from the multiple group CFA using the raw and the transformed data are presented in Table 6.

Table 6. Results from multiple group CFA at age 10/11 for 2 cohorts (NSHD and NCDS).

	Model	RMSEA	CFI	TLI	ΔRMSEA	ΔCFI	ΔTLI
Configural	Raw	0.106	0.993	0.972			
	Transformed	0.106	0.993	0.972			
Metric	Raw	0.077	0.994	0.986	-0.029	0.001	0.014
	Transformed	0.097	0.990	0.977	-0.009	-0.003	0.005
Scalar	Raw	0.344	0.806	0.710	0.238	-0.187	-0.262
	Transformed	0.198	0.936	0.904	0.092	-0.057	-0.068
Partial Scalar*	Raw	0.071	0.994	0.988	-0.035	0.001	0.016
	Transformed	0.089	0.990	0.980	-0.017	-0.003	0.008

* Intercepts of Gq (maths) and Gc^2 (reading) tests freely estimated across cohorts.

The configural and metric models fit the data well, which indicates that the variances and covariances of the latent general ability factor can be compared across NSHD and NCDS. However, the scalar model resulted in a substantial worsening of model fit. An inspection of the modification indices indicated this poor fit was due to the equality constraints

placed on the intercepts of the Gq (mathematics) and the Gc^2 (vocabulary and reading comprehension) measures. Freeing these parameters resulted in a large improvement in fit (ΔRMSEA, ΔCFI and ΔTLI now within acceptable range); therefore partial measurement invariance was supported. We also tested the multiple group CFA using the transformed data as shown in Table 6. This did not improve the model fit; the results using the raw and transformed data were very similar. As such, it is possible to compare the latent means of the general ability factor across the two cohorts; however the unequal intercepts of the mathematics and crystallised ability tests (Gc^2) will introduce an element of bias into the results when comparing the NSHD with the NCDS. Despite this bias, the overall model fit and equality of the other parameter estimates suggests that comparisons of the means and variances of the latent cognitive ability across the two cohorts are reasonable according to current guidelines (Little, 2013).

The unstandardised parameter estimates of the partial scalar model (including fixed and freely estimated parameters) are presented in Figure 6. Differences in latent means were examined across the groups by fixing the mean and variance of the latent factor to 0 and 1 respectively in the NSHD (the reference group) and freely estimating these parameters in the NCDS. On average there were significantly higher scores on the latent cognitive ability in the NSHD (0.51), when compared to the NCDS.

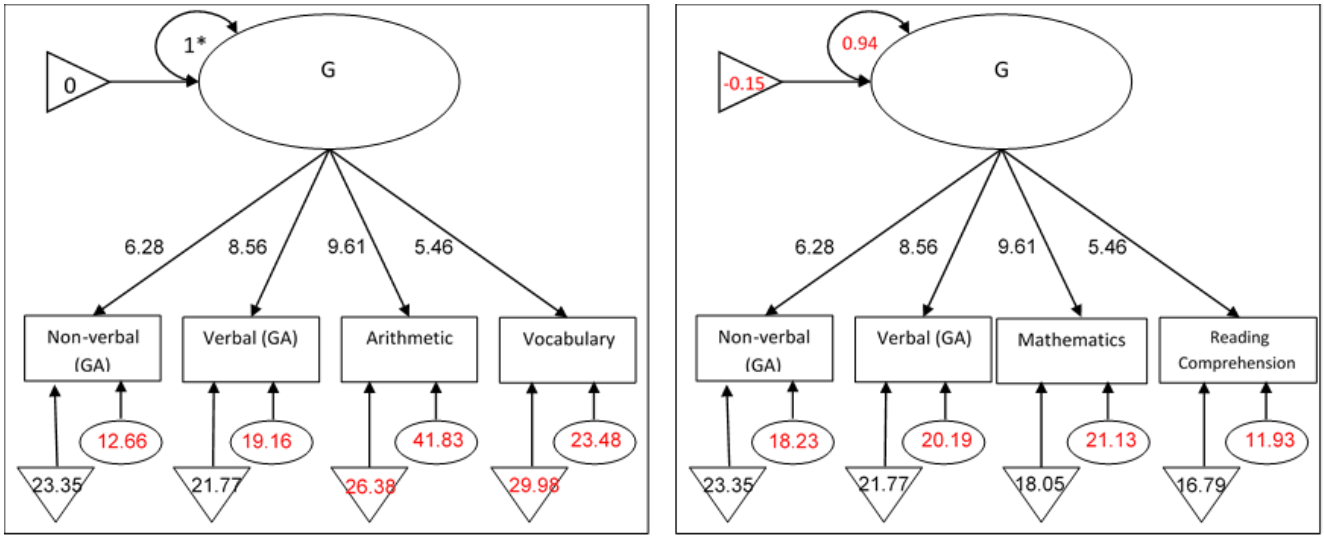

Figure 6. Unstandardised parameter estimates of partial invariance model. Parameters in black were fixed/held equal across groups. Parameters in red were freely estimated across the two cohorts.

4.2 Testing for measurement invariance in midlife (age 46-53 years)

Several comparable tests were administered across the NSHD, NCDS and BCS70 when participants were aged between 46 and 53 years old (Table 7). The common tests were verbal fluency/executive function (animal naming), immediate and delayed verbal memory (word list recall tests) and visual processing speed (letter search task).

Table 7. Comparable constructs in mid-life (age 46-53)

	NSHD (Age 53)	NCDS (Age 50)	BCS70 (Age 46)
General Ability	National Adult Reading Test (NART)		
Verbal Memory	Immediate and Delayed Verbal Learning/ Word List Recall Test	Immediate and Delayed Verbal Learning/ Word List Recall Test	Immediate and Delayed Verbal Learning/ Word List Recall Test
Verbal fluency /executive function	Verbal Fluency (animal naming)	Verbal Fluency (animal naming)	Verbal Fluency (animal naming)
Visual Processing speed	Timed Letter Search/Letter Cancellation Test	Timed Letter Search/Letter Cancellation Test	Timed Letter Search/Letter Cancellation Test

The animal naming and letter cancellation tests were administered in a similar manner in all three cohorts. For the animal naming test, respondents were given one minute to name as many animals as they could think of. For the letter cancellation tests, participants were presented with blocks of letters, and were asked to read through the blocks from left to right, crossing out the 'Ws' and 'Ps' as they read, as quickly and accurately as possible. Search speed was calculated by summing the total number of words scanned, including both target and non-target words[4].

The immediate and delayed memory trials differed slightly across the cohorts. In the NSHD, participants were shown a list of 15 words at a rate of one word every two seconds. They were then asked to write down as many words as they could recall. This trial was done a total of three times, and a total score was calculated as the sum of the words correctly recalled over the three trials. In both the NCDS and BCS70, participants were played an audio recording of 10 words (one word every 2 seconds) and were then given two minutes to orally recount as many as they could recall. Only one trial was

[4] Alternative forms of scoring have been used in the literature e.g. (Davis et al., 2016; Silverwood et al., 2014)

administered. To make these variables more comparable, we used the first trial only from the NSHD (wlt199),and collapsed scores on this variable by recoding scores of greater than 10 to exactly 10. This placed the variable on a similar 0-10 metric that was comparable with the variables in the NCDS and BCS70. We did this for both the immediate and delayed conditions.

A breakdown of the precise variables used in our psychometric analyses are presented in Table 8.

Table 8. Variables used in cross-cohort psychometric analyses in midlife (age 46-53)

Measure	Cohort	Variable	Harmonisation	N	Mean (SD)	Range
Word List Recall Test	NSHD	wlt199 (immediate memory – 1st trial)	Scores > 10 recoded to a value of 10	2,909	5.80 (2.01)	0-10
		wlt499 (delayed memory – 1st trial)	Scores > 10 recoded to a value of 10	2,292	7.99 (2.05)	0-10
	NCDS	N8CFLISN (immediate)	-	9,648	6.54 (1.48)	0-10
		N8CFLISD (delayed)	-	9,591	5.41 (1.84)	0-10
	BCS70	B10CFLISN (immediate)	-	8,501	6.61 (1.44)	0-10
		B10CFLISD (delayed)	-	8,494	5.47 (1.81)	0-10
Animal naming	NSHD	anin	-	2,949	23.56 (6.91)	1-62
	NCDS	N8CFANI	-	9,648	22.28 (6.30)	0-65
	BCS70	B10CFANI	-	8,498	23.63 (6.19)	1-70
Letter cancellation	NSHD	CANSP99 (Search speed)	-	2,932	281.07 (76.08)	64-591
	NCDS	N8CFRC (Search speed)	-	9,442	334.10 (88.83)	84-780
	BCS70	B10CFRC (Search speed)	-	8,242	346.45 (84.77)	28-780

In each of the cohorts, four variables loaded onto a general ability factor: i) animal naming, ii) letter search speed, iii) immediate recall, iv) delayed recall (Figure 7). The results from the multiple group CFA are presented in Table 9.

Table 9. Results from multiple group CFA across mid-life.

Model	RMSEA	CFI	TLI	ΔRMSEA	ΔCFI	ΔTLI
Configural	0.077	0.980	0.939			
Metric	0.054	0.980	0.970	-0.023	0.000	-0.031
Scalar	0.209	0.550	0.550	0.132	0.430	0.389
Partial Scalar*	0.081	0.948	0.933	0.004	0.032	0.006

* Intercepts of letter cancellation and delayed recall tests freely estimated across cohorts

The configural and metric models fit the data well, which indicates that the variances and covariances of the latent general ability factor can be compared across all cohorts. However, the scalar model resulted in a considerable worsening of model fit. An inspection of the modification indices suggested that this poor fit was due to the equality constraints placed on the intercepts of the delayed memory and letter cancellation tests in the NSHD cohort. Freeing these parameters resulted in a large improvement in fit (ΔRMSEA and ΔTLI now within acceptable range); therefore partial measurement invariance was supported. As such, it is possible to compare the latent means of the general ability factor across all three cohorts. However, the unequal intercepts of the letter cancellation and delayed memory tests will introduce an element of bias into the results when comparing the NSHD with the NCDS/BCS70. This may be due to the impact of methodological factors (e.g. slight differences in tests/administration procedures). Despite this bias, the overall model fit and equality of the other parameter estimates suggests that comparisons of the means and variances of the latent cognitive ability across all three cohorts are justifiable according to current guidelines (Little, 2013). Full scalar invariance between the NCDS and BCS70 was supported, which was unsurprising since the same tests were administered in a consistent format across these two cohorts.

As such, comparisons of the latent variable can validly be made across these two populations.

The unstandardised parameter estimates of the partial scalar model (including fixed and freely estimated parameters) are presented in Figure 7. Differences in latent means were examined across the groups by fixing the mean and variance of the latent factor to 0 and 1 respectively in the first group (NSHD) and freely estimating these parameters in the second and third groups (NCDS and BCS70). Thus, compared with the reference group (NSHD), latent means were higher in the NCDS (0.41) and BCS70 (0.47). The statistical significance of these differences was determined by calculating *Z*-scores for each mean difference by dividing the estimate by its standard error, with resultant values of ± 1.96 reflecting a statistically significant difference at $p<0.05$. The standardised *Z*-scores for the NCDS and BCS70 were 14.37 and 16.50 respectively, indicating significantly higher scores on the latent cognitive ability variables in the later cohorts.

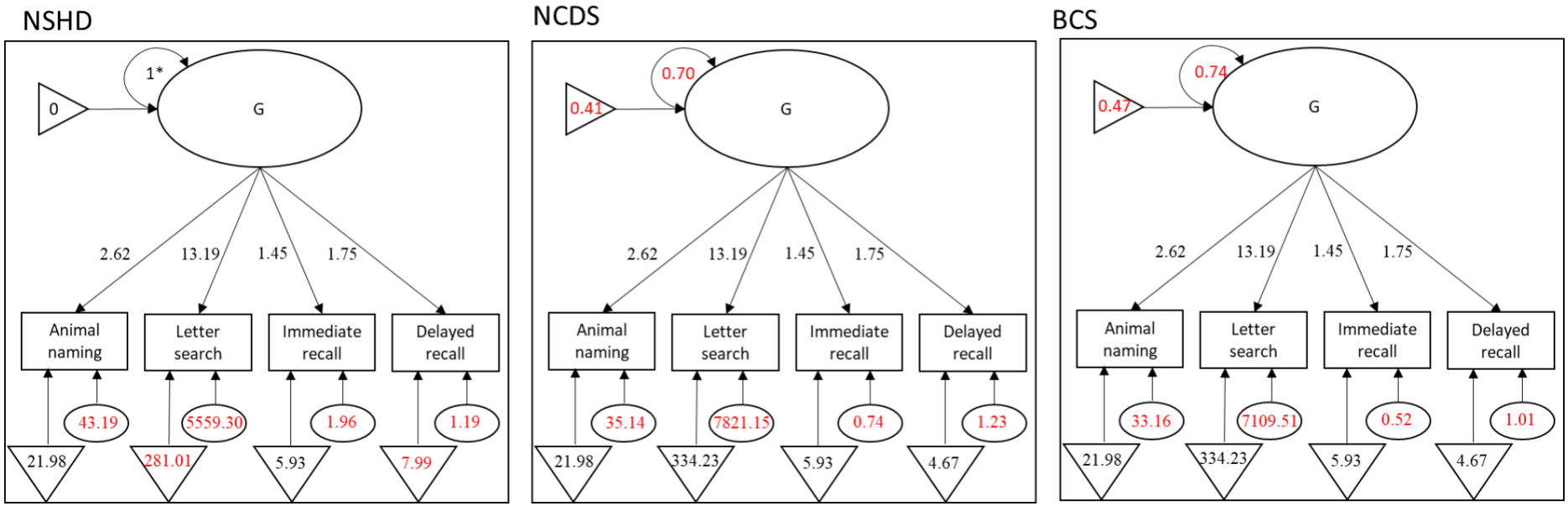

Figure 7. Unstandardised parameter estimates of partial invariance model. Parameters in black were fixed/held equal across groups. Parameters in red were freely estimated across groups.

4.3 Testing for measurement invariance in adulthood in the NSHD

To test for measurement invariance within a particular cohort over time, at least three common measured indicators (i.e. cognitive tests) are required at each assessment wave to ensure that models converge and are identified (Little, 2013). As can be seen in Appendix I, there were no instances where three or more identical tests were administered at multiple assessments within any of the cohorts. Even at the broader conceptual level, it was difficult to identify the requisite number of conceptually similar cognitive tests to test the equivalence of these measures. As such, in the present section we outline an example of a strategy to harmonise measures of cognitive ability within a particular cohort (in this case the NSHD) when both common and unique cognitive tests have been administered across time.

In recent years, attempts have been made to use latent variable models to construct harmonised scores based on both common and unique indicators (i.e. measures or tests). Examples can be found in both the mental health (Tyrell, Yates, Widaman, Reynolds, & Fabricius, 2019) and cognitive ability literature (Gross et al., 2015). This approach is broadly in line with traditional measurement invariance approaches. However, it incorporates both common and unique indicators in the measurement model, placing equality constraints on the common indicators only (Tyrell et al., 2019). The rationale is that placing equality constraints on the common measures serves to anchor the tests to a common metric, with unique indicators providing additional information relating to the underlying trait being measured (Bauer & Hussong, 2009; Tyrell et al., 2019).

As can be seen in Table 10, two tests were administered consistently throughout adulthood in the NSHD: a test of verbal memory (Word List Learning) and a test of visual processing speed (Timed Letter Search). No other tests were administered consistently, rather a variety of additional cognitive skills were assessed intermittently as participants aged (e.g. motor speed, executive function, reaction time).

Table 10. Measures administered across adulthood in NSHD

	Age 43	Age 53	Age 60-64	Age 68-70
Visual Processing Speed	Timed Letter Search/Letter Cancellation Test	Timed Letter Search/Letter Cancellation Test	Timed Letter Search/Letter Cancellation Test	Timed Letter Search/Letter Cancellation Test
Verbal Memory	Verbal Learning/ Word List Recall Test	Verbal Learning/ Word List Recall Test	Verbal Learning/ Word List Recall Test	Verbal Learning/ Word List Recall Test
Visual Memory	Visual memory test			
Motor Speed and Praxis	Peg test			Finger Tapping Test
Delayed Verbal Memory		Verbal Learning/ Word List Recall Test (delayed condition)		
General Ability		National Adult Reading Test (NART)		ACE-III
Verbal Fluency/Executive Function		Animal Naming Test		
Reaction Time			Simple and Choice Reaction Time Test	
Prospective Memory		Prospective Memory Test		

We tested a longitudinal structural equation model (SEM) incorporating both the common and unique tests as measured indicators. We placed equality constraints on the common indicators in line with traditional measurement invariance testing. In other words, we held factor loadings equal (i.e. metric invariance), followed by intercepts (i.e. scalar invariance) and examined whether this resulted in a worsening of overall model fit. A graphical illustration of the tested SEM model is presented in Figure 8. In line with standard longitudinal CFA, we allowed for the residuals amongst the common indicators to correlate over time, and we included correlations amongst the latent general ability factors. For identification and to enable the comparison of latent means, the mean and variance of the latent cognitive ability at T1 were fixed to 0 and 1 respectively.

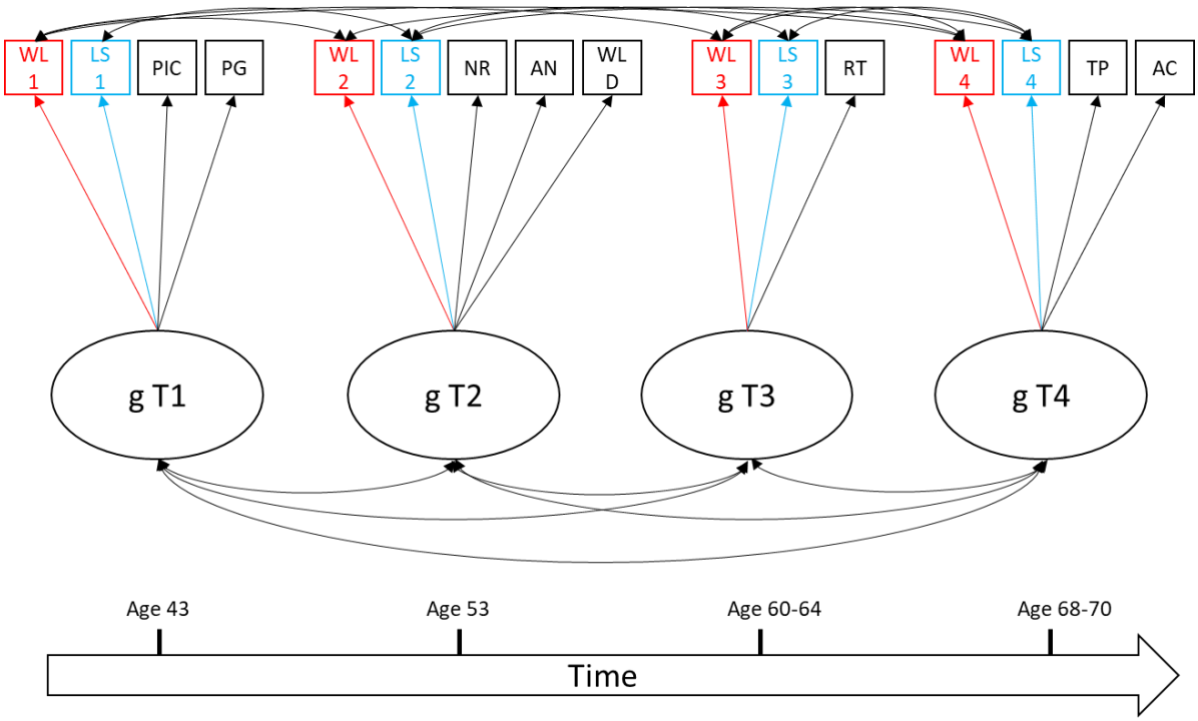

Figure 8. Graphical illustration of SEM model. Equality constraints placed loadings and intercepts of word list recall test (WL; red) and timed letter search (LS; blue). PIC = Visual memory test; PG = Peg test; NR = National Adult Reading Test (NART); AN = Animal naming test; WLD = Word List Recall Test (delayed condition); RT = Simple and choice reaction time test; TP = Finger tapping test; AC = ACE-III.

Fit statistics for this model are presented in Table 11. The configural model (i.e. no equality constraints) fit the data well. Holding the factor loadings of the common indicators equal over time (i.e. metric invariance) did not result in a significant worsening of overall fit. Fit statistics fell below acceptable levels when equality constraints were placed on the intercepts of the two common tests, and an inspection of the modification indices found this worsening in fit was largely attributed to the constraint on the timed letter search. Freeing this parameter across time resulted in an acceptable model.

Table 11. Results from multiple group CFA across mid-life.

Model	RMSEA	CFI	TLI	ΔRMSEA	ΔCFI	ΔTLI
Configural	0.042	0.941	0.918	-	-	-
Metric	0.041	0.941	0.921	0.001	<0.000	0.003
Scalar	0.076	0.792	0.729	0.034	0.149	0.189
Partial Scalar*	0.041	0.941	0.921	0.001	<0.000	0.003

* Intercept of timed letter search freely estimated across cohorts

Latent mean scores (T1 as reference group) and standardised correlations among the latent general cognitive ability factor (best fitting model) are presented in Table 12. High correlations indicate a high level of stability over time in terms of general cognitive ability.

Table 12. Correlations and latent means of general cognitive ability factor over time

	T1	T2	T3	T4
T2	0.92			
T3	0.91	0.95		
T4	0.87	0.91	0.94	
Means (SE)	0	-0.25 (0.03)	-0.3 (0.03)	-0.86 (0.05)

Using T1 (43 years) as a reference point, latent means were lower at later assessment waves. The statistical significance of these differences was determined by calculating Z-scores for each mean difference by dividing the estimate by its standard error, with resultant values of ± 1.96 reflecting a statistically significant difference at $p<0.05$. The standardised Z-scores for the later assessment waves were -8.51, -9.30, and -19.26 respectively, indicating significantly lower scores on the latent cognitive ability factor in the later assessment waves.

5 Conclusions and recommendations

This report explored the feasibility of retrospectively harmonising the measures of cognition that are available in five British birth cohorts, to maximise the comparability of these measures across the studies and over time.

In this final section we provide a summary and some general guidance on how to identify and harmonise variables based on the data available in the British birth cohorts. Again, we recommend researchers follow the broad guidelines outlined by Fortier et al. (2017).

1. *Establishing your research question*: The first key step is to firmly establish your research question, as this will impact all subsequent steps, particularly the level of harmonisation required.

2. *Assembling pre-existing knowledge*: Next, it is important to familiarise yourself with the cognitive measures that are available in the cohorts. For this we refer researchers to our companion report (Moulton et al., 2020).

3. *Identifying a harmonisable pool of cognitive tests*: When attempting to retrospectively harmonise different measures, the exact number of harmonisable tests will vary depending on the number of cohorts and/or assessment waves that are relevant to your research question. As discussed in Section 3, there are few occasions in which the same cognitive tests are available within or across the cohorts. Therefore, it may be that researchers attempt to harmonise tests that are conceptually similar (i.e. represent the same underlying cognitive ability/skill). The tables available in Appendix 1 of this report can help researchers identify conceptually similar tests. It is also important to be aware of the various features of the cognitive tests administered in the cohorts (e.g. design, method of delivery) as these can impact the suitability of a test for harmonisation. As demonstrated in Section 4, the more consistent the measures, the easier it is to retrospectively harmonise tests across cohorts.

4. *Processing the data*: If the aim is to compare mean scores of cognitive function over time or across cohorts, the data will need to be converted to a common scoring metric. We provide examples of this in Table 4 and Table 8.

5. *Estimating the quality of harmonised variables using a latent variable modelling approach*: After completing the previous steps, it is important to establish the measurement equivalence of your tests (i.e. confirm they are assessing the same underlying construct and to the same degree). As discussed in Section 2.4, metric invariance (i.e. equality constraints placed on loadings) establishes whether the same underlying construct (g) is being assessed by your set of tests, and is important to establish if you wish to compare regression coefficients within or across cohorts. Scalar invariance (equality constraints placed on intercept parameters) tests whether the underlying level of the test can be considered equivalent across groups. In other words, individuals from two different cohorts who have the same level of '*g*' will demonstrate the same score on a scalar invariant cognitive test. This is required in order to make valid comparisons of mean-levels of cognition at different time points or across cohorts. It is important to establish measurement equivalence even when identical measures are administered within/across cohorts, to ensure there are no systematic differences in measurement error due to age/cohort.

   We found that it was possible to establish metric invariance even when different tests were administered across cohorts (Section 4.1). However, scalar invariance was difficult to establish unless highly consistent tests were used across studies (Section 4.2).

6. *Disseminating and preserving final harmonisation products*: 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 two methods of doing this: i) simultaneous estimation (i.e. include latent variables in your model using SEM), and ii) produce and analyse factor scores.

Dissemination is also important, for transparency and to allow other research to replicate and/or adapt harmonisation approaches. We therefore encourage researchers to provide detailed descriptions of their harmonisation strategies, share their code, and where possible make their harmonised variables available to others (see Appendix II for syntax used in this report).

6 Appendix I. Tables of overlapping measures and cognitive constructs in the British birth cohorts

Notes

1. See Table 1 for exact age ranges.

2. The below tables detail measures that were administered to entire cohorts only (i.e. cognitive tests administered solely to targeted sub-samples were not considered for harmonisation due to their smaller sample sizes and lack of generalisability).

3. We focus only on measures that were administered to the cohort members; any measures administered to the cohorts' parents, children of the cohort members or other parties were not included.

4. We focus only on measures designed specifically to assess theoretically defined cognitive abilities (e.g. fluid reasoning, working memory, lexical knowledge, verbal comprehension); tests used to assess basic levels of skills (e.g. basic adult literacy) were not included.

5. Tests are categorised according to the main cognitive domain they are reported to assess. In practice, most cognitive tests require a range of cognitive abilities to complete.

Table 13. Cognitive abilities assessed in NSHD

Age in years: Test:	8	11	15	26	43	53	60-64	68-70
Reading Comprehension	Gc/Grw							
Word Reading	Grw	Grw						
Vocabulary	Gc	Gc						
Picture Intelligence	Gf							
General Ability Test (verbal and non-verbal)		G/Gc/Gf						
Arithmetic Test		Gq						
Alice Heim Group Ability Test			G/Gc/Gf					
Watts-Vernon Reading Test			Gc/Grw	Gc/Grw				
Mathematics Test			Gq					
Verbal Learning/Word List Recall					Glr	Glr	Glr	Glr
Long Term Recall					Glr			
Visual Memory					Glr/Gv			
Timed Letter Search/Letter Cancellation					Gv/Gs	Gv/Gs	Gv/Gs	Gv/Gs
Motor Speed and Praxis					Gp			
National Adult Reading Test (NART)						Gc/Grw		
Verbal Fluency (animal naming)						Glr		
Prospective Memory						NA		
Delayed Verbal Memory						Glr		
Reaction Time Test							Gt	
Finger Tapping Test								Gp
Addenbrooke's Cognitive Examination (5 domains)								G

Table 14. Cognitive abilities assessed in NCDS

Age in years: **Test:**	**7**	**11**	**16**	**50**
Southgate Group Reading Test	Gc/Grw			
Problem Arithmetic Test (NFER Devised)	Gq			
Copying Designs Test (CDT)	Gc	Gc		
Human Figure Drawing (HFD)	Gv			
General Ability Test (Verbal and Non-Verbal)		G/Gc/Gf		
Reading Comprehension Test (NFER)		Gc	Gc	
Arithmetic- Mathematics Test (NFER)		Gq		
Mathematics Test (NFER)			Gq	
Verbal Fluency (Animal Naming) Test				Glr
Verbal Learning/Word List Recall				Glr
Timed Letter Search/Letter Cancellation				Gv/Gs

Table 15. Cognitive abilities assessed in BCS70

Age in years: Test:	5	10	16	42	46-7
(Schonell) Reading test	Gc/Grw				
English Picture Vocabulary Test (EPVT)/ Pictorial Language Comprehension Test (PLCT)	Gc	Gc			
Copying Designs Test (CDT)	Gv				
Human Figure Drawing (HFD)	Gv				
Complete a Profile test (CPT)	Gv				
Edinburgh Reading Test (SV-ERT)		Gc/Grw	Gc/Grw		
Friendly Maths Test		Gq			
Spelling Dictation task (SDT)		Grw			
(Word) Similarities (BAS)		Gc			
Word Definitions (BAS)		Gc			
Recall of Digits (BAS)		Gsm			
Matrices (BAS)		Gf	Gf		
Vocabulary test (APU)			Gc	Gc	
Arithmetic test (APU)			Gq		
Spelling test			Grw		
Verbal Fluency (Animal Naming)					Glr
Verbal Learning / Word List Recall					Glr
Timed Letter Search / Letter Cancellation					Gv/Gs

Table 16. Cognitive abilities assessed in ALSPAC

Age in years: Test:	7.5	8.5	9	10	11.5	12.5	13	15.5	17.5
Basic Reading	Gc/ Grw								
Phoneme Deletion Task	Gc/ Grw								
Spelling Task	Grw		Grw						
Letter Decision Task	Gs/ Gv								
Motor Ability Task	Gp								
Wechsler Intelligence Scale for Children (WISC-III)		G							
Object Assembly (WISC-III)		Gf/ Gs							
Coding (WISC-III)		Gv/ Gs							
Block Design (WISC-III)		Gs/ Gv							
Picture Arrangement (WISC-III)		Gc/ Gf/ Gv							
Picture Completion (WISC-III)		Gv/ Gc							
Information (WISC-III)		Gc							
Comprehension (WISC-III)		Gc							
Arithmetic (WISC-III)		Gq							
Vocabulary (WISC-III)		Gc							
Similarities (WISC-III)		Gc							
DANVA: Faces Subtest		Gkn							
TEACh: Selective Attention and Motor Control: Sky Search		Gs/ Gps/ Gv/ Gsm			Gs/Gps/Gv/Gsm				
TEACh: Dividing Attention (Dual Task)		Gsm/ Gs/ Gps/ Gv/ Ga			Gs/Gps/Gv/Gsm/Ga				

Age in years: Test:	7.5	8.5	9	10	11.5	12.5	13	15.5	17.5
TEACh: Attentional control (Opposite Worlds)		Gs/ Gsm			Gs/Gsm				
Listening Comprehension		Gc/ Glr							
Oral Expression		Gc							
Short-term Memory (Nonword Repetition)		Gsm							
Articulatory Skills		Ga							
Word and Nonword Reading Test			Gc/ Grw						
Neale Analysis of Reading Ability (NARA II)			Gc/ Grw						
Sentence Decision Task			Gc/ Grw						
Working Memory (Counting Span Task)				Gsm/Gv					
Inhibition (Stop-Signal) Task				Gt				Gt	
Higher Conceptual Reasoning (Bike Drawing)					G				
Phonological Awareness (Spoonerisms)						NA			
Tests of Word Reading Fluency (TOWRE)						Gc/Grw	Gc/Grw		
Motor Skill and Movement Test						Gps			
Reaction Time							Gt		
Wechsler Abbreviated Scale of Intelligence (WASI)								G	
Vocabulary (WASI)								Gc	

Age in years: **Test:**	**7.5**	**8.5**	**9**	**10**	**11.5**	**12.5**	**13**	**15.5**	**17.5**
Matrix Reasoning (WASI)								Gf/Gv	
N-Back Task (Working Memory)									Gsm
Go No Go (Affective Go/No-Go Task)									NA
Probability Learning and Reversal Task									Gsm

Table 17. Cognitive abilities assessed in MCS

Age in years: **Test:**	**3**	**5**	**7**	**11**	**14**	**17**
Bracken School Readiness	Gc (Gq/Gv)					
Naming Vocabulary (BAS II)	Gc	Gc				
Pattern Construction (BAS II)		Gv	Gv			
Picture Similarities (BAS II)		Gf				
Word Reading (BAS II)			Gc/Grw			
Progress in Maths (NFER, adapted)			Gq			
Verbal Similarities (BAS II)				Gc		
Cambridge Gambling Task (CGT; CANTAB)				Gt/Gs	Gt/Gs	
Spatial Working Memory Task (SWM; CANTAB)				Gsm		
Vocabulary test (Applied Psychological Unit (APU))					Gc	
Number Analogies test (CAT3)						Gq

Table 18. Comparable constructs at age 5

	BCS70 (Age 5)	**MCS (Age 5)**
Gc (Crystallised ability)	English Picture Vocabulary Test (EPVT)	Naming Vocabulary (BAS II)
Gc/Grw (Crystallised ability/Reading & writing)	(Schonell) Reading test	
Gf (Fluid ability)		Picture Similarities (BAS II)
Gv (Visual processing)	Copying Designs Test (CDT) Human Figure Drawing (HFD) Complete a Profile Test (CPT)	Pattern Construction (BAS II)

Table 19. Comparable constructs at age 7/8

	NSHD (Age 8)	NCDS (Age 7)	ALSPAC (Age 7.5)	ALSPAC (Age 8.5)	MCS (Age 7)
Gf (Fluid ability)	Picture Intelligence			Object Assembly (WISC-III)	
Gc (Crystallised ability)	Vocabulary			Picture Completion (WISC-III) Information (WISC-III) Comprehension (WISC-III) Vocabulary (WISC-III) Similarities (WISC-III) Listening Comprehension Oral Expression	
Gc/Gf (Crystallised ability/Fluid ability)				Picture Arrangement (WISC-III)	
Gc/Grw (Crystallised ability/Reading & writing)	Reading Comprehension	Southgate Group Reading Test	Basic Reading Phoneme Deletion Task		BAS II Word Reading
Grw (Reading & writing)	Word Reading		Spelling Task		
Gq (Quantitative knowledge)		Problem Arithmetic Test		Arithmetic (WISC-III)	NFER Progress in Maths (adapted)
Gv (Visual processing)		Copying Designs Test Human Figure Drawing			BAS II Pattern Construction

Gs/Gv (Processing speed/Visual processing)			Letter Decision Task	Coding (WISC-III) Block Design (WISC-III) TEACh (the Tests of Everyday Attention for Children): Selective Attention and Motor Control: Sky Search TEACh: Dividing Attention (Dual Task)	
Gp (Psychomotor ability)			Motor Ability Task		
Gkn (Domain-specific knowledge)				DANVA: Faces subtest	
Gsm (Short-term memory)				TEACh (the Tests of Everyday Attention for Children): Selective Attention and Motor Control: Sky Search TEACh: Dividing Attention (Dual Task) TEACh: Attentional Control (Opposite Worlds) Nonword Repetition	
Glr (Long-term storage & retrieval)				Listening Comprehension	
Ga (Auditory processing)				Articulatory Skills	

Table 20. Comparable constructs assessed at age 10-11

	NSHD (Age 11)	NCDS (Age 11)	BCS70 (Age 10)	ALSPAC (age 10*/11)	MCS (Age 11)
Gc (Crystallised ability)	General ability (NFER) Verbal Test Vocabulary	General ability (NFER) Verbal Test	Pictorial Language Comprehension Test (PLCT) (Word) Similarities (BAS) Word Definitions (BAS)		Verbal similarities (BAS II)
Gc/Grw (Crystallised ability/ Reading & writing)	Word Reading	Reading Comprehension test (NFER)	Edinburgh Reading Test (ERT) Spelling Dictation Task (SDT)		
Gf (Fluid ability)	General ability (NFER) Non-Verbal Test	General ability (NFER) Non-verbal Test	Matrices (BAS)	Higher Conceptual Reasoning (Bike Drawing)	
Gsm (Working memory)			Recall of Digits (BAS)	Working Memory (Counting Span Task)*(TEACh) – Sky task and Dividing Attention: Dual Task	Spatial working memory (CANTAB)
Gq (Quantitative knowledge)	Arithmetic Test (NFER)	Mathematics Test	Friendly Maths Test (ERT)		
Gv (Visual processing)		Copying Designs Test (CDT)			
Gt (Decision speed)				Inhibition (Stop Signal Task)*	Cambridge Gambling Task (CANTAB)
Gs (Processing speed)				(TEACh) – Attentional control: Opposite Worlds	

Table 21. Comparable constructs at age 14-16

	NSHD (Age 15)	NCDS (Age 16)	BCS70 (Age 16)	ALSPAC (Age 15)	MCS (Age 14)
Gc (Crystallised ability)	AH4 verbal ability		Vocabulary (APU)	Vocabulary (WASI)	Vocabulary (APU)
Gc/Grw (Crystallised ability/ Reading & writing)	Watts-Vernon Reading Test	Reading Comprehension (NFER)	Edinburgh Reading Test (ERT) Spelling test		
Gf (Fluid ability)	AH4 non-verbal ability		Matrices (BAS)	Matrix Reasoning (WASI)	
Gq (Quantitative knowledge)	Mathematics Test	Mathematics (NFER)	Arithmetic (APU)		
Gt (Decision speed)				Inhibition (Stop-signal) Task	Cambridge Gambling Task (CANTAB)

Table 22. Comparable constructs in mid-life (age 46-53)

	NSHD (Age 53)	**NCDS (Age 50)**	**BCS70 (Age 46)**
Gc/Grw (Crystallised ability/ Reading & writing)	National Adult Reading Test (NART)		
Glr (Long-term storage & retrieval) Verbal Memory	Immediate and Delayed Verbal Learning/ Word List Recall Test	Immediate and Delayed Verbal Learning/ Word List Recall Test	Immediate and Delayed Verbal Learning/ Word List Recall Test
Glr (Long-term storage & retrieval) Verbal fluency/ executive function	Verbal Fluency (Animal Naming)	Verbal Fluency (Animal Naming)	Verbal Fluency (Animal Naming)
Gv/Gs (Visual processing/ Processing speed) Visual Processing speed	Timed Letter Search/Letter Cancellation Test	Timed Letter Search/Letter Cancellation Test	Timed Letter Search/Letter Cancellation Test

7 Appendix II. Syntax

7.1 Stata code for converting tests to common metric/scale

Formula: gen newVar = Min_new_scale + ((Max_new_scale – Min_new_scale)/(Max_old_scale – Min_old_scale))*(oldVar- Min_old_scale)

e.g. modify variable X old variable (range 10-30, old scale) so that it has the same metric as variable Y (range 20-80, new scale)

Gen XY = 20 + ((80-20) / (30-10))*(X-10)

7.2 Mplus code configural invariance (midlife)[5]

```
title: Configural model midlife
   data:
    file is merged_NSHD_NCDS_BCS.csv;
   variable:
        names are
        case_id cohort age anin let_s let_a let_h mem_i mem_d;

       usevar are
        anin let_s mem_i mem_d;

        missing are all (-99);
        grouping = cohort (1 = NSHD 2 = NCDS 3 = BCS);

   analysis:
         estimator = MLR;

model:
    G BY  anin* let_s mem_i mem_d;
    G@1;
    [G@0];


  model NCDS:
  G BY  anin* let_s mem_i mem_d;
 [anin-mem_d];


     model BCS:
  G BY  anin* let_s mem_i mem_d;
 [anin-mem_d];

output:
   sampstat standardized modindices(ALL);
```

[5] Same code used for age 10/11 cross-cohort analysis, but with appropriate variables from those sweeps.

7.3 Mplus code metric invariance (midlife)

```
title: Metric model midlife
  data:
       file is merged_NSHD_NCDS_BCS.csv;
  variable:
       names are
       case_id cohort age anin let_s let_a let_h mem_i mem_d;

      usevar are
       anin let_s mem_i mem_d;

       missing are all (-99);
       ! categorical are all;
       !USEOBSERVATIONS are cohort== 1;
       grouping = cohort (1 = NSHD 2 = NCDS 3 = BCS);

  analysis:
       estimator = MLR;

  model:
   G BY  anin* let_s mem_i mem_d;
   G@1;
   [G@0];


 model NCDS:
! G BY  anin* let_s mem_i mem_d;
[anin-mem_d];
G*;


    model BCS:
 !G BY  anin* let_s mem_i mem_d;
[anin-mem_d];
 G*;

  output:
  sampstat standardized modindices(ALL);
```

7.4 Mplus code scalar invariance (midlife)

```
title: Scalar model midlife
  data:
       file is merged_NSHD_NCDS_BCS.csv;
  variable:
       names are
       case_id cohort age anin let_s let_a let_h mem_i mem_d;

      usevar are
       anin let_s mem_i mem_d;

       missing are all (-99);
       ! categorical are all;
       !USEOBSERVATIONS are cohort== 1;
       grouping = cohort (1 = NSHD 2 = NCDS 3 = BCS);

  analysis:
       estimator = MLR;

  model:
   G BY  anin* let_s mem_i mem_d;
   G@1;
   [G@0];

  model NCDS:
 ! G BY  anin* let_s mem_i mem_d;
 ![anin-mem_d];
 G;
 [G];

    model BCS:
  !G BY  anin* let_s mem_i mem_d;
 ![anin-mem_d];
  G;
 [G];


  output:
  sampstat standardized modindices(ALL);
```

7.5 Mplus code partial scalar invariance (midlife)

```
title: Partial scalar model midlife
  data:
        file is merged_NSHD_NCDS_BCS.csv;
   variable:
        names are
        case_id cohort age anin let_s let_a let_h mem_i mem_d;

       usevar are
        anin let_s mem_i mem_d;

        missing are all (-99);
        ! categorical are all;
        !USEOBSERVATIONS are cohort== 1;
        grouping = cohort (1 = NSHD 2 = NCDS 3 = BCS);

   analysis:
        estimator = MLR;

   model:
    G BY  anin* let_s mem_i mem_d;
    G@1;
    [G@0];

  model NCDS:
 ! G BY  anin* let_s mem_i mem_d;
 [let_s ] (a);
 [mem_d] (b);
 G;
 [G];

     model BCS:
  !G BY  anin* let_s mem_i mem_d;
 [let_s ] (a);
 [mem_d] (b);
  G;
 [G];


   output:
   sampstat standardized modindices(ALL);
```

8 References

Ackerman, P. L. (2017). Adult intelligence: The construct and the criterion problem. *Perspectives on Psychological Science, 12*(6), 987-998.

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.

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

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.

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.

Davis, D., Cooper, R., Terrera, G. M., Hardy, R., Richards, M., & Kuh, D. (2016). Verbal memory and search speed in early midlife are associated with mortality over 25 years' follow-up, independently of health status and early life factors: a British birth cohort study. *International Journal of Epidemiology, 45*(4), 1216-1225.

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

Dorans, N. J. (2018). Scores, scales, and score linking. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development* (pp. 571-605). New Jersey: Wiley.

Dunteman, G. H. (1989). *Principal Components Analysis. 1st ed.* CA: Sage Publications, Inc.

Elliott, C., Smith, P., & McCulloch, K. (1997). *British Ability Scales Second Edition (BAS II). Technical Manual*. London: Nelson.

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.

Griffith, L. E., Van Den Heuvel, E., Fortier, I., Sohel, N., Hofer, S. M., Payette, H., . . . Doiron, D. (2015). Statistical approaches to harmonize data on cognitive measures in systematic reviews are rarely reported. *Journal of Clinical Epidemiology, 68*(2), 154-162.

Gross, A. L., Power, M. C., Albert, M. S., Deal, J. A., Gottesman, R. F., Griswold, M., . . . Sharrett, A. R. (2015). Application of latent variable methods to the study of cognitive decline when tests change over time. *Epidemiology, 26*(6), 878.

Gross, A. L., Sherva, R., Mukherjee, S., Newhouse, S., Kauwe, J. S., Munsie, L. M., . . . Green, R. C. (2014). Calibrating longitudinal cognition in Alzheimer's disease across diverse test batteries and datasets. *Neuroepidemiology, 43*(3-4), 194-205.

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.

Jewsbury, P. A., Bowden, S. C., & Duff, K. (2017). The Cattell–Horn–Carroll model of cognition for clinical assessment. *Journal of Psychoeducational Assessment, 35*(6), 547-567.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*: New York: Springer.

Kolen, M. J., Tong, Y., & Brennan, R. L. (2009). Scoring and scaling educational tests. In *Statistical models for test equating, scaling, and linking* (pp. 43-58). New York: Springer.

Little, T. D. (2013). *Longitudinal Structural Equation Modeling*. New York: Guilford Press.

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods, 14*(2), 126.

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research, 29*(3), 223-236.

Moulton, V., McElroy, E., Richards, M., Fitzsimons, E., Northstone, K., Conti, G., . . . O'Neill, D. (2020). *A guide to the cognitive measures in five British birth cohort studies*. London, UK: CLOSER.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide. Eighth Edition.* . Los Angeles, CA: Muthén & Muthén.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *The American Council on Education/Macmillan series on higher education. Educational measurement* (pp. 221-262): Macmillan Publishing Co, Inc; American Council on Education.

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.

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. Kopenhagen: Nielsen & Lydiche.

Schneider, J., & McGrew, K. (2012). The Cattell-Horn-Carroll (CHC) Model of Intelligence v2. 2: A visual tour and summary. Retrieved from: http://www.iapsych.com/chcv2.pdf

Schneider, J., & McGrew, K. (2018). The Cattell–Horn–Carroll theory of cognitive abilities.

Schoon, I. (2010). Childhood cognitive ability and adult academic attainment: Evidence from three British cohort studies. *Longitudinal and Life Course Studies, 1*(3), 241-158.

Silverwood, R. J., Richards, M., Pierce, M., Hardy, R., Sattar, N., Ferro, C., . . . Teams, D. C. (2014). Cognitive and kidney function: results from a British birth cohort reaching retirement age. *PloS One, 9*(1), e86743.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman and Hall/CRC.

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.

Tyrell, F. A., Yates, T. M., Widaman, K. F., Reynolds, C. A., & Fabricius, W. V. (2019). Data harmonization: Establishing measurement invariance across different assessments of the same construct across adolescence. *Journal of Clinical Child & Adolescent Psychology, 48*(4), 555-567.

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.

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.

Yen, W., & Fitzpatrick, A. R. (2006). Item Response Theory. *Educational Measurement*, 187-220.

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.