Institute of Education

UCL

Calibration of mental health and cognitive ability measures

By Emily Gilbert, Hannah Jongsma, Vanessa Moulton, Praveetha Patalay and George B. Ploubidis

CENTRE FOR LONGITUDINAL STUDIES

UKRI Economic and Social Research Council

Summary

Research question

Understanding the comparability of the various measures used to assess adult mental health in the British birth cohorts and pilot a study to examine comparability of childhood cognitive assessments administered at age 10/11 across cohorts.

Methods

For both studies we required ethics approvals from institutional ethics committees and agreements with external fieldwork agencies for data collection, the latter causing delays to the start of data collection.

Mental health

We aimed to recruit 5000 adult participants from the general population (not cohort members) across the full adult age range. Participants completed nine different mental health measures used in the cohort studies, measuring both psychological distress and wellbeing. Data was used to examine the correlation between the different measures and will be used to examine the correspondence of the various cut-off scores used in the cohorts, as well as to provide information about their relative utility to inform measure selection for future cohort sweeps.

Cognitive ability

We aimed to recruit 100 children aged between 118 and 140 months (year 5/6) from five schools. Children were asked to complete seven cognitive ability tests used in the NSHD, NCDS, BCS and MCS, as well as the British Ability Scales (BAS) 3 core set.

The pilot set-up, fieldwork, data collection and analysis will inform the feasibility of a larger study. Results will be used to construct overall test scores, derive ability and difficulty estimates and, where possible examine measurement invariance between tests. The BAS3 will be used as a Gold Standard to calibrate the remaining measures against.

Findings

Mental health

Missing data was generally low across questionnaires (maximum 5.8%). In general, there was at least a moderate-high correlation (>0.60) between all measures. In our initial analyses we have focussed on the correlations between the different measures collected, and in depth calibration analysis is planned next. The highest correlation between different measures of psychological distress was 0.90 (between the General Health Questionnaire-12 item version and General Health Questionnaire -28 item version) and the lowest correlation was between the Malaise questionnaire and the Kessler-6 (0.62). The highest correlation between different measures of well-being was between the wellbeing subscale of the SF-36 health survey and the Warwick-Edinburgh Mental Wellbeing Scale (0.78) whereas the Office for National Statistics life satisfaction measure correlated with both these measures to a slightly lesser extent (0.67 for the SF and 0.69 for the WEMWBS).

Cognitive ability

Four schools have agreed to participate, and fieldwork is ongoing. Collected data will be processed as per the original marking instructions and delivered to CLS in digital format. On the basis of progress to date, the following issues have been noted as affecting the feasibility

of a wider calibration project. The set-up time was considerable, including agreement of contracts (two months) with the external agency, recruitment of schools, establishing copyright issues and preparation of test materials. In the pilot, we excluded tests using CAPI or external software, due to set-up time, compatibility of software, copyright and administration issues. In addition, at age 10/11 only a subsample of the cognitive tests that have previously been used in the cohorts could be administered in exactly the same way as in earlier cohorts. This is partly due to length - we needed to identify 2 hours of tests (from 5 hours in total). Some tests were lengthy (>30 minutes) and during fieldwork it was established that the administration test time was greater than the estimated test times. Some tests used outdated language, formulae and visuals. In future work, if full versions of the tests are used, they will need to be updated by expert including educational psychologists, cognitive and language scientists. Finally, using this approach, there is a finite window when one to one administration can be conducted in school, given room availability, school hours, breaks, school holidays and school commitments.

Outputs and next steps

Mental health

Outputs will include one peer review journal article setting out the results from the mental health calibration analyses. We will deposit the data collected for this project in a data repository to permit other researchers to use the data for their research if interested.

The mental health data gathered in this project will be used to examine the degree of correspondence between their respective cut-off scores and will be used to inform decisions on future measures in the cohorts and beyond.

Cognitive ability

The cognitive ability pilot learnings and data in this pilot will facilitate a feasibility assessment of a larger calibration of childhood cognitive ability measures across the British birth cohorts.

Outputs will include one peer review journal article setting out the results from this pilot cognitive calibration. Also, production of a grant application to take forward this project beyond the pilot stage

Contents

Mental health calibration project

A widely-used feature of the British birth cohorts is the wealth of mental health collected throughout the life course. Nevertheless, the measures used vary, both within and between cohorts. This project will calibrate existing adult mental health measures across national birth cohorts with each other and (where possible) with external up-to-date measures using innovative survey design and statistical modelling. The proposed work complements the existing CLOSER harmonisation project on mental health which is, where possible, creating harmonised variables using existing cohort data. In combination with findings from the current project they will increase the usability of the studies, particularly from disciplines that haven't traditionally done so.

Ethics

Ethical approval for the project was gained from the Institute of Education Research Ethics Committee prior to the commencement of data collection. The Committee reviewed the study protocols, participant-facing information sheet, debrief information and consent forms, and gave full ethical approval (reference: REC 1210).

Contracts and fieldwork agency

Given the online survey nature of the data collection, we decided to use Qualtrics as the online survey platform. Our institution has a contract with them, resulting in the provision of the online survey platform to use for this project at no extra cost. Qualtrics also has an online survey panel in the UK and given the use of their online platform to collect data we decided to use them as the fieldwork agency for recruiting participants. The statement of work with Qualtrics based on our survey design was agreed in early March. The next step was for the contracts teams at both institutions to agree on the terms of the contract. This process took more than two months with back and forth dozens of times. Despite the uncomplicated nature of project collection (all data were anonymous, there was no copyright or intellectual property issues), this was a longer than expected process which seriously strained the planned timelines for the project. (Though in the end the project has been delivered on time).

Measures being harmonised

The aim of the project was to include measures of mental health and wellbeing that are used in the various CLS cohorts (1958, 1970, Next Steps, MCS) and related birth cohorts used in cross cohort research and included in CLOSER (1946 and ALSPAC).

The measures included in this calibration project are detailed in Table 1 below. The Malaise questionnaire, Kessler-6, GHQ-12, GHQ-28, CESD, PSE and PSF assessed psychological distress, whereas the SF, ONS life satisfaction and WEMWBS assessed wellbeing. The table also provides information on the cohort and ages in which this measure has been used to-date to as rationale for the inclusion of the measure in this project.

Table 1: Measures included in the 2019 mental health calibration project

Measure	Description	Cohorts (ages) used
Malaise questionnaire	24 items covering emotional disturbance and associated somatic symptoms. No timescale indicated, binary (yes/no) response options. Total possible score range: 0-24.	NCDS[1] (age 23,33,42,50) BCS[2] (age 16,26,30,34,42, 46)
Kessler-6	6 items covering nonspecific psychological distress during the past 30 days. Responses recorded on a 5-point Likert scale. Total possible score range: 0-24.	BCS[2] (age 34) MCS[3] (parent-completion, child age: 3,5,7,11,14)
General Health Questionnaire (GHQ)-12	12 items from the original General Health Questionnaire aimed at detecting psychiatric disorders in the community. Responses recorded on a 4-point Likert scale. Timeline: the past few weeks. Total possible score range: 0-36	BCS[2] (age 16, 30) Next Steps (age 15,17,25)
General Health Questionnaire (GHQ-28	28 items from the original General Health Questionnaire aimed at detecting psychiatric disorders in the community. Responses recorded on a 4-point Likert scale. Timeline: the past few weeks. Total possible score range: 0-84.	NSHD[4] (age 53, 63, 69)
Centre for Epidemiologic Studies Depression (CESD) scale	10-item self-report measure for depressive symptoms in the past week. Responses recorded on a 4-point Likert scale. Total possible score range: 0-30.	BCS[2] (age 46)
Present State Examination (PSE)	Responses recorded on a 3-point Likert scale. Total possible score range: 0-18.	NSHD[4] (age 36)
Psychiatric Symptom Frequency Scale (PSF)	18 items designed to assess symptoms of anxiety and depression experienced over the past year in the general population. Responses recorded on a 6-point Likert scale. Total possible score range: 0-90.	NSHD1 (age 43)
Short Form Survey (SF)	10 items from the RAND Short Form Survey designed to measure quality of life. Responses recorded in quintiles (0-100%) with higher percentage representing a preferable health state. Total possible score range: 0-1,000.	NCDS[1] (age 50) BCS[2] (age 46/47) ALSPAC[5] (age 18, 22, 26)
Office for National Statistics Life Satisfaction (ONS)	'Overall, how satisfied are you with your life nowadays?' Responses recorded from 0 (not at all) - 10 (completely). Total possible score range: 0-10	NSHD[4] (age 36,63) NCDS[1] (age 23, 33, 42, 46, 50) BCS[2] (age 26,30,34,42,46) Next Steps (age 19, 25)
Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS)	14 items designed to measure positive mental health in the last two weeks. Responses recorded on a 5-point Likert scale. Total possible score range: 0-56.	NSHD[4] (age 63) NCDS[1] (age 50) BCS[2] (age 42, 46) Next Steps (age 25)

[1] National Child Development Study 1958
[2] British Cohort Study 1970
[3] Millennium Cohort Study 2000
[4] National Survey of Health and Development 1946
[5] Avon Longitudinal Study of Parents and Children 1991

Methods

The project aimed to recruit 5,000 participants. Data collection was done via Qualtrics. Participants were equally distributed across five age groups (23-30, 31-40, 41-50, 51-60 and 61-70) and were quota-sampled to reflect the general population in terms of sex, ethnicity, country of residence (England, Scotland, Wales, Northern Ireland) and highest level of education achieved. Quota percentages were derived from the 2011 census.

Procedure

Participants were informed in writing of the purpose of the study, the voluntary nature of their participation, the minimal risks involved in participating and the anonymous collection of data. Written informed consent was obtained before starting the survey. Participants were completely anonymous (no identifying information was included in the survey).

All participants were asked to complete all measures listed in Table 1, with the exception of the Present State Examination (PSE) and Psychiatric Symptom Frequency scale (PSF) which were only administered in the 31-40 and 41-50 age groups respectively (as they had only been used in the 1946 cohort at these particular ages). Completion of all measures combined took no more than 20 minutes. The order in which the measures were administered was randomised to prevent order effects. After completion of all instruments, participants were given information on how to access support for mental health difficulties.

Results

By 18th June we received a total of 4,381 (87.6% of target) responses which are analysed for this report (data collected up to this point are analysed to allow time for analysis). The remaining data are being collected and will be available shortly. Missing data due to item non response was generally low across questionnaires (below 5.8%). Whilst both the 23-30 and 31-40 year age group included 1,120 participants (25.6%), the 61-70 year old group only included 491 (11.2%) participants (to-date, the target is to get a minimum of 1000 participants in each age group). Approximately 64% of all participants were women (n=2,800). The majority of participants were of white ethnicity (n=3,957, 90.3%), with the remainder being of Asian (n=206, 4.7%), Black (n=99, 2.3%), mixed (n=95, 2.2%) and ‘other’ (n=24, 0.6%) ethnicity. Approximately 86 percent of participants lived in England (n=3,767), followed by Scotland (n=302, 6.9%), Wales (n=217, 5.0%) and Northern Ireland (n=95, 2.2%). Approximately a third of participants (n=1,680, 38.4%) was educated to GCSE-equivalent level, whereas 5.4% (n=238) had no formal educational qualifications. Approximately 24.3% (n=1,064) had at least A-level education and 25.7% (n=1,126) was educated to degree level, with a further 6.2% (n=273) educated to higher degree level.

The mean and standard deviation of all questionnaires are reported in Table 2 below. The highest correlation was reported between the GHQ-12 and GHQ-28 (0.90) and the lowest (-0.46) between the Malaise questionnaire and the ONS Life Satisfaction question. In general, there was at least a strong correlation (0.60-0.79) between all questionnaires. The highest correlation between different measures of psychological distress was 0.90 (between the GHQ-12 and GHQ-28) and the lowest correlation was between the Malaise questionnaire and the Kessler-6 (0.62). The highest correlation between different measures of well-being was between the SF and WEMWBS (0.78) whereas the ONS correlated with both these measures to a slightly lesser extent (0.67 for the SF and 0.69 for the WEMWBS).

The mean and standard deviation of all questionnaires are reported in Table 2 below.

Table 2: Mean and standard deviation of all measures, and their correlations

		Correlations									
Measure	**Mean (SD)**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1 Malaise	9.10 (6.19)	x	0.73	0.62	0.70	0.70	0.70	0.77	-0.69	-0.46	-0.54
2 Kessler	9.03 (6.41)		x	0.71	0.77	0.80	0.73	0.80	-0.80	-0.61	-0.68
3 GHQ-12	14.92 (6.63)			x	0.90	0.74	0.71	0.68	-0.72	-0.54	-0.62
4 GHQ-28	30.30 (16.66)				x	0.78	0.76	0.75	-0.75	-0.55	-0.62
5 CESD	11.84 (7.01)					x	0.78	0.84	-0.83	-0.64	-0.70
6 PSE	507.44 (214.03)						x	x[1]	-0.77	-0.51	-0.57
7 PSF	5.58 (2.61)							x	-0.79	-0.57	-0.61
8 SF	7.25 (4.52)								x	0.67	0.78
9 ONS	32.89 (21.71)									x	0.69
10 WEMWBS	27.80 (11.74)										X

[1] No correlation derived as PSE and PSF were administered in different age groups.

Next steps

The data gathered in this project will be used to calibrate groups of mental health measures used in the cohorts as listed in Table 1. This work will enable a more formal comparison of mental health measures and an examination of their existing cut-offs to generate comparable thresholds for “caseness” across measures. These analyses will generate guidance on how researchers should adjust results from various measures to aid comparability. The findings from this project will also help inform the measures to include in future sweeps of the cohort studies, for instance, forthcoming adult sweeps of the Millennium Cohort Study.

Outputs and dissemination

The key audience for this work is academic users who will be able to use the calibrations, and the underlying data collected in their research, to carry out comparative work using these cohorts. Outputs will include one peer review journal article setting out the results from the mental health strand. We will deposit the data collected for this project in a data repository to permit other researchers to use the data for their research if interested.

We will organise a workshop to disseminate findings and guide researchers on how to use these data and the findings in their own research.

Cognitive measures calibration

In the British birth cohorts a wide variety of cognitive measures have been collected throughout the life course. Work investigating the extent to which these cognitive measures are comparable within and between cohorts is in its infancy. This project was a pilot study to explore the feasibility of calibrating childhood (age 10/11) cognitive tests in four British birth cohorts: The National Survey of Health and Development (NSHD, 1946), the National Child Development Study (NCDS, 1958), the British Cohort Study (BCS, 1970) and the Millennium Cohort Study (MCS, 2000). The approach taken aimed to replicate previous administration of the cognitive tests in the cohorts with a current, independent, cohort of children.

The proposed work complements the CLOSER harmonisation project on cognition which is assessing the cognitive measures and, where possible, creating harmonised variables using existing cohort data. In combination with findings from this pilot the work will increase our understanding of the feasibility of constructing comparable cognitive measures across the British birth cohort studies.

Ethics

Ethical approval for the project was gained from the Institute of Education Research Ethics Committee prior to the commencement of data collection. The Committee reviewed the study protocols, participant-facing information sheet, debrief information and consent forms, and gave full ethical approval (reference: REC 1210).

Measures being calibrated

The measures included in this project are detailed in Table 3. One major reason for choosing this particular age-group (aged 10 - 11 years or 118-140 months) was the wide range of cognitive tests administered at this age across the four cohorts. The tests included in this pilot were a subset of all measures administered at age 10/11, with at least one test administered from each of the four cohorts. Other reasons for the choice of tests are outlined under learnings in the results section. The BAS3 core set served as the Gold Standard against which to calibrate the remaining measures.

Methods

The project aimed to recruit 100 children aged between 118 and 140 months (year 5/6), spread over five schools. Data collection and fieldwork was conducted via fieldwork agency National Foundation for Educational Research (NFER) in schools. Basic demographic data and Key Stage 1 results were collected for each child; region, type and percentage of children eligible for free school meals were collected for each school.

In terms of analysis, CLS will construct overall test scores for each test, use item response theory to identify ability and difficulty estimates, measurement precision/reliability and examine measurement equivalence between tests. Additionally, the BAS3 will be used as a Gold Standard to calibrate the cognitive tests in the cohorts. The BAS3 has a standardised sample derived from a validation sample of approximately 1,500 children aged 3 to 18 representative of the UK population of children. The frame was used to establish new norms (2010-11) and identified difficulty estimates on each item of the tests. The BAS3 will be used as a benchmark to compare the children's results on the existing cognitive tests in the cohorts, to their results on the BAS3 and the external BAS3 validation sample.

Table 3: Cognitive measures included in the calibration project

	Test	Cohort	Age (months)	Mode	Duration (minutes)
1	**NFER Verbal and Non-Verbal Test**	NSHD[1] NCDS[2]	127-137 130-152	Pen & paper	30
2	**Reading comprehension test**	NCDS[2]	130-152	Pen & paper	20
3	**BAS tests** **(Word) Similarities** **Word definitions** **Recall of Digits** **Matrices**	BCS[3]	117-139	 Oral and verbal response Oral and verbal response Oral and verbal response Pen & paper	30 in total
4	**BAS II Verbal Similarities**	MCS[4]	122-148	CAPI: skip rules Oral and verbal response	<10
5	**Gold Standard: BAS3 (core set)**			Face to face	35-40

[1] National Survey of Health and Development 1946
[2] National Child Development Study 1958
[3] British Cohort Study 1970
[4] Millennium Cohort Study 2000

Procedure

The fieldwork was conducted in June 2019, post key-stage testing. Schools were informed of the purpose of the study, the voluntary nature of the children's participation, the minimal risks involved in participating and the anonymous collection of data. When schools agreed to participate, parents were approached with the same information. Written informed consent was obtained from parents, and assent from their children. Participants were completely anonymous. All children in the appropriate age bracket were eligible, regardless of any learning disability such as dyslexia.

The tests were administered under the same conditions as previously (with the exception of tests in the MCS which, at the time, was conducted at home). All tests were administered individually in schools, with the child supported by a Test Administrator (TA). Children were asked to complete all tests listed in Table 3. The exact same tests were administered in line with the original instructions. Tests and their accompanying instructions were reformatted for use, and TAs were briefed by CLS and NFER. The tests were estimated to take two hours in total per child and were split over two days, one hour per day.

Tests were administered using pen and paper, or read to the child with the verbal response recorded on paper as previously administered (see Table 3). Tests were grouped together and split into days A (test sets 1-3) and B (test sets 4-5) based on time taken to complete and the content of the tests. The order in which the tests were administered was rotated by day and tests within the day grouping (unless the original tests were conducted in a specific order e.g. BAS measures in BCS) and the order of administration was noted by the TA. The TA was also asked, where possible, to note any period effects (e.g. a child doesn't understand a word because it is no longer commonly used).

Results

Through NFER, four schools have been recruited and agreed to participate as shown in Table 4.

Table 4: Progress from NFER as of June 21st

No. of schools Recruited	No of schools with testing scheduled or completed		No. of Pupils	
	Competed	Scheduled	Tested to date	Scheduled to test
4	1	2	33	53

As of June 21st, field work is ongoing. Collected data will be processed as per the original marking instructions and data delivered to the CLS in digital format.

On the basis of progress to date, the following issues have been noted as affecting the feasibility of a wider calibration project. The set-up time was considerable, including agreement of contracts (two months) with the external agency, recruitment of schools, establishing copyright issues and preparation of test materials.

At age 10/11 across the four cohorts there were approximately five hours of cognitive measures. We were advised by NFER that at age 10/11 two hours of testing was an unreasonable burden for each child, and one hour ideal. Therefore we were restricted on including some tests which took 30 minutes or more to complete i.e. the Arithmetic Test in the NSHD, and the Edinburgh Reading Test, Pictorial Language Comprehension Test and Friendly Maths Test in BCS. In addition, some tests particularly in the earlier cohorts included questions which the current cohort of children would not understand because of changes in language and mathematical formulae e.g. Arithmetic Test in NSHD. The CANTAB tests were initially to be included in the pilot, however it was not possible to include them in the end because the software set-up previously used in the MCS was not available as the CANTAB tests have since been revised, so the current test version is not directly comparable with the version administered in the MCS. Also, the cost of copyright and administration, as well as the set-up was beyond the scope of this pilot.

School recruitment was challenging: in part because it was required to identify schools who would take part at short notice. In scaling up, recruitment of schools may prove challenging, as schools are required to commit a significant amount of time to the task. Furthermore, during the early summer, schools were administering key-stage tests, therefore the fieldwork could not take place until post internal testing, specifically in June. In the first instance we aimed to conduct 250 interviews, however with a short window to conduct the fieldwork the sample size was reduced to 100. Also, the number of children completing the tests per day with an individual TA present was constrained by breaks and school hours (original estimation 5 children per TA per day). In addition, the time taken to complete the tests varied considerably by each child, with a smaller sample the norm is more difficult to estimate. At present, the tests are on average taking 1.5 hours for group A tests and 1.1 hour for B, thus reducing the number of completes within the school timetable.

Pupils were generally enjoying the task and were keen to come back for day 2 of testing. However, some tasks were very similar, which pupils found a little distracting. The test

materials were updated, as far as possible, into a usable format. However, for the current cohort the font on some of the tests, language and understanding of particular concepts was outdated. If full versions of the tests are used, ‘expert’ educational psychologists will be needed to adapt the tests.

Next steps

The learnings and data gathered in this pilot will enable a feasibility assessment of a larger calibration of childhood cognitive tests across the birth cohorts. Due to the time it would take for an individual child to complete all tests at the appropriate age group in the cohorts and the associated commitment by schools an innovative approach is needed. For example, dependent on the final results of the pilot, an approach may include asking children different sets of tests but including test(s) which bridge across children or conducting fieldwork post reduction of the number of items per test which capture the similar amounts of information, where possible. These items could be tested on a larger sample and used to anchor retrospective tests with a gold standard test such as BAS3.This approach could also possibly eliminate the difficulty of period-appropriate tests.

Outputs and dissemination

The key audience for this work is academic users who will be able to use the calibrations, and the underlying data collected in their research, to carry out comparative work using these cohorts. Outputs will include a report, including the analysis from the pilot and further recommendations for next steps. Results dependent, one peer review journal article setting out the methodology and findings from this pilot cognitive calibration will be written. In addition, a grant application to take forward this project beyond the pilot stage will be delivered.