A latent variable modelling approach for the pooled analysis of individual participant data on the association between depression and Chlamydia infection in adolescence and young adulthood in the UK

Artemis Koukounari[1,2], Andrew J. Copas[3], Andrew Pickles[1]

[1]Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, United Kingdom

[2]Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, UK

[3]Institute for Global Health and Hub for Trials Methodology Research at the MRC Clinical Trials Unit, University College London, United Kingdom

Summary

Despite the increasing evidence of association between Chlamydia infection and depression, currently there is a paucity of research with limited scope to better understand the temporal nature of the relationship between them. We consider this problem in adolescence and young adulthood by pooled analysis of 7250 participants from the Avon Longitudinal Study of Parents and Children (ALSPAC) and the Third National Survey of Sexual Attitudes and Lifestyles (Natsal-3). We propose a latent variable modelling approach which can handle harmonization of categorical variables including ordinal measures from the two studies as well as measurement error and time trends.

Keywords: ALSPAC, chlamydia, depression, NATSAL-3, latent variable models, pooled analysis

Correspondence: Artemis Koukounari, Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, United Kingdom. (E-mail: artemis.koukounari@kcl.ac.uk ; artemis.koukounari@lshtm.ac.uk)

1. Introduction

Increasingly, biostatisticians, epidemiologists as well as social, behavioural and medical researchers are interested in leveraging the power of pooled analysis of individual participant data (Griffith *et al.* 2013, Higgins *et al.* 2001). There are many advantages to pooled analysis of individual participant data for combining data from two or more studies. These advantages include greater power to detect effects, increased sample heterogeneity, stability for studying rare outcomes and the ability to perform more sophisticated analyses than pooled analyses that rely on published aggregated results (Riley *et al.* 2010, Siddique *et al.* 2015). Nevertheless, pooled analysis of individual participant data introduces an essential challenge where most often the scales (i.e. collections of items combined into a composite score intended to reveal levels of theoretical variables not readily observable by direct means) of interest between the different studies to be combined are measured in different ways. On the other hand, to restrict the pooled analysis of individual participant data from different studies to those that have identical measurement protocols would seem unduly restrictive.

The particular substantive topic that we consider is to explore the predictive relationships between trajectories of depression during middle childhood, adolescence and early adulthood and acquisition of biologically confirmed current *Chlamydia trachomatis* infection, analysing for the first time relevant data of young people living in the UK. Despite the increased evidence of association between Chlamydia infection and depression (Doyle *et al.* 2015, Wang *et al.* 2014) currently there is a paucity of research presenting limited scope and no robust evidence to better understand the temporal nature of the relationship between these two diseases. To our knowledge, only two studies have assessed to date the longitudinal relationship between depression and sexually transmitted infections (STIs) among adolescents using repeated measurements of depression based on US data (Khan *et al.* 2009, Shrier *et al.* 2002), without though distinguishing between *C. trachomatis* and other STIs and with the

earlier of these studies using only self-reported STIs (Shrier *et al.* 2002) while the second one (Khan *et al.* 2009) has been faced with a long time gap between data collection in adolescence and young adulthood. Such research is important to inform meaningful public health interventions and prevention strategies. For instance, if depression is suggested as a predictor of infection then improved diagnosis and care for depression would be needed not only because depression constitutes an important public health concern in itself but also because addressing depression may lead to improved physical health, such as lower risk for Chlamydia infection. Alternatively, if depression is implied as an outcome of infection this could have major implications for including mental health in the management of individuals diagnosed with chlamydia.

The Avon Longitudinal Study of Parents and Children (ALSPAC) (Boyd *et al.* 2013) is highly informative for our investigation, as it involves repeated measurements of depressive symptoms from middle childhood to adolescence (approximately 18 years) and chlamydia infection as measured through nucleic acid amplification of urine specimens taken once at age approximately 17, with questions about recent sexual behaviour also asked at the same time. However, the Chlamydia prevalence in this general population sample was very low (20/2879) - even after adjustment for bias introduced by selective participation in testing. Furthermore, only 51% of those eligible for follow up attended the clinic and 60% of these provided urine specimens for testing for chlamydia infection (Crichton *et al.* 2014), most likely creating problems in the stability of complex statistical modelling estimation. Moreover, with the current available ALSPAC measures, empirical study of trajectories of depression can only get realized up to the approximate age of 18 years old, rendering the investigation of the hypothesis that depression might follow rather than precede a *C.trachomatis* diagnosis, impossible. We thus seek to handle such intricacies, by pooling and simultaneously analysing ALSPAC data and other sexual health and depression relevant data covering ages 16-24 years old from an external very detailed cross-sectional national sample such as the third British National Survey of Sexual Attitudes and Lifestyles (Natsal-3) (Erens *et al.* 2014, Mercer *et al.* 2013). An additional

methodological challenge in these circumstances is to equate different scales which purport to measure the same construct at the item level, so that the two separate studies using different scales of measurement for both the potential risk factor and health outcome of interest can be harmonised into respective common metrics.

Thus, the main goal of this paper is to explore and propose a latent variable modelling approach that addresses all of these particular issues, presuming that unobserved (latent) propensities give rise to the observed (imperfect) measurements on related questions to Chlamydia infection and two separate scales of depression, permitting pooling of individual patient data (IPD) from a longitudinal (i.e ALPSAC) and a cross sectional study (i.e. Natsal-3). We also provide Mplus codes- enhancing transparency of the selected methods and for future reference. By pooling these two studies together, Natsal-3 can be viewed to enrich the ALSPAC study, by being more diverse in terms of representation of ethnicity and geographic locations, concerning UK adolescents and young adults. Furthermore, through the proposed modelling approach we hope that maximizing us of available Chlamydia infections data and harmonizing measurement of Chlamydia and depression constructs across studies, provides a sound basis to test and generalize the effects of depression and changes in depression on acquisition of Chlamydia. This is justified even though ALSPAC is a regional study and not representative of the UK by the fact that in general the overall associations observed in a study population may not apply to every subgroup even if the study population is representative of some larger source population (Rothman *et al.* 2013). The overall effect is merely an average effect that has been weighted by the distribution of people across these subgroups. It is not representativeness of the study subjects that enhances the generalization, it is knowledge of specific conditions and an understanding of mechanism that makes for a proper generalization (Rothman *et al.* 2013). However, for us to proceed it is necessary to assume that the associations between depression trajectories and chlamydia infection are the same in the two populations sampled, an assumption that is not testable given the employed data. Nonetheless, the proposed approach does allow evaluation of *measurement*

equivalence or *measurement invariance* or *non - differential item functioning (DIF)* between shared items of the examined studies (Curran and Hussong 2009). By testing measurement invariance hypotheses, one can investigate 'whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute' (Horn and Mcardle 1992). Hence, questions about the comparability of data obtained from different studies, essential within any comparative research, can be addressed through this approach (Curran *et al.* 2008).

Latent variable models are fundamental and well known statistical methods in the psychometrics field (Groenen and Andries van der Ark 2006). Their potential to provide psychometrically sound scale scores reflecting potential differences in item functioning across different individual characteristics, studies and over time, as well as the derived parameters is what renders them as attractive methods for generating and analysing harmonized pooled datasets. Such use of latent variable models in the wider epidemiology is relatively uncommon. Full information maximum likelihood accommodates partially missing data under the Missing At Random (MAR) assumption within these models. Our approach assumes that multivariate associations among the available variables of one study can be extrapolated to the second study which by design has no data on one or more variables. We also assume that data are missing at random within studies. Both these assumptions are untestable.

The remainder of the paper is organized as follows. Section 'Data & exploratory analysis' introduces the data. Section 'Methodology' describes our latent variable modelling approach: Initially, we propose a *latent class model* for integration of relevant cross-sectional Chlamydia infection data, with two classes defined by infection status at the time of testing, or at the time of other data collection if no testing occurred, which for individuals in these data occurs once at an age between 16 and 24 years. The uninfected class will therefore include individuals who have been infected but have cleared infection spontaneously, or due to treatment, and also individuals who may become infected in the future. We then employ the IRT graded response model to get pre-calibration parameters. We subsequently extend existing *latent growth curve IRT models* (Wang *et al.* 2015) for harmonization by

integrating longitudinal and cross-sectional ordinal measurements from ages 10-18 as well as cross-sectional ordinal measurements from ages 19-24 (i.e. spanning the whole time period from middle childhood to young adulthood) for depression estimated in one stage via *anchored calibration*. We then show how to link the derived latent construct of Chlamydia infection and growth factors of the trajectories of depression, providing a pooled effect of the studied associations. For this purpose, we split time for the trajectory of depression into ages 10-16 and 16-24 years, so that the first period is prior to the chlamydia test and likely prior also to infection for most infected individuals, whilst the second will be around or after the time of the test and possibly after the time of any infections. We also provide some additional sensitivity analyses -wherever feasible- evaluating the success of the harmonization process enhancing also the validity of our findings. Our results section provides a summary of the main findings, after which we discuss these findings, the plausibility and reasonableness of implied assumptions and make recommendations for further research.

2. Data & exploratory analysis

The two studies contributing data to the current pooled analysis are described below.

2.1 Data set one-source: The Avon Longitudinal Study of Parents and Children (ALSPAC)

ALSPAC is an ongoing population-based study designed to investigate the effects of a number of factors on health and development. ALSPAC invited all pregnant women resident in the former Avon Health Authority (Bristol) in South West England with an estimated date of delivery between 1 April 1991 and 31 December 1992 to take part, resulting in a 'core' cohort of 14541 pregnancies and 13617 singletons alive at 12 months of age. Ethical approval for the study was obtained from the ALSPAC Law and Ethics Committee and local research ethics committees (Boyd *et al.* 2013). The study website

contains details of all the data that are available through a fully searchable data dictionary: www.bris.ac.uk/alspac/researchers/data-access/data-dictionary.

2.1.1 ALSPAC measures

Depressive symptoms in late childhood and adolescence

Depressive symptoms in late childhood and adolescence were assessed using at the first surveyed age the short (13-item) Mood and Feelings Questionnaire (MFQ) (Angold *et al.* 1995, Messer *et al.* 1995, Niarchou *et al.* 2015) and at subsequent ages the 16-item MFQ, to enquire about the occurrence of depressive symptoms over the past 2 weeks. The MFQ was completed by ALSPAC study members at research clinics at ages 10.5, 12 and 13 years and by postal questionnaire at approximately 16, 17 and 18 years, scoring each as 0 ("not true") to 2 ("true").

Chlamydia tests and sexual behaviour indicators in adolescence and young adulthood

9568 eligible ALSPAC participants were invited to attend a research clinic at approximately 17 years of age. Those who attended the clinic were offered biological tests for chlamydia and gonorrhoea, in partnership with the National Chlamydia Screening Programme (NCSP). All clinic participants were invited to provide urine specimens irrespective of their sexual activity and from those who did, chlamydia positives were confirmed using the Gen-Probe Aptima CT assay. They were asked to report whether they were sexually active with the question 'Have you ever had sexual intercourse with either a female (woman/girl) or a male (man/boy)? (Yes/No)'. Those answering in the affirmative, were asked to report numbers of partners in the past year and in total as well as new partners in the past year and whether they used a condom on most recent occasion. These variables were used as measures of recent sexual behaviour in the current study. Detailed descriptions of methods and numbers of participants at the research clinics have already been published (Chrichton *et al.* 2014). In all presented

models (-except for lack of direct adjustment in the last model displayed in Table 7-) we conditioned on gender and thus MAR was assumed conditionally on this covariate.

2.2. Data set two-source: Third National Survey of Sexual Attitudes and Lifestyles (Natsal-3)

Natsal-3 is a stratified probability sample survey of 15162 men and women aged 16-74 years in Britain interviewed between Sept 6, 2010, and Aug 31, 2012. The overall response rate was 57·7% and the cooperation rate was 65·8% (of all eligible addresses contacted) with a good representation of ethnicity except a slight underrepresentation of Asian men and women providing in general a nationally representative sample of the British population. Participants were interviewed using a combination of computer assisted face-to-face and self-completion questionnaires (Erens *et al.* 2014). The Natsal-3 study was approved by the Oxfordshire Research Ethics Committee A (reference: 10/H0604/27). Participants provided oral informed consent for interviews. In the current study, we are only using Natsal-3 data for the age range 16-24 years old which is also the age group targeted by the NCSP in England (Woodhall *et al.* 2016). The study website contains details of all the data that are available and ways of their acquisition for further analysis:

http://www.natsal.ac.uk/natsal-3/collaboration-opportunities.aspx

2.2.1. Natsal-3 measures

Depressive symptoms in adolescence

Depression was measured by another validated patient health questionnaire (PHQ-2) at all ages included in this study (16-24 years). The PHQ-2, comprising the first 2 items of the PHQ-9, inquires about the degree to which an individual has experienced depressed mood and anhedonia over the past two weeks scoring each as 0 ("not at all") to 3 ("nearly every day") (Kroenke *et al.* 2003).

Chlamydia tests and sexual behaviour indicators in adolescence and young adulthood

A subset of participants, including all 16–17 year olds (regardless of reported sexual activity) and 18–24 year olds who reported at least one sexual partner by the time of the interview (hereafter termed 'sexually experienced') were invited to provide a urine sample for anonymous STI testing. Participants did not receive their test results. Of all Natsal-3 respondents eligible for the urine study, 57% provided a sample. Urine samples were posted to Public Health England where they were batch-tested for chlamydia using the Aptima Combo 2 assay (Hologic Gen-Probe); positive and equivocal results were confirmed with the Aptima chlamydia monospecific assay (Sonnenberg *et al.* 2013). Similarly as in ALSPAC, Natsal-3 questions about numbers of partners as well as new partners in the past year and whether they used a condom on most recent occasion, were considered again as measures of recent sexual behaviour in the recent study. Additional recent sexual behaviour Natsal-3 measures that we included in our subsequent analyses were: questions about any STI symptoms, number of heterosexual/homosexual partners without a condom last year, any overlap between partners and whether the participants had attended an STI clinic (see Table 1 for common and non-common measures in overlapping ages with ALSPAC).

For Natsal-3, comparisons with British 2011 census data, showed that the survey achieved good representation on various characteristics including ethnicity and general health. Similarly, as with ALSPAC, in all presented models (-except for lack of direct adjustment in the last model displayed in Table 7-), we conditioned on gender and thus MAR was assumed conditionally on this covariate.

2.3. **Final pooled sample & exploratory data analysis:**

Relevant data were drawn from ALSPAC and Natsal-3, as these two studies provide overlapping variables at shared ages (i.e. 16-20 years old and Tables 1 and 2 illustrate such descriptives). With regards to data from the two considered conditions that was not common across the two studies but also included in the final pooled sample, ALSPAC provided depression longitudinal self-reported

symptoms at ages of 10, 12 and 13 years old while Natsal-3 extends the examined time period to the age of 19 up to 24 years old with both depression and Chlamydia cross-sectional data.

The constitution of the final pooled sample that we included in each stage of analyses as described in Section 3, was as follows. With regards to Chlamydia tests and sexual behaviour indicators, 7506 participants contributed cross-sectional data in total: 3,670 were drawn from ALSPAC with age ranges 16-20 years and 3,836 contributed data from Natsal-3 with age ranges 16-24 years into the LCA model for Chlamydia infection as described in Section 3.1. We have subsequently included 8,660 ALSPAC participants with longitudinal data having answered the MFQ scale with age ranging from 10-18 years, and 3,611 Natsal-3 participants having answered the PHQ-2 scale with cross-sectional data and age ranging from 16-24 years. For ALSPAC participants, 19.06 % had complete MFQ data at all ages between 10-18 years old while 70.88% had complete MFQ data within at least three-time points (i.e. examined ages). The data from the 12271 participants in total were used to harmonize and calculate the depression trajectories (Section 3.3). Out of the 7506 participants from both studies who contributed data to the LCA for Chlamydia infection, 3647 ALSPAC and 3603 Natsal-3 participants also had data on depression reported symptoms (Section 3.4).

With regards to the depressive symptoms, we recoded the two PHQ items from Natsal-3 (i.e. 'little pleasure, interest in doing things' and 'feeling down, depressed or hopeless') to have 'logically' equivalent response scales with the two MFQ items from ALSPAC (i.e. 'didn't enjoy anything' and 'teenager felt miserable or unhappy', respectively), as our selected approach requires at least one overlapping item at some time point (i.e. not the same items need to be used repeatedly to link the scales at different occasions, but some of the items still need to be repeated over time) (Griffith *et al.* 2013). Through this process, these specific PHQ and MFQ items were then *harmonised* (i.e. altered to be comparable across studies) and assumed to be equivalent across all the covered ages from the two studies in response options. However, they may still not be truly commensurate because different responses to the PHQ and MFQ items may continue to reflect factors other than actual individual

differences in depression symptoms. In fact, the assumption that all individuals interpret and respond to these PHQ and MFQ items in the same way, is more tenuous in this measurement scenario, since these items were not in fact administered in an identical format across the two studies, enhancing the potential for context effects. Thus, we tested and allowed for measurement invariance by study in these harmonised items whenever this was feasible (for more details about this see section 3.2 and 3.3 below). With regards to measures examined in the current study, the two afore mentioned harmonised MFQ and PHQ-2 items, the chlamydia biological tests as well as four sexual behaviour indicators were overlapping between the two studies at mainly during ages 16-18 as well as during ages 19-20 (with some very few from ALSPAC in the latter age range, n=195).

As mentioned at the begin of this section, descriptive statistical details for the overlapping ages and the common measures in those within the two studies and the pooled sample are outlined in Tables 1 and 2. Table 2 shows percentages and frequencies of individuals for the overlapping ages of 16-20 years old and the corresponding titles and response options of: a) the original concerned overlapping MFQ and PHQ items, b) the harmonised (i.e. recoded) PHQ items to match the 2 MFQ items and c) the final pooled sample of the 2 'common' depression items. The latter (i.e. recoded items) together with the remaining 14 non- common MFQ items are the ones included in the harmonization of the MFQ and PHQ scales as described in section 3.2. For data management/harmonization of variables as well as descriptive results, we used SAS version 9.3 (SAS Institute Inc., Cary, NC). Finally, Figure 1 describes the flow of data and the level of nesting within each of the stages of analysis explained in the next section.

3. Methodology

3.1 LCA for Chlamydia infection

First, we employed a latent class analysis (LCA) (Goodman 1974) to link relevant overlapping and non-overlapping questions as well as the results from the biological Chlamydia tests from the two

studies for Chlamydia infection at ages 16-24. LCA is a classical latent variable model which examines associations between items *J* with dichotomous or polytomous responses observed on each individual *n* that imperfectly measured a latent categorical variable *C* with *K* discrete classes *k = 1, ..., K* where class membership is unknown. In such models, the probability for each item u_j (from a set of *J* observed items) is the product of the conditional probability of u_j, given membership in class *k*, with weights defined by the class probabilities summed over the latent classes. Since the responses or non-responses to the *J* items are assumed conditionally independent given latent class membership, the probability for the vector of a response pattern $\mathbf{u}_n = \left(u_{1n,} u_{2n,\ldots,} u_{Jn}\right)$ is given by

$$P(\mathbf{u}_n) = \sum_{k=1}^{K} P(C = k) \prod_{j=1}^{J} P(u_j \mid C = k) \qquad (1)$$

LCA leads to estimating unconditional probabilities *P*(*C=k*), which are the prevalence of each class in the population (or the size of each latent class). In addition, LCA also summarizes for each response of each item u_j probabilities conditional on class membership and these are called the *item response probabilities.* Both of these sets of probabilities are the model parameters. In our case the considered items 'u' were the biological Chlamydia test and the four common sexual behaviour indicators such as: number of partners, as well as new partners in the past year, number of partners in lifetime and whether condom was used on most recent occasion. There were also some unique Natsal-3 questions about any STI symptoms, number of heterosexual/homosexual partners without a condom last year, any overlap between partners and finally attendance of an STI clinic (see Table 1 for descriptive statistics and available measures) (see also B1. Path Diagram in Supplementary Information for relevant model).

In all models for the latent categorical variable, we assumed two latent classes *C* as 'infected' and 'not infected' with Chlamydia (for further details with regards to which models we compared among those with two classes please see Table A1 in the Supplementary Information). We also fixed

values for the item response probabilities of the urine test as outlined in constrained LCA by Goodman (Goodman 1974). Such a decision was based on our a priori knowledge of the high values of specificity and sensitivity of the Chlamydia biological test (i.e. we assumed that both these probabilities were 1, i.e. that the biological test was the 'gold standard' diagnostic tool). Through this approach, we pre-specified the Chlamydia biological result test to be the latent class when data on this variable were available. By estimating the remaining item response probabilities for the remaining questions, we still allowed for measurement error in the rest of the questions mentioned above, to be taken into account from these models. LCA regression assumes *non-differential measurement* where, within a latent class, observed responses and covariates are independent. To detect possible differential measurement by gender, we tested for lack of conditional independence. The probability for the vector of a response pattern $\mathbf{u}_n$ is given in such a scenario generally by

$$P(\mathbf{u}_n \mid X) = \sum_{k=1}^{K} P(C = k) \prod_{j=1}^{J} P(u_j \mid C = k, X) \qquad (2)$$

, where *X* denotes the number of modelled covariates and in our case it represents the single effect of gender. Since relevant questions were worded and administered in similar ways between the two studies, we have assumed study measurement invariance in these models. This decision was based on the fact that -as we initially explored latent class models within each study separately as recommended by (Curran and Hussong 2009)- such an analysis did imply measurement invariance between the two studies (results not presented). The current approach classified all ALSPAC and Natsal-3 participants into two latent classes (i.e. even those with partially missing data on the Chlamydia biological result or other self-related reported symptoms). To tally the fundamental LCA assumption of conditional independence (i.e. that observed variables are independent within latent classes), items that might have been alternative measures of the same basic construct or might have measured closely related traits, were not allowed to be included in the models simultaneously. Examples of such items follow:

'number of partners in lifetime' and 'number of partners in the past year'; 'number of heterosexual/homosexual partners without a condom last year' and 'use of condom at most recent occasion'; 'number of partners in the past year' and 'number of new partners in the past year'.

In general, as mentioned above, a few competing latent class models were compared in order to determine which indicators best described the data in a parsimonious way and whether measurement invariance hypotheses-described above- were valid-(see Table A1 in the Supplementary Information). We used the Bayesian information criterion (BIC) (Schwarz 1978) and model interpretability to compare different models with different indicators and hypotheses. Parameter estimates from the optimal model provided a basis on which we could describe the prevalence of each latent class. We acknowledge that the current approach assumes that the relationship between the sexual behaviour variables and the latent variable is the same in the two populations studied, and since only one is measured in ALSPAC it is not possible to test this in the data. Allocation to the latent classes will be based on the available response pattern of an individual. Furthermore, previous methodological research has suggested that estimated parameters from these latent variable models of the items are not seriously affected by incorrectly assuming that nonresponse is ignorable, unless the non-ignorability is strong (Kuha *et al.* 2018).

3.2 IRT for obtaining pre-calibration parameters of equating depression scores

We next acquired depression scores that would be ultimately *equated* (i.e. used and interpreted interchangeably/converted to common valid composites) as captured by two different questionnaires for the depression related symptoms from the two studies. As the responses to depression self-reported symptoms involve not only ordinal categories but also repeated measurements, as inherited from the ALSPAC study design, induced dependencies within persons over time and consequently increased computational complexities must be taken into account in the final fitted models. Our solution to this problem was to obtain precalibration parameters from an appropriate IRT model. Those precalibration

parameters were then used as fixed values in longitudinal IRT equating models linked to second order Latent Growth Curve models. Those fixed values (i.e. they are not estimated), were set equal for the same anchor items over time/different ages, an assumption of longitudinal measurement invariance in the Latent Growth Curve second-order IRT models.

A basic IRT model assumes a one-dimensional continuous latent variable θ representing the trait that predicts the probability of a certain response y_j on a particular item j. Item parameters determine the exact relationship between the latent trait and the probability of the response to a particular item (Hambleton *et al.* 1991). IRT linkage of common items, also known as anchor items, can be viewed as a restriction function on the joint parameter space of the questionnaires to be equated (von Davier and von Davier 2007). For our problem we chose to examine this for all the parameters of an IRT model fitted to the data from the two studies of interest here. Following the *calibration sample strategy* as outlined in (Curran *et al.* 2008, Curran *et al.* 2014) we thus formulated a calibration sample in which all possible participant ages were represented from both studies but without initially including any repeated measurements. We randomly selected approximately equal numbers of individuals who had a single observation from the set of all available 16 polytomous MFQ items and recoded (or harmonised) PHQ-2 items for depression from both studies (n=7406; with 51.3 % from ALSPAC and 48.7 % from Natsal-3). For example, if a given individual was assessed a total of five times with MFQ, we randomly selected one of these five MFQ assessments; if another individual was assessed two times by MFQ, we randomly selected one of these two MFQ assessments. Responses were coded 'missing' for the items on the test form that each set of examinees did not take. The numbers of participants distributed across all the ages in the 2 studies and included in the calibration sample are presented in Table 3. We then fitted a graded response IRT model (Bock and Moustaki 2007, Samejima 1969) to the 16 polytomous items of the calibration sample. In a graded response IRT model an item has m_j ordered categories and the examinee is permitted to respond to only one of the categories. For dichotomous response patterns the probability of a response y_{nj} on a particular item *j (j*

= 1. . . J), of a given person, *n (n = 1 . . . N),* under the IRT two-parameter logistic model is defined as

$$P_j(\theta_n) = P(y_{nj} = 1 \mid \theta_n) = \{1 + \exp[-a_j(\theta_n - b_j)]\}^{-1} \qquad (3)$$

where θ_n is the latent trait/score or factor of a person *n*, α_j and b_j are the *discrimination* and *difficulty* parameters respectively for item *j*. The discrimination parameter value of an item indicates how strong the relationship is between the latent trait and the item response variable, and is therefore similar to a factor loading in factor modeling. The so-called difficulty parameter provides information about the general probability of a positive response to a particular item, and is very similar to the threshold parameter in liability models. Because the latent scores θ are estimated conditional on the item parameters for the administered items, the scoring process becomes independent of the particular items in the test.

The item response function for the incorrect response is $Q_j(\theta) = 1 - P_j(\theta)$. Just as there are two item response functions for a dichotomous item it is possible to specify m_j category response functions for each graded response item. However, such category functions do not have a consistent form and Samejima 1969 defined the boundary response function to represent the cumulative probability $P_{jk}^*(\theta)$ of a response category above *k.* Boundary response functions have a consistent form for a given item and may be characterized by a discrimination parameter a_j and m_j-1 difficulty parameters which are ordered. $P_{jk}(\theta)$ is defined in terms of $P_{jk}^*(\theta)$, where $P_{jk}^*(\theta)$ represents the cumulative probability of a response above category *k* as follows:

$$P_{j1}(\theta) = 1 - P_{j1}^{*}(\theta); \qquad \text{when } 1 < k < m_j \qquad (4)$$

$$P_{jk}(\theta) = P_{j(k-1)}^{*}(\theta) - P_{jk}^{*}(\theta) \qquad (5)$$

$$P_{jm_j}(\theta) = P_{j(m_j-1)}{}^{*}(\theta) \qquad \text{when } k = m_j \qquad (6)$$

The logistic form of the boundary response function is given by

$$P_{jk}^{*}(\theta) = \{1 + \exp[-\alpha_j(\theta_n - b_{jk})]\}^{-1} \qquad (7)$$

An important assumption of this form of the graded response model is that the reasoning process is homogeneous throughout the set of response categories for an item. Samejima 1969 interpreted this to mean that a_j is constant for all categories in equation 7 and such an assumption is only specific to the item. Within the graded response IRT model, we also allowed the latent variable mean to vary by study, continuous age (including linear and quadratic trends) and gender. The effects of age across the continuum were moderated by study membership (see *AgexStudy* effect under the Heading *Factor mean/Covariate effect* of Table 4). For our DIF test (i.e. when an item is suspected to perform differently in the groups) we allowed a study effect on the one harmonised PHQ item (2. *Young person has felt unhappy/miserable in the last two weeks*). In other words, we allowed the variable study to have direct influence on the item in question, thereby making its threshold (-*difficulty* in the IRT framework-) parameter be different in the two studies. We did not allow for DIF by age because then the necessity of the hypothesis of longitudinal measurement invariance in the next considered models would be violated introducing modelling and interpretation challenges of the depression equated scores such as: was the change over time not solely a change of level of depression but also a change in the nature of depression. We compared BIC with and without the covariates on latent variable mean as well as the one harmonised item (-the latter comparison gave evidence of DIF), to develop one final measurement model (see B2. Path Diagram in Supplementary Information for relevant model).

3.3 Longitudinal IRT models and second order Latent Growth Curve models for equating depression scores as well as expressing their growth trajectories

As mentioned in the previous section, *equating* is a process that permits the comparison of scores obtained from different questionnaires. IRT equating methods provide a linear transformation of person and item parameters, and the coefficients of this function are called *equating coefficients* (Battauz 2015). Fixing those equating coefficients to known values as estimated from model described in equations 4-7 and beyond (i.e. *anchored calibration*), allows us to fit a single model that combines longitudinal (i.e. multidimensional) IRT ordinal measurement models at the first level with Latent Growth Curve modelling at the second level.

Within the first level, the boundary response function of category *k* is given by

$$P_{jk}^{*}(\theta_{nt}) = \{1 + \exp[-\alpha_{jt}(\theta_{nt} - b_{jkt})]\}^{-1} \qquad (8)$$

Where θ_{nt} is the latent trait of person *n* at time (age) *t* and α_{jt} and b_{jkt} denote the discrimination and difficulty parameters for item *j* and time (age) *t* (i.e. corresponding loadings and thresholds in the structural equation modelling (SEM) framework). We performed longitudinal IRT equating by assuming *longitudinal measurement invariance*; we achieved this by fixing the parameters of the same anchor items to be equal over time/different ages. Only for the harmonised item 2. *Young person has felt unhappy/miserable in the last two weeks'* did we allow different parameters to be estimated freely for the age interval 19-24 years old. Through this approach we attempted to reflect study measurement invariance, as assigned during the previous analysis stage- although within this specific age interval we also assumed longitudinal measurement invariance (i.e. the parameters were the same within the 19-24 years old age group, but different from the considered younger age group 10-18). Modal a posteriori or empirical Bayes modal (Rabe-Hesketh and Skrondal 2004) estimates of the latent traits

of depression θ_{nt} were obtained through these longitudinal IRT equating models. More precisely, this approach yielded depression IRT factor scores for each person at each examined age (not just those of the calibration sample and even if all the participants of the pooled sample did not have originally available data at the examined ages). The computation of these factor scores entailed the combination of the likelihood function of the given item parameters as computed in Model of equations (4) – (7) and beyond and the vector of observed item responses for an ALSPAC or Natsal-3 participant with a prior distribution of the latent depression trait (this is assumed as a multivariate normal in the IRT framework) to estimate the posterior distribution of the latent traits of depression (i.e. the predicted factor scores).

Within the second level, we estimated a growth trajectory on the latent traits of depression θ_{nt} - which means that the latter were modelled as dependent variables of one or more growth factors; that is, intercept growth factor and slope growth factor(s) (Kohli and Harring 2013). Essentially, we specified a piecewise linear model (Bollen and Curran 2006) which allowed two separate slopes to be fitted to repeated observations - occurring before and after the age of 16 years old- based on a number of factors (see. B3. Path Diagram in Supplementary Information for relevant model). First and foremost, 16 years old- is the age where Chlamydia infection information started to be available in both studies. In order to examine predictive relationships of depression to Chlamydia, it would not thus be possible to assume an alternative age as a knot point for the nature of our specific problem with the data we have available. Furthermore, the age of 16 years old may represent a critical time period for heightened vulnerability to depression and thus selection of this age is not an unreasonable assumption for the developmental changes of depression we are aiming to study (Hankin *et al.* 1998). In a previous study (Edwards *et al.* 2014) using ALSPAC depression MFQ scores derived from summing all items at each age, highlighted higher age differences at the age of 16. Such a fact also renders the selected knot point a reasonable modelling approach for the relationship we are aiming to explore.

This growth trajectory is

$$\theta_{nt} = \pi_{0n} + \pi_{1n} t_{ng} \delta_{ng} + \pi_{2n} t_{ng} (1 - \delta_{ng}) + \varepsilon_{ng} \qquad (9)$$

, where π_{0n}, π_{1n} and π_{2n} are individual intercept and slope parameters (i.e. latent factors or growth factors). Those are assumed to deviate from the average intercept β_0, and slopes β_1 and β_2 via the following relations: $\pi_{0n} = \beta_0 + u_{0n}$, $\pi_{1n} = \beta_1 + u_{1n}$ and $\pi_{2n} = \beta_2 + u_{2n}$ where *u*'s stand for disturbance terms from a normal distribution with a zero mean vector and a covariance matrix **Φ**. Furthermore, in formula (9), t_{ng} denotes the time of measurement *g* for person *n* relative to 16 years old, δ_{ng} is an indicator which equals to 1 for the ages before 16 years old and 0 otherwise and finally ε_{ng} are residual terms. The residual term ε_{ng} is assumed to be normally distributed with a zero mean vector and a covariance matrix $\mathbf{\Psi_n}$. The subscript *n* in $\mathbf{\Psi_n}$ indicates that the covariance matrix is subject-dependent and, thus, allows for missing data. The matrix $\mathbf{\Psi_n}$ is presumed to be diagonal and the residuals are independent between measurements with constant variance across time.

Because we assume the anchor item parameters are given, no additional constraints are needed to ensure model identifiability. To conclude, the estimated parameters included a) IRT parameters α and *b* only for the response categories of the harmonised item 2.*Young person has felt unhappy/miserable in the last two weeks'* for the age interval 19-24 years old, as explained above b) mean intercept β_0, and slopes, β_1 and β_2, (latent factors or growth factors), the variances and covariances of the intercept and the slopes as well as the residual variance. Such a model had the benefit of analysing growth using latent constructs which are disattenuated from measurement error that would be present when analysing only one of the repeated manifest scale values or even some aggregate across scales. Despite the fact that we had a number of common items shared among adjacent time points/ages and the local item dependence might be violated, we were not able to incorporate additional specific latent factors which would account for residual dependence above and beyond the primary

due to convergence difficulties (Cai 2010). In other words, fitting the complete model with 15 nuisance factors (i.e. 15 is the number of common items) proved to be computationally very challenging and thus this issue was not pursued further.

3.4 LCA with covariates; associations between depression growth factors and Chlamydia infection in adolescence and young adulthood

Having built the measurement or classification model (the LCA as described in section 3.1) for the pooling of the Chlamydia infection data, we then related the class membership (i.e. infected/non infected with Chlamydia) to explanatory/external variables (i.e. growth factors of the trajectories of depression as derived in equation (9) as well as the study effect). In other words, we included auxiliary information into the final mixture model as derived in equation 2 in the form of covariates (also called predictors or independent variables). Recent methodological work has provided a framework for avoiding the measurement parameter shift problem; namely, the *three-step method* for estimating the effects of covariates and distal outcomes in mixture models (Asparouhov and Muthén 2014, Vermunt 2010). This can be viewed as a bias-adjusted three-step method due to the inclusion of information on the measurement error inherent in the allocated class variable following modal-assignment. Under modal assignment, individuals are predicted to be in the latent class for which they have the highest posterior class membership probability (i.e. the probability of membership in class k given an observed response pattern $\boldsymbol{u}$). The three-step method is a sequential-step method which fixes the measurement parameters of the latent class model with covariates at values from the unconditional latent class model. Three sequential modelling stages are involved: (a) estimating the unconditional mixture model (as described in equation (9)), (b) assigning individuals to latent classes using modal class assignment (in Mplus this variable is automatically created using the SAVEDATA command with the option SAVE=CPROB-as described in section 3.1) and (c) estimating a mixture model with measurement parameters that are fixed at values that account for the measurement error in the class assignment given in Table A2 in Supplementary Information (i.e. current analysis). Once the model in

the third stage is specified, auxiliary information is included in this model in the traditional regression modelling fashion (see B4. Path Diagram in Supplementary Information for relevant model). Nevertheless, in some situations when the auxiliary variables are included, in the final stage of the three-step approach the latent class variable can shift substantially and still invalidate the results. Another approach based on the work of (Bolck *et al.* 2004), called the BCH method might avoid shifts in latent class in the final stage to which the three-step method is susceptible. In its final stage the BCH method uses a weighted multiple group analysis, where the groups correspond to the latent classes, and thus the class shift is not possible because the classes are known. The BCH method uses weights w_{ik} which reflect the measurement error of the latent class variable. In the estimation of the auxiliary model, the i-th observation in class/group k is assigned a weight of w_{ik} and the auxiliary model is estimated as a multiple group model using these weights. However, the main drawback of the BCH method is that it is based on weighting the observations with weights that can take negative values (Asparouhov and Muthén 2014).

For our problem, we attempted to implement the BCH method in MPlus and estimate the effect of the latent class variable on the covariates of interest via two runs. In the first run we estimated the latent class measurement model as described in section 3.2 and saved the BCH weights. Unfortunately, for the second class those weights were negative. Thus, we did not proceed with a second run where we would include auxiliary information in the form of covariates (also called predictors or independent variables) into the final mixture model, using those BCH weights. Instead we based the remaining of our analysis and conclusions solely on the results of the three-step method and the sensitivity analysis as outlined below.

3.5 Sensitivity Analyses

To evaluate the quality of obtaining precalibration parameters from the selected IRT model as described in Section 3.2 we cross-validated those with a new randomly selected calibration sample from the 2 studies. For further details see Supplementary Information, Section A3.1.

Subsequently, to test the validity of the association between trajectories before age 16 and Chlamydia infection, from the harmonization effort among the two studies, we explored this relationship with ALSPAC data only. For further details see Supplementary Information, Section A3.2.

4. Results

4.1 LCA for Chlamydia infection

LCA Model for Chlamydia infection yielded item response probabilities and the plot presented in Figure 2 representing these parameters, was initially used to interpret, label and validate the two assumed latent classes (in this model no measurement invariance on any of the questions was tested). For latent class 1, we fixed the first threshold of the biological Chlamydia test to be 15; this yield a probability of 1 for observed negatives and thus through this way we fixed the specificity of this test to be 1 or 100 % (-in the plot it is shown the complementary probability result of 0 within latent class 1 which we labelled as the 'Uninfected'-). Similarly, for latent class 2, we fixed the first threshold of the biological Chlamydia test to be -15 which yield a probability of 0 for observed negatives and its complementary probability of 1 for observed positives; through this approach we fixed the sensitivity of this test to be 1 or 100 % and we thus labelled this latent class as the 'Infected'. For the latter (i.e. 'Infected' latent class), as expected, all the items and their response probabilities which corresponded on risky sexual behaviour questions, were much higher than in the 'Uninfected' class (see Figure 2 and Model 1 in Supplementary Information). Subsequently, BIC for partial measurement invariance tests indicated that the item response probabilities for the questions with regards to 'number of partners without a condom in the last year' and any 'overlap between partners' to be slightly higher for females rather than males in the 'Infected' latent class (Model 2 in Supplementary Information). The results of this latter model were later used in Step 4. The 'Infected' class from this Model consisted of 1.72 % (129/7506) of the pooled sample.

4.2 IRT for obtaining the precalibration parameters for equating of depression scores

Table 4 contains estimates from the graded response IRT model to the calibration sample. The standardized factor loadings show a strong relationship with the hypothesized latent trait of depression for all fifteen items except for item 4. *Young person has felt very restless in the last two weeks* which is estimated to have the lowest standardized loading. Specifically, positively worded items (i.e., items 14-16), showed lower factor loadings (i.e., discriminated best at lower levels of the depression trait). Two thresholds were estimated since there were three responses and if divided by unstandardized factor loadings, one gets so-called IRT difficulty parameters (where higher values mean more difficult items). As expected, difficulties varied in all 16 items. Factor means of depression varied significantly by study, continuous age (including linear-non significant- and quadratic –significant- trends) and gender. The effects of age across the continuum were moderated by study membership (see *AgexStudy* effect under the Heading *Factor mean/Covariate effect* of Table 4). To help with interpretation, Figures 3 and 4 display model-implied conditional mean plots across age and stratified by covariates for the depression factor from the calibration cross-sectional ALSPAC and Natsal-3 data. Figure 3 presents the model implied mean IRT depression across age conditioned on study membership in the calibration sample. ALSPAC participants had significantly lower mean levels of IRT depression scores compared to Natsal-3 participants across overlapping ages. Figure 4 presents the model implied mean IRT depression scores across age conditioned on gender in the calibration sample. Males had slightly higher levels of depression up to age 16 while females had slightly higher levels of depression after this age. We also allowed for DIF for the harmonised item 2. *Young person has felt unhappy/miserable in the last two weeks* by the study effect-which we consider critical in order to account for the complex nature of this pooled analysis, permitting us to test questions about study comparability. More specifically, since this effect was significant, we allowed for different thresholds of the harmonised item between studies.

4.3 Latent Growth Curve second-order IRT model of depression scores

Table 5 shows estimates from the second-order Latent Growth Curve model augmented by the longitudinal IRT equating model. We can see that the mean IRT depression score at age 16 was estimated to be 7.704. The first and second slopes were estimated as 1.056 and 0.308 respectively and such results indicate that on average there were significant increases of IRT depression scores both between the ages of 10-16 as well as the ages of 16-24 years old. To test for potential gender and study differences in initial levels and rates of change of the IRT depression scores over time, we also included interaction terms of the growth factors (i.e. the intercept and the two slopes) and these two covariate effects. Such additions led to significant improvement of model fit as judged by BIC (Schwarz 1978), thus indicating that gender and study differences would need to be incorporated in the final model. Females at 16 years old reported on average lower levels of depression by 0.910 compared with males at 16 years old. The slope of the first linear piece was also lower for females compared with males (meaning on average smaller changes for females than males during the ages of 10-16 years old) while the slope of the second linear piece was higher for females compared with males (meaning on average greater changes for females than males during the ages of 16-24 years old). Natsal-3 participants yielded on average significantly lower levels of IRT depression scores compared to ALSPAC participants at 16 years old. Although IRT depression scores were available for all 12271 participants, we have not allowed for different slopes of the first linear piece between studies since initially we had no data for PHQ-2 items for that age period for Natsal-3. The slope of the second linear piece was lower for Natsal-3 compared to ALSPAC participants (which means on average lower changes for scaled depression scores of Natsal-3 compared to ALSPAC during the ages 16-24 years old). The fixed effects only reflect the average trajectory pooling over all individuals, and the random effects reflect individual variability around these mean values. In Table 6, variance components for intercept and two slopes show that there is substantial individual variability around all components of the depression trajectory. Positive and negative covariances mean patterns of fanning out and in respectively: for

instance, those values representing associations between the estimates of the intercept and slope 1 of IRT depression scores indicated that individuals with high age 16 scaled depression scores increased more (or decreased less) from ages 10-16 years than individuals with lower age 16 scores [estimate (est)=1.182, p-value<0.001]. In a similar way, the negative covariance in this same model between the intercept and slope 2 would imply that on average individuals with high age 16 depression scores increased less (or decreased more) from ages 16-24 years than individuals with lower age 16 scores (est=-0.140, p-value<0.001). Covariance between changes from ages 10-16 and ages 16-24 was also negative which suggests that changes during these ages are negatively associated (est =-0.077, p-value<0.001). Finally, we conducted a series of outlier detection analysis using graphical representations of the scores as a function of specific person covariates (Figures 5 and 6). Such graphical analyses indicated that most of observations were not potentially particularly aberrant or extremely outlying.

4.4 LCA with covariates; associations between Chlamydia infection and depression trajectories

Table 7 shows the covariate results for the LCA model initially estimated as described in Section 3.1, regressed on the growth factors of IRT equated scores as well as a study dummy covariate. Those falling in the latent class 'infected with Chlamydia' were significantly more likely to have experienced larger changes in depression between the ages 10-16 (OR=4.007, p-value =0.045). Depression at 16 or changes in depression between the ages 16-24 were not significantly associated with Chlamydia infection. Those in the latent class of Chlamydia infection were more likely to be Natsal-3 participants; in other words, higher Chlamydia infection was estimated in the Natsal-3 participants than the ALPSAC ones.

4.5 Sensitivity analyses

There were not big differences in the displayed coefficients from the same IRT model (as described in Section 3.2) fitted to different pre-calibration samples (compare relevant estimates in Table 4 and Table A.4 in Supplementary Information). Such coefficients would then be fixed for common item parameters in the longitudinal IRT equating models linked to second order Latent Growth Curve models (as described in Section 3.3). Thus, the harmonized depression scores deriving from these longitudinal IRT equating models, would not change dramatically either.

A positive association between the changes of MFQ scores during 10-16 years and increased odds to being infected with Chlamydia as determined by the urine test in the ALSPAC data only (n=2776) was found at the approximate age of 17 years old (OR=2.054, $p<0.001$, Table A.5 in Supplementary Information).

5. Discussion

In the current paper we have established comprehensive latent variable models that incorporate measurement information about study and group membership in terms of demographic characteristics such as age and gender for the pooled analysis of the association between depression and current Chlamydia infection in adolescence and young adulthood. We combined for the first time two very rich datasets in terms of sample sizes and wealth of relevant variables from the UK. More precisely, we have initially identified potential item pools between ALSPAC and Natsal-3 studies allowing for some broader and rigorous tests for the psychometric properties of the key theoretical studied constructs of depression and current Chlamydia infection in an aggregated sample of 12271 participants for the depressive symptoms and a subsample of those from 7506 participants with regards to biological Chlamydia tests and sexual behaviour indicators.

To our knowledge, this methodological application is novel within the pooled analysis framework, providing the means to a cumulative style to scientific enquiry, comparing results across the two examined studies here, through the direct analysis of their primary individual participant data building on the pioneering methodological work of some relatively recent studies (Curran and Hussong 2009, Curran, *et al.* 2008, Curran, *et al.* 2014, Flora *et al.* 2008, Hussong *et al.* 2013, Hussong *et al.* 2008, McArdle *et al.* 2009). We demonstrated empirically for the first time that latent class modelling can also be a useful measurement model in pooled analysis when the desired harmonised construct is of categorical nature. Having pre-fixed the item response probabilities of the biological Chlamydia test, we ensured that the latent classes of infected and not infected were perfectly measured- that is observed for those with available data on the biological Chlamydia test. For the participants who had missing data on the biological Chlamydia test, the final LCA model still assigned them to the latent classes of infected and not infected by maximizing the likelihood function of the incomplete observed data. This approach resulted in classifying 129 individuals with very good class separation (entropy = 0.956- entropy with values approaching 1 indicate clear delineation of classes (Celeux and Soromenho 1996) as infected with Chlamydia - including the 82 individuals originally coming from the two studies- increasing the power of subsequent analysis. We believe that in our problem we do not have large uncertainty about the LCA estimates since in addition to a good entropy, the sample size was also large in the fitted model. Subsequently, by fitting an IRT graded response model to a calibration sample we allowed the relevant MFQ and harmonised PHQ-2 items to be differently related to the underlying construct of depression - including DIF by study on one item. This is a substantial improvement over other potential methods used for scoring such as proportion, sum or z- score (Curran, *et al.* 2008, Gorter *et al.* 2016, Gorter *et al.* 2015, Griffith, *et al.* 2013, Gross *et al.* 2015). Next, IRT longitudinal equating through incorporation of known item parameters for the harmonization and scoring of longitudinal and cross-sectional ordinal data together with latent growth models such as a piecewise linear model, were demonstrated that they can be implemented simultaneously. The fact that we did

not have to separately estimate IRT depression scores for different ages and then incorporate them in latent growth models, minimizes the within-person dependence created by ignoring subjects in a two-stage approach, allowing a possible gain in statistical efficiency from using this simultaneous approach (McArdle *et al.* 2009). One additional benefit of piecewise linear models is that they allow simultaneous evaluation of change at different developmental stages –here, depression during adolescence and young adulthood- and the influence of covariates may vary in the different periods - such as gender as well as study effects on the growth factors of the depression trajectory in the current study (Koukounari *et al.* 2017, Li *et al.* 2001). Exploration of the trajectories using only the ALSPAC data examined, could be one way of identifying the fixed transition point, yet in our problem we choose to determine this theoretically as indicated in (Bollen and Curran 2006). Also, in order to examine predictive relationships of depression to Chlamydia, it would not be possible to assume an alternative age as a knot point for the nature of our specific problem with the data we have available. Only if we had longitudinal data on Chlamydia infection for instance in the ALSPAC study, we could then have tested through fit criteria different piecewise linear models placing the transition point at different ages. Finally, we were also interested in investigating whether the latent classes of Chlamydia infection differed with respect to the mean of the growth factors of the trajectories of depression. To materialize this aim we used a sequential-step method (Asparouhov and Muthén 2014, Vermunt 2010) in order to estimate a more advanced secondary model that included a latent class variable. Through this process, we were more in line with our measurement wishes-that the MFQ and PHQ-2 items contributing to the IRT depression scores as well as their growth factors summarizing their trajectories, would not influence the measurement of the latent class (i.e. current Chlamydia infection). In addition, a more easily computed complex SEM including growth factors from a piecewise linear model regressed on a latent class variable was enabled. Nevertheless, it should be noted here that we have not included directly gender as a moderator in the association between depression and Chlamydia despite the existence of biological differences between males and females that tend to manifest with puberty and

that could affect biological susceptibility for this association. The reasons for such a decision were that we had included gender as a moderator for the trajectories of depression and thus such differences were taken into account in the intercept and slope of the corresponding model. We had also included gender in the LCA, testing if this covariate might influence the measurement process directly.

In general, the strengths of our approach are amplified by the fact that by pooling two studies, we were able to cover the broader developmental period of 10-24 years old for depression and 16-24 years for Chlamydia infection. This would not have been possible if the analysis was conducted separately for each of the two studies. We hope that the provided computer codes in commercial software such as Mplus add to the transparency of our selected approaches and can motivate and help other researchers in similar and broader applications of pooled analysis.

In terms of substantive findings, IRT depression scores were lower at the age of 16 years old, with lower changes of those scores during the ages 16-24 years for Natsal-3 participants compared to ALSPAC ones. For females at 16 years as well during ages of 10-16 years old, on average, smaller changes of IRT depression scores were estimated compared to males. During the ages of 16-24 years old, on average, greater changes of IRT depression scores were estimated for females compared to males. The gender differences in the depression trajectories agrees with reviews of relevant studies' epidemiological findings - indicating higher rates of depression in females detected at mid-puberty through adult life, as opposed to a male preponderance until early adolescence (Edwards, *et al.* 2014 , Piccinelli and Wilkinson 2000 , Weller *et al.* 2006). Chlamydia infection was estimated to be higher in Natsal-3 than in ALPSAC. Those currently infected with Chlamydia (at an age between 16 and 24) were significantly more likely to have experienced larger changes in depression between the ages 10 and 16. However depression at 16 and changes in depression between the ages 16-24 were not significantly associated with current Chlamydia infection. Moreover, to enhance the validity of our finding for an association between trajectories before age 16 and Chlamydia infection, from this harmonization effort among the two studies, we provided in Section 4.5 as well as the Supplementary

Information (Section A3.2) results having explored this relationship with ALSPAC data only. This approach still yielded a positive association between the changes of MFQ scores during 10-16 years and an increased likelihood to being infected with Chlamydia at the approximate age of 17 years old (OR=2.054). These observations support and enhance prior findings that depression appears to predict sexually transmitted infections among adolescents and young adults (Hallfors *et al.* 2005, Khan, *et al.* 2009, Nuttbrock *et al.* 2013 , Shrier, *et al.* 2002). Possible mechanisms that might explain how changes in depression between 10-16 years can lead to Chlamydia infection in late adolescence/young adulthood could involve sexual risk taking and/or inflammation markers. Causal mediation modelling (Imai *et al.* 2010) to accurately explore such mechanisms within the ALSPAC data merits further investigation. Due to the cross-sectional nature of the Natsal-3 data, one cannot directly assess whether the association between Chlamydia and depression between the two studies differs, but analysis of ALSPAC data alone, as mentioned before, did broadly support the findings. Further discussion of substantive conclusions regarding the association between depression and current Chlamydia infection in adolescence and young adulthood as well as differences in findings between Natsal-3 and ALSPAC studies was not an aim of the current paper.

Our findings need to be considered in light of some limitations. The current selected latent variable modeling approach makes strong assumptions about each involved question/item contributing to the constructs of current Chlamydia infection and depression and presumes that these assumptions- where measurement invariance was assumed or tested and established-hold across populations. We try to improve the fit of these models by testing and relaxing such assumptions on some of the parameters of these models through checking information criteria. With strong data, these assumptions are often unnecessary and in case of constructs composed of many questions, violations of these assumptions may not have critical consequences or inferences about outcome scores. With weak, poorly connected data, as we encountered in this study the assumptions of these models are much needed as they act as a substitute for the data. Our approach assumes that multivariate associations among the available

variables of one study can be extrapolated to the second study which by design has no data on one or more variables. We also assume that data are missing at random within studies. Both these assumptions are untestable.

Incorporating the available weights for urine non response would have perhaps provided a more principled approach for the MAR assumption in Natsal-3. As the models used for the harmonization process were complicated enough and the sensitivity analysis by including only ALSPAC data provided similar results, we decided not to use these weights further in our analysis but we do recommend further research of exploration of inclusion of sampling and non-response weights in similar analyses. Furthermore, the MAR assumption might have been undermined by the fact that we only conditioned upon gender in the involved harmonization models. Within ALSPAC for instance, refusing to provide a urine sample was also associated with lower educational attainment at age 16 and measures of family socioeconomic disadvantage (Crichton, *et al.* 2014). Various observed variables were also associated with refusing to give a urine sample in Natsal with those being different among men and women (Erens *et al.* 2013). More precisely, for men these variables included: ethnicity, highest educational qualification, marital status, number of opposite- or same-sex partners without a condom in the last year, same-sex experience (ever), attended a sexual health clinic (ever), overlap in partners in the last 5 years, injected non-prescribed drugs (ever). For women variables associated with urine response were: region, total number of lifetime opposite- and same-sex partners, heterosexual anal sex (last 5 years), blood test for HIV (ever), same sex experience (ever), other people were present in the household during the Natsal-3 interview, children aged 6-15 were present or passing through during the interview. Developing and interpreting a model for chlamydia and depression trajectory separately for young men and young women conditioning on so many different variables, some of which are behavioural, would be complex but could be an area for further work as it would make the MAR assumption more plausible.

Moreover, the obtainment of reliable equatings and the calculation of valid scale scores through the proposed process for pooled analysis does not depend only on overlap of ages between the considered studies but also on measurement overlap (i.e. common items/questions) for the considered constructs. For instance in the current paper, the final equated coefficients that were implemented to the MFQ and PHQ-2 questions in model described in Section 3.3 were tested through fit criteria and limited DIF tests as described in Section 3.2 without having tested the assumed IRT model initially in each single study as recommended by (Curran and Hussong 2009). This was due to the fact that the Natsal-3 study had only 2 items available from PHQ-2 questionnaire for depression. Having had a larger number of common items would have allowed us to at least first fit IRT models separately within each study to check about potential fluctuations of relevant coefficients for similar items and such a fact eventually could perhaps attenuate the effect of potential fluctuations on a subset of item parameters during the equating process (Battauz 2015). Nevertheless, sensitivity analysis involving randomly drawing a new calibration sample and fitting to it an identical IRT model as the one described in Section 3.2, suggested no great fluctuations in the pre-calibration parameters as obtained from this model. Such a result strengthens the validity of our findings from the harmonized depression scores deriving from the longitudinal IRT equating models (Supplementary Information Section A3.1).

The possibility of conducting a *bridging study* (Hussong *et al*. 2013) is recommended as another avenue for future research in order to validate and improve the current harmonization process with regards to depression scores (as previously described in sections 3.2 and 3.3). The basic idea would be to embark on a new primary data collection for the express purpose of linking together the MFQ and PHQ-2 measures used in ALSPAC and Natsal-3 studies originally. Pooling the data from the two original studies and the new bridging study, one would then have the opportunity to construct an improved commensurate measure of depression. A bridging study would involve recruiting new participants, ideally from a similar population as that sampled in the contributing studies intended for the pooled analysis described in this paper. Not all of the original 16 MFQ items would have to be

administered; a subset of the MFQ items would be sufficient (reducing participant fatigue and eliminating redundancies between items). We propose this new primary data collection to consider the PHQ-9 (Kroenke, *et al.* 2003, Kroenke *et al.* 2001) instead of the PHQ-2 and nine similarly worded questions from the long version (i.e. 33 questions) of the MFQ child version administered to adolescents and young adults perhaps within a UK sexual health clinic- to ensure robustness to departures from current selected models assumptions. Further methodological work such as a simulation study adjusted to the scenario of equating MFQ and PHQ questionnaires, would be necessary too in order to determine the most optimal required sample size in this new bridging study. Finally, once ALSPAC releases additional MFQ data up to the age of 25 years old, such data should be also included to improve the presented models.

To allow a simpler and clearer demonstration of our methodology we did not use information available in Natsal-3 on treatment for depression or diagnosis of chlamydia or other STIs. Using Natsal-3 data over a wider age range an analysis found little association between treatment for depression and STI diagnosis (Field *et al.* 2016). The relationship between depression and chlamydia infection would be better understood based on cohort studies with repeated testing for chlamydia infection and questions concerning chlamydia testing, diagnosis and treatment as well repeated measurement of depression and questions concerning diagnosis and treatment of depression. Such studies would in particular allow greater examination of the change in mental health following chlamydia infection or diagnosis than was permitted in this work. Despite these limitations, we believe that the current findings set a worthy firm basis for future research on specific pathways to good physical and mental health outcomes from middle childhood to early adulthood.

Acknowledgments

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory

technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The UK Medical Research Council and Wellcome (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC.

Natsal-3 is a collaboration between University College London (London, UK), the London School of Hygiene and Tropical Medicine (London, UK), NatCen Social Research, Public Health England (formerly the Health Protection Agency), and the University of Manchester (Manchester, United Kingdom). We also thank the Natsal-3 study participants, the team of interviewers from NatCen Social Research, operations, and computing staff from NatCen Social Research. Natsal-3 was supported by grants from the Medical Research Council [G0701757]; and the Wellcome Trust [084840]; with contributions from the Economic and Social Research Council and Department of Health. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the article; and decision to submit the article for publication. This publication is the work of the authors and Artemis Koukounari (AK), Andrew J. Copas (AJC) and Andrew Pickles (AP) will serve as guarantors for the contents of this paper. This research was specifically funded by a grant funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley National Health System (NHS) Foundation Trust and King's College London.

AK is grateful to Drs Peter White and Nigel Field as well as Professor Pam Sonnenberg for the completion of this work. The manuscript benefitted greatly from comments provided by three referees.

References

Angold, A., Costello, E. J., Messer, S. C., Pickles, A., Winder, F. and Silver, D. (1995) The development of the short questionnaire for use in epidemiological studies of depression in children and adolescents. *Int J Methods Psychiatr Res,***5** 237-249.

Asparouhov, T. and Muthén, B. O. (2014) Auxilliary Variables in Mixture Modeling: Three-Step Approaches Using Mplus. *Struct Equ Modeling,* **21** 329-341.

Battauz, M. (2015) Factors affecting the variability of IRT equating coefficients. *Stat Neerl,***69** 85–101.

Bock, R. D. and Moustaki, I. (2007) Item response theory in a general framework In *Handbook of Statistics, Psychometrics* eds C. R. Rao and S. Sinharay), pp. 475-476, North-Holland, Amsterdam, The Netherlands.

Bolck, A., Croon, M. and Hagenaars, J. (2004) Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Anal,***12** 3–27.

Bollen, K. A. and Curran, P. J. (2006) *Latent Curve Models. A structural Equation Perspective. An Introduction to Latent Variable Growth Curve Modeling. Concepts, Issues and Applications,* New Jersey: Psychology Press Taylor and Francis Group.

Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., Molloy, L., Ness, A., Ring, S. and Smith, G. D. (2013) Cohort Profile: The 'Children of the 90s'-the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol.,***42** 111-127.

Cai, L. (2010) High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika,***75** 33-57.

Celeux, G. and Soromenho, G. (1996) An entropy criterion for assessing the number of clusters in a mixture model. *J. Classification,***13** 195-212.

Crichton, J., Hickman, M., Campbell, R., Heron, J., Horner, P. and Macleod, J. (2014) Prevalence of Chlamydia in Young Adulthood and Association with Life Course Socioeconomic Position: Birth Cohort Study. *Plos One*,**9** e104943.

Curran, P. J. and Hussong, A. M. (2009) Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychol Methods*,**14** 81-100.

Curran, P. J., Hussong, A. M., Cai, L., Huang, W., Chassin, L., Sher, K. J. and Zucker, R. A. (2008) Pooling data from multiple longitudinal studies: the role of item response theory in integrative data analysis. *Dev Psychol*,**44** 365-380.

Curran, P. J., McGinley, J. S., Bauer, D. J., Hussong, A. M., Burns, A., Chassin, L., Sher, K. and Zucker, R. (2014) A Moderated Nonlinear Factor Model for the Development of Commensurate Measures in Integrative Data Analysis. *Multivariate Behav Res*,**49** 214-231.

Doyle, C., Swain, W. A., Ewald, H. A., Cook, C. L. and Ewald, P. W. (2015) Sexually Transmitted Pathogens, Depression, and Other Manifestations Associated with Premenstrual Syndrome. *Hum Nat*,**26** 277-291.

Edwards, A. C., Joinson, C., Dick, D. M., Kendler, K. S., Macleod, J., Munafò, M., Hickman, M., Lewis, G. and Heron, J. (2014) The association between depressive symptoms from early to late adolescence and later use and harmful use of alcohol. *Eur Child Adolesc Psychiatry*,**23** 1219-1230.

Erens, B., Phelps, A., Clifton, S., Hussey, D., Mercer, C. H., Tanton, C., Sonnenberg, P., Macdowall, W., Copas, A. J., Field, N., Mitchell, K., Datta, J., Hawkins, V., Ison, C., Beddows, S., Soldan, K., Coelho da Silva, F., Alexander, S., Wellings, K. and Johnson, A. M. (2013) National Survey of Sexual Attitudes and Lifestyles 3 Technical Report Volume 1: Methodology.

Erens, B., Phelps, A., Clifton, S., Mercer, C. H., Tanton, C., Hussey, D., Sonnenberg, P., Macdowall, W., Field, N., Datta, J., Mitchell, K., Copas, A. J., Wellings, K. and Johnson, A.

M. (2014) Methodology of the third British National Survey of Sexual Attitudes and Lifestyles (Natsal-3). *Sex Transm Infect*,**90** 84-89.

Field, N., Prah, P., Mercer, C. H., Rait, G., King, M., Cassell, J. A., Tanton, C., L., H., Mitchell, K. R., Clifton, S., Datta, J., Wellings, K., Johnson, A. M. and Sonnenberg, P. (2016) Are depression and poor sexual health neglected comorbidities? Evidence from a population sample. *BMJ Open*,**6** e010521.

Flora, D. B., Curran, P. J., Hussong, A. M. and Edwards, M. C. (2008) Incorporating Measurement Non-Equivalence in a Cross-Study Latent Growth Curve Analysis. *Struct Equ Modeling*,**15** 676-704.

Goodman, L. A. (1974) Exploratory Latent Structure-Analysis Using Both Identifiable and Unidentifiable Models. *Biometrika*,**61** 215-231.

Gorter, R., Fox, J. P., Apeldoorn, A. and Twisk, J. W. (2016) Measurement model choice influenced randomized controlled trial results. *J Clin Epidemiol* pii: S0895-4356(0816)30192-30195.

Gorter, R., Fox, J. P. and Twisk, J. W. (2015) Why item response theory should be used for longitudinal questionnaire data analysis in medical research. *BMC Med Res Methodol*,**30** 55.

Griffith, L., Van Den Heuvel, E., Fortier, I., Hofer, S., Raina, P., Sohel, N., Payette, H., Wolfson, C. and Belleville, S. (2013) Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data Meta-Analysis. In *AHRQ Methods for Effective Health Care*, Rockville MD (US) Agency for Healthcare Research and Quality.

Groenen, P. J. and Andries van der Ark, L. (2006) Visions of 70 years of psychometrics: the past, present, and future. *Stat Neerl*, **60** 135-144.

Gross, A. L., Power, M. C., Albert, M. S., Deal, J. A., Gottesman, R. F., Griswold, M., Wruck, L. M., Mosley, T. H. J., Coresh, J., Sharrett, A. R. and Bandeen-Roche, K. (2015) Application

of Latent Variable Methods to the Study of Cognitive Decline When Tests Change over Time. *Epidemiology,* **26** 878-887.

Hallfors, D. D., Waller, M. W., Bauer, D., Ford, C. A. and Halpern, C. T. (2005) Which comes first in adolescence--sex and drugs or depression? *Am J Prev Med,***29** 163-170.

Hambleton, R. K., Swaminathan, H. and Rogers, H. J. (1991) *Fundamentals of item response theory*, Newbury Park, CA: Sage.

Hankin, B. L., Abramson, L. Y., Moffitt, T. E., Silva, P. A., McGee, R. and Angell, K. E. (1998) Development of depression from preadolescence to young adulthood: emerging gender differences in a 10-year longitudinal study. *J Abnorm Psychol,***107** 128-140.

Higgins, J. P., Whitehead, A., Turner, R. M., Omar, R. Z. and Thompson, S. G. (2001) Meta-analysis of continuous outcome data from individual patients. *Stat Med* **20**, 2219-2241.

Hussong, A. M., Curran, P. J. and Bauer, D. J. (2013) Integrative data analysis in clinical psychology research. *Annu Rev Clin Psychol,***9** 61-89.

Hussong, A. M., Flora, D. B., Curran, P. J., Chassin, L. A. and Zucker, R. A. (2008) Defining risk heterogeneity for internalizing symptoms among children of alcoholic parents. *Dev Psychopathol,***20** 165-193.

Imai, K., Keele, L. and Tingley, D. (2010) A general approach to causal mediation analysis. *Psychol Methods,***15** 309-334.

Khan, M. R., Kaufman, J. S., Pence, B. W., Gaynes, B. N., Adimora, A. A., Weir, S. S. and Miller, W. C. (2009) Depression, Sexually Transmitted Infection, and Sexual Risk Behavior Among Young Adults in the United States. *Arch PediatrAdolesc Med,***163** 644--652.

Kohli, N. and Harring, J. R. (2013) Modeling growth in latent variables using a piecewise function. *Multivar Behav Res,***48** 370-397.

Koukounari, A., Stringaris, A. and Maughan, B. (2017) Pathways from maternal depression to young adult offspring depression: an exploratory longitudinal mediation analysis. *Int J Methods Psychiatr Res,***26**.

Kroenke, K., Spitzer, R. L. and Williams, J. B. (2003) The Patient Health Questionnaire-2 - Validity of a two-item depression screener. *Medical Care,***41** 1284-1292.

Kroenke, K., Spitzer, R. L. and Williams, J. B. (2001) The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med,***16** 606-613.

Kuha, J., Katsikatsou, M. and Moustaki, I. (2018) Latent variable modelling with non-ignorable item nonresponse: Multigroup response propensity models for cross-national analysis. To appear in *J Roy Stat Soc, Series A*.

Li, F., Duncan, T. E., Duncan, S. C. and Hops, H. (2001) Piecewise growth mixture modelling of adolescent alcohol use data. *Struct Equ Modeling,***8** 175-204.

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P. and Meredith, W. (2009) Modeling Life-Span Growth Curves of Cognition Using Longitudinal Data With Multiple Samples and Changing Scales of Measurement. *Psychol Methods,***14** 126-149.

Mercer, C. H., Tanton, C., Prah, P., Erens, B., Sonnenberg, P., Clifton, S., Macdowall, W., Lewis, R., Field, N., Datta, J., Copas, A. J., Phelps, A., Wellings, K. and Johnson, A. M. (2013) Changes in sexual attititudes and lifestyles in Britain through the life course and over time: findings from the National Surveys of Sexual Attitudes and Lifestyles (Natsal). *Lancet* **382** 1781-1794.

Messer, S. C., Angold, A., Costello, E. J., Loeber, R., VanKammen, W. and StouthamerLoeber, M. (1995) Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents: Factor composition and structure across development. *Int J Methods Psychiatr Res,***5** 251-262.

Niarchou, M., Zammit, S. and Lewis, G. (2015) The Avon Longitudinal Study of Parents and Children (ALSPAC) birth cohort as a resource for studying psychopathology in childhood and adolescence: a summary of findings for depression and psychosis. *Soc Psychiatry Psychiatr Epidemiol,***50** 1017-1027.

Nuttbrock, L., Bockting, W., Rosenblum, A., Hwahng, S., Mason, M., Macri, M. and Becker, J. (2013) Gender abuse, depressive symptoms, and HIV and other sexually transmitted infections among male-to-female transgender persons: a three-year prospective study. *Am J Public Health,***103** 300-307.

Piccinelli, M. and Wilkinson, G. (2000) Gender differences in depression. *Br J Psychiatry,***177** 486-492.

Rabe-Hesketh, S. and Skrondal, A. (2004) *Generalized Latent Variable Modeling. Multilevel, Longitudinal and Structural Equation Models*.

Riley, R. D., Lambert, P. C. and Abo-Zaid, G. (2010) Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* **340** c221.

Rothman, K. J., Gallacher, J. E. and Hatch, E. E. (2013) Why representativeness should be avoided. *Int J Epidemiol,***42** 1012-1014.

Samejima, F. (1969) Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17*.

Schwarz, G. E. (1978) Estimating the dimension of a model. *Ann Stat* **6** 461–464.

Shrier, L. A., Harris, S. K. and Beardslee, W. R. (2002) Temporal associations between depressive symptoms and self-reported sexually transmitted disease among adolescents. *Arch Pediat Adol Med,***156** 599-606.

Siddique, J., Reiter, J. P., Brincks, A., Gibbons, R. D., Crespi, C. M. and Brown, C. H. (2015) Multiple imputation for harmonizing longitudinal non-commensurate measures in individual participant data meta-analysis. *Stat Med,***34** 3399-3414

Sonnenberg, P., Clifton, S., Beddows, S., Field, N., Soldan, K., Tanton, C., Mercer, C. H., da Silva, F. C., Alexander, S., Copas, A. J., Phelps, A., Erens, B., Prah, P., Macdowall, W., Wellings, K., Ison, C. A. and Johnson, A. M. (2013) Prevalence, risk factors, and uptake of interventions for sexually transmitted infections in Britain: findings from the National Surveys of Sexual Attitudes and Lifestyles (Natsal). *Lancet*,**382** 1795-1806.

Vermunt, J. K. (2010) Latent Class Modeling with Covariates: Two Improved Three-Step Approaches. *Polit Anal*,**18** 450-469.

von Davier, M. and von Davier, A. A. (2007) A unified approach to IRT scale linking and scale transformations. *Methodology*,**3** 115–124.

Wang, C., Kohli, N. and Henn, L. (2015) A Second-Order Longitudinal Model for Binary Outcomes: Item Response Theory Versus Structural Equation Modeling. *Struct Equ Modeling*,**23** 455-465.

Wang, X., Zhang, L., Lei, Y., Liu, X., Zhou, X., Liu, Y., Wang, M., Yang, L., Zhang, L., Fan, S. and Xie, P. (2014) Meta-analysis of infectious agents and depression. *Sci Rep*,**4** 4530.

Weller, E. B., Kloos, A., Kang, J. and Weller, R. A. (2006) Depression in children and adolescents: Does gender make a difference. *Curr Psychiatry Rep* **8** 108-114.

Woodhall, S. C., Soldan, K., Sonnenberg, P., Mercer, C. H., Clifton, S., Saunders, P., da Silva, F., Alexander, S., Wellings, K., Tanton, C., Field, N., Copas, A. J., Ison, C. A. and Johnson, A. M. (2016) Is chlamydia screening and testing in Britain reaching young adults at risk of infection? Findings from the third National Survey of Sexual Attitudes and Lifestyles (Natsal-3). *Sex Transm Infect*,**92** 218-227.

Figure 1. ***Flow Chart of data within each stage of analysis***

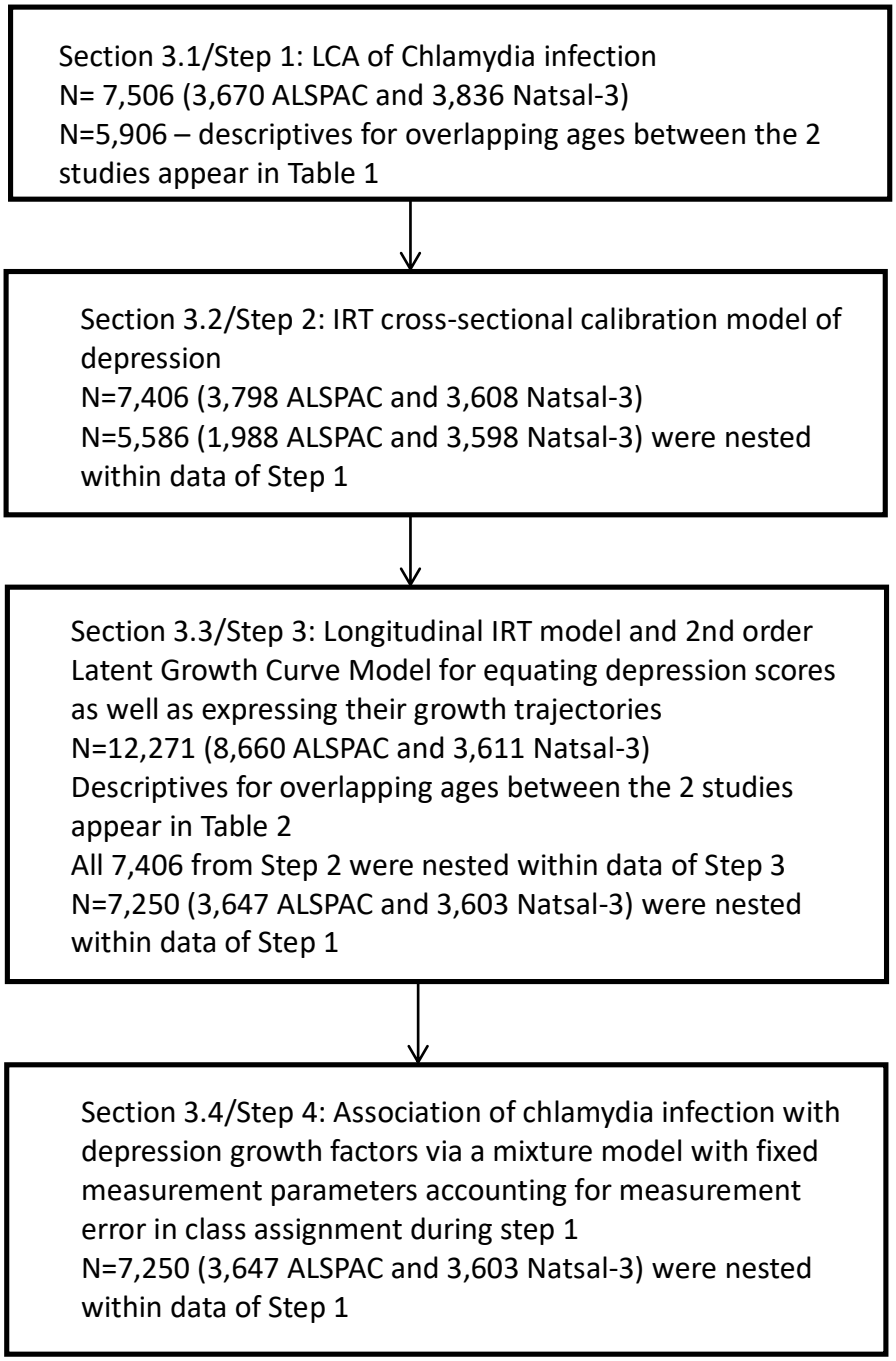


Figure 2. ***Item Response Probabilities from Latent Class Model during Step 1-n=7506***

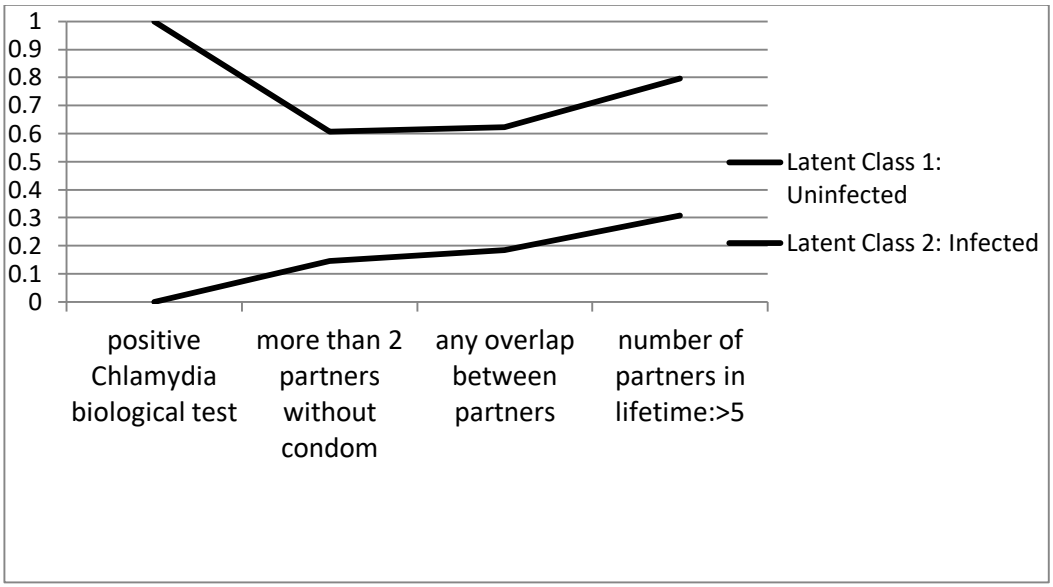


Figure 3. ***Mean IRT depression scores by Age and Study during Step 2 -n=7406 (calibration sample)***

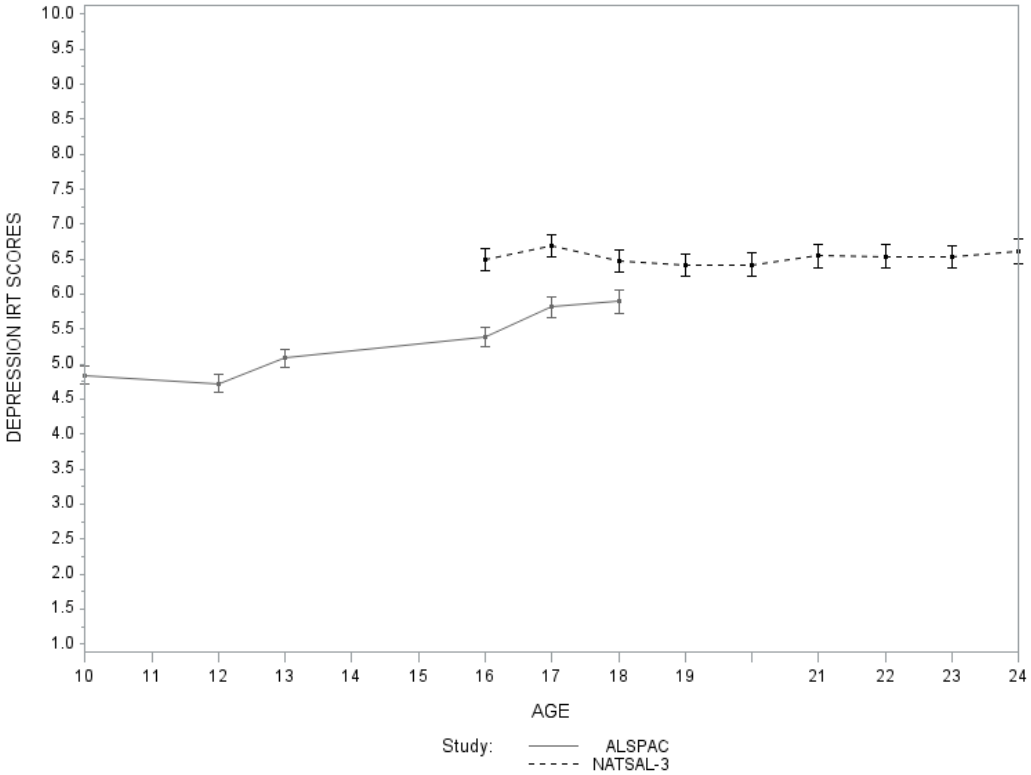


Figure 4. ***Mean IRT depression scores by Age and Gender during Step 2-n=7406 (calibration sample)***

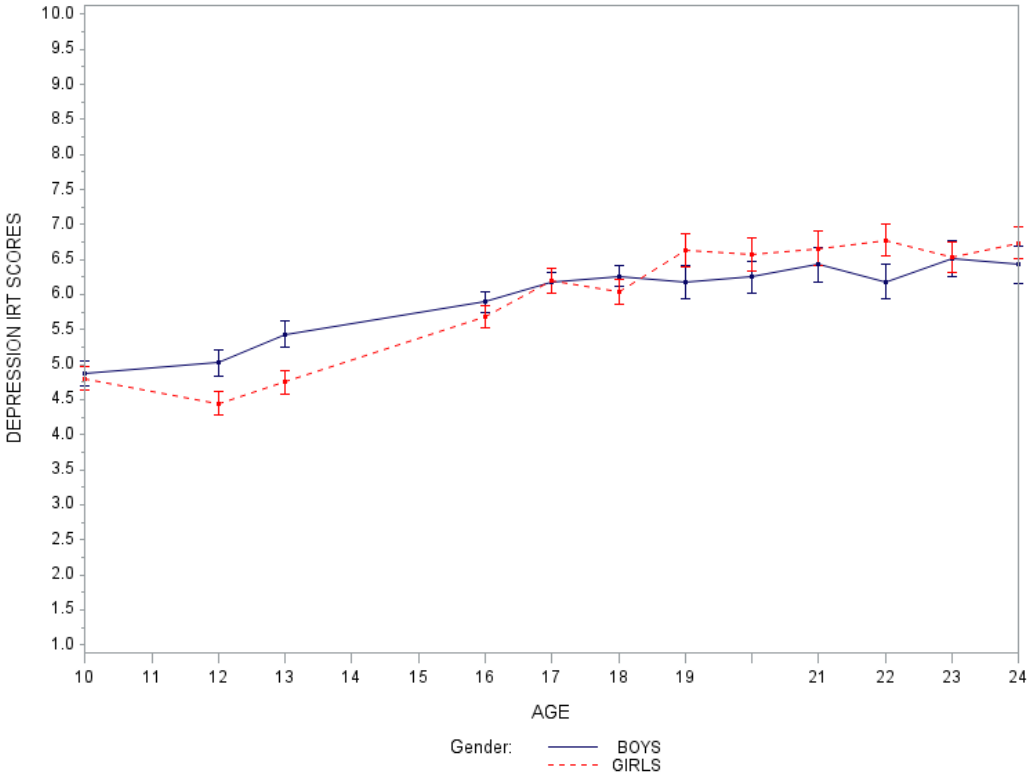
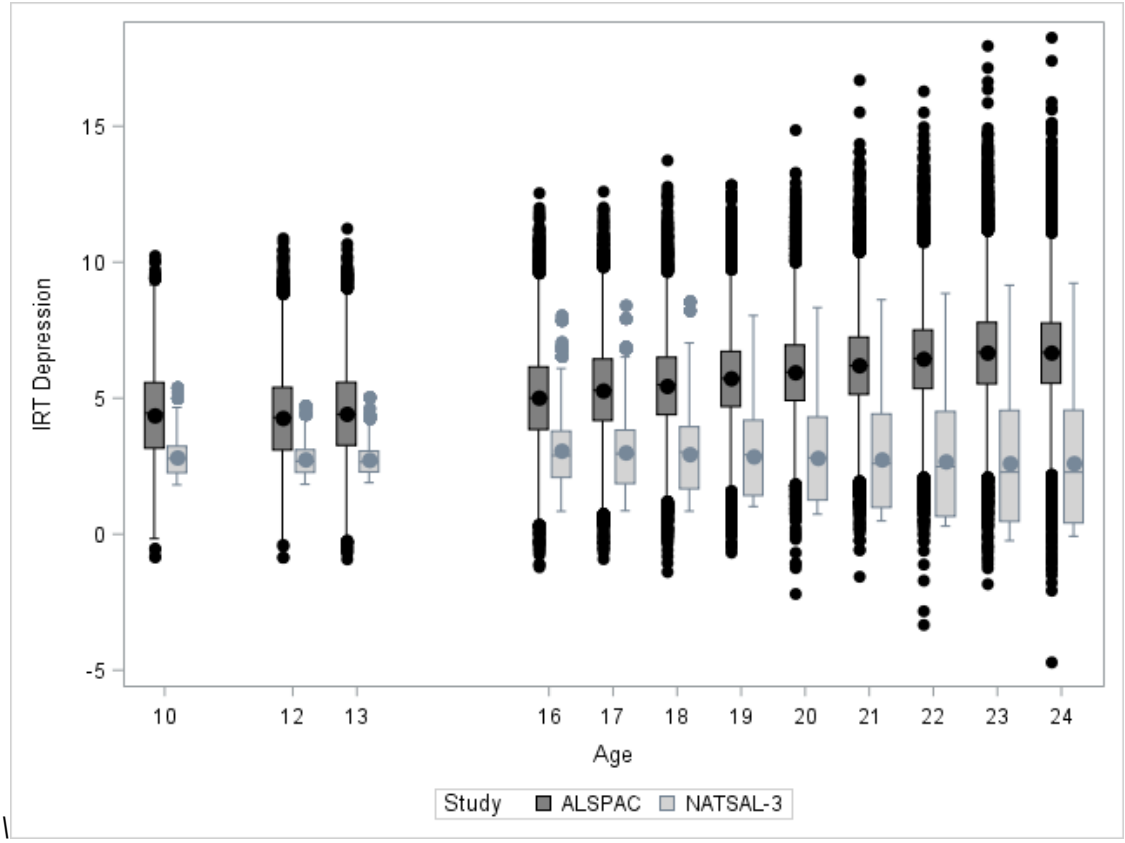


Figure 5. ***Box plots for IRT Depression scores by Age and Study during Step 3 -n=12271 (longitudinal data)***



\

Figure 6. ***Box plots for IRT Depression scores by Age and Gender during Step 3 -n=12271 (longitudinal data)***

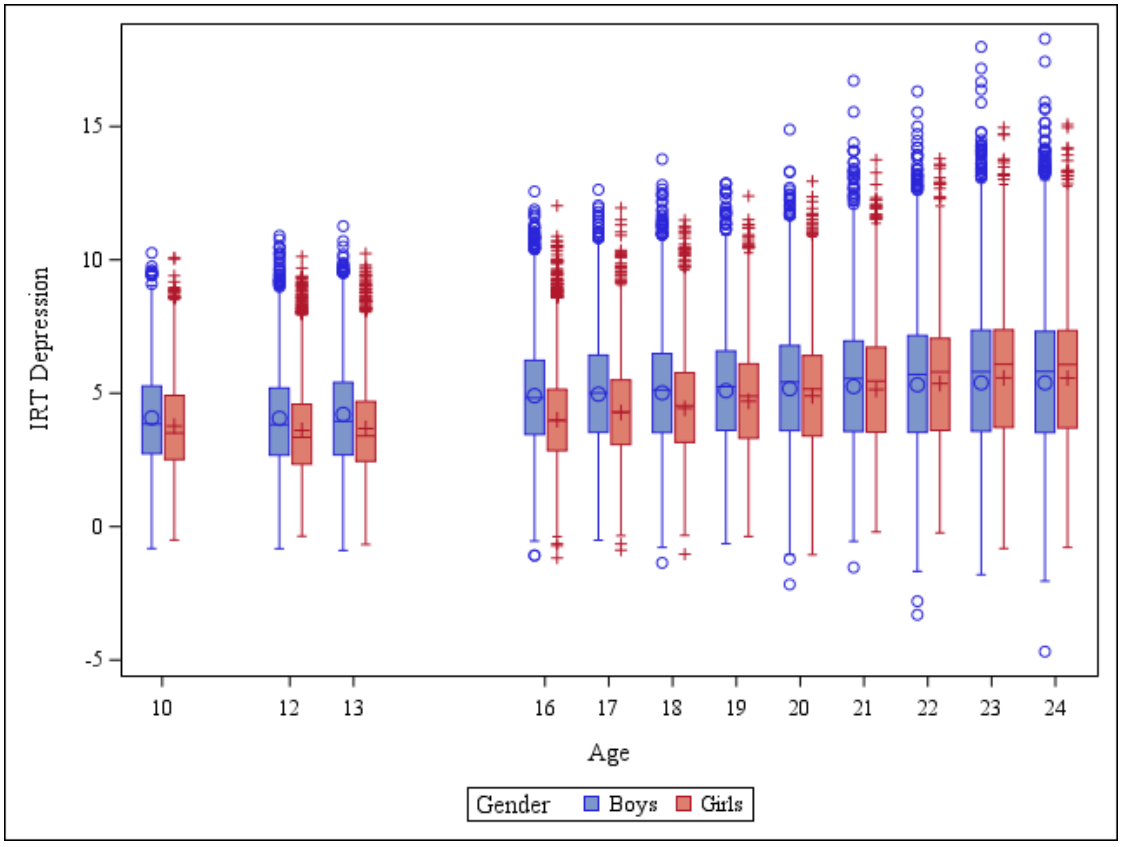


Table 1: ***Pooled Data Analysis for overlapping Chlamydia tests and Sexual Behaviour Indicators Sample Description by Study (i.e. 16-20 years)***

	ALSPAC	**Natsal-3**	**Pooled Sample**
Chlamydia test urine % positive	0.7 n=2881	2.5 n=1159	1.2 n=4040
Number of heterosexual &/or homosexual partners without a condom, last year % 1 % >=2	NA	33.7 16.0 n=2195	33.7 16.0 n=2195
Any overlap between partners (i.e. sex with one then another then the first again) in last 5 years %Yes	NA	16.7 n=2222	16.7 n=2222
Number of partners in lifetime % 1 % 2 % 3-4 % >=5	38.9 20.2 19.8 20.9 n=2388	18.3 11.7 14.7 27.9 n=2226	29.0 16.1 17.3 24.3 n=4614
Any STI Symptom in the last month (if had any lifetime partners) %Yes	NA	20.0 n=1614	20.0 n=1614
STD clinic attendance, ever attended a sexual health (GUM) clinic (if had any lifetime partners) %Yes	NA	38.8 n=1619	38.8 n=1619
Used condom at most recent occasion (if had sex in last 4 weeks) %Yes	50.8 n=2381	37.2 n=1075	46.6 n=3456
New Partners Last Year % 1 % >=2	49.8 24.9 n=2382	26.1 20.0 n=2211	38.4 22.5 n=4593
Number of Partners Last Year % 1 % 2 % 3-4 % >=5	57.7 19.3 12.6 6.1 n=2393	36.9 13.6 11.2 7.5 n=2212	47.7 16.6 11.9 6.8 n=4605
%Male	44.4 n=3670	46.3 n=2236	45.1 n=5906
Mean Age (Std Dev)	17.9 (0.5) n=3670	17.9 (1.4) n=2236	17.9 (0.9) n=5906

NA=Not available-these measures were only recorded in Natsal-3 and not in ALSPAC. STI= Sexually Transmitted Infections; STD=Sexually Transmitted Disease; GUM= genitourinary medicine (often used more restrictively as alternative to sexually transmitted disease clinic); Std Dev=Standard Deviation

Table 2: ***Pooled Data Analysis for overlapping ALSPAC/MFQ and Natsal-3/PHQ-2 items for Depression construct (i.e. 16-20 years)***

		Age in years: % (n)				
Source	**Questions**	**16** ◦	**17**ǂ	**18**◊	**19**	**20**
Original ALSPAC/MFQ	*Did not enjoy anything* 1: sometimes 2:true	 15.3 (744) 3.3 (160)	 23.2 (984) 2.5 (107)	 22.8 (735) 7.6 (243)	NA	NA
	Teenager felt miserable or unhappy 1: sometimes 2:true	 49.7 (2418) 21.5 (1045)	 56.4 (2398) 13.5 (574)	 47.2 (1520) 25.0 (805)	NA	NA
Original Natsal-3/PHQ-2	*Little pleasure, interest in doing things* 2: several days 3: more than half days 4: nearly every day	 24.1 (92) 2.9 (11) 5.2 (20)	 29.4 (127) 6.5 (28) 5.1 (22)	 24.0 (109) 6.2 (28) 4.0 (18)	 24.9 (97) 5.1 (20) 4.4 (17)	 24.6 (96) 5.9 (23) 3.9 (15)
	Feeling down, depressed or hopeless 2: several days 3: more than half days 4: nearly every day	 33.5 (128) 5.5 (21) 3.9 (15)	 32.2 (139) 7.4 (32) 5.6 (24)	 29.9 (136) 7.9 (36) 3.1 (14)	 29.0 (113) 5.6 (22) 4.6 (18)	 29.4 (115) 7.7 (30) 2.8 (11)
Pooled Data Analysis						
Harmonised/recoded PHQ-2 items to match the 2 MFQ items	*Did not enjoy anything* 1: sometimes 2:true	 24.1 (92) 8.1 (31)	 29.4 (127) 11.6 (50)	 24.0 (109) 10.1 (46)	 24.9 (97) 9.5 (37)	 24.6 (96) 9.7 (38)
	Teenager felt miserable or unhappy 1: sometimes 2:true	 33.5 (128) 9.4 (36)	 32.2 (139) 13.0 (56)	 29.9(136) 11.0 (50)	 29.0 (113) 10.3 (40)	 29.4 (115) 10.5 (41)
Final pooled sample 2 depression items	*Did not enjoy anything* 1: sometimes 2:true	 15.9 (836) 3.6 (191)	 23.7 (1111) 3.4 (157)	 23.0 (844) 7.9 (289)	 24.9 (97) 9.5 (37)	 24.6 (96) 9.7 (38)
	Teenager felt miserable or unhappy 1: sometimes 2:true	 48.5 (2546) 20.6 (1081)	 54.2 (2537) 13.5 (630)	 45.0 (1656) 23.3 (855)	 29.0 (113) 10.3 (40)	 29.4 (115) 10.5 (41)

◦Answers to these questions for ALSPAC/MFQ items were provided through questionnaires sent out to children when they were approximately 16 years old

ǂ Answers to these questions for ALSPAC/MFQ items were provided through questionnaires provided to young people who were approximately 17 years of age. Answers were primarily completed/provided by attendees at clinic but it may have been completed online at home and by some participants who did not attend clinic.

◊ ǂ Answers to these questions for ALSPAC/MFQ items were provided through questionnaires sent out to young people who were approximately 18 years of age

Table 3: ***Number of participants in the calibration sample per study and age***

	AGE												
	10	12	13	16	17	18	19	20	21	22	23	24	***Total***
STUDY													
ALSPAC, N (%)	638 (100)	663 (100)	732 (100)	662 (63.41)	590 (57.73)	512 (52.95)	0	0	0	1 (0.25)	0	0	**3798**
NATSAL-3, N (%)	0	0	0	382 (36.59)	432 (42.27)	455 (47.05)	390 (100)	391 (100)	370 (100)	405 (99.75)	410 (100)	373 (100)	**3608**
Total	**638**	**663**	**732**	**1044**	**1022**	**967**	**390**	**391**	**370**	**406**	**410**	**373**	**7406**

Table 4: ***IRT model pre-calibration parameters of equating depression scores from both studies -n=7406 (displayed estimates are then used in anchored calibration)***

Item Description	***Factor Loading (SE)***	***Standardized Factor Loading (SE)***	***Threshold (SE)*** 1^{st} 2^{nd}
1. Young person has not enjoyed anything in the last two weeks	1.000 (0.000)	0.766 (0.010)	7.434 (0.561) 10.136 (0.577)
2. Young person has felt unhappy/miserable in the last two weeks *Study covariate effect (DIF)*	1.114 (0.049)	0.801 (0.011)	2.179 (0.602) 5.867 (0.616) -2.764 (0.113)
3. Young person has felt so tired they sat around and did nothing in the last two weeks	0.535 (0.027)	0.537 (0.016)	2.606 (0.311) 5.241 (0.324)
4. Young person has felt very restless in the last two weeks	0.459 (0.025)	0.480 (0.018)	2.306 (0.270) 4.977 (0.283)
5. Young person felt they were no good anymore in the last two weeks	1.818 (0.092)	0.908 (0.007)	12.447 (1.081) 15.948 (1.129)
6. Young person has cried a lot in the last two weeks	0.944 (0.047)	0.747 (0.013)	6.741 (0.552) 8.898 (0.569)
7. Young person has found it hard to think properly/concentrate in the last two weeks	0.635 (0.031)	0.604 (0.015)	2.935 (0.361) 5.968 (0.376)
8. Young person has hated themselves in the last two weeks	1.679 (0.087)	0.894 (0.008)	12.190 (1.010) 14.936 (1.047)
9. Young person has felt they were a bad person in the last two weeks	1.026 (0.051)	0.774 (0.013)	7.619 (0.609) 10.281 (0.635)
10. Young person has felt lonely in the last two weeks	1.277 (0.059)	0.836 (0.009)	7.654 (0.727) 10.932 (0.754)
11. Young person has felt nobody really loved them in the last two weeks	1.407 (0.071)	0.859 (0.010)	10.312 (0.839) 12.972 (0.870)
12. Young person thought they could never be as good as other kids in the last two weeks	1.145 (0.054)	0.806 (0.011)	7.644 (0.659) 10.338 (0.680)
13. Young person has felt they did everything wrong in the last two weeks	1.392 (0.069)	0.856 (0.009)	9.800 (0.820) 12.961 (0.856)
14. Young person has been having fun in the last two weeks	-0.516 (0.034)	-0.524 (0.022)	-7.386 (0.371) -4.141 (0.330)
15. Young person has felt happy in the last two weeks	-0.566 (0.033)	-0.559 (0.020)	-6.738 (0.368) -4.064 (0.346)
16. Young person has enjoyed doing lots of things in the last two weeks	-0.525 (0.031)	-0.531 (0.020)	-6.253 (0.344) -3.706 (0.324)

Factor mean/Covariate effect	***Estimate (SE)***	***Standardized Estimate (SE)***
Study membership (reference category: ALSPAC)	5.813 (1.065)	2.691 (0.488)
Age	0.077 (0.084)	0.036 (0.039)
Age^2	0.013 (0.005)	0.006 (0.002)

AgexStudy	-0.291 (0.062)	-0.134 (0.028)
Gender (reference category:males)	-0.127 (0.057)	-0.059 (0.026)

SE=standard error

Table 5: ***Fixed Effects for the Piecewise Linear Growth Model fitted to the IRT Depression Scale scores including covariate effects of gender and study membership (n=12271)***

Effect	***Estimate (SE)***	***p-value***
Intercept (at 16 years old)	7.704 (0.150)	<0.001
Slope 1 (changes during ages 10-16)	1.056 (0.083)	<0.001
Slope 2 (changes during ages 16-24)	0.308 (0.056)	<0.001
Effect of gender on mean intercept*	-0.910 (0.065)	<0.001
Effect of gender on slope 1	-0.477 (0.055)	<0.001
Effect of gender on slope 2	0.167 (0.021)	<0.001
Effect of study membership‡ on mean intercept	-1.571 (0.115)	<0.001
Effect of study membership on slope 2	-0.316 (0.046)	<0.001

*Gender reference category: males; ‡Study membership reference category: ALSPAC

Table 6: ***Random Effects for the Piecewise Linear Growth Model fitted to the IRT Depression Scale scores (n=12271)***

Variance	***Estimate (SE)***	***p-value***
Intercept (at 16 years old) σ^2_I	3.472 (0.123)	<0.001
Slope 1 (changes during ages 10-16) σ^2_{S1}	1.289 (0.073)	<0.001
Slope 2 (changes during ages 16-24) σ^2_{S2}	0.164 (0.008)	<0.001
Covariance		
Intercept - Slope 1 σ_{IS1}	1.182 (0.082)	<0.001
Intercept - Slope 2 σ_{IS2}	-0.140 (0.028)	<0.001
Slope 1- Slope 2 σ_{S1S2}	-0.077 (0.020)	<0.001

Table 7: ***Effects of Growth Factors of IRT depression scaled scores (i.e. Intercept at 16 years old and Slopes representing changes between ages 10-16 & 16-24 years old) and study on the latent class of Chlamydia infection (n=7250)***

Infected with Chlamydia latent class	***Effect***	***OR***	***95 % Confidence Intervals***	***p-value***
	Intercept (at 16 years old)	0.850	0.537 to 1.344	0.485
	Slope 1 (changes during ages 10-16)	4.007	1.030 to 15.585	0.045
	Slope 2 (changes during ages 16-24)	0.457	0.159 to 1.312	0.146
	Study (Ref: ALSPAC)	11.280	3.372 to 37.726	<0.001

OR: Odds Ratio