National Institute of Economic and Social Research

# WHAT YOUNG ENGLISH PEOPLE DO ONCE THEY REACH SCHOOL-LEAVING AGE: A CROSS-COHORT COMPARISON FOR THE LAST 30 YEARS

Jake Anders & Richard Dorsett

NIESR Discussion Paper No. 454

Date: 29 October 2015

National Institute of Economic and Social Research

## About the National Institute of Economic and Social Research

The National Institute of Economic and Social Research is Britain's longest established independent research institute, founded in 1938. The vision of our founders was to carry out research to improve understanding of the economic and social forces that affect people's lives, and the ways in which policy can bring about change. Seventy-five years later, this remains central to NIESR's ethos. We continue to apply our expertise in both quantitative and qualitative methods and our understanding of economic and social issues to current debates and to influence policy. The Institute is independent of all party political interests.

National Institute of Economic and Social Research
2 Dean Trench St
London SW1P 3HE
T: +44 (0)20 7222 7665
E: enquiries@niesr.ac.uk
niesr.ac.uk
Registered charity no. 306083

This paper was first published in October 2015

# What young English people do once they reach school-leaving age: a cross-cohort comparison for the last 30 years

Jake Anders & Richard Dorsett

## *Abstract*

This paper examines how young people's early transitions into the labour market have changed between cohorts born in 1958, 1970, 1980, and 1990. We use sequence analysis to characterise transition patterns and identify three distinct pathways in all cohorts. An 'Entering the Labour Market' group has declined significantly in size (from 91% in the earliest cohort, to 37% in the most recent), an 'Accumulating Human Capital' group has grown in its place (from 4% to 51%), but also a 'Potential Cause for Concern' group has grown alongside this, reaching 12% in the most recent cohort. These trends appear to reflect behavioural rather than compositional changes. Females and those who are from a non-white ethnic background have gone from being more likely to be in the 'Potential Cause for Concern' group, to being less likely. Coming from a low socio-economic status background has remained a strong predictor of having a transition of this type across all four cohorts. These early transitions are important, not least since we show they are highly predictive of longer-term outcomes.

## *Acknowledgements*

This research was funded by the UK Department for Business Innovation and Skills (BIS) and the Centre for Learning and Life Chances in Knowledge Economies and Societies, an ESRC-funded Research Centre (grant reference ES/J019135/1). We are grateful to Vahé Nafilyan and Laura Kirchner Sala of the Institute of Employment Studies for preparing the data. Matt Bursnall at BIS and Paolo Lucchino at NIESR provided helpful comments. The paper was presented at BIS, a meeting of the Jacob's Foundation PATHWAYS programme, and the UCL Institute of Education's Centre for Longitudinal Studies Cohort Studies Conference 2015. The usual disclaimer applies.

## *Contact details*

Jake Anders (j.anders@niesr.ac.uk), National Institute of Economic and Social Research, 2 Dean Trench Street, London SW1P 3HE.

## Non-Technical Summary

This paper examines how young people's early transitions from school to work have changed between cohorts born in 1958, 1970, 1980, and 1990. It also looks at how young people's characteristics predict how successful this transitions will be, and whether this has changed between the four cohorts.

We use sequence analysis to compare and quantify the differences between young people's month by month activities from the September after their 16th birthday (i.e. when they are no longer in compulsory education) for a period of 29 months. We use the measures of similarity obtained in this way to distinguish three broad groups of individuals:

1. An "Entering the Labour Market" group, who move quickly from school into work without completing much education beyond that which is compulsory.
2. An "Accumulating Human Capital" group, who remain in full time education throughout the 29 month period that we observe them.
3. A "Potential Cause for Concern" group, who appear to leave education but without successfully moving into stable employment. They may move in and out of work, report being consistently unemployed, or report being economically inactive.

The 'Entering the Labour Market' group has declined significantly in size (from 91% in the earliest cohort, to 37% in the most recent), while the 'Accumulating Human Capital' group has grown (from 4% to 51%). However, the 'Potential Cause for Concern' group has also grown, from 4% in the first cohort to 12% in the most recent one.

Females and those who are from a non-white ethnic background have gone from being more likely to be in the 'Potential Cause for Concern' group than males and those from a non-white ethnic background, to being less likely. However, coming from a low socio-economic status background has remained a strong predictor of having a transition that we characterise as being a "Potential Cause for Concern" across all four cohorts. We show these early experiences to be important predictors of longer-term outcomes.

# 1. Introduction

In recent years there has been growing concern about the number of young people failing to make a successful transition from education into employment. Increasingly, this appears to be a structural, rather than cyclical, problem. We see evidence of this from that fact that although youth unemployment in the UK was falling in the late 1990s and early 2000s, it started rising again as early as 2004, long before the general downturn in the economy (OECD, 2008). This is an important issue, not least because making a successful transition from education into the labour market is important for young people's long-term economic success; periods of unemployment during these early years may have long-term scarring effects on later employment on earnings prospects (Arulampalam, 2001; Gregg, 2001; Gregg & Tominey, 2005).

In this paper, we examine how early transitions have changed over the last thirty years. We focus in particular on the group of young people whose early experiences are a potential cause for concern in the sense that they neither continue in education nor do they find stable employment. We assess the changing size of this group and examine the extent to which it is possible to predict, on the basis of characteristics at the time of reaching school-leaving age, which individuals may experience a difficult transition from school to work.

Our approach is to use sequence analysis (Abbott, 1995) to quantify the similarity between individuals' transitions over a period of 29 months from the September following their 16$^{th}$ birthday. Previous research has shown that young people's transitions into work may be highly differentiated (Fergusson, Pye, Esland, McLaughlin, & Muncie, 2000). Sequence analysis provides a means of comparing the full detail of individuals' labour market trajectories. This permits a fuller comparison than the more usual methods of studying labour market states at single point in time (Andrews & Bradley, 1997) or specific changes in states (Berrington, 2001). In taking this approach, we build on previous research that uses sequence analysis to study young people's transitions from education into the labour market (Anyadike-Danes & McVicar, 2005, 2010; Dorsett & Lucchino, 2014; Halpin & Chan, 1998; Martin, Schoon, & Ross, 2008; Quintini & Manfredi, 2009).

The major contribution of this paper is that it uses detailed survey data based on four birth cohorts, each roughly a decade apart. Previous research (Schoon, McCulloch, Joshi, Wiggins, & Bynner, 2001) has examined how transitions have changed between individuals born in 1958 and individuals born in 1970. We extend this to include also individuals born in 1980 and individuals born in 1990. This provides a major update to the existing empirical literature. By focusing on these more recent cohorts, we are able to consider individuals for whom the school to work transitions are relatively recent (at the time of writing). More specifically, the transitions of the 1980 and 1990 cohorts took place in the 1996-1999 and 2006-2009 periods, respectively, while the transitions of the 1970 and 1958 cohorts took place in the 1974-77 and 1986-89 periods, respectively. Ours is the first study to conduct

cross-cohort analysis using sequence analysis over such an extended period and, by using more recent data, the results are more closely related to the present-day labour market.

This paper proceeds as follows. We first describe the datasets used in this analysis (Section 2). In Section 3, we describe our methodological approach. Our analysis results in the identification of a typology of transition pathways, discussed in Section 4, along with an account of how pathways have changed over time. We also examine the extent to which it is possible to use characteristics at age 16 to predict which young people will experience transitions that are a potential cause for concern (in the sense of not being characterised by either education participation or stable employment). We consider the extent to which these relationships have changed over the four cohorts analysed in this paper. In Section 5, we extend the horizon over which individuals' transitions are considered, examining the extent to which the early transitions that we have considered are predictive of longer-term transitions, up to approximately age 24. Section 6 concludes.

## 2. Data

Our analysis uses information on month-on-month transitions for young people in England, starting in the September following their 16$^{th}$ birthdays and continuing for 29 months. This is dictated by the nature of the available data, which are drawn from for four birth cohort surveys:

- The National Child Development Study (NCDS) is a longitudinal survey of all individuals born in one week in 1958. Background variables (used later to predict transitions) were taken from interviews with the participant and their parents at age 16 (NCDS Sweep 3, 1974) and activity histories assembled using recall interviews at age 23 (NCDS Sweep 4, 1981). The analysis sample has around 6,000 individuals.
- The British Cohort Study (BCS) is a longitudinal survey of individuals born in 1970. Background variables were taken from interviews with the participant and their parents at age 16 (BCS Sweep 4, 1986). Activity histories were assembled primarily using recall interviews at age 26 (BCS Sweep 5, 1996). The analysis sample contains around 8,600 individuals.
- Cohort 8 of the Youth Cohort Study (YCS) is a longitudinal survey of individuals born in 1980. Background variables were taken from interviews with the participant at age 16 (YCS Cohort 8, Sweep 1, 1996), who also provided information about their parents. Activity histories were constructed using annual interviews between ages 17 and 19. The analysis sample has around 8,700 individuals.
- The Longitudinal Study of Young People in England (LSYPE) is a longitudinal survey of individuals born in 1989-90. Background variables were taken from interviews with the participant and their parents up to and including age 16 and activity histories were constructed using annual interviews between ages 17 and 19 (LSYPE Waves 1-5, 2005-2010). The analysis sample has around 9,350 individuals.

Our methodological approach requires complete, month-by-month, activity histories without any gaps. With the YCS and LSYPE, activity histories were provided with the dataset. However, with the NCDS and BCS, these needed to be constructed using the recall questions about young people's activities, along with their start and end dates. Constructing these histories required some data cleaning. This involved reconciling overlapping activity spells and, in the case of the NCDS, imputing education as the status where this was not recorded in the data but could be safely assumed.[1] Furthermore, we avoided dropping individuals missing a small number of months' activities by filling in gaps where activity status was unknown. Where there was a gap of a single month, this was imputed to have the same status as the subsequent month. Where there was a gaps of two months and the same activity was recorded before and after the gap, the missing two months were imputed to also have that same status.

While these steps reduced the number of observations that were dropped, there was still some loss of sample. In the NCDS, there are partial activity histories for 9,697 individuals but the analysis (i.e. full activity history) sample is only 8,356. For the BCS, there was a partial history for 9,760 individuals but the analysis sample is 9,518. In the YCS, the respective figures are 9,265 and 8,682; and in the LSYPE they are 9,371 and 9,347. In view of this, we carry out a sensitivity analysis to assess the extent to which sample reduction is likely to influence our results (Appendix A). As is inevitable in longitudinal surveys, our sample is also affected by attrition. We deal with this using an inverse probability weighting strategy, applying our own weighting scheme for the NCDS and BCS, and provided weights for the YCS and LSYPE.

The monthly histories distinguish between four activity status types: employment; education; unemployment; and other inactive. This exception is the LSYPE for which unemployment and 'other inactive' are combined into a single NEET ('not in education, employment or training') status.

Other than sample loss as discussed above, the other concern with the data is that the activity histories rely on the recall of survey respondents. Paull (2002) finds evidence of recall bias in similar longitudinal data, noting that this is more likely among younger respondents and those with the most transient employment histories. As indicated above, the NCDS and BCS rely on quite long recall periods, while the problem is much reduced with the YCS and LSYPE since recall is only over the period of a single year. As such, we should bear in mind the potential increase in recorded short spells in YCS/LSYPE compared to NCDS/BCS that may be driven not by a change in behaviour, but by a change in data collection. Less worrying in a comparative sense is Paull's suggestion of bias among younger

[1] For example, we made use of a separate variable reporting school leaving date and also used characteristics such as young people's highest educational qualification reported by age 23 to impute earlier education status.

respondents; because all cohorts consider the same age group, any such bias should affect all datasets equally.

Table 1 summarises the analysis sample, showing the size of each of the cohorts, along with mean levels of those characteristics later used to predict transitions pathways. There is a good balance of the genders in the BCS, YCS and LSYPE, but males are somewhat over-represented in the case of the NCDS, suggesting that perhaps more of the female participants have been excluded from the analysis due to missing labour market histories over the period. The proportion of the sample from a minority ethnic group also changes between the cohorts, but this seems more likely to be tracking the changing ethnic composition of the population of England over this period. Similarly we see the increased levels of parental education across the cohorts, with a rising proportion of parents having completed a degree.

Table 1 also provides an indicator of the proportion of individuals who experience a NEET spell at some point during our period of analysis. We see that 15% of the NCDS sample experienced being NEET for at least one month. This figure drops to 10% among the BCS and then rises to 25% among the YCS and 26% among the LSYPE. Overall, therefore, we see a long-term increase in the proportion of individuals who will be NEET at some point. However, given the caution above about shorter recall periods in the YCS and LSYPE potentially increasing the reporting of short spells, we might be concerned that this is, at least in part, driven by differences in data, rather than capturing a real change. In the next section, we describe how we use sequence analysis to compare individuals' transition patterns. As a preliminary comment, we note that, among other advantages, this alleviates the problem of differential reporting of short spells since a change in just one month does not greatly affect the similarity between two individuals' sequences.

<<< Table 1 around here>>>

## 3. Methods

Our analytical approach involves three steps and is conducted separately for each cohort. First, we compare individuals' transition experiences. We do this using sequence analysis to produce a measure of dissimilarity for each pair of individuals in the dataset. Second, we perform cluster analysis using these dissimilarity measures in order to group together individuals sharing broadly similar experiences. Third, we look at predictors of cluster membership using logistic regression. We discuss the methods employed for these three steps in turn.

### *Comparing individuals' transition experiences*

Sequence analysis, also known as optimal matching,[2] provides a means of assessing the similarity of activity histories. Although originally used to compare DNA sequences (Durbin, Eddy, Krogh, & Mitchison, 1998), it is increasingly being used in other applications (see Martin & Wiggins, 2011, for a review). The aim of sequence analysis is to quantify the difference between successive pairs of sequences. The most popular methods of doing this are of two broad types.[3]

The first of these is to calculate the minimum number of insertions and deletions that need to be made in order to transform one sequence into another. In this setting, that means adding in, or removing, months of doing a particular activity into the sequence that the individual actually experienced. An example of transforming one individual's sequence into another's using an insertion and deletion is shown in Figure 1 Panel A. An important characteristic of methods of this type is that time can become 'warped' (Lesnard, 2006; Martin & Wiggins, 2011); that is, specific points in time no longer match up with one another across sequences as some periods are stretched by insertions and others are compressed by deletions. While this may not be an issue in some applications, avoiding 'warping' of the timeline in sequences seems particularly important in this application, since young people's transitions are often influenced by specific fixtures in calendar time (such as the start/end of the academic year). In Figure 1 Panel A, we can see that the insertion and deletion ('indel') operations act to change the calendar time over which Sequence B takes place (aligning it to Sequence A).

<<< Figure 1 around here >>>

The other main class of methods calculates the number of substitutions from one state to another that need to be made to transform one sequence into another. In this setting, that means replacing the activity that an individual did in a given month with a different activity, such as substituting being in work for being unemployed. An example of this approach is demonstrated in Figure 1 Panel B. This approach maintains timelines within a sequence, but this might in some circumstances exaggerate differences between sequences that are actually quite similar but slightly offset. Consider the example in Figure 1, which needs just one deletion and one insertion, compared with three substitutions. In addition, this approach requires all sequences to be exactly the same length, whereas allowing insertions at the end of sequence might offer some flexibility to include cases where individuals drop out of the sample during the period being analysed. In our case, as mentioned above, avoiding time-warp is desirable so we adopt this second approach.

[2] Optimal matching in this sense should not be confused with the identically named technique within the propensity score matching literature. Partly in order to avoid this ambiguity we use the term sequence analysis throughout.

[3] Combinatorial approaches, such as those outlined by Elzinga (for example Elizinga & Liefbroer, 2007), are another alternative.

To arrive at a measure of dissimilarity using this method, the 'cost' (that is, how much of a change to the sequence we should consider ourselves to have made when we replace a month's activity with a different activity) associated with each substitution must be specified. The simplest approach would be to count the number of substitutions carried out and set this as the dissimilarity measure. However, this ignores the possibility that some substitutions might involve a qualitatively more extreme change in status (for instance, altering a month's status from 'other inactive' to employment may constitute more of a change than a switch from 'other inactive' to unemployment). We can take account of the likelihood that some pairs of states are more similar to one another than others. We do this by setting the cost of substituting between two very similar states to be lower than that of substituting between two very different states. The dissimilarity measure is then the total cost of all substitutions involved.

The relevant question is how to set these costs. In some applications it is meaningful for the analyst to provide a matrix of substitution costs, based on prior knowledge of relevant differences between the states. Indeed, in a previous analysis of school to work transitions Anyadike-Danes and McVicar (2005) base their substitution matrix "on the degree of attachment to the labour market of the different activities" (p. 515), but note that their results are very robust to specifying alternative substitution costs, including setting them all equal.

Arbitrary choice of substitution costs is one of the criticisms most often levelled at applications of sequence analysis (Wu, 2000).To attempt to identify appropriate substitution costs, we use the probability of transition between the two states being substituted. The less likely a transition between two states, the greater is the cost associated with a substitution of these two states. We allow these probabilities (and therefore costs) to vary over time, based on a moving average of the probability of transition in the months around the point in time at which the substitution is made. This method is referred to as calculating the Dynamic Hamming Distance (DHD) between two sequences (Lesnard, 2006).

In summary, in this application, the cost of a substitution is allowed to vary depending on which substitution being made (e.g. a different cost to substitute employment for education than to substitute unemployment for inactive) and to vary across time rather than being fixed throughout the entire period (e.g. a different cost to substitute education for employment in January than in July). We implement this approach using the TraMineR package for R (Gabadinho, Ritschard, Müller, & Studer, 2011).

### *Identifying groups of individuals with similar trajectories*

Using the dissimilarity measures calculated through sequence analysis, we then carried out cluster analysis in order to group together sets of sequences that were similar. Technically, we used the non-hierarchical k-medoids/Partitioning Around Medoids (PAM) method of cluster analysis (Kaufman & Rousseeuw, 1987). The non-hierarchical approach does not impose the same constraints on cluster formation as hierarchical approaches, while k-

medoids rather than k-means is more robust to outliers. However, a criticism that is levelled at this method is that it requires the analyst to choose the desired number of clusters; choosing an inappropriate number may result in unreliable results. We discuss our approach to choice of number of clusters below. PAM begins by randomly choosing the requested number of 'medoids', which are actual individuals within the dataset. All other individuals are then assigned to the cluster of the medoid to which they are most similar. There is then an iterative process of swapping current individuals selected as medoids with other potential candidates, with swaps being made where this reduces within cluster variance, until no further swaps that reduce variance are available.

We started the analysis with two clusters and then repeated, adding one more cluster each time until there were twenty. In selecting which of these to use as our preferred cluster solution, we were guided by the average silhouette distance (Rousseeuw, 1987) as a primary diagnostic. This is reported for each of the requested solutions and for each of the four cohorts is shown in Figure 2. However, we also used some qualitative assessment of the sequences found within each cluster. Nevertheless, in all cases the average silhouette distance of the solution used is above 0.7, which is the "rule of thumb" for the resulting cluster solution indicating that a strong structure has been found, suggested by Kaufman and Rousseeuw (1990, p. 88). Ultimately, we settle on seven cluster solutions in all of the datasets analysed.

<<< Figure 2 around here >>>

### *Predicting who will experience a pattern of transitions that may be a cause of concern*

Being able to predict what kind of transition to the labour market individuals are likely to have, before this process has begun, is of clear potential use to policy makers. It potentially makes it easier to target support on those likely to be at risk of experiencing a pattern of transitions that might be a cause for concern. This work is in a similar spirit to that of Caspi, Entner Wright, Moffitt, and Silva (1998), who use childhood characteristics to model the probability of experiencing unemployment during the transition into the labour market. We use logistic regression models in order to assess how well age 16 characteristics that are common to the four datasets can predict outcomes. The logistic regression model we estimate has as its dependent variable whether an individual is identified as being in a cluster that may be a 'Potential Cause for Concern'[4] (see next section).

Since the aim is to examine change between cohorts, we only make use of variables that can be derived to be comparable across cohorts. It is important to note that we make no claim

[4] We focus here on the predictors of being a "Potential Cause for Concern", but results from a multinomial logistic regression model comparing membership of all three groups (i.e. "Entering the Labour Market", "Accumulating Human Capital", and "Potential Cause for Concern") are reported in Appendix B.

that the associations found are causal (especially as there are relatively few available control variables to include in the regression models). The predictors we include in these models are gender, ethnicity (a dichotomous variable of white or non-white), highest parental education (specifically having achieved A-Levels or degrees), housing tenure (specifically social renting or owner occupation), whether living with just one parent, and whether an individual's household is workless.

We estimate four separate models on the four datasets. Comparing these models provides evidence on how the roles of different predictive factors change, or remain the same, over time. In addition, we estimate a combined model on the pooled sample of all four datasets, which allows us to formally test whether the differences between these models are statistically significant from one another. Finally, we assess the relative importance of cohort and characteristic influences by predicting group membership for members of the LSYPE cohort using the relationships estimated for members of the NCDS cohort.

# 4. Results

## *Cluster solutions*

We categorise the seven clusters identified in each cohort's transitions into three broader groups which we label as follows:

- 'Entering the Labour Market' includes individuals who make a relatively early entry into the labour market, leaving education and finding a job before or within the period of analysis
- 'Accumulating Human Capital' includes individuals who remain in education throughout the period of analysis and are, hence, likely to have received higher education ahead of their labour market entry.
- 'Potential Cause for Concern' includes individuals whose experience includes extended periods of unemployment or economic inactivity.

While clusters that fit into these three groupings exist in each of the four cohorts, the relative sizes of these groupings have changed dramatically over time. In order to show this, we present 'index plots' of young people's labour market states over the periods considered. An index plot is a month-by-month representation of the sequence, where each horizontal line represents one young person's transition, with changes in colour showing changes in labour market state. Showing an index plot for a whole cohort is rather impenetrable, but showing plots for the clusters identified above gives a useful overview of the transitions experienced by individuals in the cluster.

Shown first, in Figure 3, is the 'Entering the Labour Market' group. These are clusters in which visual inspection reveals individuals who are either in employment throughout the period considered or who enter employment straight after education. The exception is the second LSYPE cluster which is a little more ambiguous (note that it is rather small). This

group has diminished significantly between the cohorts, from over 90% in the earliest to under 40% in the most recent. In addition, for those who do still follow this route, a visual inspection of the individual transitions that make up these clusters over the four cohorts suggests that earlier entry into the labour market may have become a less stable path with increasing evidence of short spells of unemployment.[5]

Table 2 reveals the extent to which the composition of the 'Entering the Labour Market' group has changed over time. In the NCDS, the characteristics of individuals in this group essentially mirror those of the population as a whole, as one might expect for a group that makes up over 90% of the sample. However, by the BCS cohort, some differences have started to be evident. Most notably, young people whose parents' hold a degree make up only 3% of the group, compared to 5% of those in the population as a whole. The under-representation in this group of individuals with highly-educated parents persists in later cohorts too. Likewise, while there is little difference between the 'Entering the Labour Market' group and the rest of the cohort in the proportion of those who are non-white in the NCDS, by later cohorts a large gap has opened with young people with a non-white ethnic background under-represented in this group.

<<< Figure 3 around here >>>

<<< Table 2 around here >>>

By contrast to the overall decline in the 'Entering the Labour Market' group, the size of the 'Accumulating Human Capital' group, shown in Figure 4, has grown significantly across the cohorts, from 4% in the earliest, to around 50% in the most recent. These are clusters in which individuals remain in education throughout the period of analysis, and the growth reflects increases in both further and higher education across the cohorts analysed. Perhaps unsurprisingly given the large increase in the size of the group, there have been differences in the average characteristics of individuals in this group. For example, the proportion of young people whose ethnicity is not white has increased from 1 per cent in the NCDS (similar to the population as a whole) to 19 per cent in the LSYPE (compared to 14 per cent in the population as a whole).

<<< Figure 4 around here >>>

Lastly, the size of the 'Potential Cause for Concern' group, shown in Figure 5, has also grown, although less dramatically than the 'Accumulating Human Capital' group, from 5% in the earliest cohort to 12% in the most recent. Unlike the other groups, whose respective fall and rise are relatively evenly spread through time, the growth of this group was concentrated between the 1980-born cohort and the 1990-born cohort. This group contains clusters in which individuals spend extended periods in inactivity or unemployment,

[5] It is also possible, though, that some of this effect is explained by under-reporting of short spells in the NCDS/BCS, as discussed earlier.

seemingly not managing to settle into a job or any kind of educational activity throughout this period of their lives. We should note that, for some, particularly where we see inactivity rather than unemployment, a transition of this type might be an active decision, for example individuals who become homemakers. As such, we should not necessarily regard all individuals in this group as a cause for concern.

Relatedly, we also see a change in the behaviour of those who go straight from education into extended inactivity (predominantly young women, especially in earlier cohorts). In earlier cohorts, individuals who experience this kind of transition move into inactivity at around age 16. However, by the later cohorts, otherwise similar looking transitions show individuals moving into inactivity at around age 18, suggesting that such individual are more likely to receive two additional years of education in later cohorts than they were in earlier ones. There may well be benefits for these individuals from the additional human capital they gain from these two years.

It is encouraging to note that these findings accord with those of previous analyses of this time. Most directly, while our Potential Cause for Concern group makes up 5% of the YCS cohort (born in 1980) and 12% of the LSYPE cohort (born in 1990), the size of this group in the analysis by Dorsett and Lucchino (2014) falls somewhere in between (10%), for a sample born between 1975 and 1988. Similarly, the growth in the size of the Accumulating Human Capital group tracks the well-documented trend towards increased levels of post-compulsory education.

<<< Figure 5 around here >>>

### *Predicting difficult transitions*

Table 3 reports the estimation results from separate logistic regression models of membership of a cluster in the 'Potential Cause for Concern' group.[6] The changing influence of gender and ethnicity are the most striking results, with both moving from being a significant predictor in one direction in the earliest cohort to being a significant predictor in the opposite direction by the most recent. First, in the case of ethnicity, individuals born in 1958 who are of non-white ethnicity have nearly three times the odds of being a potential cause for concern than their white peers. By contrast, in the 1990 cohort, individuals of non-white ethnicity instead have 30% lower odds of being a potential cause for concern, compared to their white counterparts. Similarly, males born in 1958 have just a fifth of the odds of being a potential cause for concern that females from a similar background would be predicted to have, while for the cohort born in 1990 males have 20% higher odds than females.

[6] We also fitted a single logistic regression model on the pooled sample from all cohorts, including a cohort regressor and all predictors interacted with these. These replicated the results obtained from the separate models, but allowed for inference testing of the differences between the influences of characteristics in each cohort. These are not reported in this paper but are available on request.

<<< Table 3 around here >>>

The individual coefficients on each of our proxies for socio-economic status (SES) are not straightforward to interpret in isolation, nor do they form any particularly obvious patterns. This partially reflects the changing importance of factors such as housing tenure as indicators for SES. Instead, to illuminate the combined role of SES, Table 4 presents the predicted probability of an individual being a 'Potential Cause for Concern' by gender, ethnicity and two combinations of the other model characteristics chosen to be an example of a 'high SES' individual and a 'low SES' individual. A 'high SES' individual is from a two-parent household, where at least one parent works, at least one parent holds a degree, and their house is owner-occupied. Conversely, a 'low SES' individual is from a lone parent, workless household, where the parent's highest qualification is below A-Level and their home is socially rented. Taken as a whole, these combinations remain indicative of advantage and disadvantage across all four cohorts.

Table 4 shows that the increase in the proportion of young people in clusters categorised as 'Potential Cause for Concern', differs across ethnic/gender combinations. White females have a 6.8% probability of being a 'Potential Cause for Concern' in the NCDS, compared with a 1.5% probability for white males and 17.3% for non-white females. By the time of the LSYPE cohort, white females have a 10% probability of being a 'Potential Cause for Concern', slightly lower than the 11.8% probability for white males and higher than the probability for non-white females, which has fallen significantly to 7.2%.

The most obvious message from the predicted probabilities is that, throughout this period, young people from more advantaged backgrounds have been less likely than those from less advantages backgrounds to experience a transition considered a 'Potential Cause for Concern'. There is evidence of this gap widening over time; from 5.0 percentage points in the NCDS to 17.2 percentage points in the LSYPE.

<<< Table 4 around here >>>

An interesting question is whether changes in the size of the 'Potential Cause for Concern' group are due to cross-cohort differences in composition or to cross-cohort changes in the influence of background characteristics. To explore this, we used the coefficients estimated using the NCDS to predict how group membership among individuals in the LSYPE would be had the influence of background characteristics not changed since the time of this first cohort. These predicted probabilities are reported in Table 5, in a similar way to those reported in Table 4. In each combination of ethnicity and gender we can compare young people's probabilities of being in a 'Potential Cause for Concern' cluster in the NCDS, the LSYPE, and the LSYPE if the probabilities are affected by characteristics in the same way as

they were in the NCDS cohort.[7] For each ethnicity/gender combination, a comparison of the NCDS row with the 'NCDS associations/LSYPE cohort' row shows how changing composition over time affects the predicted probabilities. Similarly, a comparison of the LSYPE row with the 'NCDS associations/LSYPE cohort' row shows the changing influence of background characteristics, assuming composition is fixed.

Looking at the comparison of the NCDS probabilities with those of NCDS association on the LSYPE cohort we find that, across the full sample, the results suggest that the change in composition would be expected to, if anything, reduce the probability of being a 'Potential Cause for Concern' from 2.9% to 2.1%. The biggest difference due to changing composition is among non-white females. In particular, those in 'Low SES' group see their probability of being a potential cause for concern fall from roughly 29% to 8% (among the 'High SES' group, there is a slight increase).

This implies that it is the change in the influence of background characteristics that is primarily responsible for the growth in this group. The second comparison (of the LSYPE row with the 'NCDS associations/LSYPE cohort' row) provides more detail; applying the NCDS associations to the LSYPE cohort predicts 2.1% will be a 'potential cause for concern', whereas in fact 10.4% are. As such, it is the change in the relationship between characteristics and cluster memberships, rather than changes in the composition of the cohorts, that explains the growth in the proportion classified as a 'Potential Cause for Concern'. The increased probability of being a potential cause for concern is seen across all ethnicity/gender combinations for both high and low SES groups. However, it is among the low SES groups that the most dramatic differences are seen.

<<< Table 5 around here >>>

## 5. Extending sequence analysis to age 24

A possible reservation about the results discussed so far is that they may not warrant particular attention since what is more important is how the school to work transitions plays out in the longer run. Such a view may be justified if these early patterns to not persist. However, if they are predictive of transitions over a longer period, their importance is greatly increases.

In order to explore this, we also carried out an analysis of sequences beginning 30 months after turning 16 (i.e. following the end of the period we have been considering so far) up to approximately age 24, and compared the resulting groupings to those for the first 29 months post-16. This is only possible for the two datasets where the data are available: the NCDS and the BCS. We carry out sequence and cluster analysis on the same basis as was done for the earlier time period analyses, except that it starts 30 months after the

[7] The NCDS and LSYPE rows in Table 5 contain identical results to corresponding rows in Table 4 but are included for convenience of comparison.

September following their 16th birthday and continues for 69 months.[8] This time we use 14 cluster (rather than 7 cluster) solutions, reflecting the greater heterogeneity possible within longer sequences. Again, our choice of a 14 cluster solution is primarily on the basis of average silhouette distances.

We once again aggregate these clusters into our three broad groupings: Entering the Labour Market, Accumulating Human Capital and Potential Cause for Concern. One particular challenge with conducting extended sequence analysis on the NCDS is the quality of the monthly activity data available particularly once we extend to age 24. The NCDS appears to have a rather systematic problem with gaps between different spells, which results in the loss of a substantial number of individuals from our analysis, reducing the sample size from 8,372 to 6,122. This loss seems concentrated among individuals in the 'Entering the Labour Market' group, and we suspect that this is responsible for inflating the size of the 'Accumulating Human Capital' grouping compared to that estimated in the shorter analysis. Consequently, there is a concern about the ability of the NCDS to support the longer-run analysis. The BCS analysis does not suffer from the same problem; extending to age 24 reduces the sample size only marginally (from 9,518 to 9,419). In view of this, we feel more confident about the BCS results.

In order to learn more about the relationship between the two sets of categorisations, we cross-tabulate the groupings in to which individuals are placed in the shorter- (29 month) and longer-term (98 month) analyses. Considering first the NCDS, we see that a majority of individuals in the short-term groupings remain in the same grouping on the basis of the extended sequence analysis. There is also, for example, some movement from 'Entering the Labour Market' into the 'Potential Cause for Concern', where this was not evident in the earlier sequences; similarly some individuals initially classified as 'Potential Cause for Concern' have seen a recovery by this later time period, but overall there is a strong correlation between the two sets of groupings. We should also note that the 'Potential Cause for Concern' category grows primarily from individuals that were previously characterised as being 'Entering the Labour Market' and very few from the 'Accumulating Human Capital' grouping. In the BCS, the picture is much the same, except for the much reduced size of the missing category, as discussed in the introduction to this section.

<<< Table 6 around here >>>

<<< Table 7 around here >>>

What do we learn from this? Those who are a potential cause for concern in the earlier analysis are likely still to be considered a potential cause for concern on the basis of the longer run analysis: in the NCDS 85.9% of those deemed to be Potential Cause for Concern on the early basis (and for whom we can derive a longer run grouping) are placed in this group over the longer term; in the BCS the comparable figure is 70.0%. In addition, as one might expect, the longer analysis also picks up an additional number of cases that we deem to be a potential cause for concern, on the basis of their trajectories post-29 months.

[8] In addition, we carried out the same analysis over the whole time period (i.e. both the initial 29 months and the following 69 months) and achieved similar results to those reported later in this section.

However, we next explore whether this changes the risks of various observable characteristics associated with being a potential cause for concern.

Reassuringly, we find a fairly similar pattern in the odds of being a potential cause for concern in this analysis as we did in the 29-month analysis, although there are unexpected or surprisingly insignificant results associated with a few characteristics for the NCDS. Nevertheless, we conclude that this suggests that while the sequence and cluster analyses themselves do not necessarily pick up all the individuals who are a potential cause for concern in this shorter timeframe, our shorter-run analysis nevertheless identifies the observable groups that are likely to be at greater risk.

<<< Table 8 around here >>>

## 6. Conclusions

In this paper we have used sequence analysis in order to analyse young people's transitions into the labour market and how these have changed over the past forty years or so. The advantage of sequence analysis is that it allows us to consider young people's transition patterns as a whole, rather than concentrating on specific individual transitions and their timings. Through this approach, we have shed new light on the large structural changes in young people's entry to the labour market, highlighting the increasing prevalence of transitions that may give us some cause for concern as well as how the groups that experience these difficult transitions have changed in some ways and remained similar in others over our period of analysis.

Unsurprisingly, we find a substantial shift away from early labour market entry towards gaining significant amounts of additional education or training before entering a job. However, in addition we have documented a rise in the proportion of successive cohorts that experience a transition that seems to be a 'Potential Cause for Concern', with prolonged or numerous spells not in education, employment or training. This group has grown in size from 5% of the sample born in 1958 to 12% of the sample born in 1990, with pretty much all of this growth concentrated between the 1980- and 1990-born cohorts.

Focussing on the 'Potential Cause for Concern' group, there are two particularly striking results. First, females have gone from being more likely than males to be members of this group in early cohorts to being less likely by the more recent cohorts. One reason for this is likely to be a decline in the proportion of young women who choose to move quickly from education into an extended period of inactivity associated with homemaking or starting a family. Alongside this we find that individuals who move from education into long-term inactivity have become more likely to remain in education for two additional years (leaving education at age 18, rather than age 16) before entering inactivity.

Second, we find that young people from a non-white ethnic background go from being more likely than whites to experience a transition that is a 'Potential Cause for Concern' to being less likely. Across this period the non-white population of England has grown significantly in size, has diversified and has become more established. We suspect that all three of these

facts have contributed to the relative improvement in the probability that individuals of non-white ethnicity experience transitions likely to be precursors of future economic prosperity.

In addition, we find that socioeconomic status, as captured through a combination of indicators, remains a powerful predictor of young people's chances of experiencing a transition that is a 'Potential Cause for Concern'. This is unsurprising, but underlines that it is among those from disadvantaged backgrounds where there has been the greatest increase in difficult transitions.

Lastly, we assess the extent to which the patterns seen in these early years predict longer-term outcomes. The fact that we find a high degree of correlation suggests that those likely to face ongoing difficulties in the labour market are often identifiable at a very early stage. This points to the importance of early transitions.

**Table 1. Descriptive statistics of each cohort**

| | NCDS | BCS | YCS | LSYPE |
|---|---|---|---|---|
| *N* | 8,356 | 9,518 | 8,682 | 9,347 |
| *Male* | 0.57 | 0.49 | 0.51 | 0.48 |
| *Non-White* | 0.01 | 0.02 | 0.09 | 0.14 |
| *Single parent family* | 0.07 | 0.04 | 0.15 | 0.25 |
| *Parent has A Levels (no degree)* | 0.11 | 0.05 | 0.07 | 0.22 |
| *Parent has a degree* | 0.01 | 0.05 | 0.07 | 0.17 |
| *Home owner occupied* | 0.31 | 0.31 | 0.80 | 0.74 |
| *Home socially rented* | 0.42 | 0.05 | 0.15 | 0.19 |
| *Living in workless household* | 0.06 | 0.03 | 0.08 | 0.13 |
| *Ever NEET?* | 0.15 | 0.10 | 0.25 | 0.26 |

**Notes:** NCDS results weighted using author's own attrition weighting scheme. No weights applied to BCS analysis, as number excluded due to attrition was too small to model. YCS and LSYPE analysis weighted using dataset-provided attrition weights.

**Table 2. Descriptive statistics for identified groups within each cohort**

| NCDS | Entering the Labour Market | Accumulating Human Capital | Potential cause for concern | Overall |
|---|---|---|---|---|
| *N* | 7,110 | 852 | 394 | 8,356 |
| *Proportion* | 0.91 | 0.04 | 0.05 | 1.00 |
| *Male* | 0.59 | 0.57 | 0.23 | 0.57 |
| *Non-White* | 0.01 | 0.01 | 0.02 | 0.01 |
| *Single parent family* | 0.07 | 0.04 | 0.11 | 0.07 |
| *Parent has A Levels (no degree)* | 0.10 | 0.28 | 0.06 | 0.11 |
| *Parent has a degree* | 0.01 | 0.08 | 0.00 | 0.01 |
| *Home owner occupied* | 0.31 | 0.51 | 0.13 | 0.31 |
| *Home socially rented* | 0.43 | 0.14 | 0.53 | 0.42 |
| *Living in workless household* | 0.06 | 0.03 | 0.10 | 0.06 |
| **BCS** | **Entering the Labour Market** | **Accumulating Human Capital** | **Potential cause for concern** | **Overall** |
| *N* | 6867 | 2282 | 369 | 9518 |
| *Proportion* | 0.72 | 0.24 | 0.04 | 1.00 |
| *Male* | 0.50 | 0.48 | 0.34 | 0.49 |
| *Non-White* | 0.02 | 0.05 | 0.04 | 0.02 |
| *Single parent family* | 0.03 | 0.04 | 0.05 | 0.04 |
| *Parent has A Levels (no degree)* | 0.04 | 0.10 | 0.01 | 0.05 |
| *Parent has a degree* | 0.03 | 0.14 | 0.01 | 0.05 |
| *Home owner occupied* | 0.28 | 0.44 | 0.11 | 0.31 |
| *Home socially rented* | 0.05 | 0.03 | 0.10 | 0.05 |
| *Living in workless household* | 0.03 | 0.03 | 0.07 | 0.03 |
| **YCS** | **Entering the Labour Market** | **Accumulating Human Capital** | **Potential cause for concern** | **Overall** |
| *N* | 2,956 | 5,408 | 318 | 8,682 |
| *Proportion* | 0.40 | 0.55 | 0.05 | 1.00 |
| *Male* | 0.49 | 0.51 | 0.55 | 0.51 |
| *Non-White* | 0.03 | 0.13 | 0.07 | 0.09 |
| *Single parent family* | 0.16 | 0.14 | 0.17 | 0.15 |
| *Parent has A Levels (no degree)* | 0.04 | 0.09 | 0.07 | 0.07 |
| *Parent has a degree* | 0.03 | 0.09 | 0.08 | 0.07 |
| *Home owner occupied* | 0.75 | 0.85 | 0.64 | 0.80 |
| *Home socially rented* | 0.20 | 0.10 | 0.28 | 0.15 |
| *Living in workless household* | 0.08 | 0.08 | 0.15 | 0.08 |
| **LSYPE** | **Entering the Labour Market** | **Accumulating Human Capital** | **Potential cause for concern** | **Overall** |
| *N* | 2,994 | 5,399 | 954 | 9,347 |
| *Proportion* | 0.37 | 0.51 | 0.12 | 1.00 |
| *Male* | 0.48 | 0.48 | 0.52 | 0.48 |
| *Non-White* | 0.07 | 0.19 | 0.13 | 0.14 |
| *Single parent family* | 0.27 | 0.22 | 0.36 | 0.25 |
| *Parent has A Levels (no degree)* | 0.20 | 0.25 | 0.16 | 0.22 |
| *Parent has a degree* | 0.12 | 0.22 | 0.12 | 0.17 |
| *Home owner occupied* | 0.74 | 0.79 | 0.52 | 0.74 |
| *Home socially rented* | 0.19 | 0.15 | 0.36 | 0.19 |
| *Living in workless household* | 0.10 | 0.13 | 0.26 | 0.13 |

**Notes:** NCDS results weighted using author's own attrition weighting scheme. No weights applied to BCS analysis, as number excluded due to attrition was too small to model. YCS and LSYPE analysis weighted using dataset-provided attrition weights.

**Table 3. Estimated odds ratios of an individual's membership of a cluster in the "potential cause for concern" group from cohort-specific logistic regression models**

| | NCDS | BCS | YCS | LSYPE |
|---|---|---|---|---|
| *Non-White* | 2.880** | 1.560 | 0.746 | 0.697*** |
| | (2.477) | (1.447) | (-1.048) | (-3.271) |
| *Male* | 0.209*** | 0.483*** | 1.206 | 1.210** |
| | (-9.663) | (-6.455) | (1.309) | (2.341) |
| *Workless Household* | 1.457 | 1.727** | 1.650** | 1.753*** |
| | (1.183) | (2.410) | (2.143) | (4.864) |
| *Lone parent* | 1.268 | 1.737* | 0.846 | 1.250** |
| | (0.739) | (1.885) | (-0.807) | (2.271) |
| *Socially Rented* | 0.955 | 2.179* | 1.660 | 1.209 |
| | (-0.185) | (1.810) | (1.385) | (1.165) |
| *Owner Occupier* | 0.363*** | 0.489* | 0.736 | 0.508*** |
| | (-3.446) | (-1.697) | (-0.923) | (-4.361) |
| *Parental A-Levels* | 0.795 | 0.224** | 0.768 | 0.819* |
| | (-0.798) | (-2.549) | (-0.760) | (-1.824) |
| *Parental Degree* | 0.289* | 0.362** | 2.087** | 0.957 |
| | (-1.691) | (-2.217) | (2.225) | (-0.365) |
| *N* | 8356 | 9505 | 8682 | 9144 |

**Notes:** Models also include regional dummy variables and missing variable dummies for the variables above. NCDS results weighted using author's own attrition weighting scheme. No weights applied to BCS analysis, as number excluded due to attrition was too small to model. YCS and LSYPE analysis weighted using dataset-provided attrition weights. T statistics reported in parentheses. Stars indicate statistical significance: * p=0.10; ** p=0.05; *** p=0.01.

**Table 4. Predicted probability of membership of a cluster in the "potential cause for concern", by SES, gender and ethnicity in four cohorts**

| White Male | Low SES | High SES | Overall | N |
|---|---|---|---|---|
| *NCDS* | 2.8 | 0.3 | 1.5 | 4,338 |
| *BCS* | 6 | 0.5 | 1.3 | 4,564 |
| *YCS* | 11.6 | 5.9 | 4.1 | 3,429 |
| *LSYPE* | 29.5 | 10.7 | 11.8 | 3,161 |
| **White Female** | **Low SES** | **High SES** | **Overall** | **N** |
| *NCDS* | 12.3 | 1.3 | 6.8 | 3,844 |
| *BCS* | 11.6 | 1.1 | 2.6 | 4,706 |
| *YCS* | 9.8 | 4.9 | 3.4 | 4,555 |
| *LSYPE* | 25.7 | 9 | 10 | 3,201 |
| **Non-White Male** | **Low SES** | **High SES** | **Overall** | **N** |
| *NCDS* | 7.8 | 0.8 | 4.2 | 103 |
| *BCS* | 9 | 0.8 | 2 | 113 |
| *YCS* | 8.9 | 4.5 | 3.1 | 283 |
| *LSYPE* | 22.6 | 7.7 | 8.5 | 1,291 |
| **Non-White Female** | **Low SES** | **High SES** | **Overall** | **N** |
| *NCDS* | 28.8 | 3.7 | 17.3 | 71 |
| *BCS* | 9 | 1.6 | 4 | 122 |
| *YCS* | 7.5 | 3.7 | 2.6 | 415 |
| *LSYPE* | 19.5 | 6.5 | 7.2 | 1,491 |
| **Overall** | **Low SES** | **High SES** | **Overall** | **N** |
| *NCDS* | 5.5 | 0.5 | 2.9 | 8,356 |
| *BCS* | 8.5 | 0.8 | 1.8 | 9,505 |
| *YCS* | 10.5 | 5.3 | 3.7 | 8,682 |
| *LSYPE* | 26.6 | 9.4 | 10.4 | 9,144 |

Notes: Predicted probabilities from underlying regression models reported in Table 3. Models also include regional dummy variables and missing variable dummies for the variables above. NCDS results weighted using author's own attrition weighting scheme. No weights applied to BCS analysis as number excluded due to attrition was too small to model. YCS and LSYPE analysis weighted using dataset-provided attrition weights. 'High SES' individual is from a two parent household, where at least one parent works, at least one parent holds a degree, and their house is owner occupied. 'Low SES' individual is from a lone parent, workless household, where the parent's highest qualification is below A-Level and their home is socially rented. 'Overall' are predictions based on the complete sample, not a weighted average of the 'Low SES' and 'High SES' predictions.

**Table 5. Predicted probability of membership of a cluster in the "potential cause for concern", by SES, gender and ethnicity for cohort born in 1989/90 and for same cohort assuming same influence of characteristics as that seen for cohort born in 1958.**

| **White Male** | **Low SES** | **High SES** | **Overall** | **N** |
|---|---|---|---|---|
| *NCDS* | 2.8 | 0.3 | 1.5 | 4,338 |
| *NCDS associations/LSYPE cohort* | 2.5 | 0.4 | 0.8 | 3,161 |
| *LSYPE* | 29.5 | 10.7 | 11.8 | 3,161 |
| **White Female** | **Low SES** | **High SES** | **Overall** | **N** |
| *NCDS* | 12.3 | 1.3 | 6.8 | 3,844 |
| *NCDS associations/LSYPE cohort* | 13.0 | 2.0 | 4.2 | 3,201 |
| *LSYPE* | 25.7 | 9.0 | 10.0 | 3,201 |
| **Non-White Male** | **Low SES** | **High SES** | **Overall** | **N** |
| *NCDS* | 7.8 | 0.8 | 4.2 | 103 |
| *NCDS associations/LSYPE cohort* | 7.9 | 1.2 | 2.4 | 1,291 |
| *LSYPE* | 22.6 | 7.7 | 8.5 | 1,291 |
| **Non-White Female** | **Low SES** | **High SES** | **Overall** | **N** |
| *NCDS* | 28.8 | 3.7 | 17.3 | 71 |
| *NCDS associations/LSYPE cohort* | 7.9 | 6.3 | 12.6 | 1,491 |
| *LSYPE* | 19.5 | 6.5 | 7.2 | 1,491 |
| **Overall** | **Low SES** | **High SES** | **Overall** | **N** |
| *NCDS* | 5.5 | 0.5 | 2.9 | 8,356 |
| *NCDS associations/LSYPE cohort* | 7.0 | 1.0 | 2.1 | 9,144 |
| *LSYPE* | 26.6 | 9.4 | 10.4 | 9,144 |

Notes: Predicted probabilities from underlying regression models reported in Table 3. Models also include regional dummy variables and missing variable dummies for the variables above. Missing value dummies are set to zero. NCDS results weighted using author's own attrition weighting scheme. LSYPE analysis weighted using dataset-provided attrition weights. 'High SES' individual is from a two parent household, where at least one parent works, at least one parent holds a degree, and their house is owner occupied. 'Low SES' individual is from a lone parent, workless household, where the parent's highest qualification is below A-Level and their home is socially rented. 'Overall' are predictions based on the complete sample, not a weighted average of the 'Low SES' and 'High SES' predictions.

**Table 6. NCDS: Cross-tabulation of groupings on basis of 16-18 sequence analysis and of groupings on basis of 18-25 sequence analysis**

| | 18-24 Groupings | | | | |
|---|---|---|---|---|---|
| **16-18 Groupings** | **ELM** | **AHC** | **PCC** | **Missing** | **Total (freq.)** |
| ELM | 61.9 | 1.1 | 12.4 | 24.6 | 7,110 |
| AHC | 13.7 | 41.7 | 2.1 | 42.5 | 852 |
| PCC | 8.6 | 0.5 | 55.6 | 35.3 | 394 |
| Missing | 75.0 | 0.0 | 25.0 | 0.0 | 16 |
| Total | 54.6 | 5.2 | 13.4 | 26.9 | 8,372 |

Notes: ELM = Entering the Labour Market; AHC = Accumulating Human Capital; PCC = Potential Cause for Concern. Reporting row proportions, except for the final (total) column, which reports frequencies.

**Table 7. BCS: Cross-tabulation of groupings on basis of 16-18 sequence analysis and of groupings on basis of 18-25 sequence analysis**

| | 18-24 Groupings | | | | |
|---|---|---|---|---|---|
| **16-18 Groupings** | **ELM** | **AHC** | **PCC** | **Missing** | **Total (freq.)** |
| ELM | 81.3 | 5.7 | 12.2 | 0.9 | 6,867 |
| AHC | 6.8 | 87.2 | 4.4 | 1.6 | 2,282 |
| PCC | 23.0 | 6.8 | 69.4 | 0.8 | 369 |
| Total | 61.1 | 25.3 | 12.6 | 1.0 | 9,518 |

Notes: ELM = Entering the Labour Market; AHC = Accumulating Human Capital; PCC = Potential Cause for Concern. Reporting row proportions, except for the final (total) column, which reports frequencies.

**Table 8. Estimated odds ratios of an individual's membership of a cluster in the "potential cause for concern" group from cohort-specific logistic regression models – Sequences to age 25**

| | NCDS | BCS |
|---|---|---|
| *Non-White* | 1.551 | 1.184 |
| | (1.232) | (0.722) |
| *Male* | 0.143*** | 0.223*** |
| | (-13.951) | (-19.089) |
| *Workless Household* | 1.417 | 1.966*** |
| | (1.519) | (4.456) |
| *Lone parent* | 1.079 | 1.222 |
| | (0.375) | (1.042) |
| *Socially Rented* | 0.932 | 1.577* |
| | (-0.378) | (1.895) |
| *Owner Occupier* | 0.657** | 0.495*** |
| | (-2.118) | (-3.191) |
| *Parental A-Levels* | 0.648** | 0.533*** |
| | (-2.563) | (-3.037) |
| *Parental Degree* | 0.377** | 0.238*** |
| | (-2.303) | (-5.020) |
| *N* | 6122 | 8574 |

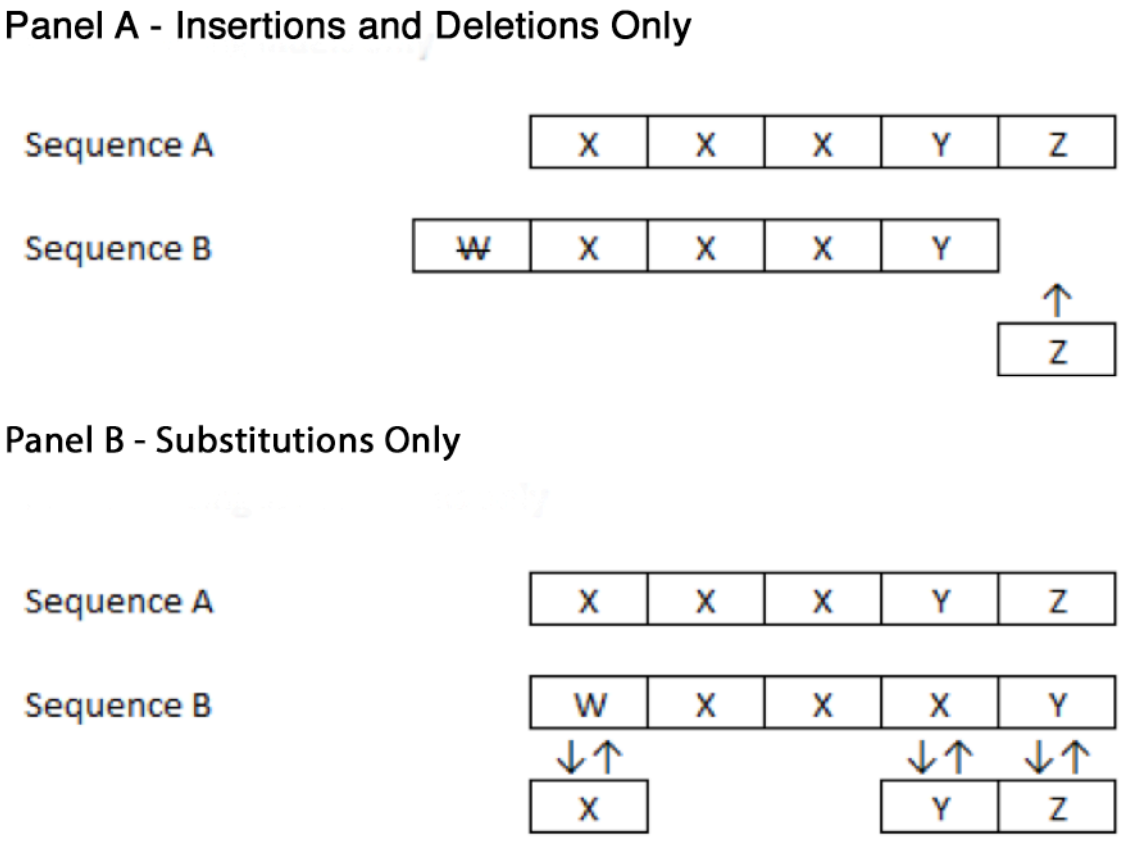
**Notes:** Models also include regional dummy variables and missing variable dummies for the variables above. NCDS results weighted using author's own attrition weighting scheme. No weights applied to BCS analysis, as number excluded due to attrition was too small to model. T statistics reported in parentheses. Stars indicate statistical significance: * p=0.10; ** p=0.05; *** p=0.01.

## Table 9. Average marginal effects from multinomial logit model of three cluster groupings.

| | NCDS-ELM | NCDS-AHC | NCDS-PCC | BCS-ELM | BCS-AHC | BCS-PCC | YCS-ELM | YCS-AHC | YCS-PCC | LSYPE-ELM | LSYPE-AHC | LSYPE-PCC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Non-White** | -0.049** | 0.002 | 0.047** | -0.191*** | 0.171*** | 0.019* | -0.351*** | 0.356*** | -0.005 | -0.211*** | 0.237*** | -0.025** |
| | (-2.309) | (0.197) | (2.500) | (-6.927) | (6.706) | (1.727) | (-11.423) | (12.017) | (-0.403) | (-12.609) | (15.145) | (-2.277) |
| **Male** | 0.071*** | -0.001 | -0.070*** | 0.023** | 0.004 | -0.027*** | -0.019 | 0.011 | 0.008 | -0.001 | -0.018 | 0.019** |
| | (8.609) | (-0.298) | (-9.036) | (2.505) | (0.455) | (-6.229) | (-1.534) | (0.860) | (1.302) | (-0.115) | (-1.556) | (2.328) |
| **W'less H/hold** | -0.023 | 0.006 | 0.017 | -0.018 | -0.002 | 0.020** | -0.012 | -0.010 | 0.022** | -0.127*** | 0.071*** | 0.056*** |
| | (-1.312) | (0.603) | (1.176) | (-0.730) | (-0.075) | (2.404) | (-0.455) | (-0.374) | (2.090) | (-6.204) | (3.488) | (4.855) |
| **Lone parent** | -0.003 | -0.008 | 0.011 | -0.037 | 0.017 | 0.020* | -0.022 | 0.029 | -0.007 | 0.056*** | -0.079*** | 0.023** |
| | (-0.153) | (-0.901) | (0.745) | (-1.537) | (0.748) | (1.880) | (-1.116) | (1.491) | (-0.798) | (3.950) | (-5.433) | (2.294) |
| **Socially Rented** | 0.014 | -0.012 | -0.002 | 0.081** | -0.109*** | 0.029* | 0.085** | -0.110*** | 0.024 | -0.002 | -0.017 | 0.019 |
| | (1.017) | (-1.445) | (-0.158) | (2.202) | (-3.159) | (1.817) | (2.170) | (-2.804) | (1.478) | (-0.087) | (-0.579) | (1.163) |
| **Owner Occupier** | 0.018 | 0.027*** | -0.045*** | -0.022 | 0.048* | -0.025* | -0.073** | 0.086** | -0.013 | -0.043* | 0.112*** | -0.069*** |
| | (1.200) | (3.689) | (-3.391) | (-0.750) | (1.765) | (-1.649) | (-2.097) | (2.517) | (-0.911) | (-1.686) | (4.260) | (-4.410) |
| **Parental A-Levels** | -0.039*** | 0.048*** | -0.009 | -0.150*** | 0.200*** | -0.049** | -0.072* | 0.081** | -0.009 | -0.030** | 0.050*** | -0.020* |
| | (-2.926) | (10.748) | (-0.743) | (-6.294) | (12.162) | (-2.304) | (-1.875) | (2.149) | (-0.612) | (-2.099) | (3.455) | (-1.801) |
| **Parental Degree** | -0.053 | 0.096*** | -0.043 | -0.255*** | 0.281*** | -0.026 | -0.152*** | 0.117*** | 0.035** | -0.133*** | 0.134*** | -0.000 |
| | (-1.625) | (15.987) | (-1.304) | (-12.044) | (17.954) | (-1.580) | (-4.054) | (3.188) | (2.490) | (-8.250) | (8.430) | (-0.029) |
| **N** | 8356 | 8356 | 8356 | 9518 | 9518 | 9518 | 8682 | 8682 | 8682 | 9144 | 9144 | 9144 |

**Notes:** ELM stands for "Entering the Labour Market", AHC stands for "Accumulating Human Capital" and PCC stands for "Potential Cause for Concern". Models also include regional dummy variables and missing variable dummies for the variables above. NCDS results weighted using author's own attrition weighting scheme. No weights applied to BCS analysis, as number excluded due to attrition was too small to model. YCS and LSYPE analysis weighted using dataset-provided attrition weights. T statistics reported in parentheses. Stars indicate statistical significance: * p=0.10; ** p=0.05; *** p=0.01.

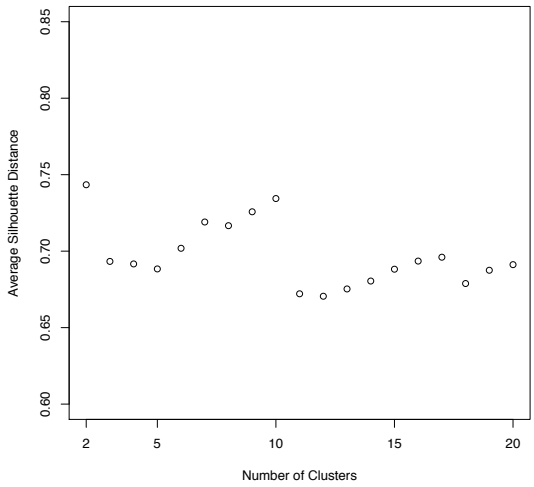
**Figure 1. Example of substitutions carried out to transform Sequence A into Sequence B**



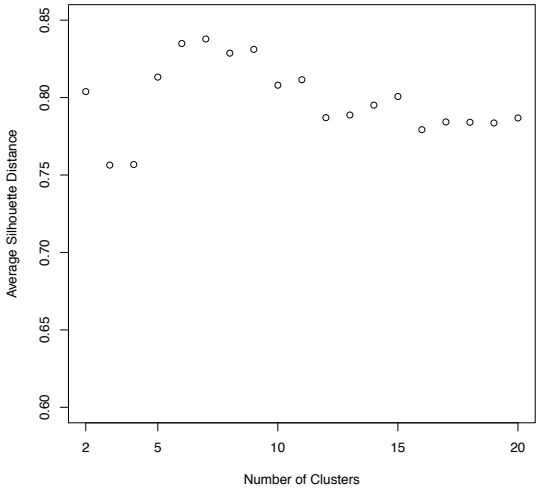

.

## Figure 2. Average silhouette distance of the cluster solutions

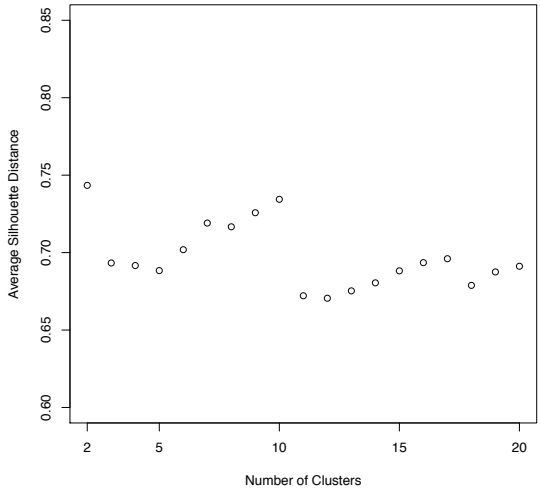
NCDS:



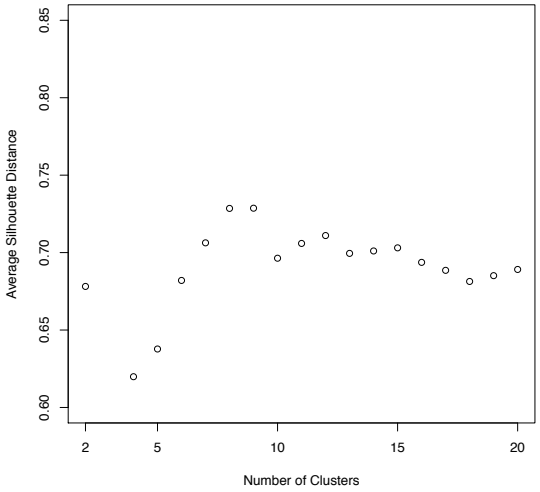

BCS:




YCS:




LSYPE:




Notes: Graphs report the average silhouette distance for each cluster solution from two to twenty clusters in each cohort. Graphs share common axes to allow comparison of the average silhouette distances in different datasets. Rule of thumb suggested by Kaufman and Rousseeuw (1990, p. 88) for "reasonable structure" is greater than 0.5 and for "strong structure" is greater than 0.7.

**Figure 3. Plots of young people's individual transitions in four cohorts between the September following their 16th birthday and 29 months later: clusters placed in the "entering the labour market" group**

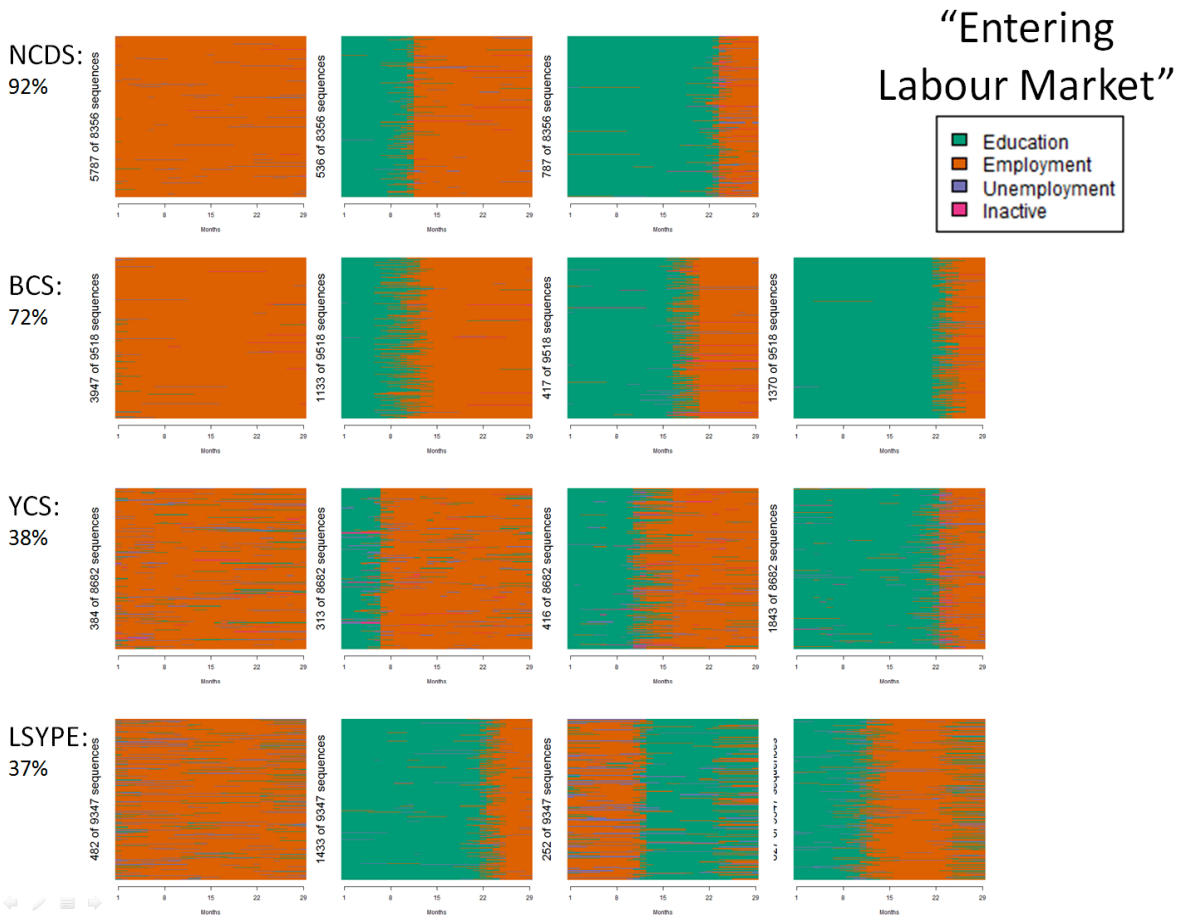


**Figure 4. Plots of young people's individual transitions in four cohorts between the September following their 16th birthday and 29 months later: clusters placed in the "accumulating human capital" group**

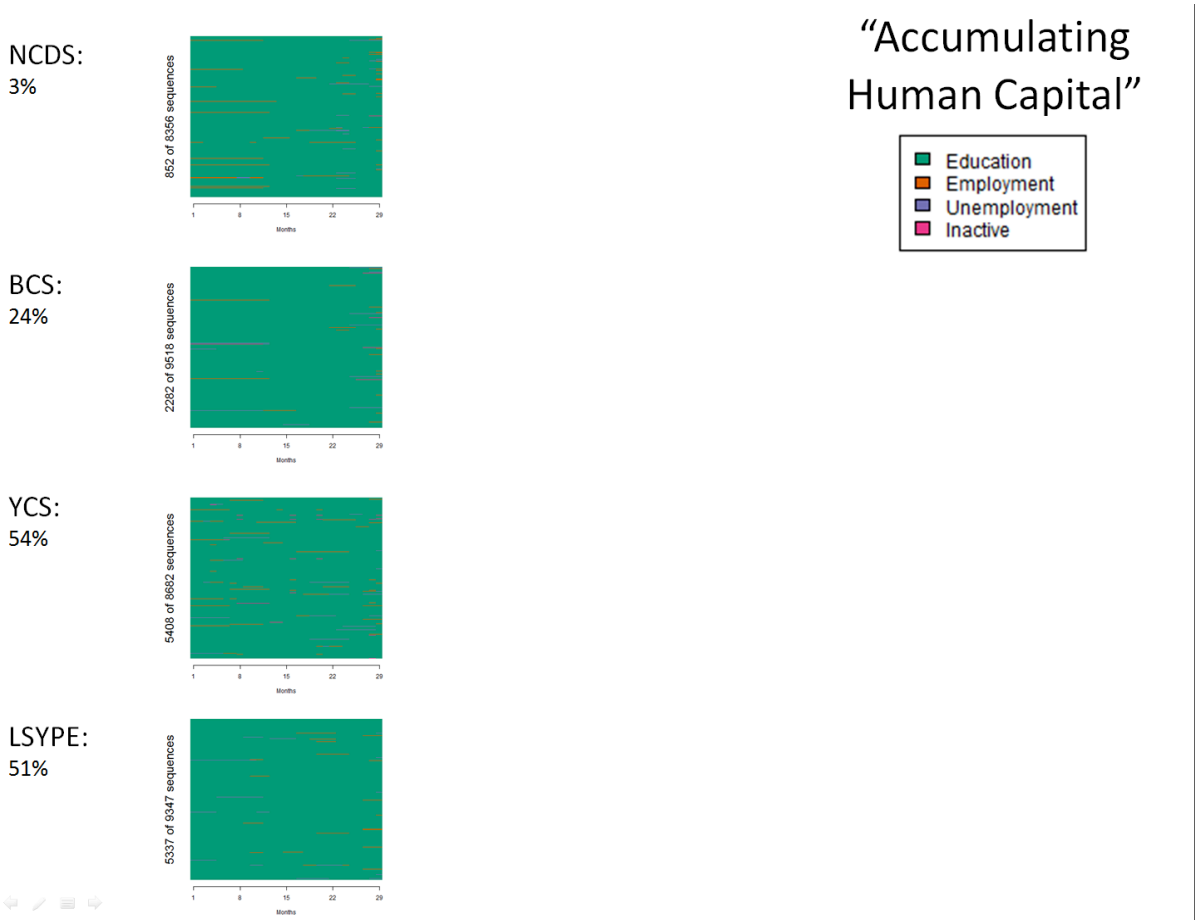
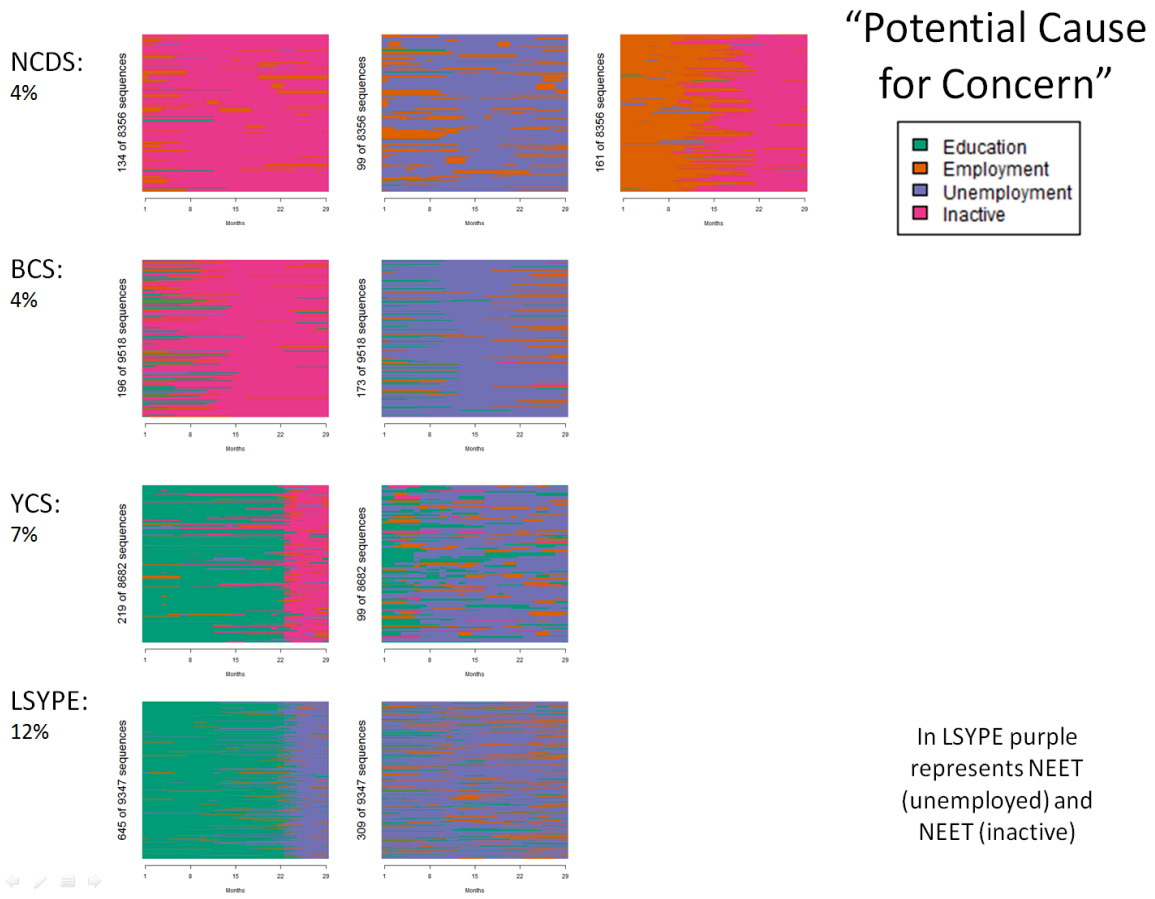


**Figure 5. Plots of young people's individual transitions in four cohorts between the September following their 16th birthday and 29 months later: clusters placed in the "potential cause for concern" group**



## Appendix A: Treating missing as its own state in sequence analysis

There are several methods for handling missing data in sequence analysis. One approach is to use imputation methods (Halpin, 2012, 2013). Another is to treat "missing" as a state in itself, situated equally distant from included substantive states, such that the sequence analysis algorithm is indifferent about substituting from a missing state to any of the others (Gabadinho, Ritschard, Studer, & Müller, 2011, pp. 55-61). As this method does not seem to have been extensively studied in the literature we do not use it as our primary method of analysis. Instead we test the robustness of our treatment of missing data in the main body of this paper (i.e. complete case analysis) to this alternative method, as stability between these two methods gives us additional confidence that our results are not being affected by a sample biased by excluding individuals with any missing outcomes data.

We continue to exclude individuals who attrit from the sample, using the same weighting strategy as that deployed in the main body of the paper. For the NCDS we now analyse 9,697 cases, rather than the 8,356 complete cases (i.e. 116% of the complete-case sample); for the BCS, the figures are 9,760 and 9,495 (103%); for the YCS, they are 9,265 and 8,682 (107%) and for the LSYPE we analyse 9,371 cases, rather than 9,347 cases (slightly over 100%).

We do not find that performing sequence analysis in this alternative way makes a substantive difference to our results, even in the NCDS cohort. In all cases other than the NCDS, the results do not even alter the general look of our preferred cluster solution: the additional cases with missing data are simply placed into the existing group that they most closely resemble, as is the intuition of this method.

In the case of the NCDS, we do see one alternative cluster formed from the missing cases: this appears to consist of a group of individuals who have missing data at the beginning of the period, followed by a move into employment. The missing data would appear to be due to the inconsistency between young people's reported school leaving age and the first other activity recorded, as discussed in section 2. Nevertheless, the fact that individuals within this cluster all move into employment after this period of missing data suggests it still fits naturally into this cohort's dominant group "Entering the Labour Market", rather than being a heterogeneous group in which the main similarity is that individuals experience missing data.

As a result of the predominant stability in the clusters that emerge, the remaining small changes make only very small differences to the relative size of the three groups, the composition of these groups, and the predictive power of young people's age 16 characteristics for the risk of experiencing a transition that is a "potential cause for concern". We view this stability as reassuring about the robustness of the findings of our main analysis.

## Appendix B: Results of multinomial logit regressions of group membership

In addition to the logistic regression models of young people's odds of being in a cluster deemed 'Potential Cause for Concern', rather than any other, we also estimate multinomial logit models of young people's odds of being in each of the three groups in which we place clusters. Since there is no natural baseline group in this setting, we report the average marginal effects of the various characteristics in the model on the probability of being in each of the three groups, rather than the more traditional approach of reporting relative risk ratios compared to one specific category. These results are reported in Table 9.

The results for being in a cluster that is 'Potential Cause for Concern' broadly replicate the results that we found from our simple logistic regression models, as should be the case. Indeed it would be a concern if this were not the case. As such, we again see females go from being more likely to be in this group to being less likely to be in the group, while those from non-white ethnic backgrounds follow the same trajectory. However, this model also allows us to discuss any changes in the predictors of membership of a cluster in which individuals are 'Entering the Labour Market' or 'Accumulating Human Capital'. We consider these in turn.

Throughout all cohorts, individuals from non-white ethnic backgrounds are statistically significantly less likely to be in clusters we categorise as 'Entering the Labour Market'. In general this association becomes larger over time, from a 5 percentage point reduction in probability in the NCDS to a 36 percentage point reduction in the YCS, although it weakens somewhat in the LSYPE back to a 24 percentage point reduction. In earlier cohorts, the NCDS and BCS, male participants are statistically significantly more likely to make transitions characterised as 'Entering the Labour Market' as female participants; this weakens between these two cohorts and the gap has closed to statistical insignificance by the YCS cohort and extremely close to zero in the LSYPE cohort.

It is noticeable, but hardly surprising, that parental education is statistically significantly associated with a participant experiencing an 'Accumulating Human Capital' transition in all four cohorts. In addition, in all cases the association between a parent having achieved A-Levels and membership of this grouping is weaker than the association between a parent having achieved a degree and group membership. The associations are similar in the NCDS, YCS and LSYPE: 5-8 percentage point increase associated with parental A-Levels and 10-13 percentage point increase associated with parental degree. However, the BCS is something of an outlier in this respect, with a significantly larger estimated association than for the other three cohorts. Throughout all the cohorts there is no statistically significant association between gender and experiencing a transitions characterised by 'Accumulating Human Capital', while coming from a non-white ethnic background and experiencing a transition of this type are statistically significantly positively correlated in all cohorts other than the NCDS.

## Bibliography


Abbott, A. (1995). Sequence Analysis: New Methods for Old Ideas. *Annual Review of Sociology, 21*, 93-113. http://dx.doi.org/10.1146/annurev.so.21.080195.000521

Andrews, M., & Bradley, S. (1997). Modelling the Transition from School and the Demand for Training in the United Kingdom. *Economica, 64*(255), 387-413. http://dx.doi.org/10.2307/2554833

Anyadike-Danes, M., & McVicar, D. (2005). You'll never walk alone: Childhood influences and male career path clusters. *Labour Economics, 12*, 511-530. http://dx.doi.org/10.1016/j.labeco.2005.05.008

Anyadike-Danes, M., & McVicar, D. (2010). My Brilliant Career: Characterizing the Early Labor Market Trajectories of British Women From Generation X. *Sociological Methods & Research, 38*(3), 482-512. http://dx.doi.org/10.1177/0049124110362968

Arulampalam, W. (2001). Is Unemployment Really Scarring? Effects of Unemployment Experiences on Wages. *The Economic Journal, 111*(475), F585-F606. http://dx.doi.org/10.1111/1468-0297.00664

Berrington, A. (2001). Transition to Adulthood in Britain *Transitions to Adulthood in Europe* (Vol. 10, pp. 67-102): Springer Netherlands. Retrieved from http://dx.doi.org/10.1007/978-94-015-9717-3_4

Caspi, A., Entner Wright, B. R., Moffitt, T. E., & Silva, P. A. (1998). Early Failure in the Labor Market: Childhood and Adolescent Predictors of Unemployment in the Transition to Adulthood. *American Sociological Review, 63*(3), 424-451.

Dorsett, R., & Lucchino, P. (2014). Explaining patterns in the school-to-work transition: An analysis using optimal matching. *Advances in Life Course Research, 22*(1-14). http://dx.doi.org/10.1016/j.alcr.2014.07.002

Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge, UK: Cambridge University Press

Elizinga, C. H., & Liefbroer, A. C. (2007). De-standardization of Family-Life Trajectories of Young Adults: A Cross-National Comparison Using Sequence Analysis. *European Journal of Population, 23*, 225-250. http://dx.doi.org/10.1007/s10680-007-9133-7

Fergusson, R., Pye, D., Esland, G., McLaughlin, E., & Muncie, J. (2000). Normalized dislocation and new subjectivities in post-16 markets for education and work. *Critical Social Policy, 20*(3), 283-305. http://dx.doi.org/10.1177/026101830002000302

Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software, 40*(4), 1-37. http://www.jstatsoft.org/v40/i04

Gabadinho, A., Ritschard, G., Studer, M., & Müller, N. S. (2011). Mining sequence data in R with the TraMineR package: A user's guide: University of Geneva, Switzerland. Retrieved from http://mephisto.unige.ch/pub/TraMineR/doc/TraMineR-Users-Guide.pdf

Gregg, P. (2001). The Impact of Youth Unemployment on Adult Unemployment in the NCDS. *The Economic Journal, 111*(475), F626-F653. http://dx.doi.org/10.1111/1468-0297.00666

Gregg, P., & Tominey, E. (2005). The wage scar from male youth unemployment. *Labour Economics, 12*, 487-509. http://dx.doi.org/10.1016/j.labeco.2005.05.004

Halpin, B. (2012). Multiple Imputation for Life-Course Sequence Data *Sociology Working Paper*. Limerick, Ireland: Department of Sociology. Retrieved from http://hdl.handle.net/10344/3639

Halpin, B. (2013). Imputing Sequence Data: Extensions to initial and terminal gaps, Stata's mi *Department of Sociology Working Paper Series*. Limerick, Ireland: University of Limerick. Retrieved from http://www.ul.ie/sociology/pubs/wp2013-01.pdf

Halpin, B., & Chan, T. W. (1998). Class Careers as Sequences: An Optimal Matching Analysis of Work-Life Histories. *European Sociological Review, 14*(2), 111-130. http://www.jstor.org/stable/522630

Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of Medoids *Statistical Data Analysis Based on the L1–Norm and Related Methods* (pp. 405-416): North-Holland

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Chichester: Wiley Inter-Science

Lesnard, L. (2006). Optimal Matching and Social Science *Série des Documents de Travail du CREST*. Paris, France: Institut National de la Statistique et des Etudes Economiques. Retrieved from https://halshs.archives-ouvertes.fr/halshs-00008122/

Martin, P., Schoon, I., & Ross, A. (2008). Beyond Transitions: Applying Optimal Matching Analysis to Life Course Research. *International Journal of Social Research Methodology, 11*(3), 1-21. http://dx.doi.org/10.1080/13645570701622025

Martin, P., & Wiggins, R. D. (2011). Optimal Matching Analysis. In M. Williams & W. P. Vogt (Eds.), *The SAGE Handbook of Innovation in Social Research Methods* (pp. 385-408). London: SAGE Publications Ltd.

OECD. (2008). *Jobs for Youth - United Kingdom*. Paris, France: Organisation for Economic Cooperation and Development

Paull, G. (2002). Biases in the Reporting of Labour Market Dynamics *IFS Working Paper Series*. London, UK: Institute for Fiscal Studies. Retrieved from http://dx.doi.org/10.1920/wp.ifs.2002.0210

Quintini, G., & Manfredi, T. (2009). Going Separate Ways? School-to-Work Transitions in the United States and Europe *OECD Social, Employment and Migration Working Paper*. Paris, France: Organisation for Economic Cooperation and Development

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65. http://dx.doi.org/10.1016/0377-0427(87)90125-7

Schoon, I., McCulloch, A., Joshi, H. E., Wiggins, R. D., & Bynner, J. (2001). Transitions from school to work in a changing social context. *Young, 9*(1), 4-22. http://dx.doi.org/10.1177/110330880100900102

Wu, L. L. (2000). Some Comments on Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect. *Sociological Methods and Research, 29*(1), 41-64. http://dx.doi.org/10.1177/0049124100029001003