ROCKWOOL Foundation Berlin
Institute for the Economy and the Future of Work (RFBerlin)

DISCUSSION PAPER SERIES 164/26

Predicting Labor Force Types

Rui Castro, Jiyoung Kim, Fabian Lange, Jérôme Larivière, Markus Poschke

www.rfberlin.com JUNE 2026

Predicting Labor Force Types

Authors

Rui Castro, Jiyoung Kim, Fabian Lange, Jérôme Larivière, Markus Poschke

Reference

JEL Codes: J0, J21, J64

Keywords: Labor force attachment, Clustering, Type prediction, Health, Early intervention

Recommended Citation: Rui Castro, Jiyoung Kim, Fabian Lange, Jérôme Larivière, Markus Poschke (2026): Predicting Labor Force Types. RFBerlin Discussion Paper No. 164/26

Imprint

RFBerlin
ROCKWOOL Foundation Berlin –
Institute for the Economy
and the Future of Work

Gormannstrasse 22, 10119 Berlin
Tel: +49 (0) 151 143 444 67
E-mail: info@rfberlin.com
Web: www.rfberlin.com

Predicting Labor Force Types*

Rui Castro Jiyoung Kim Fabian Lange Jérôme Larivière Markus Poschke

June 2026

Abstract

A small group of people accounts for a large majority of flows between labor market states and of spells in un- and non-employment. In this paper, we ask whether it is possible to identify those weakly attached to the labor market during their prime working-age years using information available early in their lives.

First, we use information on labor force transitions between ages 30 and 50 contained in the long panel provided by the NLSY79 to identify those weakly connected to the labor market during their prime age. To do so, we use k-means clustering on moments describing observed spells in employment, unemployment, and non-employment between 30 and 50. This points to a group of less attached individuals who are disproportionally female, less educated, and in poor health.

In a second step we predict, using information collected at various points before age 30 – which we do not use in clustering – whether individuals will turn out to belong to the weakly attached type in their prime age. We find that information from ages 22 to 29 allows predicting membership of the low-attachment group with high precision. Particularly influential is information on early labor market experiences and health.

The fact that we can predict weak and strong labor market attachment during prime working age using variables observed in individuals' twenties suggests the presence of persistent heterogeneity that shapes labor market experiences throughout the life cycle.

Keywords: Labor force attachment, Clustering, Type prediction, Health, Early intervention.
JEL Codes: J0, J21, J64.

*This work was supported by funding from the Social Sciences and Humanities Research Council of Canada (Award 435-2022-0136). We thank feedback from participants at a Rockwool Foundation Master Class in Berlin and of participants of the Society of Labor Economics Annual conference. The authors are entirely to blame for any errors and omissions. The authors used generative artificial intelligence (AI) tools to assist with selected coding tasks in this project, including drafting and refining code for data processing and analysis. All AI-generated code was reviewed, tested, and, where necessary, modified by the authors to ensure accuracy, reproducibility, and consistency with the study's methodology. The authors take full responsibility for the integrity of the code, results, and conclusions presented in this paper.

Introduction

Full employment remains the goal of much policy making. Increasing employment, preferably stable employment, promises to reduce poverty, increase tax revenue, and relieve stress on welfare states. Employment provides individuals with income and with increased opportunities to participate in society. Employment loss and repeated, extended spells of non-employment by contrast jeopardize these benefits.

Exposure to periods of non-employment thus constitutes a major risk individuals face. However, it remains poorly understood how large this risk is, and how it is distributed in the population. Recent work suggests that it is distributed very unevenly: a minority of the population accounts for a large majority of spells in unemployment and non-participation in the labor market, as well as for a large majority of transitions between employment and non-employment. For example, Morchio (2020) shows that 10% of the population account for two-thirds of time spent unemployed. Gregory, Menzio, and Wiczer (2025) find that 17% of the population that is at least partially attached to the labor market spends almost eight times more time in non-employment than the rest. Hall and Kudlyak (2022) and Ahn, Hobijn, and Sahin (2023) find similar patterns.[1]

These findings suggest that there is a subset of the population that has difficulties attaching to the labor market. It stands to reason that effectively designed labor market policies should target such individuals to help them achieve better outcomes. Yet, the existing literature provides hardly any guidance on how to identify these individuals. In this paper, we ask whether it is possible to identify members of this unattached group early in their lives, when careers are still malleable and interventions more likely to succeed. We answer this question with a prediction exercise that uses information on young adults to predict their attachment to the labor market during prime age (30-50). The results of the prediction exercise help us understand what events and characteristics prevent individuals from attaching to the labor market.[2]

We find that it is possible to predict attachment with high precision for a substantial share of the population. Besides the implications for targeting early life interventions, our findings also illustrate the level and persistence of heterogeneity in labor market experiences, as well as increasing our understanding of the characteristics of high- and low-attachment groups.

The first step of our analysis is to use the long panel provided by the National Longitudinal Survey of Youth 1979 (NLSY79) to identify those prime-age individuals who are very weakly connected to the labor market. We construct individual-level moments that describe respondents labor force histories based on the employment surveys completed between ages 30 and 50. Using clustering methods on these moments, we characterize the heterogeneity in labor force histories and identify the weakly attached. Those persistently employed or persistently out of work throughout the observation window can be trivially classified as either strongly or weakly attached to the labor market. The remaining individuals are classified through a k-means clustering procedure. We then inspect the characteristics of the different clusters and group them into two

[1] In a chapter of the Handbook of Labor Economics (Castro, Lange, and Poschke, 2025), three among us discuss this body of work in detail, report similar patterns from Canadian data, and explore ramifications for related topics, like recall and earnings losses from displacement.

[2] We do not have answers for the very different question of which policies might help attach individuals to the labor market. However, information on the primary predictors of low attachment to the labor market suggests which obstacles prevent individuals from attaching to the labor market and thus need addressing.

broad types - weakly and strongly attached - with preassigned individuals incorporated accordingly.

This procedure requires specifying the number of clusters, a choice that is to some extent arbitrary, but that can deliver different representations of the heterogeneity in the population. What matters for our purposes is that the group identified to be weakly attached is stable irrespective of the number of clusters. We find that as the number of clusters increases to 4 or beyond, we get fairly consistent assignment of individuals to the low type – the focus of our analysis. Thus, for parsimony, we restrict ourselves to four clusters for the k-means algorithm and six clusters overall, including the two pre-assigned groups of persistently employed or non-employed individuals.

Of these six clusters, two (almost) always work; a third cluster experiences few, mostly short non-employment spells and is likewise employed for most of the prime age years. 78% of the population belong to these highly attached clusters. The remainder of the population belongs to three clusters that spend 40% or more of their time non-employed. Such individuals are thus characterized by low attachment to the labor market. Two of these clusters experience very long non-employment spells and make up about half of the weakly attached population. The other half of the weakly attached population (11% of the total population) experience a sequence of short employment and non-employment spells.

We find that members of the low-attachment type are disproportionately female, non-white, and less educated. They are also much less healthy, both during their prime-age and, strikingly, also when young (22 to 29). Those who repeatedly report health limitations when under 30 have a very high propensity of low labor market attachment in their prime age years.

In a second step, we predict prime-age type using variables gathered prior to age 30 that were not used in clustering. We ask which variables allow us to predict attachment successfully, how much information is required, and how well we can predict future membership in the unattached group based on information available at different ages.

An important question is how to determine success in predicting group membership. Conventional statistical measures, in particular Pseudo-R^2 measures, although popular in the literature on discrete prediction problems, are hard to interpret and do not typically align with the questions policymakers face. We propose a highly stylized policy problem that allows us to determine the appropriate metric to use: precision (the fraction correctly predicted to be unattached) in the subset of the population most likely to belong to the unattached group. This policy problem presupposes the existence of an intervention that exclusively benefits low-attachment individuals. If the intervention is costly, treatment will be rationed to those most likely to benefit from the program. Thus, the intervention will target those most likely to be weakly attached later in life. Precision in predicting true type for those most likely to be weakly attached determines how many treated individuals actually benefit from the program, and thus is crucial for determining how large the program should be.

Overall, our results show that it is possible to predict with a high degree of precision who will be unattached during their prime working age using only information contained in the NLSY when individuals are young. When we use time-invariant demographics alone, we find that precision for the 5% of the population most likely to belong to the unattached type is 27% for men (28% for women). Adding contextual information available by age 22 raises precision to 55% (57%). Additional years of health and labor market

information in an individual's twenties further increase precision at the 5% level to around 70%. Delaying the intervention to collect more information during an individuals' twenties thus allows targeting the policy more precisely, but might come at the expense of intervening later in life, when benefits from the intervention might be lower.

Information on health status and labor market outcomes during an individual's twenties are the primary drivers of this information gain. The relative importance of these two predictors varies with the size of the target group: by age 29 and for the 1-2% most likely to be unattached, early-career health limitations are the most informative predictor for women, and the second most informative for men. When broadening the group to 5–10%, employment histories including occupations and industry become the most important predictor for both men and women, accounting for around 53% of predictive power.

To summarize, we find that it is possible to predict low attachment with high precision, in particular for the most at risk group, using information observed when individuals are in their twenties. The ability to predict who will be weakly attached allows targeting potential interventions. Our results demonstrate that this is indeed possible. In addition, our results highlight and support that weak labor market attachment is indeed a persistent feature of individuals across their adult life and they demonstrate that health is associated with it. They also inform how one thinks about risk in the labor market: clearly, it is unevenly distributed in the popuation and it is possible to predict who bears most of it. Our findings thus not only directly inform policy, but also have implications for a wide range of future work on labor market dynamics and labor market risk.

Related literature. This paper contributes to four literatures. It is most closely related to a recent literature on heterogeneity in labor force transitions. It also relates to a more methodological literature on clustering methods as a tool for understanding heterogeneity in the labor market and related areas. Third, it relates to a literature relating early experiences in the labor market and in life more generally to later labor market outcomes. Finally, it speaks to a policy-oriented literature on targeting interventions in active labor market policy.

A recent literature has studied heterogeneity in labor force transitions and has provided clear evidence on the presence of such heterogeneity. Core contributions in this literature include Morchio (2020), Shibata (2019), Hall and Kudlyak (2022), Ahn et al. (2023) and Gregory et al. (2025). This literature has been reviewed in Castro et al. (2025). This work has shown how conceptualizing heterogeneity in labor force transitions, typically in terms of a Hidden Markov Model, helps to understand both the cross-sectional distribution of labor market flows and aspects of the cyclicality of labor market states. This approach naturally connects to recent work in macroeconomics that emphasizes the participation margin and the importance of three-state worker flows for business-cycle fluctuations (Elsby et al., 2015; Krusell et al., 2017).

Our work advances this literature in several ways. First, we use the wealth of the information in the NLSY 1979 to contextualize the heterogeneity in labor force transitions in the population. This reveals important differences across gender and emphasizes that health plays an important role in shaping careers. In contrast to claims in the literature (Hall and Kudlyak, 2022; Gregory et al., 2025), we also show that

it is indeed possible to predict with substantial precision who will be weakly attached, particularly when concentrating on those in the population most at risk. We can do so thanks to our use of the precision in predicting true type in a part of the population, rather than using measures applying to the entire population such as the pseudo-R^2. One way to interpret this is to acknowledge that both statements are correct: it is possible to predict who will be weakly attached if focusing on the at-risk population, but not for the entire population.

To the best of our knowledge, this is also the only paper in this literature on labor force heterogeneity and types that connects prime-age labor market experiences to those early in life. This connection reveals the strong relationship between early-life health and labor market experiences and attachment, especially among those most at risk. Maybe surprisingly, our findings suggest that education, cognitive and non-cognitive skill measures as well as family background measures are relatively weak predictors of future labor force attachment. These findings can provide a crucial input to future analyses of the welfare implications of risk and optimal policy design.

We also contribute to methodological work on the suitability of type-based approaches for representing heterogeneity. Key contributions here are Bonhomme and Manresa (2015) and Bonhomme et al. (2022), who analyze how to use discrete heterogeneity – like clusters – as a dimension reduction device, even when underlying heterogeneity may not be discrete. We discuss the details of dimension reduction and how it affects the choice of the number of clusters in our setting. In particular, we discuss in detail how the description of heterogeneity in the labor market varies with the number of clusters. As in the analysis of health types by Borella et al. (2025), our focus is on characterizing the types themselves, and identifying the factors associated with them.

Third, our work connects to a large causal literature relating very specific early-life experiences to later labor market outcomes. Almond and Currie (2011a,b) and Almond et al. (2018) synthesize evidence that events before age five rival schooling in explaining adult outcomes, and Black et al. (2007) provide within-twin causal evidence that birth weight affects adult education and earnings. Work-limiting health conditions of the kind we measure here have also been documented to shape labor supply and earnings over the life cycle. Currie and Madrian (1999) review the literature on health and labor market outcomes in detail. They note a close link with participation, but also remark that research mostly focuses on older workers' health. A closely related literature – much of it based on the NLSY79 – documents substantial labor market returns to cognitive and non-cognitive skills measured in young adulthood (Heckman et al., 2006; Lindqvist and Vestman, 2011; Heckman and Kautz, 2012). The framework of Cunha and Heckman (2007), emphasizing self-productivity and dynamic complementarity in skill formation, rationalizes why early-life shocks persist throughout the life cycle.

A complementary literature documents that adverse labor market events in early adulthood also leave durable scars. Using the NLSY79, Kahn (2010) shows that men who graduate from college in a recession suffer persistent wage and occupational losses. Oreopoulos et al. (2012) and Schwandt and von Wachter (2019) extend this finding to broader samples and document larger scars for less advantaged entrants. Arulampalam (2001) and Gregg and Tominey (2005) establish wage scarring from unemployment spells experienced in young adulthood.

Our finding of persistent types, which are either already predictable in an individual's twenties or related to events occurring at that time, aligns with this work. Our analysis goes beyond the existing literature in that we consider a broad measure of labor market attachment as an outcome, rather than focusing specifically on wages. Moreover, the predictive power of early life events for prime age attachment that we document likely reflects a combination of the informative content of early life events about latent type, as well as their potential effect on later employment. While the causal literature just cited tends to focus exclusively on the latter, the broader picture we provide is important for assessing the full level of heterogeneity in the population.

A fourth, more policy-oriented literature on targeting in active labor market policy (ALMP) directly motivates our prediction exercise. Whereas the literatures above ask why workers do or do not attach to the labor market, this literature asks who would benefit most from a given intervention. Black et al. (2003) document that statistical profiling of reemployment services improves outcomes, and Behaghel et al. (2014) find substantial heterogeneity in returns to job-search counseling. Card et al. (2018) and Le Barbanchon et al. (2024) synthesize the broader ALMP evidence. Kitagawa and Tetenov (2018) provide a formal empirical welfare-maximization framework for choosing whom to treat. We contribute by quantifying how precisely prime-age labor-market type can be predicted in the twenties, and how prediction quality varies with the age at which targeting decisions are made.

The paper proceeds as follows. Section 1 introduces the NLSY79 and discusses how we select the analysis sample and how the employment histories are constructed. The k-means clustering algorithm and our procedure for selecting the number of clusters are discussed in Section 2. In Section 3, we describe how types differ and describe the relationship between attachment and health. Section 4 describes our main results on predicting types. Section 5 discusses how our results can help policy intervention, and concludes.

1 Data

We begin with an overview of the NLSY79 survey, followed by our sample selection criteria, and data preparation procedures. Appendix A provides more detail on our choices made in the process of selecting and preparing the analysis sample.

1.1 The NLSY79 Survey

The NLSY79 is a nationally representative longitudinal survey that began in 1979 with a sample of 12,686 individuals aged 14-22 at the time of the initial interview. The original NLSY79 sample consists of three main components: a cross-sectional sample of 6,111 respondents designed to be representative of the non-institutionalized civilian youth population, supplemental over-samples totaling 5,295 respondents that include Black, Hispanic, and economically disadvantaged white youth, and a military over-sample of 1,280 respondents enlisted in the military as of September 1978.

In 1994, the NLSY switched from an annual to a biannual survey thereafter. Data collection for this panel continues to this day. The respondents were born between 1957 and 1964, providing us with detailed

information about the labor market experiences of these birth cohorts through their youth and their prime age, which we define as 30 to 50. For our analysis, the survey's primary advantage lies in combining detailed information obtained during the initial rounds with comprehensive data on employment histories in later years. In each survey, respondents retrospectively report their labor force status and detailed information on up to five jobs for each week since the last interview. This retrospective approach, while subject to potential recall bias, generates an exceptionally detailed weekly employment history spanning several decades of respondents' working lives.

1.2 Sample Selection

We impose sample restrictions to ensure data quality and analytical coherence. First, we exclude respondents from the military over-sample and the over-sample of economically disadvantaged white youth, which were discontinued in 1984 and 1990 respectively. Since our analysis focuses on labor force attachment patterns during prime working years, these early discontinuations render these samples unsuitable for our purposes. This leaves us with 9,964 respondents from the cross-sectional sample and the Black and Hispanic over-samples.

Second, 460 or 4.6% of this group never report data after age 30. Of these, 128 passed away and the remainder attrite for reasons unknown. Third, we drop 1,091 respondents because they were not interviewed in five consecutive years between ages 30 and 50. This criterion ensures that we maintain reasonably complete labor force status histories during the prime working years that are central to our analysis. Again, mortality contributes to this group since 223 respondents passed away prior to 45, but most (868) attrite or fail to respond for long periods for reasons unknown. After applying these restrictions, our final analytical sample contains 8,413 respondents. Weighted by the sample weights provided by the NLSY79 to account for the complex sampling design, these 8,413 make up 77.8% of the union of the cross-sectional and Black and Hispanic over-samples.

In any long-run survey such as the NLSY79, attrition and non-response is a potential concern. Bick et al. (2024) investigate how much selective attrition by race, gender, and education has impacted the representativeness of the NLSY79 in the 40 years up until the 2020 wave. They also compare distributions of earnings, hours, and employment with those from the March CPS to evaluate how much attrition related to unobservable outcomes has affected the survey. Based on these comparisons they "conclude that, four decades on, labor market outcomes in the NLSY79 remain broadly representative of individuals in their birth cohort" (Bick et al. (2024), page 3). One of the reasons the NLSY79 does so well on this dimension is that it has an unusually high retention rate, probably because of continued efforts to contact the original respondents even after they have missed multiple rounds of interviews. Bick et al. (2024) report that after accounting for mortality, the participation rate in the 2020 survey (excluding the discontinued samples) still amounts to an astonishing 74.5 percent.[3] Even so, we adjust the sample weights to account for potential differential attrition that leads to the sample at age 22 being different from the sample used to cluster by

[3]Their work builds on and extends the work of MaCurdy et al. (1998) and Aughinbaugh et al. (2017). The latter also report that the NLSY79 is broadly representative even after attrition.

labor force attachment.[4]

1.3 Constructing Labor Force Status Histories

The NLSY79's detailed employment histories are based on respondents' retrospective reports in each survey round, which cover the period since the last survey the individual responded to. These histories provide weekly information on labor force status which can take the values employed, unemployed or out of the labor force. The employer identifiers further allow us to measure job-to-job transitions. We use these histories to construct the moments to describe individual heterogeneity in labor force histories observed during prime age (30-50).

Reported histories are not always complete, with some weeks missing or containing ambiguous labor force states. When a gap is shorter than 52 weeks, we impute the labor force status history during the gap. For this, we leverage the employer identifiers and the labor force history sequences immediately before and after each gap. For example, for gaps lasting less than 4 weeks, we examine the employment status and employer information on both sides of the gap. If the same employer appears before and after the gap, we assume the employment relationship continued across the missing period. If employers differ on either side of a gap or if the gap involves a transition between employment and non-employment, we use pattern matching with similar observed labor force status spells from the same individual and comparable respondents to impute the most likely employment status for each missing week. For gaps longer than a year, we retain the individual in the sample, but compute the moments used for clustering using only the periods excluding the gap. This approach eliminates minor gaps in employment histories while maintaining the integrity of the underlying employment patterns, allowing us to retain a large population for clustering.[5]

Our final analytical sample consists of 8,413 respondents. For 79.9% of these we have complete histories during their prime age years. An additional 14.9% have gaps of no more than one year, for which we have imputed the missing data. For an additional 4.1%, labor force status gaps do not extend beyond two years. Thus, the vast majority of our sample (98.9%) has complete or almost complete labor force histories for ages 30-50. These histories serve as the foundation for our analysis of labor force attachment types, allowing us to classify individuals based on their observed patterns of employment stability, job mobility, and labor force participation over two decades of their prime working years.

1.4 Prediction Variables

Once we assigned an individual to the attached or unattached type using information on their prime age labor market histories, we rely on variables observed prior to age 30 to predict which type they will belong to. We include demographic variables, information on parental background and education, skill measures, as well as information on health, incarceration and children.

Demographic variables include gender, race/ethnicity (Black, Hispanic, and other), and birth cohorts for

[4] To do so, we estimate a Probit predicting whether an individual observed at age 22 enters into the clustering sample. We then adjust the sampling weights using the inverse probability of attrition. Controls used for this are all variables observed at age 22 including race, gender, education, health, labor force participation, and cognitive and non-cognitive skill measures.

[5] Appendix A.2 describes this imputation process in more detail.

1957-1964. We measure geography with indicators for the Census region (North-East, North Central, South, and West), for rural vs urban residency, and for whether residing outside of, or in the central or non-centrals district of a metropolitan statistical area (MSA). For parental background, we use employment for both mother and father in 1979, parental education, and indicators for the presence of one or both of the the biological father and mother at age 14.

We measure education using enrollment and highest education credential by ages 22-29 with indicators for high school graduation, some college, and college graduation. We further employ a cognitive skill measure from the Armed Forces Qualification Test score administered in 1979, normed to yield age-adjusted percentile scores. Non-cognitive skills are proxied using the Rotter Locus of Control and the Rosenberg Self-Esteem variable. Social skills are proxied by four measures, self-reported sociability, retrospective sociability at six asked in 1985, the number of extracurricular clubs joined in high school as in Deming (2017), and high school sports participation.

Health plays an important role in our analysis. We utilize reports from the different survey waves for whether respondents report that a health condition limits the kind or amount of work they can perform, or prevents them from working altogether.

Finally, we have indicators for whether the individual was incarcerated at the time of each interview and for the number and age at which individuals had children.

In addition to the variables observed prior to age 30 and the labor force status variables observed between age 30 and 50, we also report contextual variables observed between age 30 and 50. These include the health variables described above, but also a more detailed set of variables describing physical and mental health at age 40 and 50. In addition, we use data on the number of children and whether the individual spent time in prison to provide further context on the lives of our respondents.

Finally, we also describe how types differ in their economic outcomes, in particular, wages, hours worked, and earnings. We adopt the approach of Bick et al. (2025) to handle wage outliers. At the lower end, wages below one half of the federal minimum wage are recoded to one half of the federal minimum wage. At the upper end, we assume that division bias from misreported hours gives rise to the top 0.1% of wages and therefore set hours, wages, and earnings for these observations to missing. In addition, we set to missing any observations with annual hours below 200. Real wages are constructed using the CPI and reported in 2022 dollars.

Summary statistics for the prediction and the contextual variables will be presented and discussed in Section 3 below.

2 Measuring Heterogeneity in the Labor Force

2.1 k-means clustering

We measure individual-level labor market types by applying the k-means clustering algorithm to our sample of prime age labor market histories. The clusters thus capture the heterogeneity with respect to transitions across labor force states in our data.

The algorithm maps each individual $i \in \{1, 2, \ldots, I\}$ to a cluster k, $k \in \{1, 2, \ldots, K\}$, given a set of individual observations of variables $\mathbf{x}_i = \{x_{i,1}, \ldots x_{i,N}\}$. In our case, these variables are individual-level moments we construct based on the observed labor force status histories. These moments are chosen to characterize the long labor force status histories obtained from the panel. k-means clustering thus requires the analyst to first decide on the variables to use for clustering, and second to specify the number of clusters K to use to summarize the heterogeneity in the population.

Once $\mathbf{x}_i$ and K have been specified, the algorithm minimizes the sum of within-cluster Euclidean distances from the centroid of the cluster by assigning each individual i to a cluster k. Define $\boldsymbol{\mu}_k$ as the average of $\mathbf{x}_i$ for all individuals assigned to cluster k. The objective function that the algorithm minimizes is then

$$\min_{k(i)} \sum_{k=1}^{K} \sum_{i \in k} (\mathbf{x}_i - \boldsymbol{\mu}_k)'(\mathbf{x}_i - \boldsymbol{\mu}_k), \tag{1}$$

where $k(i)$ is an assignment of individuals i to clusters k.

When clustering with K clusters, we refer to the clusters as $K.1, K.2, ..., K.K$. Thus, when we for example refer to cluster 4.3, it indicates that we allowed for four clusters in the algorithm, and are referring to the third of these four clusters. In addition to these clusters assigned by the k-means algorithm, we also have two preassigned clusters containing individuals that are either always working or almost always non-employed. We cannot include these in the clustering exercise as not all moments required can be constructed for them. We refer to these preassigned groups as $K.0$ and $K.K+1$. We sort clusters $\{K.0, K.1, ..., K.K, K.K+1\}$ in ascending order by the fraction of time spent non-employed.

2.2 Variables used to describe labor force histories

We use five variables ($N = 5$) to describe the dynamics of employment and non-employment for each individual. Indexed by n, they are for each individual, using the individual's prime age labor force history:

- $n \in \{1, 2\}$: Average durations of employment (E) and non-employment (NE) spells in the individual's history.
- $n \in \{3, 4\}$: Fraction of weeks spent out of the labor force (OLF) or unemployed (U).
- $n = 5$: Number of jobs relative to the number of quarters employed (1 quarter=13 weeks).

All moments are standardized as $x_{i,n}/\text{std}(x_{i,n})$, i.e. relative to the population dispersion.[6]

These moments capture different, complementary aspects of labor force attachment. Clearly, moments 3 and 4 are direct measures of non-attachment. In addition, the first moment captures the stability of employment spells, and the fifth the stability of individual employment relationships. The second moment helps to distinguish individuals with long non-employment spells from those with recurrent short spells. Jointly, these moments allow us to partition the population into a distinct set of clusters.[7]

[6] This is to avoid assigning a higher weight to some moments simply based on the units of measurement. In Table 1, we show each moment in its original units.

[7] The moments we use are very similar to those used by Gregory et al. (2025).

2.3 Preassigned Types

We can compute the full set of moments only for those who experience periods of both employment and non-employment. If instead an individual experiences only one status – for example, because they are employed in all periods, so that average non-employment duration cannot be computed – then we exclude them from the clustering exercise. Those always employed are instead assigned to cluster $K.0$. Those with cumulative total employment short of 2 years or with a non-employment spell exceeding 10 years are assigned to cluster $K.K+1$. The size of these preassigned groups is independent of the choice of how many clusters K to employ in the analysis. 15.9% of our (weighted) sample belongs to $K.0$ and 7.6% to $K.K+1$.[8]

Table 1 summarizes the moments we use for clustering, both in the whole population and among the clustered individuals, i.e. excluding the preassigned groups.

	Population	K.0 (15.9%)	Clustered (76.4%)	K.K+1 (7.6%)
Mean duration of E spells	366.3	1031.5	255.9	82.0
	(338.1)	(35.1)	(189.0)	(103.5)
Mean duration of NE spells	70.4	-	43.1	491.2
	(162.2)		(63.5)	(330.8)
Fraction of time spent OLF	15.8	-	12.8	78.0
	(24.6)		(17.6)	(17.8)
Fraction of time spent U	3.5	-	4.0	5.3
	(6.7)		(6.7)	(10.4)
Jobs per quarter spent in E	0.16	0.04	0.14	0.61
	(0.51)	(0.04)	(0.12)	(1.73)

Standard deviations in parenthesis. E is employment, NE is non-employment, OLF is out of the labor force, U is unemployment. Duration of spells are in weeks. K.0 and K.K+1 are pre-assgined types. K.0 do not have NE spells. K.K+1 either have longest NE spell longer than 10 years or have cumulative prime age employment less than 2 years.

Table 1: Clustering moments

2.4 The number of clusters K^*

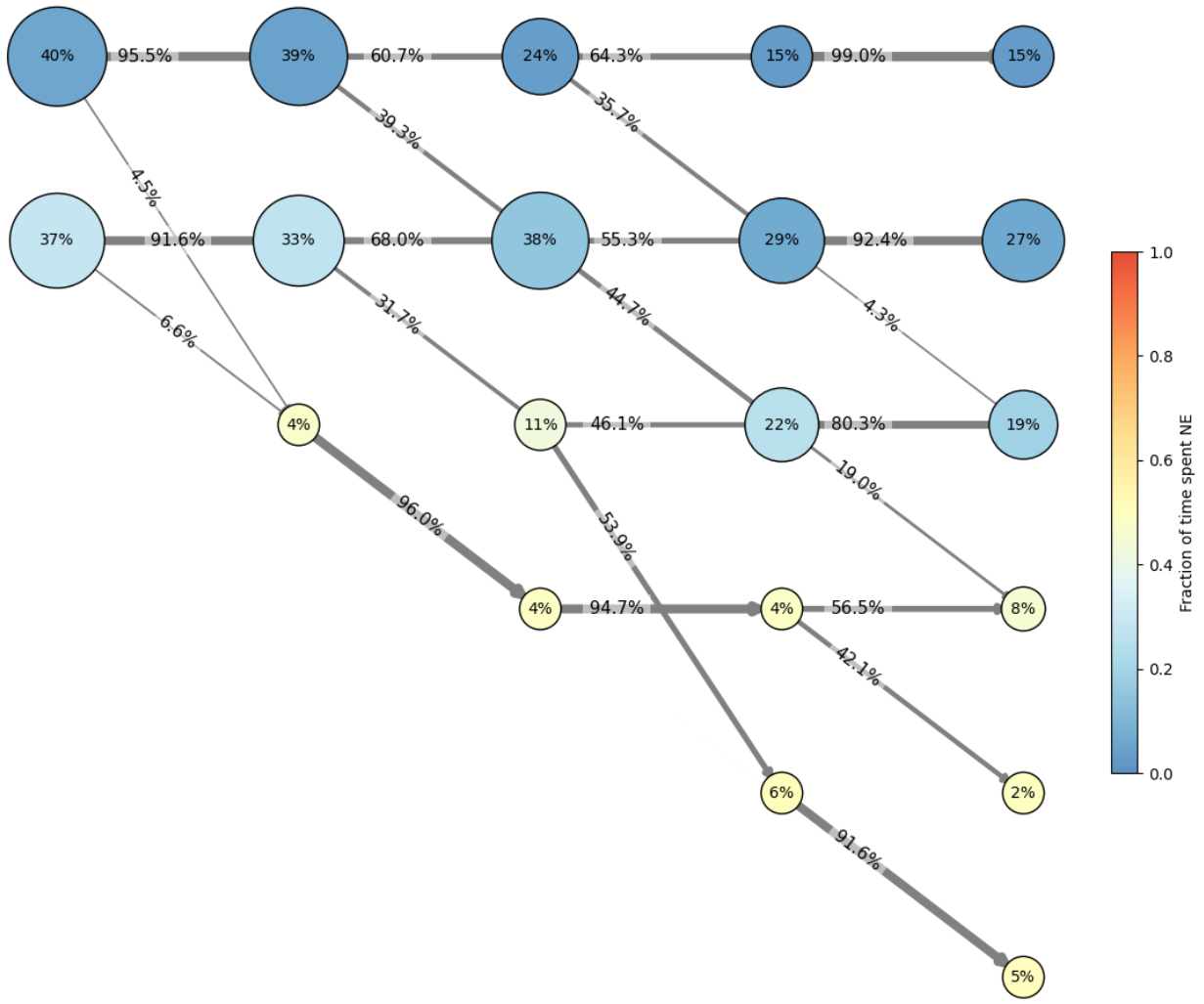
The next step is to choose the number of clusters K^*. Clearly, we need K^* to be such that we are adequately describing the heterogeneity in labor force histories in the data. As is standard in the literature, this needs to be balanced against the fact that, as K increases, we may be overfitting the data. Available methods for selecting the optimal number of clusters (e.g. Wang, 2010) are designed to guard against this problem.

Our specific application introduces an additional requirement for choosing K^*. Our ultimate goal is a more parsimonious binary classification into weak and strong attachment. As we will see, simply specifying two clusters does not achieve this. We therefore require a set of K^* clusters that (i) yields a subset of clusters that can be coherently labeled as low-attachment, and (ii) does so in a parsimonious manner, without imposing fine distinctions among individuals unrelated to the goal of identifying weak attachment.

To balance these objectives, we develop an informal procedure which lets the number of clusters increase until there is a clear, stable differentiation between low and high attachment types, and stops once additional

[8] We have experimented with different preassignment rules based on non-employment. Stricter rules result in individuals with significant attachment to be pre-assigned to the non-attached cluster. For details see Appendix B.1.

clusters only provide further differentiation within the high- and low-attachment groups. This process clearly involves researcher judgment. We therefore describe it in detail, to enable the reader to critically assess our choices. Our preferred specification turns out to feature 4 clusters (in addition to the 2 prespecified ones). As it turns out, statistical methods designed to penalize overfitting such as Wang's consistency criterion also favor four clusters.[9]



Note: Branches appear if $> 1\%$ of previous bubble and of the population

Figure 1: Assignment to clusters across number of clusters K

[9]Wang's consistency criterion splits the sample into a training and a validation half. The training half is further subdivided and the algorithm is run on both of these sub-samples producing two sets of cluster centroids. Wang's consistency criterion is obtained as the percentage of the validation sample assigned to the corresponding clusters. A higher number on this criterion suggests a stable ordering of clusters independent of sampling with well delineated clusters. Low values suggest that the cluster centroids move widely across the two training sub-samples and are highly unstable. Although it pursues a different objective compared to our approach for judging whether the algorithm describes economically relevant features of the data, both approaches yield the same optimal number of clusters ($K^* = 4$). See Appendix B.2.2. Further, the elbow method examines the change in the objective criterion of the K-means algorithm, the within sum of squares, as K increases. It looks for the K beyond which the decline in this objective slows markedly. This approach likewise favors $K = 4$.

Figure 1 describes the sequential clustering exercise, and Tables 2 and 3 report the means of the moments within each cluster.[10] Each column in the figure refers to the result from clustering with a given pre-specified K. Within each column, the clusters are sorted in ascending order by the fraction of time in non-employment, as also indicated by the shading of the bubbles. We refer to specific clusters as $K.i$, where i indexes clusters and K denotes the number of clusters used in each specific clustering exercise. The edges represent the shares of a cluster $K.i$ that belong to cluster $(K+1).j$. For example, when we allow for two clusters, 40% of the population belong to cluster 2.1. When we consider three clusters instead, 95.5% of those in 2.1 now belong to cluster 3.1.[11]

Just two clusters are not sufficient to capture the diversity in labor force histories in the sample. This becomes apparent once an additional third cluster is allowed for. When $K = 3$, we find an additional cluster 3.3 with very low attachment, characterized by very long durations (see Table 2). This group comprises 4.2% of the population. They work on average a little more than half their prime-age years and their spells, whether in or out of work, are very long, lasting almost 5 years on average. This cluster is very distinct from clusters 3.1 and 3.2, as well as from the originating 2.1 and 2.2. Since identifying it requires at least three clusters, we conclude that two clusters are not sufficient. Looking forward, we observe that the group captured by cluster 3.3 is very stable as the number of clusters increases further, with 96% of individuals in cluster 3.3 belonging to 4.4, and 95% of these belonging to 5.4. Thus, there is clearly a group of individuals with low attachment and very long durations. Once out of employment, this group does not tend to search for work but rather spends most of this time out of the labor force.

Once we allow for a fourth cluster, we find that the additional cluster (cluster 4.3), which accounts for 11% of the sample, is entirely drawn from cluster 3.2. This group spends nearly 40% of its time non-employed and is very distinct from 4.1 and 4.2, who both are employed around 90% or more of their prime-age years. It is also very distinct from 4.4 in that its spells, in employment or non-employment, are much shorter. With four clusters, we thus obtain a description of the labor market with two clearly distinct low-attachment groups, in addition to the preassigned group 4.5. The latter essentially do not work at all, while the other two groups spend 40-50% of time not working. Among those two, 4.3 has much shorter durations than 4.4. Three clusters were not sufficient for obtaining this clear partition, since cluster 3.2 contained large shares of both high- and low-attachment individuals.

As we move to five clusters, the clusters 5.1, 5.2 and 5.4 closely resemble 4.1, 4.2 and 4.4, respectively. In addition, we observe that 4.3 splits roughly in two, with one half establishing the new cluster 5.5. Its members are very similar to those of 4.3, spending about half their prime-age years non-employed, with short spells in both employment and non-employment. The other half of 4.3 merges with part of 4.2 into cluster 5.3. Table 4 shows the means of the clustering moments for these groups. Both groups drawn from 4.3 spend significantly more time non-employed than those in 5.3 who hail from 4.2 (33 and 51% as opposed to 23%). They also have significantly shorter employment durations, and a larger number of different jobs. Overall, we conclude that those in 5.3 drawn from 4.3 are better described as less attached to the labor force,

[10] To facilitate the interpretation, the tables display OLF/NE and U/NE as well as NE/total-observed-weeks separately, instead of simply reporting the two actual clustering moments OLF/total weeks and U/total weeks, which imply the three moments shown in the table.

[11] To keep the graph readable, we suppress outflows accounting for less than 1% of a given cluster.

while those in 5.3 who are drawn from 4.2 appear quite strongly attached, so that cluster 5.3 straddles the line between low and high attachment groups. It therefore does not help our analysis, and we thus stop at $K^* = 4$ clusters.

To further validate our judgment we note that, when moving to six clusters, we find again a very strong separation between high and low attachment clusters, with those in clusters 6.4 to 6.6 spending 45-50% of their prime age years non-employed. These groups make up about 15% of the total population. As the attached/unattached partition here closely resembles that obtained with four clusters, we prefer the more parsimonious specification with $K^* = 4$.

In summary, once we have 4 or more clusters, we find a significant population with low labor force attachment. The dividing line between being weakly or strong attached to the labor market appears quite clearly with 4 or 6 clusters, whereas one cluster straddles this line when we use 5 clusters. For parsimony, we focus on the case with $K^* = 4$ in the following, and refer to clusters 4.0 to 4.2 as the highly attached types and clusters 4.3 to 4.5 as the low-attachment ones.

Clustering design	$K=2$		$K=3$			$K=4$			
	2.1	2.2	3.1	3.2	3.3	4.1	4.2	4.3	4.4
Population share	39.9	36.5	38.7	33.5	4.2	23.6	38.1	10.6	4.1
Clustering moments									
Mean duration of E spells	386.1	113.8	376.5	108.7	317.5	473.2	169.0	66.4	305.0
Mean duration of NE spells	33.3	53.7	21.8	41.6	249.9	21.0	31.4	52.9	251.7
Fraction of time spent NE	6.0	28.7	4.8	27.0	47.4	3.1	14.9	42.3	48.5
OLF/NE	73.9	76.6	68.2	74.6	90.5	67.5	71.3	77.2	90.5
U/NE	26.1	23.4	31.8	25.4	9.5	32.5	28.7	22.8	9.5
Jobs per quarter spent in E	0.07	0.22	0.07	0.23	0.09	0.06	0.14	0.37	0.09

Duration of spells are in weeks. E is Employment, NE is non-employment, OLF is out of the labor force, U is unemployment. Individuals without NE spells, with NE spells exceeding 10 years, or with total E under 2 years are excluded from clustering. Population shares are in the whole population. Clustering designs are denoted by $K.i$, where K denotes the number of clusters, and i indexes clusters (ordered by Fraction of time spent NE). Moments refer to cluster means.

Table 2: Clustering moments for $K = 2, 3, 4$

Clustering design	$K=5$					$K=6$					
	5.1	5.2	5.3	5.4	5.5	6.1	6.2	6.3	6.4	6.5	6.6
Population share	15.2	29.5	22.1	3.9	5.7	15.1	27.4	19.2	7.9	1.6	5.3
Clustering moments											
Mean duration of E spells	550.0	246.4	109.8	302.7	55.8	552.4	250.0	108.0	169.8	405.5	55.6
Mean duration of NE spells	24.1	21.7	42.2	257.0	61.1	23.7	19.5	27.0	127.1	353.5	57.7
Fraction of time spent NE	2.9	7.0	25.3	48.6	50.6	2.8	6.4	19.5	45.0	49.5	50.2
OLF/NE	72.1	67.9	73.3	90.4	78.7	71.8	66.2	65.9	87.2	93.1	77.8
U/NE	27.9	32.1	26.7	9.6	21.3	28.2	33.8	34.1	12.8	6.9	22.2
Jobs per quarter spent in E	0.05	0.09	0.2	0.09	0.45	0.05	0.09	0.21	0.14	0.07	0.46

See Table 2.

Table 3: Clustering moments for $K = 5, 6$

Clustering design	$K.i = 4.3$		$K.i = 4.2$
	5.3	5.5	5.3
Population share	4.9	5.7	17.0
Clustering moments			
Mean duration of E spells	78.9	55.8	118.6
Mean duration of NE spells	43.3	61.1	41.0
Fraction of time spent NE	32.7	50.6	22.8
OLF/NE	74.4	78.7	72.4
U/NE	25.6	21.3	27.6
Jobs per quarter spent in E	0.28	0.45	0.18

See Table 2.

Table 4: Assignment for $K.i = 4.2$, 4.3 under $K = 5$

Clusters	4.0	4.1	4.2	4.3	4.4	4.5	Population
Population share	15.9	23.6	38.1	10.6	4.1	7.6	100
Clustering moments							
Mean duration of E spells	1031.5	473.2	169.0	66.4	305.0	82.0	366.3
Mean duration of NE spells	0.0	21.0	31.4	52.9	251.7	491.2	70.4
Fraction of time spent NE	0.0	3.1	14.9	42.3	48.5	83.3	19.2
OLF/NE	-	67.5	71.3	77.2	90.5	93.6	81.9
U/NE	-	32.5	28.7	22.8	9.5	6.4	18.1
Jobs per quarter spent in E	0.04	0.06	0.14	0.37	0.09	0.61	0.16

Note: See Table 2.

Table 5: Population distribution under preferred clustering design ($K^* = 4$)

2.5 Heterogeneity in Labor Force Dynamics

Having argued for using four clusters $K^* = 4$ plus the two preassigned groups, we briefly review how labor force dynamics differ across these clusters. For convenience, we reproduce the moments for each cluster in Table 5 with the average in the population in the last column.

Clusters are labeled in order of the fraction of time spent in non-employment. That fraction averages around 19% in the population. Three clusters (4.0 to 4.2) exhibit lower average time in non-employment than this, ranging from 0 in 4.0 to 14.9% in 4.2. The remaining clusters are characterized by much longer time in non-employment. Even in cluster 4.3, mean time in non-employment is almost three times that in cluster 4.2. It is due to this stark difference in the non-employment propensity that we will below refer to clusters 4.0 to 4.2 as strongly attached to the labor market, and clusters 4.3 to 4.5 as weakly attached.

Time non-employed can take the form of unemployment or can be spent out of the labor force. The share of non-employment time in unemployment – an indicator of attachment – decreases monotonically across across clusters, from about a third in cluster 4.1 to less than ten percent in clusters 4.4 and 4.5.

The average duration of non-employment spells also increases monotonically across clusters. It is much below the population average of 70 weeks in clusters 4.0 to 4.2, close to the average in 4.3, and several times higher than the average in 4.4 and 4.5. The picture is more nuanced for employment duration, which exceeds the population average of 366 weeks only in clusters 4.0 and 4.1. It is very short in 4.3 and 4.5, and intermediate in 4.2 and 4.4.

Overall, one could informally summarize population heterogeneity as follows. There are two very strongly attached clusters, 4.0 and 4.1, with negligible time spent out of work, long employment spells, and short unemployment spells. These two clusters account for about 40% of the population. Individuals in cluster 4.2 also spend little time out of work, but have shorter employment spells and more frequent and slightly longer non-employment spells. Adding this cluster, the strongly attached part of the population comes to 78%.

In addition, there are three clusters of only weakly attached individuals. By construction, those in cluster 4.5 work very little. Individuals in clusters 4.3 and 4.4 work a bit more than half the time, but exhibit very different dynamics. Those in cluster 4.3 experience unstable employment, as they go through a succession of short spells of employment and non-employment, and also change jobs frequently when employed. Those in cluster 4.4, in contrast, have very slow dynamics, with long employment durations when working, but also long non-employment spells. The two stable low-attachment clusters, 4.4 and 4.5, jointly account for 11.7% of the population. Adding the cluster with unstable dynamics, 4.3, brings the total size of the low-attachment groups to 22.3%.

Appendix D.1 shows the 95% confidence intervals for Tables 2, 4, and 5 obtained using the Bootstrap with 1,000 replications. These are generally quite tight. For our preferred specification with four clusters, none of the confidence intervals for the cluster centroids overlap.

3 Who Are the Unattached Types?

In this section, we describe how different contextual social variables obtained from the NLSY vary across the clusters identified above, beginning with demographics, education and skill measures as well as measures of fertility and incarceration, followed by measures of health, and subsequently wages and earnings. A key advantage of the NLSY over administrative data sources is that it significantly expands the set of contextual variables beyond basic demographic indicators.[12]

3.1 Demographics, education, incarceration, and fertility

Table 6 reports cluster means of the contextual variables characterizing the types during prime age, for women and men separately. From the full sample we notice clear gender differences in cluster composition, with the less attached labor force types much more likely to be female. Women make up between 63% and 71% of the low-attached clusters, but only 30-52% of the high-attached ones.

Next, focus on the variables in the top block of Panels A and B, which are observed early in life and remain fixed thereafter. Among these, we note that the information available in administrative data sets is typically restricted to very basic demographic indicators, namely race in addition to gender.

Several patterns emerge. First, the race gradient found in almost all other social data: Blacks and Hispanics are overrepresented among those weakly attached to the labor market during their prime age, a pattern driven almost entirely by men.

Second, we observe the expected gradient in education. High school dropouts account for 3-9% of the highly attached types among women, and 6-15% among men. These figures are about twice as large for the low-attachment types, 16-22% among women and 24-29% among men. High school graduates and those with some college are more evenly distributed across types. College graduates, by contrast, are much more likely to be of the highly attached type, particularly among men.

Third, there is a steep gradient in measures of cognitive and non-cognitive skills across types, particularly regarding cognitive skills among men, with those in the least attached groups scoring several standard deviations below those in the highly attached ones. Only social skills seem not to vary systematically across attached and unattached groups.

We now turn to the variables in the bottom block (starting with Jail) of each panel in Table 6, which are observed at different points in life. The table shows the fraction of individuals who report having been in jail at the time of the interview at least once between age 22 and 50. This share is on average 1% for women and 8% for men. However, it is about one third among men with low attachment to the labor market, compared to 0-6% for the highly attached.

Finally, fertility (number of children) at age 50 is also significantly related to labor force attachment, but with different signs by gender. Women in the low-attached clusters have had on average over two children, whereas this figure is only 1.45–1.95 for highly attached women. The opposite pattern emerges among men, with those in clusters 4.4 and 4.5 having a lower number of children compared to the highly attached. We thus see traditional gender roles asserting themselves in terms of labor market attachment.

[12] See Appendix C for detailed information on how these variables were constructed.

Table 7 shows means of the time-varying contextual variables by cluster measured earlier in life (ages 22-29). The associations present later in life start to assert themselves early on. We see that 16-24% of men exhibiting low attachment during prime age have spent some time in jail early in life, compared to only 0-3% among the highly attached, whereas we observe no jail time for women independent of type. Women who will turn out to have low attachment later in life have higher fertility when young, although the difference is less quantitatively significant compared to the number of children at age 50. In contrast, we see no relation between early-life fertility and prime-age attachment among men.

3.2 Health

At age 40, the NLSY surveys respondents in detail on their health, which permits calculating both a mental and a physical health score. We use these measures in addition to the prevalence ("frequency") and incidence ("ever") of health limitations measured at each interview.

Table 6 shows a strong health gradient in labor force attachment. The mean physical health score among the less attached men lies a third to half a standard deviation below their highly attached counterparts, with an even larger gap of more than a standard deviation for those in cluster 4.5. Among women, this gap is between two-thirds and a full standard deviation. The gap in mental health is almost as large as that in physical health among men, and about half as large among women.[13] In line with this, 14–36% of prime age men in the less attached clusters report that health currently limits their work, and 51–78% report this ever being the case, compared to only 2–7 and 8–29%, respectively, among the highly attached. The gap is only slightly smaller among women.

Returning to Table 7, we see that health limitations are less frequent among the young. In a given period between ages 22 to 29, 3.9% of men report a health limitation, compared to 7% in prime age. For women, the corresponding shares are 6.4% and 9%. Nevertheless, the health gradient across clusters already appears at ages 22 to 29. At those ages, 21–49% of those in the less attached clusters report having been limited at least once by their health in the kinds or amounts of work they were able to perform, compared to only 10–28% in the highly attached clusters. On average, they reported such limitations in 5–18% of their interviews, compared to only 2–6% for the highly attached. That is, while it is not unusual for the young overall to report a health-related work limitation at some point in time – for example, 4% of men report ever having been prevented from work by health during ages 22 to 29 – health limitations affecting work are much more common in the less attached clusters. This is particularly evident in cluster 4.5. Almost half of the men in this group have experienced a limitation at least once between ages 22 and 29, and a third have been prevented from work by health at least once in that age range.

The health gradient across clusters differs somewhat by gender. While the overall mean difference between the more and less attached clusters is similar among men and women, men in cluster 4.5 are particularly likely to report a health limitation when young: they do so in 18% of interviews during ages 22 to 29, and 30% of them report ever having been prevented from work in those years – a factor 9 or 30, respectively,

[13] Interestingly, not only women in the lowest attachment cluster 4.5, but also those belonging to the unstable type 4.3 have very low mean health scores. This is not the case for women in cluster 4.4, who have much more stable employment histories, and also much less pronounced among men in cluster 4.3. It suggests that health problems may drive employment instability in this group of women, but less so for men with unstable histories.

Clusters	4.0	4.1	4.2	4.3	4.4	4.5	Average
Full Sample							
Population Share	0.16	0.24	0.38	0.11	0.04	0.08	1.00
Share Women	0.30	0.43	0.52	0.63	0.68	0.71	0.50
Panel A: Women							
Population Share	0.10	0.21	0.40	0.14	0.06	0.11	1.00
Black	0.16	0.14	0.14	0.16	0.16	0.17	0.15
Hispanic	0.06	0.06	0.09	0.07	0.09	0.10	0.08
High School Dropout	0.03	0.05	0.09	0.16	0.18	0.22	0.11
High School Graduate	0.42	0.40	0.47	0.44	0.42	0.40	0.44
Some College	0.32	0.25	0.24	0.23	0.23	0.19	0.24
College Graduate	0.24	0.30	0.20	0.17	0.17	0.19	0.22
Cognitive Skills	0.27	0.19	−0.04	−0.18	−0.17	−0.26	−0.01
Non-Cognitive Skills	0.13	0.03	0.02	0.01	−0.03	−0.12	0.02
Social Skills	−0.10	−0.00	0.03	0.09	−0.02	−0.02	0.01
Jail	0.01	0.00	0.01	0.01	0.01	0.03	0.01
Children at 50	1.45	1.65	1.95	2.14	2.17	2.48	1.94
Health Limitation (frequency)	0.02	0.03	0.07	0.16	0.10	0.24	0.09
Health Limitation (ever)	0.10	0.18	0.31	0.55	0.41	0.59	0.33
Physical Health Score (age 40)	0.27	0.22	0.06	−0.55	0.04	−0.80	−0.07
Mental Health Score (age 40)	0.08	0.09	−0.08	−0.29	−0.10	−0.49	−0.10
Panel B: Men							
Population Share	0.22	0.27	0.36	0.08	0.03	0.05	1.00
Black	0.07	0.12	0.17	0.21	0.35	0.31	0.15
Hispanic	0.05	0.07	0.08	0.12	0.16	0.13	0.08
High School Dropout	0.06	0.09	0.15	0.29	0.24	0.29	0.13
High School Graduate	0.37	0.41	0.47	0.47	0.49	0.48	0.43
Some College	0.20	0.22	0.21	0.16	0.14	0.16	0.20
College Graduate	0.37	0.28	0.16	0.08	0.13	0.07	0.23
Cognitive Skills	0.46	0.19	−0.13	−0.43	−0.58	−0.74	0.03
Non-Cognitive Skills	0.03	0.07	−0.03	−0.02	0.08	−0.34	0.00
Social Skills	0.06	0.04	−0.03	0.07	−0.09	−0.16	0.01
Jail	0.00	0.02	0.06	0.33	0.29	0.35	0.08
Children at 50	1.88	1.92	1.68	1.87	1.50	1.55	1.79
Health Limitation (frequency)	0.02	0.03	0.07	0.14	0.21	0.36	0.07
Health Limitation (ever)	0.08	0.15	0.29	0.51	0.56	0.78	0.25
Physical Health Score (age 40)	0.28	0.18	0.06	−0.14	−0.36	−1.22	0.07
Mental Health Score (age 40)	0.29	0.20	0.11	−0.20	−0.12	−0.86	0.10

Note: Jail is an indicator for any incarceration between ages 22–50. Health Limitation is a yearly indicator for a health-related work limitation, either the kind or the amount of work, including no work, between ages 22–29; "frequency" is the fraction of survey years affected, and "ever" denotes at least one such year. Physical and Mental Health Scores are answers to a comprehensive health assessment by age 40, reported in standard deviations.

Table 6: Prime-Age Contextual Variables by Cluster and Gender (Ages 30–50)

more than those in cluster 4.0. These factors are only 4 and 10 among women. This is despite the fact that in the population, health limitations are more frequent among women than men. Men in cluster 4.3, in

Clusters	4.0	4.1	4.2	4.3	4.4	4.5	Average
	Panel A: Women						
Population Share	0.10	0.21	0.40	0.14	0.06	0.11	1.00
Jail	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Children at 30	0.99	1.07	1.38	1.56	1.39	1.47	1.31
Health Limitation (frequency)	0.03	0.04	0.06	0.09	0.06	0.13	0.06
Health Limitation (ever)	0.13	0.17	0.28	0.40	0.25	0.41	0.27
Health Limitation: 1 year	0.08	0.12	0.19	0.23	0.15	0.16	0.17
Health Limitation: 2 years	0.03	0.03	0.05	0.09	0.06	0.09	0.06
Health Limitation: 3+ years	0.03	0.02	0.04	0.09	0.06	0.18	0.06
Prevented from Work (frequency)	0.00	0.01	0.02	0.04	0.02	0.06	0.02
Prevented from Work (ever)	0.02	0.04	0.13	0.21	0.13	0.23	0.12
Prevented from Work: 1 year	0.02	0.04	0.11	0.15	0.13	0.12	0.09
Prevented from Work: 2 years	0.00	0.00	0.02	0.04	0.01	0.06	0.02
Prevented from Work: 3+ years	0.00	0.00	0.00	0.01	0.02	0.07	0.01
	Panel B: Men						
Population Share	0.22	0.27	0.36	0.08	0.03	0.05	1.00
Jail	0.00	0.02	0.03	0.18	0.16	0.24	0.04
Children at 30	0.94	0.96	0.94	1.16	0.91	0.94	0.96
Health Limitation (frequency)	0.02	0.02	0.04	0.05	0.07	0.18	0.04
Health Limitation (ever)	0.10	0.09	0.16	0.26	0.23	0.49	0.16
Health Limitation: 1 year	0.08	0.06	0.08	0.15	0.08	0.19	0.09
Health Limitation: 2 years	0.01	0.02	0.04	0.06	0.04	0.09	0.03
Health Limitation: 3+ years	0.01	0.01	0.04	0.03	0.13	0.22	0.04
Prevented from Work (frequency)	0.00	0.00	0.01	0.01	0.02	0.11	0.01
Prevented from Work (ever)	0.01	0.02	0.04	0.07	0.12	0.30	0.04
Prevented from Work: 1 year	0.01	0.02	0.03	0.05	0.11	0.07	0.03
Prevented from Work: 2 years	0.00	0.00	0.00	0.02	0.03	0.10	0.01
Prevented from Work: 3+ years	0.00	0.00	0.00	0.01	0.02	0.13	0.01

Note: Jail is an indicator for any incarceration between ages 22–29. Health Limitation is a yearly indicator for a health-related work limitation, either the kind or the amount of work, including no work, between ages 22–29; "frequency" is the fraction of survey years affected, and "ever" denotes at least one such year.

Table 7: Contextual Variables by Cluster and Gender (Ages 22–29)

contrast, report health limitations much less frequently than women in that cluster.

Table 7 also reports to what extent individuals report health limitations persistently. It shows the distribution of the frequency of any reported health limitation between ages 22 and 29 among those who answer the question in all eight interviews. Again, we observe a strong gender difference, with women reporting significantly more often to be limited or prevented in their ability to work. Nevertheless, the table shows, for both genders, that the share of individuals who report three or more times that they were either limited or prevented from working is very small.[14]

However, this small group of young individuals who repeatedly report health limitations exhibits a much

[14]This is the case despite some evidence of persistence. For example, the share of individuals reporting a limitation twice is larger than what would be expected if the realization of health limitations was independent.

lower probability of being closely attached to the labor market during prime age. Figure 2 shows, for women and men respectively, the probability of belonging to the least attached clusters as a function of the number of instances in which respondents aged 22 to 29 reported being limited in work or prevented from work. There is a strong gradient. For instance, among men, the probability of belonging to the less attached clusters is significantly larger for those who reported a health limitation once when young, and rises to more than 80% among those who report a limitation in three or more of the eight interviews. Again, the gradient is particularly strong among men. As Appendix Figures 7 and 8 show, these patterns are very similar as a function of the number of times a person has been limited (but not prevented) in work.

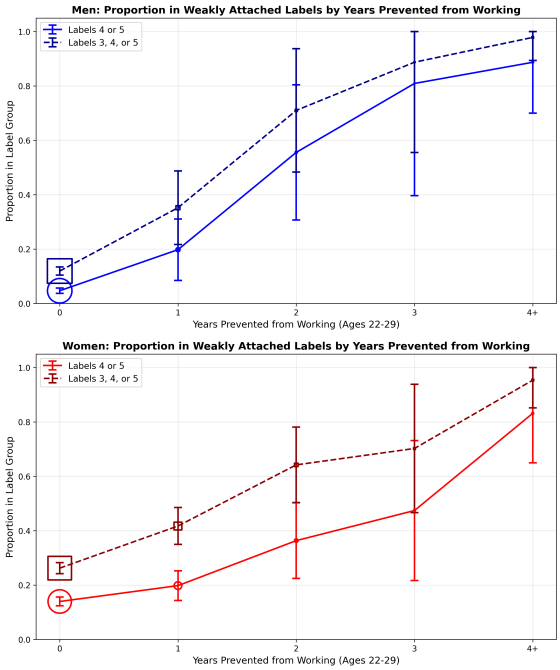


Figure 2: Weak Attachment by Years Prevented from Work: Men (left) and Women (right)

To summarize, although repeated health limitations in young adulthood are rare in the population, they are strongly associated with weak labor market attachment. This suggests that early-onset health limitations are a significant predictor of labor market type, a relationship we explore further below.[15]

3.3 Earnings, wages and employment by cluster

While our primary focus is heterogeneity in prime-age labor force status, we now document how worker types differ in earnings and wages, and how these differences are already apparent at earlier ages.

Table 8 presents mean yearly earnings, hours worked, and wages during prime age, disaggregated by cluster, both for the pooled sample and separately by gender. Mean earnings, hours, and wages are conditional on positive hours and earnings. The bottom row in each panel shows the share of respondents with

[15] What are these health limitations reported by the young? Between 1979 and 1982, the NLSY asked individuals reporting health limitations a set of follow-up questions to identify what conditions gave rise to the limitation. In about two thirds of cases, the conditions that are reported as limiting or preventing work are chronic conditions, most often musculoskeletal conditions. Mental health is very rarely reported as the condition limiting or preventing individuals from working.

positive hours and earnings. By construction, more attached types are much more likely to work. Not by construction, but consistent with our expectations, we observe that they also earn significantly higher wages when working, and work longer hours. In combination, these differences lead to very large gaps in earnings across types. These mostly reflect wages, and to a lesser extent hours worked.[16]

Clusters	4.0	4.1	4.2	4.3	4.4	4.5	Population
Panel A: Women							
Earnings	65,037	58,987	41,564	26,625	39,732	32,965	46,461
Hours Worked	2,148	2,042	1,834	1,532	1,807	1,572	1,877
Real Wage	31.5	29.6	23.2	18.9	24.1	22.0	25.3
Share Employed	0.969	0.947	0.861	0.685	0.536	0.259	0.782
Panel B: Men							
Earnings	108,751	87,332	62,195	38,784	53,161	30,917	78,725
Hours Worked	2,475	2,389	2,203	1,934	2,065	1,769	2,297
Real Wage	44.4	37.3	29.1	21.5	27.7	17.7	34.6
Share Employed	0.956	0.936	0.896	0.754	0.574	0.275	0.876

Note: Earnings are annual, in 2022 dollars. Real wage is the hourly wage in 2022 dollars. Hours Worked are annual hours. Share Employed is the fraction of individuals with positive hours in a given year. Statistics are weighted means over ages 30–50.

Table 8: Earnings, Hours, Wages, and Employment by Cluster and Gender (Ages 30–50)

Table 9 shows the same variables for ages 22 to 29. The wage, hours and earnings patterns observed in prime age already appear at younger ages, although they are much less pronounced.

Clusters	4.0	4.1	4.2	4.3	4.4	4.5	Population
Panel A: Women							
Earnings	35,906	32,050	26,278	20,211	28,806	25,886	28,002
Hours Worked	1,794	1,729	1,597	1,410	1,631	1,492	1,620
Real Wage	19.6	18.4	16.5	14.6	17.6	17.0	17.2
Share Employed	0.931	0.882	0.808	0.707	0.660	0.614	0.793
Panel B: Men							
Earnings	49,934	45,413	38,454	26,816	35,787	21,579	41,660
Hours Worked	2,035	2,016	1,897	1,657	1,827	1,473	1,932
Real Wage	24.3	22.1	20.2	16.8	18.6	14.9	21.3
Share Employed	0.916	0.887	0.883	0.818	0.746	0.579	0.871

Note: Earnings are annual, in 2022 dollars. Real wage is the hourly wage in 2022 dollars. Hours Worked are annual hours. Share Employed is the fraction of individuals with positive hours in a given year. Statistics are weighted means over ages 22–29.

Table 9: Earnings, Hours, Wages, and Employment by Cluster and Gender (Ages 22–29)

To get a more detailed sense of how differences in labor market outcomes emerge early in life, Figure 3

[16] Mean wages and hours worked are lowest in clusters 4.5 and 4.3. The unstable cluster 4.3 thus again breaks monotonicity, as with women's health.

plots the share of young respondents working between ages 22 and 29 by cluster. (Appendix Figures 5 and 6 show the same by gender.) For the highly attached, employment rates are high and stable, at 85 to 90%. For those with low attachment, employment rates are ten to twenty percentage points lower at age 22. This gap remains stable for cluster 4.3, and grows with age for cluster 4.5 and for women of cluster 4.4. This early divergence foreshadows the much lower prime-age employment rates of the unattached, and suggests that early employment histories are a strong predictor of labor market type, as we document in the next section.

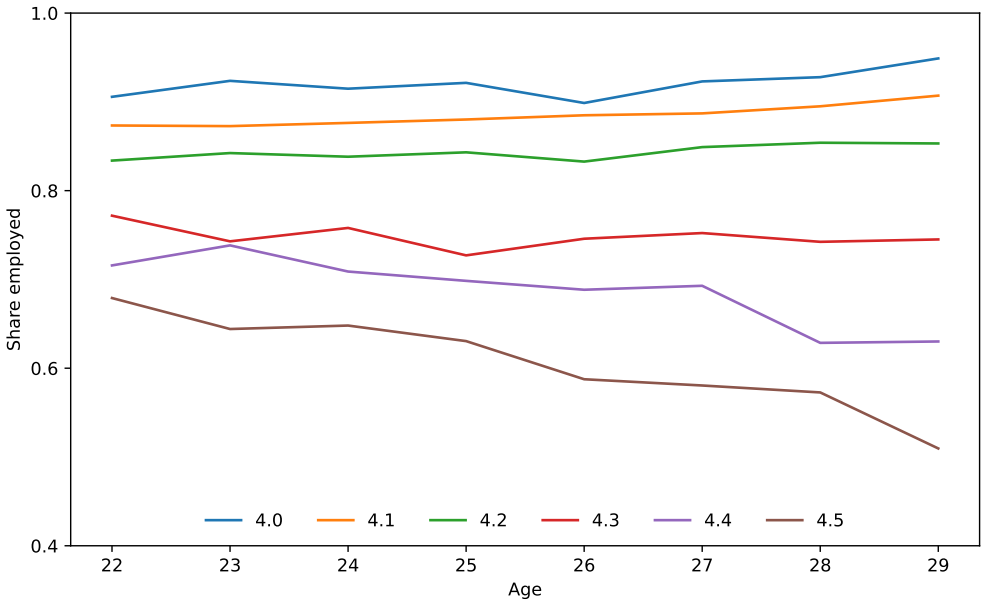


Figure 3: Share working by age and cluster, women and men, ages 22 to 29

4 Predicting Types

Having described labor market types and how they differ, we next ask whether an individual's eventual attachment type during prime working age can be predicted early in life. For this purpose, we aggregate individuals into two types, a highly attached type comprised of clusters 4.0, 4.1, and 4.2 (denoted $\tau_i = 0$), and a low-attached type combining 4.3, 4.4, and 4.5 (denoted $\tau_i = 1$). This grouping follows the description of clusters in Section 2.5 above, which showed a clear break in attachment, in particular as measured by time spent non-employed, between clusters 4.2 and 4.3.[17] Given our proposed aggregation, the low-attached type, referred to simply as "unattached," accounts for 22% of the whole population.

We then estimate several models for τ_i using a vector of controls $\mathbf{X}_i$. The models include logit, Lasso logit, random forest, and gradient boosting. We evaluate each model's success in predicting $\tau_i = 1$. In this paper, we report the results from Lasso logit, but those from other models are similar. We report results from this prediction exercise using 5 fold cross-validation thus ensuring that our results are not driven by overfitting the data (See online Appendix D.1.).

Before we turn to present these results, we however need to decide upon a metric to assess prediction quality. The metric chosen is a precision among those most likely to be of the low attached type. We discuss the choice of htis metric next.

4.1 Choosing a metric

The traditional approach to rank in-sample prediction quality among linear models, based on the mean squared error, is equivalent to ranking by the R^2. In nonlinear models such as ours, researchers often turn to a measure like the pseudo-R^2 to determine the quality of a prediction model. The pseudo-R^2 is based on the fit of the model across the entire sample. This might be appropriate for applications that rest on understanding the heterogeneity in labor force attachment across the entire population. Gregory et al. (2025) for example rely on the pseudo-R^2 when they find that little of the heterogeneity in labor force histories using clustering methods is explained by demographics. Their choice of the pseudo-R^2 as their metric is motivated by their application: a business cycle model trying to understand employment dynamics in the population as a whole. However, policymakers who consider policies targeting at-risk populations through, for example, early-life training or mentoring programs, do not care about the properties of the prediction model for the entire population. Rather, what matters for them is the precision of the prediction for those most at risk. It is this perspective that we adopt in this paper.

To fix ideas, consider a population of size 1 that belongs to one of two types, attached and unattached. Consider further a stylized policy that yields a benefit b for each unattached type that is treated, while it imparts no benefit to stable types. Assume that a policymaker has to assign individuals to treatment in a policy of program size $m \in [0, 1]$. The policymaker assigns individuals sequentially, based on a prediction model, starting with those most likely to be unattached.

[17] Any aggregation of six clusters into two groups necessarily suppresses nuances. Nevertheless, we believe our grouping to be useful, as all three low-attachment groups may stand to benefit from policy interventions motivating our analysis.

The benefit of such a program is

$$u = m \cdot p(m, I) \cdot b.$$

The number of unattached types treated is equal to the program size m times $p(m, I)$, the fraction of the treated population that truly belongs to the unattached type. This fraction is known as *precision*.

The precision of a prediction model is a function of the program size m and the information set I used to predict individual types. In general, $p(m, I)$ will decrease with program size, since as the program expands, the marginal individual is increasingly less likely to be of the unattached type. It will increase with additional information, since a larger information set (weakly) improves predictive ability.

In practice, since labor market policies such as training or mentoring programs rarely target a large share of the population, we focus on precision for relatively small values of m. Our benchmark is $m = 0.05$, although we examine sensitivity with respect to this choice.

We are particularly interested in how precision changes with I as information accumulates early in the individuals' life-cycle, conditional on m. This will inform on the trade-off policymakers face between waiting for additional information to arrive against the potential costs of delaying program intervention, for example because a given intervention might be less effective when targeting older individuals.

For comparison, we also report a traditional measure of fit, the McFadden Pseudo-R^2, which is based on the fit of the model's prediction for the entire population.[18]

4.2 Model performance and the value of information accruing with age

We now explore to what extent it is possible to predict individuals' type using information on the early years of their career.

To do so, we fit an L1-penalized logit (Lasso) predicting whether an individual belongs to the weakly or strongly attached type, expanding the conditioning set in additive blocks of variables, without interactions. This allows us to trace how predictive power accumulates as new categories of early-career information are introduced. We adopt the Lasso as our main specification because the unregularized logit becomes increasingly prone to overfitting as the number of regressors grows. For comparison, the appendix reports the same exercise using the unregularized logit and two non-linear benchmarks (random forest and gradient boosting), which relax the additivity of the linear specification and capture interactions, all evaluated out-of-sample using 5-fold cross-validation.

Our baseline predicts low attachment simply using demographic variables, namely gender, race-ethnicity, and birth year. These are the variables typically available in administrative data sets, such as the LEHD used by Gregory et al. (2025), which contain long labor market histories for a wide range of individuals but very few contextual variables. Our baseline thus uses models from the literature as our benchmark. We then progressively add the wealth of early life-cycle contextual information the NLSY79 contains. This allows us to isolate the variables most relevant for predicting low attachment as well as the ages when additional

[18] The McFadden Pseudo R^2 is given by $R^2 = \left(1 - \frac{\mathcal{L}_1}{\mathcal{L}_0}\right)$ where $\mathcal{L}_1$ is the log-likelihood for the model with controls and $\mathcal{L}_0$ for a model with an intercept only.

Table 10: Out-of-sample predictive performance by age and gender

	Men			**Women**			
	Pseudo R^2	AUC	Precision at 5%	Pseudo R^2	AUC	Precision at 5%	# Vars
Panel A: Age 22							
Demographics	0.02	0.60	0.27	-0.00	0.50	0.28	3
+Health	0.02	0.60	0.27	0.00	0.52	0.41	7
+Employment Histories	0.09	0.71	0.51	0.03	0.61	0.49	15
+Occupations + Industries	0.09	0.71	0.52	0.03	0.61	0.50	34
+Jail	0.09	0.71	0.52	0.03	0.61	0.50	35
+Cog+NCog+Education	0.11	0.73	0.51	0.03	0.62	0.57	43
+Family Variables+Local Conditions	0.12	0.73	0.53	0.03	0.62	0.56	70
+Children	0.12	0.73	0.55	0.03	0.62	0.57	73
Panel B: Age 29							
Demographics	0.02	0.60	0.27	-0.00	0.50	0.28	3
+Health	0.06	0.64	0.47	0.02	0.56	0.61	35
+Employment Histories	0.23	0.81	0.72	0.13	0.74	0.71	97
+Occupations + Industries	0.23	0.81	0.73	0.13	0.74	0.71	115
+Jail	0.24	0.81	0.76	0.13	0.74	0.71	116
+Cog+NCog+Education	0.24	0.81	0.77	0.13	0.74	0.75	124
+Family Variables+Local Conditions	0.24	0.81	0.75	0.13	0.74	0.73	151
+Children	0.23	0.81	0.74	0.13	0.74	0.73	153
N	4,138			4,275			

Notes: Each row reports out-of-sample predictive performance from an L_1-penalized (lasso) logit model of low-attachment status ($\tau_i = 1$ for clusters 4.3–4.5) on the listed covariate blocks, entered cumulatively, estimated separately by gender using variables observed by age 29. All metrics are computed on pooled out-of-fold predictions from nested cross-validation: a 5-fold stratified outer split for evaluation, with the penalty strength λ re-selected by 5-fold cross-validation over a 10-point grid within each outer training fold. Pseudo R^2 is McFadden's pseudo-R^2. AUC is the area under the ROC curve. Precision at 5% is the share of true low-attachment individuals among the 5% of each gender subsample with the highest predicted probability. # Vars is the number of candidate covariates supplied to the lasso. IPW attrition-corrected sample weights are used throughout.

information is particularly valuable in predicting attachment.[19]

From Table 10 it is clear that, with basic demographics alone, the overall fit as represented by the Pseudo-R^2 is poor with 0.02 for males or even a slight negative value for women. This echoes the perceived difficulty in predicting latent labor market types typically reported in the literature (Shibata, 2019; Hall and Kudlyak, 2022; Gregory et al., 2025; Ahn et al., 2023). By contrast, concentrating on the tail of the distribution which we believe to be most policy-relevant, i.e. the individuals most likely to be unattached, we observe that even with just the most basic demographics, precision is fairly high, at 0.27 and 0.28 for men and women respectively. Thus, part of the low ability to predict labor market types reported in the literature derives from the metric used in the literature: it is difficult using basic demographics to capture the heterogeneity in labor force histories across the entire population.

Precision increases rapidly once we include additional controls. To illustrate, we present results for a specific ordering of co-variates; a systematic assessment of the marginal contribution of each co-variate group, averaging over all possible orderings, is provided by the Shapley–Shorrocks decomposition in the next subsection.

Generally, we tend to find that the history of employment by a given age as well as health are the most useful predictors. At age 22, employment and health histories together raise precision at m=0.05 to 0.51 and 0.49 respectively. With the full set of controls, precision by age 22 reaches 0.55 and 0.57.[20] Given the young age at which we undertake this prediction exercise, we find this to be a reasonably high quality of the prediction.

Naturally, the estimates get ever more precise as we add variables collected up to age 29. Health is a particularly useful predictor at this point, adding 20 p.p. to the precision of the baseline model for men and 33 p.p. for women. Again, employment histories are also useful predictors, contributing another 25 and 10 p.p. to the respective precision by gender. These two sets of variables together raise precision by age 29 by about 43 p.p. to over 70% from the around 28 p.p. obtained using only demographics. The contribution of the other controls by age 29 (and also age 22) over and above is relatively minor, raising precision by only a few percentage points.

The information that accrues between the ages of 22 and 29 adds 19 and 16 p.p. of precision to the information available by age 22. Overall, a set of variables that is fairly easy to collect allows identifying future less attached types with reasonable precision by age 22, and with great precision by age 29.

Figure 4 shows how the precision for the full model varies across the entire support of m for the model at age 22 and at age 29. [21] The figure suggests that additional information accruing with age delivers sizeable increases in precision when m is less than about 0.5. It also illustrates how precision drops off rapidly as m increases, especially at age 22. The information available at age 22 allows identifying those loosely attached to the labor market with high precision only if m is less than 0.1 or so. At age 29, we can achieve similar precision for program sizes including 20–30% of the population.

[19] We again refer the reader to Appendix C for detailed variable descriptions.

[20] Adding covariates need not improve out-of-sample fit: noisy or correlated regressors inflate variance, a concern that the ℓ_1 penalty in Lasso logit attenuates but does not eliminate.

[21] The figure plots expected precision over the relevant region, computed as $\sum_i \hat{p}_i$ from a Lasso logit estimated out-of-sample with 5-fold cross-validation.

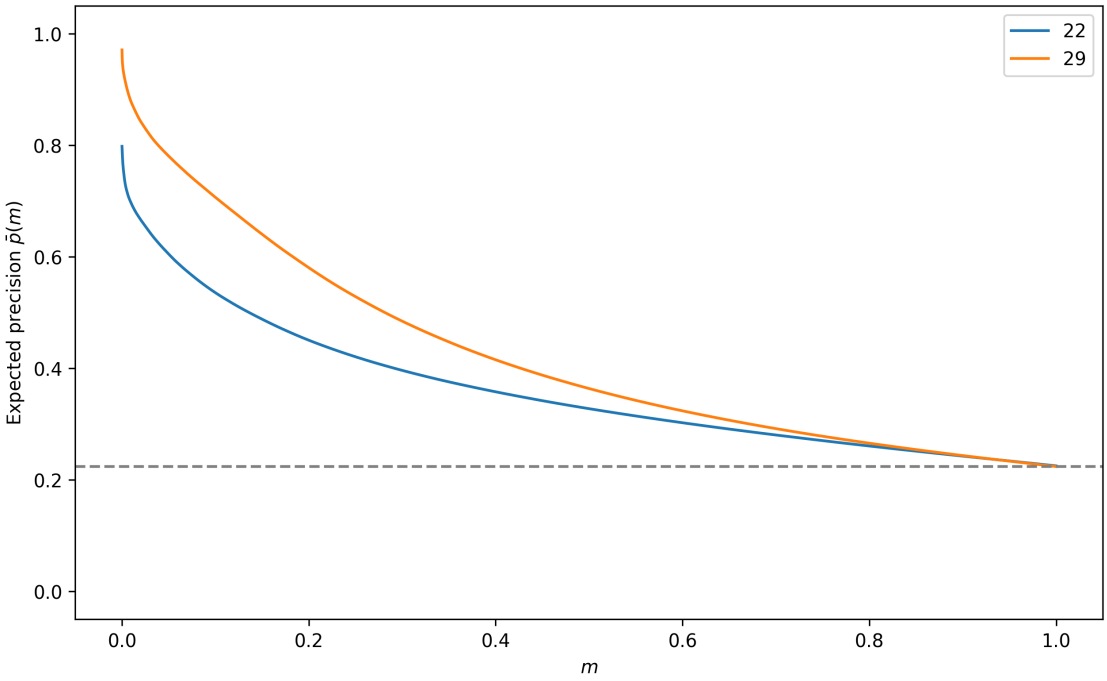


Figure 4: Precision as m varies: Full model

4.3 Shapley-Shorrocks decomposition

We use the Shapley-Shorrocks decomposition to assess the marginal contribution of each of the seven different groups of covariates to predicting type.[22] For every subset of co-variates, we compute three measures of fit, McFadden's pseudo-R^2, the Area under the Curve (AUC), and the precision p for several levels of m at age 29. Intuitively, the Shapley value of a group of co-variates gives its average marginal contribution to a measure of fit, averaging over all possible permutations of the order in which groups of co-variates can be included.[23] By construction, the Shapley values of all groups of co-variates sum to one. Thus, the Shapley value for a group can be interpreted as the fraction of the total prediction improvement (from the baseline of demographics only to the full model) attributable to that group. It thus provides an order-agnostic and additive attribution of model performance to co-variate groups.[24]

A few of the co-variate blocks warrant explicit definition given their importance; precise variable construction is detailed in Appendix C. The employment-history block is a year-by-year panel covering every

[22] We include demographics in every specification and do not separately evaluate their contribution.

[23] Formally, the Shapley value of block i of covariates out of the set N for a given measure of fit $M(\cdot)$ is

$$\phi_i \;=\; \sum_{S \subseteq N \setminus \{i\}} \underbrace{\frac{|S|!\,\big(|N|-|S|-1\big)!}{|N|!}}_{\text{Shapley weight } w(|S|)} \big[\, M(S \cup \{i\}) - M(S) \,\big]\,. \tag{2}$$

Intuitively, ϕ_i is the average marginal gain from adding block i to a model that already contains a random subset S of the other blocks; the combinatorial weight $w(|S|)$ is the probability that exactly the blocks in S precede i in a uniformly random permutation of N.

[24] To implement the calculation, we compute Shapley values exactly by evaluating $M(S)$ for all $2^{|N|} = 128$ subsets S (no Monte Carlo). For each S we fit a weighted L1-penalized logit (Lasso) with frequency weights, obtain in-sample predicted probabilities $\hat{p}$, and compute the two measures of fit.

age from 22 to the conditioning age. At each age 22–29, the Employment Histories block records: weeks of employment, the share of weeks spent out of work pooling unemployment and out-of-the-labor-force , the number of distinct jobs held, the weekly wage, a flag for missing wage, an indicator for school enrollment and its missing-flag, and an annual employment indicator. The occupation-and-industry block stores a single value per age: at each age we identify the individual's main occupation and main industry, defined as those in which they have spent the most weeks cumulatively up to that age, and enter them as a full set of 1-digit Census major-group dummies (10 industry categories, 8 occupation categories) along with per-age missing flags.

Results are given in Tables 11 and 12. The column "Precision at 5%" in Table 11 again shows how, for $m = 0.05$ for women, the full set of variables implies a precision of 0.73, compared to 0.28 with demographics only. Hence, the "Gain" from including all co-variates amounts to 0.44 (rounded). The following rows show the average marginal contribution of our seven groups of co-variates: Cognitive and non-cognitive skills and education; family variables and local conditions; employment histories; occupations and industries; health; children; and jail.

The table reveals that the relative contribution of different groups of co-variates varies with the size of the low-attachment group one attempts to identify. For the least attached group ($m = 0.01$), health variables are by far the strongest predictor of future low attachment for women, with employment histories, occupations and industries playing the next largest roles. (Note also the precision of 0.76 in this case!) For men, employment histories are the most important predictor, with health variables coming second. Incarceration also has significant predictive power. All young-age information jointly implies a precision of 90% for $m = 0.01$.

As m grows, employment histories become more predictive, while the predictive power of health declines. Among men, the role of jail also increases, up to $m = 0.05$. The importance of other groups of variables is mostly stable. For m of 0.05, employment histories plus occupation and industries are the most informative predictor, with incarceration and health variables coming next.[25]

This analysis shows that for the least attached part of the population, health problems when young are strongly predictive of low labor market attachment in prime age. As one attempts to identify a larger group of low-attachment individuals, their importance necessarily recedes in line with the limited prevalence of health problems among the young. This reflects our findings above that the group of individuals who repeatedly experienced health limitations is very likely to have very low attachment, but is also small, in particular among men. When considering a broader unattached group, health limitations still play a role, but so does the early-life employment history.

[25]The careful reader will note that reported full precision is not monotone with m and that children have a negative contribution for males at $m = 0.05$. Both these features are possible because of the random variation across subsamples inherent in our use of cross-validation. We evaluate the fit using cross-validation.

Table 11: Lasso — Women: Shapley–Shorrocks decomposition of predictive performance.

	R^2	AUC	Prec at 1%	Prec at 2%	Prec at 5%	Prec at 10%
Performance						
Baseline	-0.00	0.50	0.36	0.37	0.28	0.29
Full	0.13	0.74	0.76	0.82	0.73	0.68
Gain	0.13	0.24	0.40	0.45	0.44	0.39
Block shares of gain (%)						
Skills & education	6.25	9.15	5.64	14.68	16.90	16.11
Family & geography	1.33	6.54	1.33	4.92	4.69	5.73
Employment history	75.08	65.94	14.76	16.55	30.41	40.06
Occupation & industry	7.95	7.91	13.48	18.65	22.94	18.40
Health	8.53	7.86	64.00	41.44	21.12	14.61
Children	0.87	2.59	0.74	2.94	3.57	4.79
Incarceration	0.00	0.00	0.05	0.82	0.36	0.31
Total	100.00	100.00	100.00	100.00	100.00	100.00

Notes. Model: L1-penalized logit; the penalty strength is selected by internal cross-validation within each coalition and within each outer CV fold. N = 4,275. Predictors measured at age 29; outcome is an indicator for the unattached-worker type. *Baseline* includes demographics (age, ethnicity) only; *Full* adds all seven covariate blocks; *Gain* = Full − Baseline. Block shares are exact Shapley–Shorrocks values over $2^7 = 128$ coalitions, normalized by *Gain*, and sum to 100% by construction. R^2 is McFadden pseudo-R^2; AUC is the area under the ROC curve; Prec at m% is expected precision in the top m% of observations ranked by predicted probability. All metrics use pooled 5-fold stratified cross-validated probabilities (`random_state`=42). IPW attrition-corrected sample weights used throughout.

Table 12: Lasso — Men: Shapley–Shorrocks decomposition of predictive performance.

	R^2	AUC	Prec at 1%	Prec at 2%	Prec at 5%	Prec at 10%
Performance						
Baseline	0.02	0.60	0.24	0.27	0.27	0.27
Full	0.23	0.81	0.90	0.85	0.74	0.61
Gain	0.21	0.21	0.66	0.59	0.48	0.34
Block shares of gain (%)						
Skills & education	6.62	13.91	3.73	4.43	5.24	2.92
Family & geography	2.60	5.45	7.42	5.26	2.15	4.17
Employment history	66.00	58.08	37.33	44.29	43.56	54.60
Occupation & industry	4.58	7.13	16.71	14.19	9.00	6.23
Health	6.92	5.71	22.77	20.91	12.50	10.38
Children	0.31	0.98	0.44	0.59	-0.25	0.08
Incarceration	12.97	8.74	11.61	10.34	27.80	21.62
Total	100.00	100.00	100.00	100.00	100.00	100.00

Notes. Model: L1-penalized logit; the penalty strength is selected by internal cross-validation within each coalition and within each outer CV fold. N = 4,138. Predictors measured at age 29; outcome is an indicator for the unattached-worker type. *Baseline* includes demographics (age, ethnicity) only; *Full* adds all seven covariate blocks; *Gain* = Full − Baseline. Block shares are exact Shapley–Shorrocks values over $2^7 = 128$ coalitions, normalized by *Gain*, and sum to 100% by construction. R^2 is McFadden pseudo-R^2; AUC is the area under the ROC curve; Prec at m% is expected precision in the top m% of observations ranked by predicted probability. All metrics use pooled 5-fold stratified cross-validated probabilities (`random_state`=42). IPW attrition-corrected sample weights used throughout.

5 Discussion and Conclusion

Throughout this paper, we have maintained that there is value for a policy maker to be able to identify those less attached to the labor market during their prime working-age using variables already available while young. Implicit is the assumption that it is indeed desirable to increase labor force participation in the population.

This assumption, while not trivial, is pervasive in the policy discussion. The Conference Board for instance argues that (Committee for Economic Development of the Conference Board, 2019, p. 5):

> A growing labor force has been a significant contributor to past US economic growth. More workers can lead to more production, more wages, and more consumption. By contrast, slower labor force growth will pose a challenge for American businesses dependent on the talent available to them when they compete in the global marketplace. And, with fewer workers to support a growing number of retirees, an aging population and slowing labor force growth will also place more strain on the nation's ability to meet its commitments to seniors while also supporting younger families and funding investments that bolster future economic growth.

Similar statements from policy makers on both sides of the political spectrum are common.

Our analysis demonstrates that it is actually possible to identify early on, and with a high degree of precision, a substantial group in the population that will be only loosely attached to the labor market when reaching prime age. For a number of individuals, labor force attachment is a persistent trait which is already observable early on. For others, it only emerges gradually but is nonetheless rooted in characteristics and circumstances present early in life.

The reasons why individuals fail to form strong attachment to the labor market are likely very diverse. Some might lack the cognitive or non-cognitive skills that enable them to invest into their careers. Such non-cognitive skills may include the ability to resist temptation and have the self-discipline required to fit into a modern workplace. Other explanations might relate to labor or credit market frictions that prevent individuals from taking advantage of opportunities. For yet others, weak attachment might simply be a matter of preference.

Our prediction framework provides observables to guide policymakers in identifying at-risk individuals. In doing so, we also have in mind a fundamental trade-off underlying such interventions. According to our results, waiting for more information to arrive significantly improves the ability to properly identify the target population. However, we believe there is also a cost in terms of reduced malleability in waiting for additional information. That is, either due to psychological effects or to deep scarring effects, later interventions are likely to be less effective. By providing an assessment of the prediction gain associated with waiting for additional information, we hope to provide policymakers a key input to be balanced against reduced malleability.

While it is beyond the scope of this, or any single paper to provide a definitive answer on the determinants of weak prime-age attachment, a persistent finding in our work is that young adult health does seem to play an important role. Even though most young individuals do not persistently report a health-induced work

limitation, those that do are much more likely to subsequently become weakly attached. For the most at-risk group (top 2% of the population most likely to become weakly attached), we find using Shapley-Shorrock decompositions that health is as important as early employment histories in predicting weak attachment, and far more important than family background, education, or occupation-industry histories, or incarceration. Even when the at-risk group is expanded to the 10% least likely to be attached, we still find that early-life health limitations are about as important as the combination of education and cognitive and non-cognitive skills among women, and more than three times as important among men.

At the same time that our findings highlight the central role of health, they also raise questions for future research about which specific health conditions are important, how exactly they affect an individual's working career and labor market attachment, and whether policy interventions can be effective in face of such health conditions. Answering these questions clearly requires detailed information about an individual's health status. We nevertheless interpret our findings as not only pointing to the area where further information needs to be gathered, but also offering potential grounds for early career interventions, such as providing disability accommodations.

Our core finding that differences in attachment are highly persistent and emerge early on also has important implications for how one thinks about risk in the labor market, and about institutions designed to insure against this risk. This is another area where further research is needed.

References

Ahn, H. J., B. Hobijn, and A. Sahin (2023, May). The Dual U.S. Labor Market Uncovered. NBER Working Papers 31241, National Bureau of Economic Research.

Almond, D. and J. Currie (2011a). Human capital development before age five. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 4B, Chapter 15, pp. 1315–1486. Elsevier.

Almond, D. and J. Currie (2011b). Killing me softly: The fetal origins hypothesis. *Journal of Economic Perspectives 25*(3), 153–172.

Almond, D., J. Currie, and V. Duque (2018). Childhood circumstances and adult outcomes: Act II. *Journal of Economic Literature 56*(4), 1360–1446.

Arulampalam, W. (2001). Is unemployment really scarring? effects of unemployment experiences on wages. *The Economic Journal 111*(475), F585–F606.

Aughinbaugh, A., C. R. Pierret, and D. S. Rothstein (2017). Attrition and its implications in the national longitudinal survey of youth 1979. *JSM Proceedings. Alexandria, VA: American Statistical Association*.

Behaghel, L., B. Crépon, and M. Gurgand (2014). Private and public provision of counseling to job seekers: Evidence from a large controlled experiment. *American Economic Journal: Applied Economics 6*(4), 142–174.

Bick, A., A. Blandin, and R. Rogerson (2024). After 40 years, how representative are labor market outcomes in the nlsy79? Technical report, National Bureau of Economic Research.

Bick, A., A. Blandin, and R. Rogerson (2025). Hours Worked and Lifetime Earnings Inequality. *Working Paper*.

Black, D. A., J. A. Smith, M. C. Berger, and B. J. Noel (2003). Is the threat of reemployment services more effective than the services themselves? evidence from random assignment in the UI system. *American Economic Review 93*(4), 1313–1327.

Black, S. E., P. J. Devereux, and K. G. Salvanes (2007). From the cradle to the labor market? the effect of birth weight on adult outcomes. *Quarterly Journal of Economics 122*(1), 409–439.

Bonhomme, S., T. Lamadon, and E. Manresa (2022). Discretizing unobserved heterogeneity. *Econometrica 90*(2), 625–643.

Bonhomme, S. and E. Manresa (2015). Grouped patterns of heterogeneity in panel data. *Econometrica 83*(3), 1147–1184.

Borella, M., F. Bullano, M. De Nardi, B. Krueger, and E. Manresa (2025). Health inequality and health types. *The Econometrics Journal 28*(3), 341–384.

Card, D., J. Kluve, and A. Weber (2018). What works? a meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association 16*(3), 894–931.

Castro, R., F. Lange, and M. Poschke (2025). Labor force transitions. In *Handbook of Labor Economics*, Volume 6. Elsevier.

Committee for Economic Development of the Conference Board (2019). Growing the american workforce: Bolstering participation is critical for us competitiveness and economic strength.

Cunha, F. and J. J. Heckman (2007). The technology of skill formation. *American Economic Review Papers and Proceedings 97*(2), 31–47.

Currie, J. and B. C. Madrian (1999). Health, health insurance and the labor market. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 3C, Chapter 50, pp. 3309–3416. Elsevier.

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics 132*(4), 1593–1640.

Elsby, M. W., B. Hobijn, and A. Şahin (2015). On the importance of the participation margin for labor market fluctuations. *Journal of Monetary Economics 72*, 64–82.

Gregg, P. and E. Tominey (2005). The wage scar from male youth unemployment. *Labour Economics 12*(4), 487–509.

Gregory, V., G. Menzio, and D. Wiczer (2025, March). The alpha beta gamma of the labor market. *Journal of Monetary Economics 150*(C), 103695.

Hall, R. E. and M. Kudlyak (2022). Churn and stability: The remarkable heterogeneity of flows among employment, job search, and non-market activities in the us population. Technical report, Federal Reserve Bank of San Francisco.

Heckman, J. J. and T. Kautz (2012). Hard evidence on soft skills. *Labour Economics 19*(4), 451–464.

Heckman, J. J., J. Stixrud, and S. Urzua (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics 24*(3), 411–482.

Kahn, L. B. (2010). The long-term labor market consequences of graduating from college in a bad economy. *Labour Economics 17*(2), 303–316.

Kitagawa, T. and A. Tetenov (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica 86*(2), 591–616.

Krusell, P., T. Mukoyama, R. Rogerson, and A. Şahin (2017). Gross worker flows over the business cycle. *American Economic Review 107*(11), 3447–3476.

Le Barbanchon, T., J. Schmieder, and A. Weber (2024). Job search, unemployment insurance, and active labor market policies. In *Handbook of Labor Economics*, Volume 5, pp. 435–580. Elsevier.

Lindqvist, E. and R. Vestman (2011). The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment. *American Economic Journal: Applied Economics 3*(1), 101–128.

MaCurdy, T., T. Mroz, and R. M. Gritz (1998). An evaluation of the national longitudinal survey on youth. *The Journal of Human Resources 33*(2), 345–436.

Morchio, I. (2020). Work histories and lifetime unemployment. *International Economic Review 61*(1), 321–350.

Oreopoulos, P., T. von Wachter, and A. Heisz (2012). The short- and long-term career effects of graduating in a recession. *American Economic Journal: Applied Economics 4*(1), 1–29.

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.

Schwandt, H. and T. von Wachter (2019). Unlucky cohorts: Estimating the long-term effects of entering the labor market in a recession in large cross-sectional data sets. *Journal of Labor Economics 37*(S1), S161–S198.

Shibata, M. I. (2019, December). Labor Market Dynamics: A Hidden Markov Approach. IMF Working Papers 2019/282, International Monetary Fund.

Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika 97*(4), 893–904.

Appendix A Data

Appendix A discusses sample selection and imputation of gaps in employment histories.

A.1 Sample Selection

1. Remove discontinued subsamples.

   The NLSY79 began in 1979 with 12,686 respondents drawn from a cross-sectional sample and three oversamples: Black and Hispanic youth, economically disadvantaged non-Black/non-Hispanic white youth, and a military sample. Two of these oversamples were subsequently discontinued by the survey administrators.

   The military oversample (1,280 respondents) was discontinued in the 1985 survey wave. Rather than dropping the entire oversample, the NLSY randomly retained 201 of these respondents and discontinued the remaining 1,079. We identify the discontinued individuals using reason-for-missing code 68 (*military sample dropped*), which applies to 1,075 respondents in the data; the remaining four are accounted for by mortality prior to the 1985 wave. The 201 retained members are kept in our sample.

   The economically disadvantaged non-Black/non-Hispanic white oversample was discontinued in the 1991 survey wave. We identify the removed individuals using codes 69 (*supplemental male poor white sample dropped*, 731 respondents) and 70 (*supplemental female poor white sample dropped*, 890 respondents). An additional 26 respondents who were deceased members of this discontinued subsample are removed via code 75.

   After excluding these 2,722 individuals, we retain the cross-sectional sample together with the Black and Hispanic oversamples, leaving **9,964 respondents**.

2. No prime-age employment history.

   Our analysis requires employment histories spanning prime age, which we define as ages 30 to 50 (1,040 weeks starting from the first week after the individual's 30th birthday). Of the 9,964 remaining respondents, 460 (4.6%) never report any employment status during this window and are therefore dropped. Of these, 128 had died before reaching age 30; the remainder attrited from the survey for unknown reasons. This leaves **9,504 respondents**.

3. Gaps in employment histories

   Even among respondents at some point between ages 30 and 50, who are observed through prime age, labor force status may be missing for long stretches of weeks when an individual skipped one or more survey waves or could not recall their status for a given period. In particular, we drop individuals if they have gaps longer than 5 years in their labor force histories. This implies that we drop 1,064 individuals who attrited prior to age 45 leaving us with 8,440 individuals. Among these, there are a further 27 individuals who have a gap exceeding 5 years prior to their last appearance in the data. We are thus left with a final sample of **8,413 respondents**.

Table 13 lists the different steps resulting in loss of sample size. For steps 2 and 3, we separately highlight those dropped from the sample because they deceased prematurely and those lost because they either stopped answering the NLSY, have 5 year interview gaps or never reported employment after age 30. The main loss of raw observations is because we drop the military and the non-Black/non-Hispanic oversample. However, we only lost 6.2% of the weighted sample during this step. We loose another 16% of the population because they either deceased prior to age 45 or did not respond to the survey after age 45.

As we discuss below, we predict attrition from the sample using observable characteristics and then reweigh the data using these attrition weights on top of the NLSY provided cross-sectional weights. Thus, we assume that conditional on the variables used to generate the attrition weights, observations are missing at random.

Reason for drop	Dropped Individuals	Remaining Sample
Starting sample		12,686
Military over-sample	1,075 (0.5 %)	11,611 (99.5 %)
non-Black/non-Hispanic subsample	1,647 (5.7 %)	9,964 (93.8 %)
Deceased prior to age 30	128 (1.0 %)	9,836 (92.8 %)
Never reports E/UE status age 30+	332 (3.4 %)	9,504 (89.4 %)
Deceased prior to age 45	223 (1.9 %)	9,281 (87.4 %)
Not reaching age 45	841 (9.4 %)	8,440 (78.1 %)
Missing interview $>$ 5 years consecutively	27 (0.3 %)	8,413 (77.8 %)

Notes: Raw counts are unweighted. Percentages in parentheses are weighted shares of the original sample, with the denominator fixed throughout at the total survey weight of the 12,686 starting respondents. Because the discontinued oversamples were assigned small survey weights relative to their headcount, their weighted shares are substantially smaller than their raw count shares (e.g., the 1,075 dropped military respondents represent only 0.5% of the weighted population). The "Remaining Sample" percentage at each row therefore reflects the share of the original population that the analytical sample continues to represent after each exclusion.

Table 13: Sample attrition from full NLSY79 cohort

A.2 Imputing short gaps in the labor force histories

For the 8,413 respondents who remain, shorter gaps in the interior of the employment history are imputed. Specifically, we impute any consecutive missing spell of 52 weeks or fewer that is preceded and followed by observed labor force status. We do not impute gaps longer than 52 weeks, nor gaps that occur at the very beginning or end of an individual's labor force history. In those cases, we simply compute the moments used for clustering based on the histories excluding the missing gaps.

Below, we show the distribution of individuals by the longest gap in their labor force histories. 79.6% have complete histories and 94.5% do not have any gaps longer than 1 year. For the 14.9% with a gap shorter than 1 year, we impute their spells following the procedure outlined below.

1. We compress each individual's week-by-week employment history into a sequence of spells — maximal runs of consecutive weeks in the same labor market state (employed in job j, unemployed, OLF, or missing), characterized by their state and duration in weeks. We then form all overlapping triplets of consecutive spells. Triplets in which the middle spell is a missing-interview period serve as imputation

	Never missing	$\leq$ 1 yr	$\leq$ 2 yrs	$\leq$ 3 yrs	$\leq$ 4 yrs	$\leq$ 5 yrs
Count	6,627	7,941	8,316	8,384	8,401	8,413
Weighted %	79.6	94.5	98.6	99.3	99.5	99.7

Each column shows the number (and weighted share) of respondents whose longest consecutive missing spell in the prime-age window falls within the stated threshold. The sample is the 8,440 respondents observed until at least age 45. Respondents whose longest gap exceeds five consecutive years are dropped, leaving 8,413 in the final analytical sample.

Table 14: Distribution of longest consecutive missing spell

targets; the reference pool from which imputed states are drawn consists of those triplets with no missing spells.

2. Before applying the general procedure, we handle [E, missing, E] triplets where the same employer appears on both sides: gaps of fewer than 4 weeks are directly assigned to that employer, while longer gaps proceed to the general procedure.

3. The general procedure applied to all remaining targets proceeds as follows.

   (a) For any target triplet, we construct the relevant subset of the pool by matching on the type of the bracketing spells (E, UE, or OLF). For example, if the sequence of spells is [E,na,U] then we only consider triples from the pool with E and U as bracketing events.

   (b) Then, within the subset of the pool applying to this observation, we find the triplet whose vector of spell durations is most similar to the target's, measured by cosine similarity. The imputed state is the middle spell of that best-matching pool triplet, subject to two constraints: employment can only be imputed if at least one bracketing spell is an employment spell (to avoid creating an employer the individual never reported), and when employment is imputed, the job code is taken from the left bracketing spell.

4. Important exceptions to the general procedure arise when the target triplet do not have at least one employment spell bracketing it. Three are two possible cases:

   (a) The target spell has the same non-employment state bracketing the missing spell, for example $[U, na, U]$. In this case, the reference pool will by construction include only $[U, OLF, U]$ or $[U, E, U]$. In these cases, we retain $[U, OLF, U]$ if the cosine match delivers it. Should the cosine match be $[U, E, U]$, then we override the match and impute U for the missing period.

   (b) The target spell has different bracketing non-employment states, for example $[U, na, OLF]$. In this case, the reference pool has, by necessity only $[U, E, OLF]$ spells. There are a total of 68 such cases. For these, we impute for the missing period the non-employment state bracketting the missing period with the longer duration.

Appendix B Decisions in Clustering

B.1 Pre-assignment rules

In Section 2 we note that we preassign individuals if they are either always working or almost always non-employed to clusters $K^*.0$ and $K^*.K^*+1$ respectively. Those assigned to $K^*.K^*+1$ includes those with at least one non-employment spell lasting 10 years or longer or with cumulative employment between 30 and 50 lasting less than 2 years.

Table 15 illustrates the effect of this pre-assignment rules. It shows how individuals with long non-employment spells are distributed across the different clusters. For example, the last row shows that 7.3% of individuals in our sample have non-employment spells lasting longer than 10 years and are thus assigned to cluster 4.5.[26] The previous row shows that 10.3% of our sample have a non-employment spell lasting longer than 8 years. For 70.8% of these, this spell is longer than 10 years. Among those with a longest non-employment spell of 8 to 10 years, 52% (=0.152/(1-0.708) belong to cluster 4.4 in our preferred specification.

Overall, the table shows that the vast majority of individuals who we pre-assigned to cluster 4.5 belong to the clusters 4.3 to 4.5 if an alternative cut-off of 6 years or more is used. Thus, varying the cut-off rule does not change the fraction of the population nor the identity of the individuals assigned to one of the weakly attached types. Alternative cut-off rules therefore will not substantively affect the results of our prediction exercise as long as we choose cut-offs of 6 or more years.

Shorter cut-off rules than that start having a more substantive impact. For example, Gregory et al. (2025) choose a cut-off of 2 years for the longest NE spell and drop all individuals with longer non-employment spells. Table 15 shows that 32% of the population experience a spell of this length or longer. It also shows that 40.4% of this group or more than 10% of the total population would be assigned a highly attached type, 4.1 or 4.2, if included in the clustering exercise. This indicates that a cutoff for pre-assignment of 2 years or even of 4 years will lead to classify a substantial fraction of the population as weakly attached even though they are actually strongly attached to the labor market. A long cutoff, such as our choice of 10 years is preferable.

Longest NE spell	Pop share	Type distribution				
		4.1	4.2	4.3	4.4	4.5
$>$ 2 yrs	32.0	4.5	35.9	22.9	12.9	23.9
$>$ 4 yrs	20.6	1.2	21.7	22.0	18.1	37.1
$>$ 6 yrs	14.5	0.1	10.5	17.1	19.9	52.4
$>$ 8 yrs	10.3	0.0	3.7	9.0	15.2	72.2
$>$ 10 yrs	7.3	0.0	0.0	0.0	0.0	100.0

Note: NE is non-employment. Pop share is the population share of individuals with NE spells longer than first column threshold. Type distribution is distribution of these individuals across clusters from our preferred clustering specification, arranged from highest to lowest labor market attachment; δ' is preassigned to individuals with either longest NE spell over 10 years, or cumulative prime employment under 2 years.

Table 15: Population and type distribution of individuals with long NE samples

[26] Another 0.3% have total employment of less than 2 years but their longest NE spell is shorter than 10 years

B.2 Decisions related to clustering

B.2.1 Random initial centroids for clusters

k-means clustering proceeds as an iterative algorithm that can be affected by how the algorithm is initialized since it does not aim for a global minimum in equation 1. Therefore we restart the k-means clustering with different random seeds and use the one that generates minimum within-cluster sum of squared distances.

B.2.2 Choosing the number of clusters

We chose the optimal number of clusters $K^* = 4$ in our main results as described in 2.4. In this section, we use other metrics such as silhouette score (Rousseeuw (1987)) and Wang's score (Wang (2010)) and show consistency of the choice. The choice $K^* = 4$ maximizes both metrics defined in equations 3 and 4.

K	2	3	4	5	6	7	8
Silhouette score	0.320	0.354	0.355	0.319	0.334	0.340	0.314
Wang's score	95.3	60.0	97.5	62.2	58.3	88.6	78.9

Table 16: Scores for different number of clusters

The Silhouette score measures how close an observation is to observations in its own cluster, and how far away it is from observations in the nearest cluster. It is computed as

$$S = \sum_{\forall i} w_i \cdot \frac{b_i - a_i}{\max\{a_i, b_i\}} \tag{3}$$

where

$$a_i = \frac{1}{|C_k| - 1} \sum_{j \in C_k, j \neq i} d(i,j)$$
$$b_i = \min_{l \neq k} \frac{1}{|C_l|} \sum_{j \in C_l} d(i,j)$$

and C_k is the cluster i is in, $|C_k|$ is the size of the cluster and $d(i,j)$ is Euclidean distance between observations i, j.

To measure Wang's score, we partition the sample into a validation set V (50%) and a training set (50%), then further split the training set into two sub-samples T_1, T_2 of equal sizes (each 25% of the total). We estimate k-means clustering independently on each subsample, yielding two sets of centroids $\boldsymbol{\mu}_k^{(1)}$ and $\boldsymbol{\mu}_k^{(2)}$. Each validation observation i is then assigned to its nearest centroid under each model $C_i^{(1)}$ and $C_i^{(2)}$. Wang's score is the share of validation observations that are assigned into the same cluster under both models.

$$W = \frac{\sum_{\forall i \in V} I[C_i^{(1)} = C_i^{(2)}] \cdot w_i}{\sum_{\forall i \in V} w_i} \times 100 \tag{4}$$

Appendix C Covariate Construction

This appendix documents, in detail, how every explanatory variable used in the prediction exercises is derived from the NLSY79. Unless stated otherwise, variables are stored exactly under the names shown here (`typewriter` font) in the full dataset. The number of missing observations refers to our final clustered sample of 8413 individuals.

C.1 Baseline Demographics

All demographic controls are binary except for the birth year.

- `Gender` **(R0214800)** — Female = 1, Male = 0.
- `Black` **(R0214700)**, `Hispanic` — Ethnicity dummies; White is the reference group.
- `BirthYear` **(R0000500)** — Birth year minus 1957, so the earliest cohort is coded 0.

C.2 Skill Measures

The model incorporates three categories of skill proxies—cognitive, non-cognitive, and social. All indices are z-standardised (mean = 0, variance = 1). The construction of the non-cognitive and social skills uses the definitions of Deming (2017).

Cognitive skills (`Cognitive_Skills`) (R0618301) Armed Forces Qualification Test, administered in 1979 and re-normed in 2006 to yield age-adjusted percentile scores. (385 missing (same amount of missing if we took the previous versions of the AFQT))

Non-cognitive skills `(Non_Cognitive_Skills)` Mean of two 1979 psychological scales. If one of the non-cognitive measures is missing, the other one is assigned.

i. Rotter Locus of Control asked in 1979 (R0153710) (0 missing). This measures the degree to which subjects attribute successes and failures to their efforts. The measure correlates with motivation and resilience.

ii. Rosenberg Self-Esteem asked in 1980 (R0304410) (279 missing). The Rosenberg scale features positively and negatively worded items such as "I feel I have a number of good qualities." and "I feel I have not much to be proud of."

Relevant for the discussion of health in our paper, both measures both predict depression and anxiety.

Social skills `(Social_Skills)` Mean of four proxies:

i. Self-reported sociability in 1981 asked in 1985 (R1774300) (276 missing)

ii. Retrospective sociability at age 6 asked in 1985 (R1774400) (276 missing (same ones as the previous question))

iii. Total number of extracurricular clubs joined in high school asked in 1984 (as in Deming, if the data is missing for the club participation it counts as 0)

iv. High-school sports participation asked in 1984

C.3 Parental Background

We use parental full- and part-time work in 1979/1980, education, and household composition when the respondent was 14 as parental background variables.

Parental employment, 1979–80

Mother: work status 1979 (R0006800, 417 missing) & 1980 (R0223200, 609 missing)

Father: work status 1979 (R0008200, 1489 missing) & 1980 (R0223600, 1652 missing)

Respondents reported, for each parent, whether they had worked for pay in the previous calendar year: `1 = all year`, `2 = part of the year`, `3 = not at all`.

We convert the two annual answers into four mutually-exclusive indicators:

1. **Cross-fill missing**: if one year is missing, copy the non-missing value; if both years remain missing, set `Missing_Work_Status` = 1 and leave the other dummies 0.
2. **Assign employment category** (reference group = `Not_Worked`):

1979 value	1980 value	Assigned dummy
1	1	`Worked_All_Year`=1
1 & 2 *or* 2 & 1		`Worked_All_Year`=1
1 & 3 *or* 3 & 1		`Worked_Part_Year`=1
2	2	`Worked_Part_Year`=1
2 & 3 *or* 3 & 2		`Worked_Part_Year`=1
3	3	`Not_Worked`=1

Parental education: *MotherEduc_ (R0006500, 539 miss.) *FatherEduc_** (R0007900, 1 271 miss.)**

Four mutually exclusive indicators—`High_School_Graduate`, `Some_College`, `College_Graduate`, and `Missing_Education`. The omitted reference category is "less than high-school graduate." Following Lange (2013), cases with unreported schooling are retained via the `Missing_Education` dummy.

Household composition at age 14 (*"With whom were you living when you were 14?"*, R0001900)

The 30 original response codes are collapsed into four non-overlapping dummies that capture the presence of *biological* parents only:

`mom_dad`	1 = both biological parents present
`mom_only`	1 = mother present, father absent
`dad_only`	1 = father present, mother absent
`neitherMom_Dad`	1 = neither biological parent present

The regression baseline is `neitherMom_Dad`. If `R0001900` is missing (12 cases), the respondent is also assigned to `mom_dad`, mirroring the convention in Altonji, Bharadwaj & Lange (2013).

C.4 Geography

For each age $X \in \{22, \ldots, 29\}$ we derive three geographic indicators from the respondent's household interview:

- **Region (`Region_X`).** Census region of current residence, coded as

$$1 = \text{NE}, \quad 2 = \text{NC}, \quad 3 = \text{South}, \quad 4 = \text{West}$$

  (NC is the omitted reference category). *Missing-value handling:* carry forward the last non-missing code; if all waves are missing (15 obs.), set `Region_X`=2 (NC). (R0216400, R0405700, R0602810, R0897910, R1144800, R1520000, R1890700, R2257800, R2445200, R2870800, R3074500, R3401200, R3656600, R4007100, R4418200, R5081200, R5166500, R6479100, R7006800, R7704100, R8496500, T0988300, T2210300, T3108200, T4112700, T5023100, T5771000, T8219100, T8788300, T9300100)

- **SMSA status (`SMSA_X`).** SMSA/central-city classification, coded as

$$0 = \text{Non\_MSA}, \quad 1 = \text{MSA}, \quad 2 = \text{MSA\_central}$$

  (MSA_central is the omitted reference category).

  Missing-value handling: carry forward the last valid code; if all waves are missing (102 obs.), set `SMSA_X`=2 (MSA_central). (R0215200, R0393520, R0647000, R0897900, R1146400, R1521700, R1892400, R2259500, R2447000, R2872800, R3076500, R3403200, R3658600, R4009100, R4420200, R5083200, R5168500, R6481300, R7009000, R7706300, R8498700, T0990500, T2212300, T3110200, T4114700, T5026000, T5774100, T8221300, T8790500, T9302300)

- **Urban status (`Urban_X`).** Urban/rural indicator, coded as

$$0 = \text{Rural}, \quad 1 = \text{Urban}$$

  (Rural is the omitted reference category). *Missing-value handling:* carry forward the last non-missing code; if all waves are missing (85 obs.), set `Urban_X`=0 (Rural). (R0215100, R0393510, R0646900, R0897800, R1146500, R1521600, R1892300, R2259400, R2446900, R2872700, R3076400, R3403100, R3658500, R4009000, R4420100, R5083100, R5168400, R6481200, R7008900, R7706200, R8498600, T0990400, T2212200, T3110100, T4114600, T5025900, T5774000, T8221200, T8790400, T9302200)

C.5 Education Status (ages 22–29)

For each age $X \in \{22, \ldots, 29\}$ we capture both enrollment and the respondent's highest credential to date. Education dummies are defined identically to the parental-schooling indicators—three mutually exclusive

flags with "high-school dropout or less" as the reference category:

- **Credential dummies** We map the May-interview `HGCREV` value at age X into three mutually exclusive indicators:

  `HSGrad_X` = 1 if highest degree is a high-school diploma/GED.

  `SomeCollege_X` = 1 if some post-secondary coursework but no bachelor's degree.

  `CollegeGrad_X` = 1 if bachelor's degree or higher.

  Missing-value handling: we carry the last non-missing `HGCREV` forward, so that any gap is filled with the most recent reported credential. We classify the the very few (20) individuals that did not report any schooling between 22 and 29 and were part of the clustering as high-school drop outs.

  (R0216701, R0406401, R0618901, R0898201, R1145001,R1520201, R1890901, R2258001, R2445401, R2871101,R3074801, R3401501, R3656901, R4007401, R4418501,R5103900, R5166901, R6479600, R7007300, R7704600,R8497000, T0988800)

- **School enrollment** – `E_X` = 1 if the respondent is currently enrolled at age X (variable `ENROLLMTRE`). When missing we looked at the highest degree completed in the lifetime and if the individual already completed it, then we assign 0 to enrollment. For the remaining missing cases, we create a flag for missingness to maintain the respondents in the sample.

  Missing-value handling: If `ENROLLMTRE` is missing but the respondent's `HGCREV` at age X already equals their maximum lifetime credential, we set `E_X`=0. In all other cases of missing `ENROLLMTRE`, we leave `E_X` as missing and introduce a companion indicator `missing_enroll_X`=1 to retain the observation.

  (R0216601, R0406501, R0619001, R0898301, R1145101, R1520301, R1891001, R2258101, R2445501, R2871201, R3074901, R3401601, R3657001, R4007501, R4418601, R5104000, R5166902, R6479700, R7007400, R7704700, R8497100, T0988900)

C.6 Health Indicators

Every survey wave of the NLSY79 includes a brief health battery that focuses on respondents' capacity to engage in paid work. Three questions are particularly relevant: (i) whether health limits the *kind* of work one can perform; (ii) whether health limits the *amount* of work one can perform; and (iii) whether health would *completely* prevent the respondent from working for pay at that moment. We translate these items, into four age–specific flags:

- **Unable to work (`prevent_working_X`).** Set to 1 for age X if the respondent answers "Yes" to "Would your health prevent you from working at a job for pay now?"
  (R0144800, R0298800, R0478100, R0776300, R1021000, R1390300, R1773200, R2140700, R2348500, R2710900, R2959000, R3270400, R3557900, R3885500, R4283900, R4961100, R5616600, R6343600, R6887200, R7597600, R8297400, T0895700, T2051500, T3022800, T3953000, T4890700, T5593800, T8088100, T8620200)

- **Limit on kind of work (`limit_kind_work_X`).** Set to 1 for age X if the respondent answers "Yes" to "Does your health limit the kind of work you can do?"
(R0144900, R0298900, R0478200, R0776400, R1021100, R1390400, R1773300, R2140800, R2348600, R2711000, R2959100, R3270500, R3558000, R3885600, R4284000, R4961200, R5616700, R6343700, R6887300, R7597700, R8297500, T0895800, T2051600, T3022900, T3953100, T4890800, T5593900, T8088200, T8620300)

- **Limit on amount of work (`limit_amount_work_X`).** Set to 1 for age X if the respondent answers "Yes" to "Does your health limit the amount of work you can do?"
(R0145000, R0299000, R0478300, R0776500, R1021200, R1390500, R1773400, R2140900, R2348700, R2711100, R2959200, R3270600, R3558100, R3885700, R4284100, R4961300, R5616800, R6343800, R6887400, R7597800, R8297600, T0895900, T2051700, T3023000, T3953200, T4890900, T5594000, T8088300, T8620400)

- **Missing Health (`missing_health_X`).** Set to 1 for age X if the respondent has no valid response on any of the three health-limitation items (due to non-response, invalid codes, or skip patterns). We apply only two simple imputations before flagging missing cases:

  (i) If `prevent_working_X`=1, we infer by design that the follow-up questions were skipped and set both `limit_kind_work_X` and `limit_amount_work_X` to 0.

  (ii) If the respondent provides a valid (0/1) answer to any one health-limitation item, we assume the other items were administered and answered "No," and set any missing values among them to 0.

  All remaining cases—where none of the three indicators is observed after these rules—are coded as `missing_health_X`=1.

C.7 Labor-Market Outcomes

To describe respondents' attachment to the labor market before 30 we construct age–specific variables, each carrying the suffix `_$X$` for ages $22 \leq X \leq 29$. They capture the extensive margin (whether the respondent worked at all), the intensive margin (how many months were spent out of work and the wage earned when employed), and the sectoral orientation (industry and occupation).

- **Employment flag (`Empl_Flag_X`).** Equals 1 if the respondent engaged in any paid work during age X, based on the annual employment–status calendar; 0 otherwise.

- **Non-employment share (`NE_fraction_X`).** Measures the proportion of the year spent out of work. Specifically, we divide the total number of non-employment weeks reported for year X by 52. Values therefore range from 0 (continuously employed) to 1 (never employed).

- **Wage with missing-data guard (`wage_empl_int_X`).** Hourly wage observed for the main job *multiplied* by `Empl_Flag_X`. The transformation sets wages to zero for non-workers rather than coding them as missing, which prevents listwise deletion in models that include wage levels.

- **Modal-industry indicator (`Industries_X`).** Equal to 1 if, at age X, the respondent is employed in the *industry* where they have accumulated the most weeks of experience up to that point; 0 otherwise. The underlying industry is identified using the cumulative weeks-by-three-digit SIC codes.
- **Modal-occupation indicator (`Occupations_X`).** Constructed analogously to `Industries_X` but based on three-digit Census occupation codes, flagging whether the respondent is currently working in their most experienced occupation.

C.8 Prison (ages 22–29)

For each age $X \in \{22, \ldots, 29\}$ we construct an "ever-incarcerated" flag:

- **Any_jail_X.** Equals 1 if the respondent reports "jail" as their `TYPE_OF_RESIDENCE` in *any* interview up to (and including) the one when they turn age X, and 0 otherwise.

  Missing-value handling: codes of –4 ("not asked") and other nonresponse are treated as non-jail (0).

  (R0188000, R0402800, R0612100, R0828400, R1075700, R1451400, R1798600, R2160200, R2369100, R2500000, R2900000, R3100000, R3500000, R3700000, R4100300, R4500300, R5200300, R5800200, R6530300, R7090700, R7800600, T0001000, T1200800, T2260700, T3195700, T4181500, T5152100, T7721800, T8331800)

C.9 Children Indicators

The NLSY79 records the cumulative number of live births each wave. From this series we derive two sets of variables: (i) the respondent's age at first birth, and (ii) age–specific family-size dummies for ages 22–29.

Age at first birth (`First_Birth_Age`). Respondent's age (in years) at the first wave when "Number of children ever born" increases from 0 to 1.

Missing-value handling: If the entire child-count series is blank, set `First_Birth_Age_Missing`=1 and leave `First_Birth_Age` blank; otherwise fill single-wave gaps as in step 3 below.

Family–size dummies (`child_a`). For each age $X \in \{22, \ldots, 29\}$ we bin the number of children 0 (no children), 1–2 children, and 3 or more children. We then one-hot encode into

$$\texttt{child\_1\_2\_}X, \quad \texttt{child\_3plus\_}X, \quad \texttt{child\_missing\_}X.$$

Missing-value handling: If age X is missing but both $X-1$ and $X+1$ report the same count, fill age X with that value. Any residual missing at age X is flagged by `child_missing_a` = 1.

Reference group: the model's baseline is "0"

R0013400, R0389000, R0414000, R0898837, R1146829, R1522036, R1892736, R2259836, R2448036, R2877500, R3076841, R3407600, R3659046, R4009446, R4444600, R5087400, R5172700, R6486300, R7014100, R7711700, R8504200, T0995900, T2217700, T3115700, T4120200, T5031400, T5779600, T8226700, T8796000

Appendix D Sampling variation and Overfitting

D.1 Inference

Since the cluster assignments are themselves derived from sample, cluster level standard errors understate the variability of the estimates. Therefore, confidence intervals estimated via bootstrap provides valid inference on cluster characteristics by propagating uncertainty. Moreover, the stability of population share (and other moments) across bootstrap samples indicates that our clustering reflects genuine structure of the data rather than an arbitrary partition of the sample.

1. Resampling with replacement
   In each bootstrap iteration, $N = 8413$ individuals were drawn with replacement from the original sample of size N with probability of drawing is proportional to the $\frac{w_i}{\sum_{\forall i} w_i}$, where w_i is sample weight of individual i.

2. Before clustering we preassigned individuals to $K.0$ or $K.K+1$ using the same rules as in the main analysis.

3. Then, for $K = 2, 3, 4, 5, 6$, we clustered the remaining samples using E avg week duration, NE avg week duration, OLF/total weeks, U/total weeks and Jobs per quarter in E.

4. After clustering, we recomputed NE/total weeks, OLF/NE and U/NE.

5. We relabeled clusters in the order of NE/total weeks.

6. We repeated the steps 1 to 4 1,000 times and computed the 2.5th and 97.5th percentiles of each moment for each type $K.i$ to construct 95% confidence intervals.

7. Point estimates presented are averages of the main sample using reweighted weights for the consistency with the main table.

D.2 Cross-Validation

A concern regarding the prediction model is that we compare model performance using a criteria that is related to the objective function that our estimator maximizes. That is, the precision we report at various times is directly related to the likelihood maximized by the logit estimator used to predict type. As the number of controls and parameters increases, the fit in sample will necessarily get better and the precision will tend to improve, even if the ability to predict out of sample will not. In other words, we might be overfitting the data, a concern that especially applies once the number of parameters grows.

To gauge out-of-sample accuracy, we run a stratified five-fold cross-validation. The full sample is randomly split into five folds. By stratifying by weakly and highly attached types, we ensure that we replicate the overall share of weakly attached workers in each fold. For each fold, we re-estimate the model on the other four folds, generate predicted probabilities for the omitted fold. We then calculate precision using the fold not used in the prediction. Averaging across the five iterations yields the mean precision and its (small) standard deviation, reported in Table 21.

Clustering design	$K=2$		$K=3$		
	2.1	2.2	3.1	3.2	3.3
Population share	39.9	36.5	38.7	33.5	4.2
	[35.2, 45.4]	[31.4, 41.2]	[33.5, 42.7]	[29.4, 38.4]	[3.4, 4.9]
Clustering moments					
E avg week duration	386.1	113.8	376.5	108.7	317.5
	[357.7, 418.9]	[109.1, 126.7]	[360.2, 412.6]	[102.4, 123.9]	[258.2, 354.3]
NE avg week duration	33.3	53.7	21.8	41.6	249.9
	[21.7, 34.1]	[47.4, 67.2]	[19.2, 21.5]	[35.3, 42.3]	[223.9, 274.6]
NE/total weeks	6.0	28.7	4.8	27.0	47.4
	[4.7, 6.0]	[24.8, 31.8]	[3.9, 5.0]	[22.5, 28.0]	[44.5, 51.4]
OLF/NE	73.9	76.6	68.2	74.6	90.5
	[68.1, 76.5]	[77.8, 80.6]	[66.6, 71.2]	[74.9, 78.0]	[91.1, 94.2]
U/NE	26.1	23.4	31.8	25.4	9.5
	[23.5, 31.9]	[19.4, 22.2]	[28.8, 33.4]	[22.0, 25.1]	[5.8, 8.9]
Jobs per quarter in E	0.07	0.22	0.07	0.23	0.09
	[0.06, 0.08]	[0.21, 0.24]	[0.06, 0.08]	[0.21, 0.25]	[0.08, 0.12]

Table 17: Bootstrap 95% CI ($K = 2, 3$)

Clusters	4.0	4.1	4.2	4.3	4.4	4.5
Population share	15.9	23.6	38.1	10.6	4.1	7.6
	[15.8, 17.6]	[22.8, 25.4]	[36.6, 40.2]	[8.1, 11.8]	[3.3, 4.6]	[6.3, 7.5]
Clustering moments						
E avg week duration	1031.5	473.2	169.0	66.4	305.0	82.0
	[1029.9, 1033.9]	[464.7, 487.1]	[163.0, 176.4]	[61.2, 75.3]	[263.9, 338.9]	[81.6, 103.6]
NE avg week duration	0.0	21.0	31.4	52.9	251.7	491.2
	[0.0, 0.0]	[17.9, 21.2]	[28.7, 32.2]	[45.0, 54.2]	[226.0, 279.8]	[454.3, 514.4]
NE/total weeks	0.0	3.1	14.9	42.3	48.5	83.3
	[0.0, 0.0]	[2.6, 3.2]	[13.5, 15.4]	[36.9, 45.1]	[45.8, 52.0]	[80.3, 83.2]
OLF/NE	nan	67.5	71.3	77.2	90.5	93.6
	[nan, nan]	[63.9, 70.4]	[71.2, 74.6]	[77.7, 82.0]	[91.2, 94.3]	[94.3, 95.9]
U/NE	nan	32.5	28.7	22.8	9.5	6.4
	[nan, nan]	[29.6, 36.1]	[25.4, 28.8]	[18.0, 22.3]	[5.7, 8.8]	[4.1, 5.7]
Jobs per quarter in E	0.04	0.06	0.14	0.37	0.09	0.61
	[0.04, 0.04]	[0.05, 0.06]	[0.13, 0.15]	[0.35, 0.42]	[0.09, 0.11]	[0.46, 0.88]

Table 18: Bootstrap 95% CI ($K = 4$)

Clusters	5.1	5.2	5.3	5.4	5.5
Population share	15.2	29.5	22.1	3.9	5.7
	[14.6, 27.8]	[1.8, 34.9]	[8.4, 36.4]	[3.3, 10.3]	[1.6, 5.9]
Clustering moments					
E avg week duration	550.0	246.4	109.8	302.7	55.8
	[405.6, 558.5]	[169.7, 971.9]	[66.1, 159.7]	[54.8, 321.9]	[50.5, 413.3]
NE avg week duration	24.1	21.7	42.2	257.0	61.1
	[16.5, 24.5]	[18.6, 72.4]	[30.9, 44.5]	[47.0, 276.5]	[55.8, 363.5]
NE/total weeks	2.9	7.0	25.3	48.6	50.6
	[2.4, 3.4]	[5.7, 11.4]	[15.3, 38.9]	[39.0, 50.8]	[47.1, 54.4]
OLF/NE	72.1	67.9	73.3	90.4	78.7
	[61.9, 76.8]	[65.0, 94.1]	[72.1, 78.1]	[78.5, 93.7]	[80.2, 96.0]
U/NE	27.9	32.1	26.7	9.6	21.3
	[23.2, 38.1]	[5.9, 35.0]	[21.9, 27.9]	[6.3, 21.5]	[4.0, 19.8]
Jobs per quarter in E	0.05	0.09	0.20	0.09	0.45
	[0.05, 0.06]	[0.03, 0.14]	[0.14, 0.40]	[0.09, 0.48]	[0.06, 0.49]

Table 19: Bootstrap 95% CI ($K = 5$)

Clusters	6.1	6.2	6.3	6.4	6.5	6.6
Population share	15.1	27.4	19.2	7.9	1.6	5.3
	[14.7, 27.2]	[1.7, 33.0]	[15.7, 32.9]	[5.2, 9.4]	[1.4, 8.8]	[1.3, 6.4]
Clustering moments						
E avg week duration	552.4	250.0	108.0	169.8	405.5	55.6
	[407.7, 559.0]	[230.3, 973.9]	[93.2, 166.8]	[59.5, 184.7]	[53.0, 445.5]	[43.2, 421.3]
NE avg week duration	23.7	19.5	27.0	127.1	353.5	57.7
	[15.5, 23.9]	[17.7, 74.9]	[21.2, 28.8]	[40.0, 136.2]	[48.1, 379.4]	[55.7, 374.3]
NE/total weeks	2.8	6.4	19.5	45.0	49.5	50.2
	[2.4, 3.0]	[5.6, 8.0]	[11.3, 22.8]	[36.2, 46.8]	[42.5, 53.1]	[48.1, 61.0]
OLF/NE	71.8	66.2	65.9	87.2	93.1	77.8
	[58.9, 76.5]	[65.3, 94.7]	[64.3, 69.5]	[75.6, 91.5]	[77.9, 96.0]	[81.1, 96.9]
U/NE	28.2	33.8	34.1	12.8	6.9	22.2
	[23.5, 41.1]	[5.3, 34.7]	[30.5, 35.7]	[8.5, 24.4]	[4.0, 22.1]	[3.1, 18.9]
Jobs per quarter in E	0.05	0.09	0.21	0.14	0.07	0.46
	[0.05, 0.06]	[0.03, 0.10]	[0.14, 0.26]	[0.13, 0.44]	[0.05, 0.49]	[0.06, 0.56]

Table 20: Bootstrap 95% $K = 6$

Table 21: Top-5% Predicted Risk: Mean Predicted Probability and Realized Rate

Specification	$\hat{p}$ (in-sample)	y (in-sample)	$\hat{p}$ (CV)	y (CV)
Demographics	0.39	0.31	0.39 (0.00)	0.30 (0.04)
+Health	0.66	0.63	0.66 (0.03)	0.59 (0.03)
+Employment Histories	0.80	0.73	0.80 (0.02)	0.67 (0.08)
+Occupations + Industries	0.80	0.75	0.81 (0.01)	0.66 (0.11)
+Jail	0.81	0.77	0.82 (0.01)	0.67 (0.10)
+Cog+NCog+Education	0.82	0.79	0.83 (0.01)	0.68 (0.07)
+Family Variables+Local Conditions	0.83	0.79	0.84 (0.01)	0.69 (0.08)
+Children	0.83	0.79	0.84 (0.01)	0.68 (0.08)
N		8,413		

Notes: Each row reports results for a nested logit specification of low-attachment status on the listed covariate blocks. Within each specification, the 5% of individuals with the highest predicted probability $\hat{p}$ are selected, and we report (i) the mean predicted probability $\bar{\hat{p}}$ within this group and (ii) the realized share of low-attachment outcomes $\bar{y}$. Columns 1–2 use predictions from a model fit on the full sample; columns 3–4 use K-fold cross-validated predictions: the sample is partitioned into K folds, and each observation's $\hat{p}_i$ is generated from a model estimated on the other $K-1$ folds and evaluated on the held-out fold. Values in parentheses are standard deviations across folds. Low-attachment workers are defined as groups 4.3 and 4.4 together with individuals with virtually no employment spells. Sample size $N = 8{,}413$.

D.3 Attrition Reweighting via IPW

To correct for non-random attrition between ages 30–50, we construct inverse-probability-of-attrition weights (IPW) in the following steps:

1. **Define the stayer indicator.** Let

$$S_i = \begin{cases} 1 & \text{if individual } i \text{ is observed in the clustering sample (ages 30–50),} \\ 0 & \text{otherwise.} \end{cases}$$

2. **Assemble baseline covariates.** The predictor vector X_i includes all variables observable by age 22, grouped into the following blocks: (i) demographics (gender, Black, Hispanic, birth year); (ii) cognitive, non-cognitive, and social skill measures, with corresponding missing-value indicators; (iii) family structure at age 14 (number of siblings, parental presence) and parental education and employment in 1979–80; (iv) education status at age 22 (enrollment, highest credential); (v) labor-market outcomes at age 22 (non-employment share, one-digit occupation, one-digit industry); (vi) geography at age 22 (Census region, MSA status, urban/rural); (vii) fertility (first birth, three-plus children) and health limitations at age 22; and (viii) any incarceration by age 22. Full variable definitions appear in Appendix C.

3. **Estimate stay-probabilities.** Fit a logistic regression

$$\widehat{\pi}_i = \Pr\big(S_i = 1 \mid X_i\big) = \frac{\exp\big(X_i'\beta\big)}{1 + \exp\big(X_i'\beta\big)},$$

   where X_i is the vector of baseline covariates.

4. **Compute raw IPW.** For each stayer ($S_i = 1$), set

$$w_i^{\text{raw}} = \frac{1}{\widehat{\pi}_i}, \quad w_i^{\text{raw}} = 0 \quad \text{if } S_i = 0.$$

5. **Normalize weights to mean=1.** To anchor the weighted sample size to the unweighted one, define

$$w_i^{\text{norm}} = \frac{w_i^{\text{raw}}}{\dfrac{1}{N}\displaystyle\sum_{j=1}^{N} w_j^{\text{raw}}}.$$

   This ensures $\frac{1}{N}\sum_i w_i^{\text{norm}} = 1$.

6. **Combine with design weights.** If d_i denotes the original sampling (design) weight, the final analysis weight is

$$w_i^{\text{final}} = d_i \times w_i^{\text{norm}}.$$

This procedure delivers consistent, population-representative cluster assignments under the standard assumption.

$$\{Y_i, \ldots\} \perp\!\!\!\perp S_i \mid X_i.$$

The correlation between the baseline survey weights and the IPW weights is high ($\rho = 0.956$), indicating that our observed covariates have limited predictive power for sample retention; consequently, the reweighting only modestly perturbs the original weights.

Appendix E Robustness Across Estimators

This subsection presents the results of the prediction exercise using different estimators. All metrics are estimated out of sample using a 5-fold cross-validation method.

E.1 Unpenalized Logit

This is a standard logistic regression estimated by maximum likelihood, with sample weights equal to the IPW attrition weights. No coefficient penalty is imposed: when all seven covariate blocks are included, the model fits roughly 150 free parameters on each gender subsample of about 4,000 individuals. As a result, this specification is especially prone to overfitting in the gender splits — visible in the collapse of the out-of-sample R^2 relative to its in-sample counterpart, and in the appearance of negative Shapley shares for blocks whose marginal contribution does not survive cross-validation.

Table 22: Logit — Women: Shapley–Shorrocks decomposition of predictive performance.

	R^2	AUC	Prec at 1%	Prec at 2%	Prec at 5%	Prec at 10%
Performance						
Baseline	-0.00	0.50	0.35	0.34	0.33	0.33
Full	0.05	0.71	0.92	0.85	0.71	0.65
Gain	0.06	0.20	0.57	0.51	0.39	0.32
Block shares of gain (%)						
Skills & education	11.06	10.71	12.59	14.34	11.72	14.52
Family & geography	-13.85	8.23	3.72	-0.27	4.37	1.70
Employment history	128.12	71.02	23.68	29.12	37.32	48.16
Occupation & industry	-3.45	6.09	14.14	18.63	22.06	18.96
Health	-18.51	3.44	46.45	34.37	21.98	14.49
Children	-3.17	0.67	0.56	4.60	2.97	2.77
Incarceration	-0.21	-0.16	-1.14	-0.79	-0.42	-0.61
Total	100.00	100.00	100.00	100.00	100.00	100.00

Notes. Model: logit estimated by maximum likelihood (unpenalized). $N = 4{,}275$. Predictors measured at age 29; outcome is an indicator for the unattached-worker type. *Baseline* includes demographics (age, ethnicity) only; *Full* adds all seven covariate blocks; *Gain* = Full − Baseline. Block shares are exact Shapley–Shorrocks values over $2^7 = 128$ coalitions, normalized by *Gain*, and sum to 100% by construction. R^2 is McFadden pseudo-R^2; AUC is the area under the ROC curve; Prec at m% is expected precision in the top m% of observations ranked by predicted probability. All metrics use pooled 5-fold stratified cross-validated probabilities (`random_state`=42). IPW attrition-corrected sample weights used throughout.

Table 23: Logit — Men: Shapley–Shorrocks decomposition of predictive performance.

	R^2	AUC	Prec at 1%	Prec at 2%	Prec at 5%	Prec at 10%
Performance						
Baseline	0.03	0.60	0.26	0.27	0.26	0.28
Full	0.11	0.77	0.86	0.75	0.65	0.54
Gain	0.09	0.17	0.60	0.48	0.40	0.26
Block shares of gain (%)						
Skills & education	9.17	15.62	4.67	8.55	5.07	5.74
Family & geography	-16.37	3.07	3.93	1.09	4.36	-3.31
Employment history	93.87	61.36	20.62	35.23	45.05	59.06
Occupation & industry	-5.15	5.35	17.98	15.58	5.45	4.07
Health	-17.39	0.36	37.78	22.83	8.35	7.54
Children	2.25	2.63	1.20	0.94	0.67	0.07
Incarceration	33.62	11.61	13.81	15.78	31.06	26.84
Total	100.00	100.00	100.00	100.00	100.00	100.00

Notes. Model: logit estimated by maximum likelihood (unpenalized). $N = 4{,}138$. Predictors measured at age 29; outcome is an indicator for the unattached-worker type. *Baseline* includes demographics (age, ethnicity) only; *Full* adds all seven covariate blocks; *Gain* = Full − Baseline. Block shares are exact Shapley–Shorrocks values over $2^7 = 128$ coalitions, normalized by *Gain*, and sum to 100% by construction. R^2 is McFadden pseudo-R^2; AUC is the area under the ROC curve; Prec at m% is expected precision in the top m% of observations ranked by predicted probability. All metrics use pooled 5-fold stratified cross-validated probabilities (`random_state`=42). IPW attrition-corrected sample weights used throughout.

E.2 Random Forest

The random forest is the `scikit-learn` `RandomForestClassifier` with 500 trees, $\sqrt{p}$ candidate features per split, a minimum leaf size of 5, bootstrap sampling, and a fixed random seed. Sample weights enter both the bootstrap draws and the impurity calculations. Out-of-sample probabilities are obtained from the same 5-fold stratified split used for the linear models and are subsequently Platt-calibrated within each fold before metric evaluation.

E.3 Gradient Boosting

The gradient-boosting model is the `scikit-learn` `HistGradientBoostingClassifier`. Hyperparameters were tuned with 50 trials of Bayesian optimization over an inner cross-validation split, yielding `learning_rate` $\approx$ 0.014, `max_iter` = 977, `max_leaf_nodes` = 7, `min_samples_leaf` = 40, and `l2_regularization` $\approx$ 0.009. Out-of-sample probabilities use the same five-fold split as the other estimators and are Platt-calibrated within each fold so that the McFadden R^2 remains comparable across models.

Table 24: Random forest — Women: Shapley–Shorrocks decomposition of predictive performance.

	R^2	AUC	Prec at 1%	Prec at 2%	Prec at 5%	Prec at 10%
Performance						
Baseline	-0.00	0.50	0.27	0.32	0.30	0.31
Full	0.14	0.75	0.86	0.84	0.79	0.67
Gain	0.14	0.24	0.58	0.52	0.49	0.37
Block shares of gain (%)						
Skills & education	6.84	10.94	22.03	17.79	11.99	13.94
Family & geography	2.22	6.28	-3.97	-0.75	3.18	2.67
Employment history	76.44	62.90	32.02	45.40	46.51	46.14
Occupation & industry	6.41	8.47	13.60	16.26	18.27	19.15
Health	8.01	10.93	34.42	24.38	18.08	13.45
Children	0.16	0.78	0.90	-1.57	1.78	5.05
Incarceration	-0.07	-0.29	1.00	-1.50	0.17	-0.40
Total	100.00	100.00	100.00	100.00	100.00	100.00

Notes. Model: random forest classifier (500 trees, `min_samples_leaf`=5, `max_features`="sqrt", `random_state`=42). N = 4,275. Predictors measured at age 29; outcome is an indicator for the unattached-worker type. *Baseline* includes demographics (age, ethnicity) only; *Full* adds all seven covariate blocks; *Gain* = Full − Baseline. Block shares are exact Shapley–Shorrocks values over $2^7 = 128$ coalitions, normalized by *Gain*, and sum to 100% by construction. R^2 is McFadden pseudo-R^2; AUC is the area under the ROC curve; Prec at m% is expected precision in the top m% of observations ranked by predicted probability. All metrics use pooled 5-fold stratified cross-validated probabilities (`random_state`=42). IPW attrition-corrected sample weights used throughout.

Table 25: Random forest — Men: Shapley–Shorrocks decomposition of predictive performance.

	R^2	AUC	Prec at 1%	Prec at 2%	Prec at 5%	Prec at 10%
Performance						
Baseline	0.02	0.59	0.24	0.23	0.27	0.29
Full	0.23	0.82	0.91	0.80	0.74	0.61
Gain	0.21	0.22	0.68	0.57	0.48	0.32
Block shares of gain (%)						
Skills & education	5.38	11.20	1.91	0.84	3.61	0.98
Family & geography	4.02	6.48	8.02	8.72	3.65	3.13
Employment history	64.55	58.48	40.65	40.55	44.20	57.34
Occupation & industry	4.92	5.65	16.78	13.64	7.17	4.20
Health	10.46	9.18	25.47	24.49	15.95	12.76
Children	-0.46	0.40	-1.68	-0.76	-0.31	-0.89
Incarceration	11.12	8.62	8.85	12.51	25.73	22.49
Total	100.00	100.00	100.00	100.00	100.00	100.00

Notes. Model: random forest classifier (500 trees, `min_samples_leaf`=5, `max_features`="sqrt", `random_state`=42). N = 4,138. Predictors measured at age 29; outcome is an indicator for the unattached-worker type. *Baseline* includes demographics (age, ethnicity) only; *Full* adds all seven covariate blocks; *Gain* = Full − Baseline. Block shares are exact Shapley–Shorrocks values over $2^7 = 128$ coalitions, normalized by *Gain*, and sum to 100% by construction. R^2 is McFadden pseudo-R^2; AUC is the area under the ROC curve; Prec at m% is expected precision in the top m% of observations ranked by predicted probability. All metrics use pooled 5-fold stratified cross-validated probabilities (`random_state`=42). IPW attrition-corrected sample weights used throughout.

Table 26: Gradient boosting — Women: Shapley–Shorrocks decomposition of predictive performance.

	R^2	AUC	Prec at 1%	Prec at 2%	Prec at 5%	Prec at 10%
Performance						
Baseline	-0.00	0.50	0.27	0.33	0.30	0.31
Full	0.12	0.73	0.88	0.84	0.74	0.68
Gain	0.12	0.23	0.61	0.51	0.45	0.37
Block shares of gain (%)						
Skills & education	4.92	9.63	18.97	13.83	12.00	13.89
Family & geography	2.09	5.94	5.76	5.56	2.62	-1.71
Employment history	80.59	67.71	30.12	33.98	45.97	51.51
Occupation & industry	3.53	7.25	13.48	15.31	15.91	15.80
Health	8.82	8.71	33.41	33.11	21.02	14.68
Children	0.04	0.75	-1.75	-1.79	2.48	5.83
Incarceration	0.00	0.00	0.00	0.00	0.00	0.00
Total	100.00	100.00	100.00	100.00	100.00	100.00

Notes. Model: histogram gradient-boosting classifier (`learning_rate`=0.0141, `max_iter`=977, `max_leaf_nodes`=7, `min_samples_leaf`=40, `l2_regularization`=0.009, `random_state`=42). N = 4,275. Predictors measured at age 29; outcome is an indicator for the unattached-worker type. *Baseline* includes demographics (age, ethnicity) only; *Full* adds all seven covariate blocks; *Gain* = Full − Baseline. Block shares are exact Shapley–Shorrocks values over $2^7 = 128$ coalitions, normalized by *Gain*, and sum to 100% by construction. R^2 is McFadden pseudo-R^2; AUC is the area under the ROC curve; Prec at m% is expected precision in the top m% of observations ranked by predicted probability. All metrics use pooled 5-fold stratified cross-validated probabilities (`random_state`=42). IPW attrition-corrected sample weights used throughout.

Table 27: Gradient boosting — Men: Shapley–Shorrocks decomposition of predictive performance.

	R^2	AUC	Prec at 1%	Prec at 2%	Prec at 5%	Prec at 10%
Performance						
Baseline	0.02	0.59	0.23	0.24	0.27	0.29
Full	0.22	0.81	0.79	0.78	0.66	0.60
Gain	0.20	0.22	0.55	0.54	0.39	0.30
Block shares of gain (%)						
Skills & education	1.09	7.82	-1.44	-1.35	-1.84	-4.23
Family & geography	3.39	7.28	2.38	4.75	3.31	3.62
Employment history	70.99	62.81	32.48	41.94	52.24	61.55
Occupation & industry	3.41	4.58	24.31	14.27	6.04	3.37
Health	8.85	7.24	35.92	22.32	13.12	12.31
Children	0.05	1.26	-4.48	-1.35	-2.36	-0.59
Incarceration	12.22	9.02	10.84	19.42	29.49	23.97
Total	100.00	100.00	100.00	100.00	100.00	100.00

Notes. Model: histogram gradient-boosting classifier (`learning_rate`=0.0141, `max_iter`=977, `max_leaf_nodes`=7, `min_samples_leaf`=40, `l2_regularization`=0.009, `random_state`=42). N = 4,138. Predictors measured at age 29; outcome is an indicator for the unattached-worker type. *Baseline* includes demographics (age, ethnicity) only; *Full* adds all seven covariate blocks; *Gain* = Full − Baseline. Block shares are exact Shapley–Shorrocks values over $2^7 = 128$ coalitions, normalized by *Gain*, and sum to 100% by construction. R^2 is McFadden pseudo-R^2; AUC is the area under the ROC curve; Prec at m% is expected precision in the top m% of observations ranked by predicted probability. All metrics use pooled 5-fold stratified cross-validated probabilities (`random_state`=42). IPW attrition-corrected sample weights used throughout.

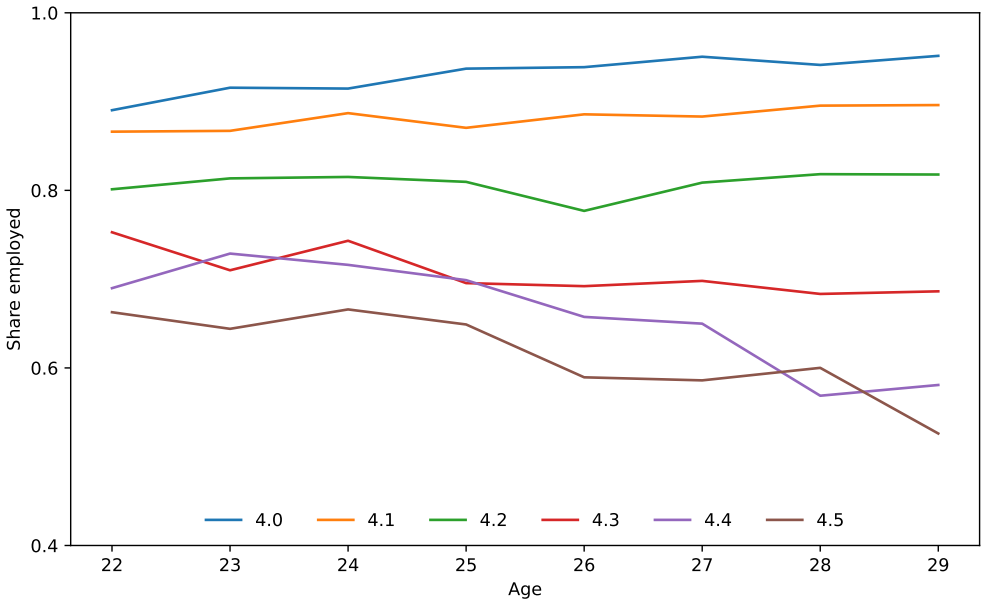


Figure 5: Share working by age and cluster, women, ages 22 to 29

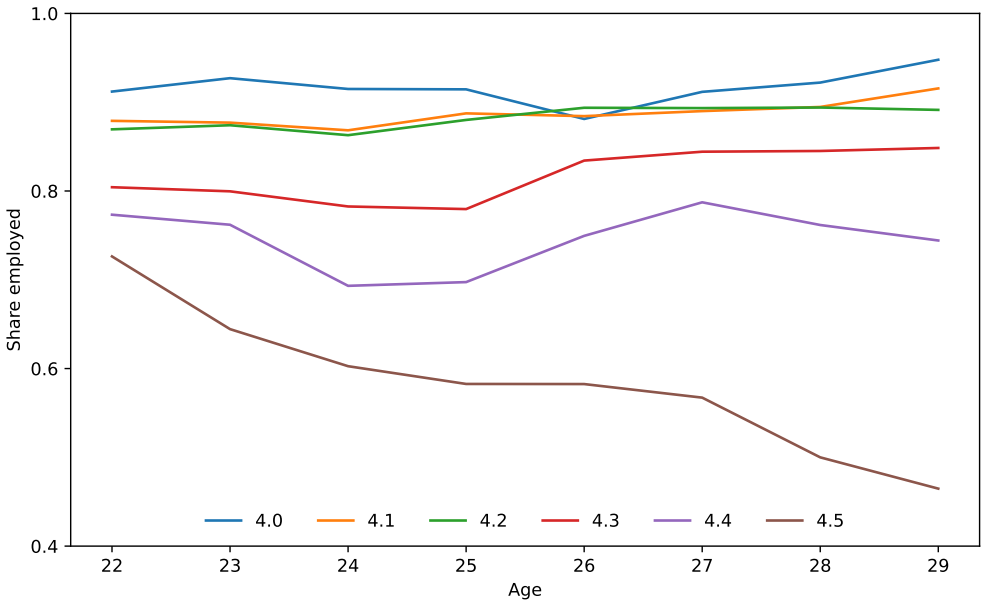


Figure 6: Share working by age and cluster, men, ages 22 to 29

Appendix F Additional Tables and Figures

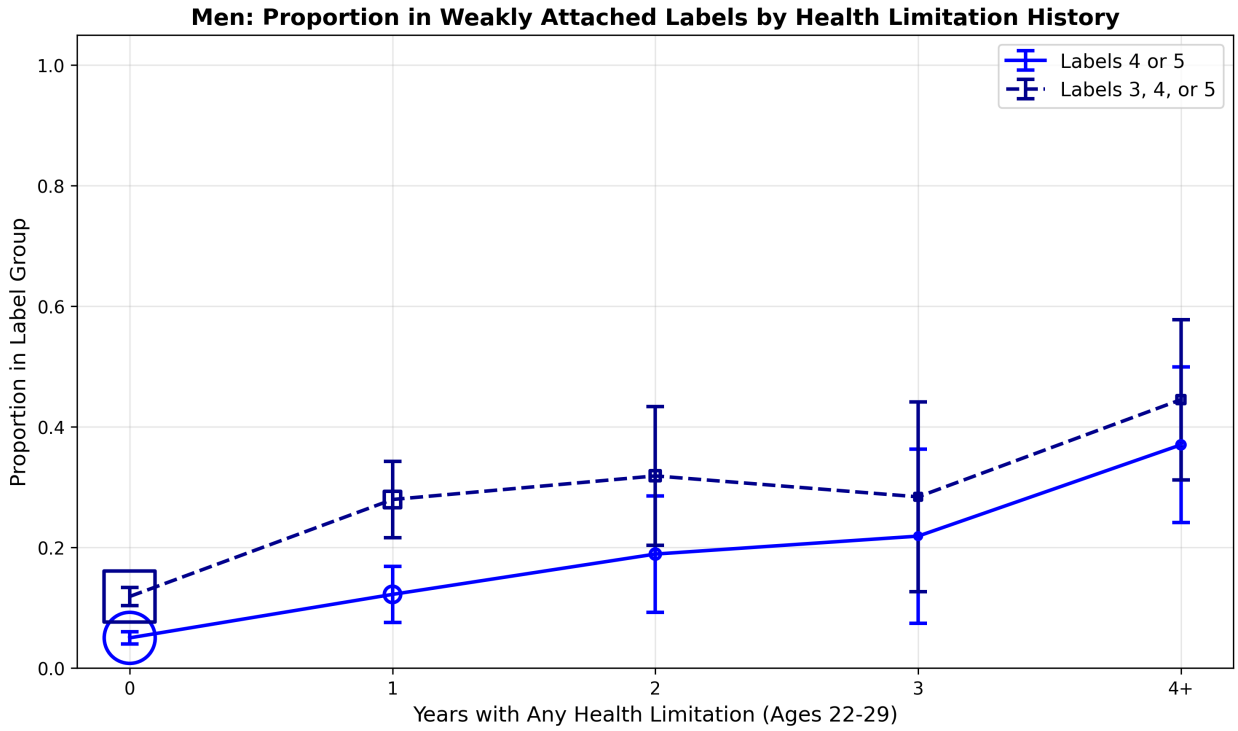


Figure 7: Weak Attachment by Years with a Health Limitation: Male

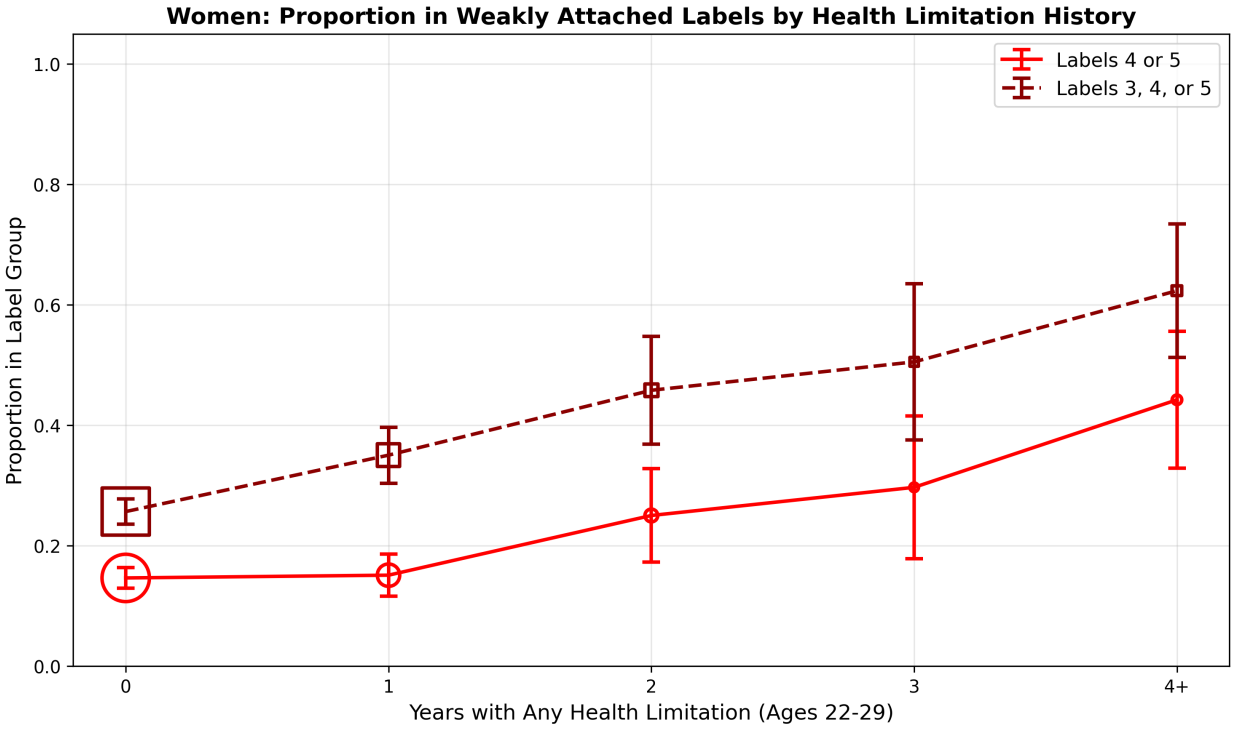


Figure 8: Weak Attachment by Years with a Health Limitation: Female