Wage & Employment Dynamics
THE WED PROJECT

WORKING PAPER

Can we identify students in ASHE?

Abstract

ASHE is a key dataset in the UK, the only one which allows long-term analysis of flows in labour market status and earnings, and hence vitally important in the understanding of low pay and wage progression. Separating out students from non-student workers therefore has considerable value. This study has tried to create a proxy for 'student working' using the ASHE dataset, and then triangulating with the Census 2011 data which has some of the same people but with an accurate marker for student status. Unfortunately, triangulating this with accurate student information on the Census suggested that our preferred method was not notably the 'best'.

Van Phan, Felix Ritchie, Damian Whittard, Lucy Stokes, John Forth and Alex Bryson

Acknowledgements: This work was produced using statistical data from the Office for National Statistics (ONS). The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates. We thank Administrative Data Research UK and the Economic and Social Research Council for funding support under Grant No. ES/T013877/1.

Contents

1. Introduction

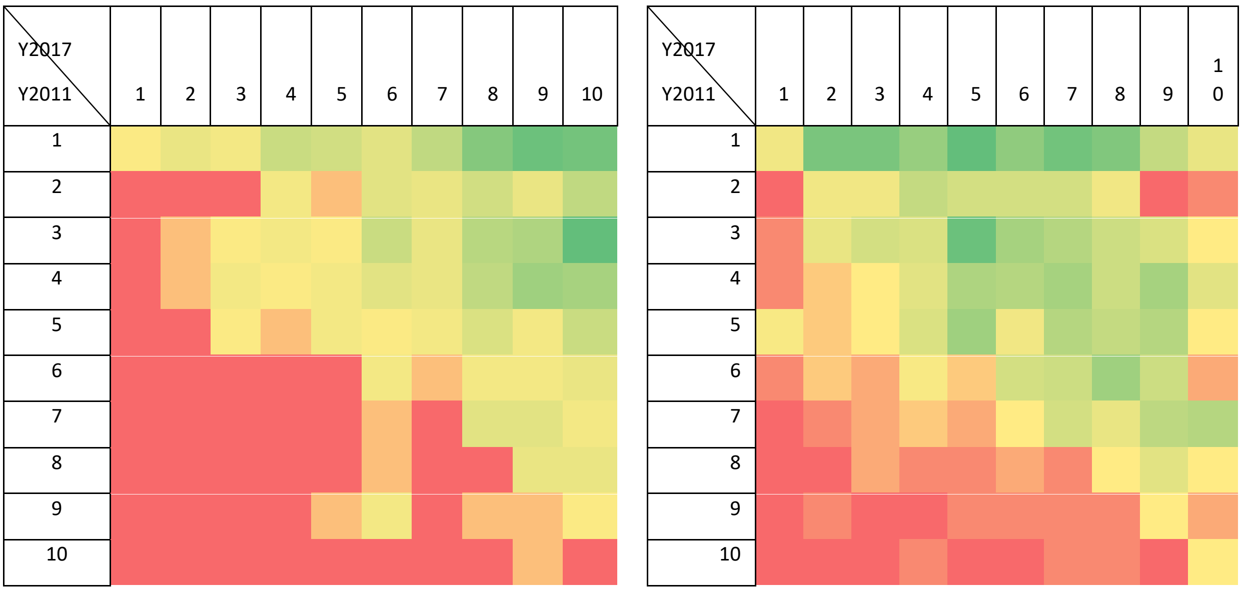
Low pay work in the UK is concentrated in a relatively small number of industries: agriculture, food processing, textiles, retail, hospitality, cleaning, social and child care, leisure, hairdressing and agency work (LPC, 2014). Most of these jobs require little or no qualifications or experience on entry, or, in the case of hairdressing, keep wages low during long periods of apprenticeship/training (Drew et al 2016). Whilst there are opportunities for advancement, often this is on a relatively small salary scale.

Retail and hospitality are characterised by a high degree of employment of students in higher education (HE). These jobs appeal to students because the working conditions complement, or can be adapted to complement, student lifestyle and timetables (Whittard et al, 2022). On the demand side, HE students are appealing to employers as they bring a range of social and technical skills (Whittard et al, 2022); they are also likely to be 18 or over, and so eligible for age-restricted roles such as bar work (Evans et al, 2022). Moreover, HE students undertaking part-time work while studying take a transactional approach to working which facilitates employers' flexible staffing plans (Evans et al 2021).

The identification of HE students in work is important for understanding the occurrence and dynamics of low pay, for two reasons.

First, when trying to explain the distribution of low pay, HE students and non-students of the same age doing the same work have very different characteristics. HE students are, by construction, better educated than the workforce on average; more likely to come from higher socio-economic classes; less likely to have childcare responsibilities; limited in the time available for work because of studies; and less attached to particular employers (Evans et al 2022). As a result, an examination of the drivers of wages in the low pay sector that does not identify students separately may find that the coefficients on many variables are biased (for example, finding that education is not a factor retail employment probability for young people).

Second, analyses of wage growth that do not take account of students are likely to mis-represent it. A low-wage student worker is likely to see a large jump in wages on reaching graduation (or may enter the job market for the first time on graduation, at an above-average salary). In contrast, a non-student is more likely to see slow, continuous wage growth, as well as more periods of unemployment.



(a) Working student (b) Non-student worker

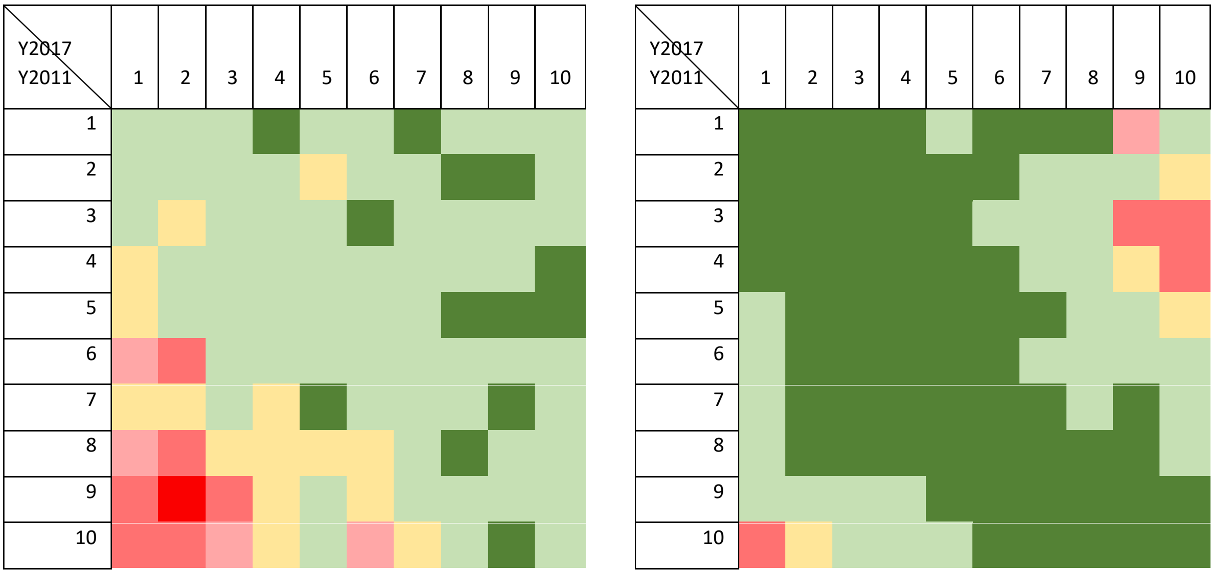
FIGURE 1 : WAGE DECILES OF WORKING STUDENTS AND NON-STUDENTS 2011 & 2017, AGED 20 IN 2011

Note: Shading indicates the scale of observations: darker green for many observations, amber for middling observations, then darker shades of red for few observations

Figure 1 shows wage growth of working students and non-students from 2011 to 2017, aged 20 in 2011 (to limit effects of late starting/finishing HE). The left-hand axis shows the distribution by income decile in 2011; the bottom axis shows the decile of the same person in 2017.

This shows that both groups move up the income distribution as they age (observations above the 45-degree line). However, students are more likely to start in a lower decile, and end up in a higher one.

Figure 2 show the same data but where the decile is calculated from that age group only (ie 18-year olds in 2011; 24-year olds in 2017).



(a) Working student (b) Non-student worker

FIGURE 2 WAGE DECILES OF WORKING STUDENTS AND NON-STUDENTS 2011 & 2017; DECILES FOR AGE GROUP 18-21

Note: Shading indicates the scale of observations: darker green for many observations, amber for middling observations, then darker shades of red for few observations

Non-student workers typically follow a 45-degree slopes; that is, they are likely to be in the same part of the income decile (relative to their age group) in their late teens and late twenties. In contrast, students start off as the lower-paid compared to their non-student peers, but have much higher wage growth post graduation.

These numbers are based on observed employees and so do not take account of sampling issues in ASHE (Forth et al., 2022a) or late graduation/post-graduate study. Nevertheless, they clearly indicate the invalidity of assuming that all young low-paid workers are the same.

This has important implications for low-wage analysis carried out using the Annual Survey of Hours and Earnings, ASHE. ASHE is the main source of low pay statistics, as its data on wages, hours, industry and occupation is thought to be of a very high standard. However, ASHE only records paid employment, and it has no information on whether an employee is a student or not.

There is therefore considerable value in being able to identify students in ASHE. This is not directly possible in ASHE alone. However, it may be possible to *infer* whether an individual is *likely* to be a student, based on their employment pattern. If the marker were reliable, it could then be used a triangulation check in analysis (‘do these results change if we allow for likely students?’).

Assessing the reliability of the marker using only ASHE data is problematic: information that is used to predict studentship cannot then be used to validate that prediction. Nevertheless, there are triangulation options that give some confidence in their accuracy.

Combining ASHE with other datasets provides more options. At present, ASHE has been linked to the 2011 Census (Forth et al., 2022b), which identifies students directly. It is then possible to compare predictions for 2011 using ASHE only with the recorded status in the 2011 Census. A second possibility is to generate a probability model for being a working student using the Census 2011, using only variables available in ASHE, and then applying that model to ASHE to see how closely the two predictors correlate. The Census prediction can also be examined in its own right as a predictor.

There are limitations to this. The Census-ASHE link is only made for those who are working in 2010-12 (and so in the ASHE files for those years; all three years are included in the match process). Hence, the linked dataset will not include students who were not working in those periods. Moreover, the match rate is approximately 65%, and likely to be lower amongst young people where addresses are more likely to change. Nevertheless, using Census for validation is likely to be the most reliable method available on this data[1].

This paper describes how a student prediction was generated in ASHE, and the validation process using ASHE alone and ASHE-Census. The next section describes how 'student' employment cycles were modelled to provide a basis for the prediction. Section 3 details the specific measures used for seven different predictions of 'studentship'. Section 4 presents results, the preferred model, and the within-ASHE validation.

Section 5 uses the linked ASHE-Census dataset to (a) examine the accuracy of the AHSE predictors (b) generate Census-based predictors, and (c) apply those predictors to ASHE and review the outcome. Section 6 concludes with recommendations for researchers.

Overall, we conclude that using the ASHE data alone does suggest that a useful, if not very accurate, predictor can be generated. However, when we compare the ASHE predictor with the true student status in the Census, the ASHE predictor is not only low-quality, but when used in statistical analysis may give contrasting results. Whilst disappointing, this does emphasise the need for an accurate student marker, something which is only likely to appear when ASHE records are matched to actual student records.

2. Student employment cycles

In this section, we propose a way to identify students by looking at their jobs over their life time.

In principle, most HE students are involved in their studies between the wages of 18-21, and they will graduate by the age of 21. In reality, there may be some students get into HE a couple of years later. Therefore, we allow 2 years in variation. This means that the period after 23 years old is known as the 'graduation' period. Meanwhile, the period before 21 years old is definitely the 'student' period while the period between 22 and 23 years old is considered as the 'transition' period.

Students tend to work in marginal, flexible, part-time and low-paid jobs alongside their studies at HE, and after their graduation, they are presumably prone to full time, and higher paid jobs. ONS has a classification system for identifying 'graduate roles', occupations which typically require a degree. We classify jobs into 2 types:

[1] Another option is HMRC data linked to ASHE. This may allow more accurate identification of students through student loan records (although this will miss the wealthier students). This option will be explored when the data becomes available in 2022.

a) 'student' jobs (which could be part-time, low-paid, or non-graduate roles, or a combination of these)
b) 'graduate' jobs (full-time, non-low-paid, graduate roles, or some combination)

We map these types of jobs to the student/transition/graduation periods to try to identify what a 'student' would look like; see Table 1, which gives a non-exhaustive summary of labour market experiences.

TABLE 1 ILLUSTRATIVE JOB PATHS FOR STUDENTS (WORKING AND ON-WORKING) AND NON-STUDENTS

Period and activity			Classification
'student' <21	'transition' 21-22	'graduation' >22	
Student job	Student job	Student job	Non-student
Student job	Student job	Graduate job	Working student
Student job	Graduate job	Graduate job	Working student
Graduate job	Graduate job	Graduate job	Non-student
No job	No job	No job	Non-student
No job	No job	Graduate job	Non-working student
No job	Graduate job	Graduate job	Non-working student
No job	Student job	Graduate job	Working student
Student job	No job	Graduate job	Working student

For example, one who has all student jobs before or at the age of 20, and has any graduate jobs at or after the age of 23 is classified as working student. It could also be the case where one who is not working during the student period (i.e. not observed in ASHE) have any graduate jobs at or after the age of 23. This is classified as non-working student.

For simplicity, we just consider these two common situations, and do not consider the mature students who may be come back to school after years in industry (see Figure 3).

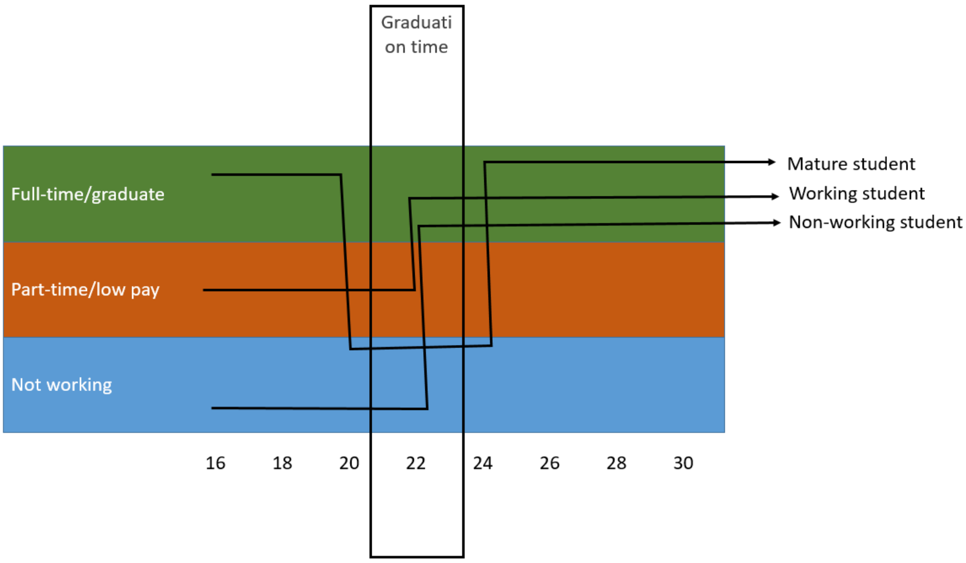


FIGURE 3 STUDENT EMPLOYMENT CYCLES

3. Studentship measurements

What defines a 'student' or 'graduate' job? There are several potential criteria: time, occupation, pay scale, or some combination of these, which have been used to map into the student employment cycles. In this section, we discuss them as follows.

TABLE 2 POTENTIAL DEFINITIONS OF 'STUDENT' AND 'GRADUATE' JOBS

Label	Definition	Criterion
1. Time		
'student' jobs	part-time jobs	Defined by the ASHE variable *fulltime*
'graduate' jobs	jobs-time jobs	
2. Occupation 1		
'student' jobs	low-paid jobs	Defined by the variable *is_LP_occ*, (a Low Pay Commission LPC definition), available in the enriched ASHE-WED dataset
'graduate' jobs	non-low paid jobs	
3. Occupation 2		
'student' jobs	non-graduate jobs	Defined by the variable *graduate_occ*, available in the enriched ASHE-WED dataset. This is based on ONS (2020).
'graduate' jobs	graduate jobs	
4. Pay		
'student' jobs	'low-paid' jobs	The dividing line is 2/3 of median earnings, derived from the calculated hourly pay rate in ASHE
'graduate' jobs	'high-paid' jobs	

In our analysis, we tested the following variables and combinations:

TABLE 3 VARIABLES AND VARIABLE COMBINATIONS USED

Variable	**Definition (from above)**
studenttime	Time
studentocc1	Occupation1
studentocc2	Occupation2
studentpay	Pay
studenttimeocc1	Time+Occupation1 (part-time AND low pay work, or not)
studenttimeocc2	Time+Occupation2 (full-time AND graduate work, or not)
studenttimepay	Time+Pay (full-time and high paid, or not)

4. Analysis and results

In this section, we first show some descriptive analysis of student jobs and graduate jobs (mentioned in Section 2). In Table 4, Panel A shows the proportion of student jobs and graduate jobs in the whole sample based on different measurements while Panel B shows these figures for the restricted sample (i.e. for main jobs only for those aged less than 30 years old). The percentage of graduate jobs varies to studentship measurements. In details, the proportion of graduate jobs defined by time, occupation 1 and pay are much higher than those defined by occupation 2 (70%, 59%, 79%, respectively, vs. 31% for panel A). Even when we use the combination definition, the combination between time+occcupation1 and time + pay criterion gives us higher figures than the combination between time and second classification (e.g. 47%, 60% vs. 26%).

TABLE 4: SUMMARY DESCRIPTIVE OF STUDENT JOBS AND GRADUATE JOBS

Panel A	**All sample**			
Criterion	**Student jobs**	**Graduate jobs**	**Neither student nor grad job**	**Total**
Time	739,460	1,763,522		2,502,982
	29.54%	70.46%		100.00%
Occupation 1	1,017,110	1,485,872		2,502,982
	40.64%	59.36%		100.00%
Occupation 2	1,736,213	766,769		2,502,982
	69.37%	30.63%		100.00%
Pay	525,034	1,977,948		2,502,982
	20.98%	79.02%		100.00%
TimeOcc1	436,283	1,181,101	885,598	2,502,982
	17.43%	47.19%	35.38%	100.00%
TimeOcc2	618,306	644,717	1,239,959	2,502,982
	24.70%	25.76%	49.54%	100.00%
TimePay	273,392	1,510,216	719,374	2,502,982
	10.92%	60.34%	28.74%	100.00%
Panel B	**Restricted sample (main job & age<30)**			
Criterion	**Student jobs**	**Graduate jobs**	**Neither student nor grad job**	**Total**
Time	199,239	403,038		602,277
	33.08%	66.92%		100.00%
Occupation 1	327,863	274,414		602,277
	54.44%	45.56%		100.00%
Occupation 2	479,310	122,967		602,277
	79.58%	20.42%		100.00%
Pay	148,101	454,176		602,277
	24.59%	75.41%		100.00%
TimeOcc1	156,767	231,597	213,913	602,277
	26.03%	38.45%	35.52%	100.00%
TimeOcc2	187,118	110,697	304,462	602,277
	31.07%	18.38%	50.55%	100.00%
TimePay	63,836	318,321	220,120	602,277
	10.60%	52.85%	36.55%	100.00%

Then we map student/graduate jobs into the student employment cycles to identify our 'potential' student markers. Table 5 displays the incident of student markers, in which column 1 is applied to main jobs for those aged less than 30 years old. The proportion of 'potential' students accounts for around 25% (lowest figure in column 1) or even up to 58% (highest figure in column 1). In addition, we divide into 2 sub sample: (i) age 18-24, and (ii) age 25-29, for further analysis. The lowest incident of student markers accounts for 18% in the younger age group (see column 2). This is similar to the ONS estimates. In details, there are, on average, around 3,500 thousand of people aged 18-24 in employment[2], 17 percent of which are in full-time education[3], during 2004-2019. Therefore, the combination of time and occupation 2 is our preferred measurement to identify 'potential' students.

[2] See https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/timeseries/ybtr/lms
[3] See https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/empoymentandemployeetypes/timeseries/agns/lms

TABLE 5: INCIDENT OF 'POTENTIAL' STUDENT MARKERS

	Proportion of student markers		
Criterion	**Sample: main job, age<30**	**Sample: main job, age 18-24**	**Sample: main job, age 25-29**
	(1)	**(2)**	**(3)**
Time	58.39	44.14	76.46
Occupation 1	51.06	41.62	63.38
Occupation 2	31.5	25.26	39.36
Pay	40.64	23.45	61.97
TimeOcc1	39.88	29.29	53.08
TimeOcc2	25.17	18.19	33.68
TimePay	36.06	20.38	55.41

Furthermore, we analyse the wage growth of working students and non-student workers. Figure 4 shows the average gross weekly earnings by age between non-students (on the left) and students (on the right) (based on 'TimeOcc2' combination classification). This shows that non-student workers are more likely to see slow, and continuous wage growth whereas working students experience a large jump in wages on reaching graduation. The similar patterns occur in the other student definition (illustrated in the appendix).

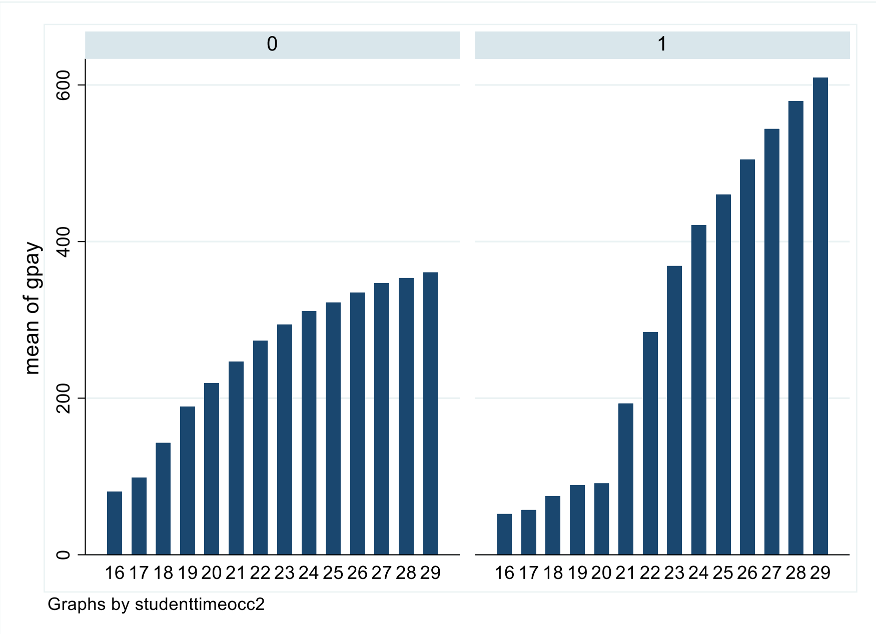


FIGURE 4: AVERAGE GROSS WEEKLY EARNINGS BY AGE BETWEEN NON-STUDENT WORKERS AND WORKING STUDENTS

5. Validation

To validate our predicted marker, we use the linked ASHE-Census dataset because there is a student marker available in the 2011 Census, indicating whether the one is a full-time student.

First of all, we simply do the cross tabulation between our predicted marker (in ASHE2011) and the actual marker in the 2011 Census for the sample of individuals aged under 30 years old, which is shown in Panel A-Table 6. In general, we find that the validation of our predicted marker seems not very strong. For example, among all 'potential' students from our markers regardless of the criterion, a minority of them are actually students. In particular, we see the biggest proportion (approximately 17%) of the consistency with the time+occupation2 definition. However, our predicted marker does well in terms of eliminating the large proportion of actual non-students.

TABLE 6: CROSS TABULATION BETWEEN OUR PREDICTED MARKER (ASHE 2011) AND THE ACTUAL MARKER IN THE 2011 CENSUS

		Panel A (sample age<30)			Panel B (sample age<30)		
Our predicted marker definition		Student marker in 2011 Census			Graduate (HE education) in 2011 Census		
		1 Yes	2 No	Total	0 No	1 Yes	Total
Time	0 No	1,457	7,316	8,773	7,723	1,050	8,773
		16.61%	83.39%	100.00%	88.03%	11.97%	100.00%
	1 Yes	2,157	11,871	14,028	8,208	5,820	14,028
		15.38%	84.62%	100.00%	58.51%	41.49%	100.00%
	Total	3,614	19,187	22,801	15,931	6,870	22,801
Occupation 1	0	1,741	7,490	9,231	8,101	1,130	9,231
		18.86%	81.14%	100.00%	87.76%	12.24%	100.00%
	1	1,873	11,697	13,570	7,830	5,740	13,570
		13.80%	86.20%	100.00%	57.70%	42.30%	100.00%
	Total	3,614	19,187	22,801	15,931	6,870	22,801
Occupation 2	0	2,262	12,095	14,357	11,981	2,376	14,357
		15.76%	84.24%	100.00%	83.45%	16.55%	100.00%
	1	1,352	7,092	8,444	3,950	4,494	8,444
		16.01%	83.99%	100.00%	46.78%	53.22%	100.00%
	Total	3,614	19,187	22,801	15,931	6,870	22,801
Pay	0	3,279	10,662	13,941	11,449	2,492	13,941
		23.52%	76.48%	100.00%	82.12%	17.88%	100.00%
	1	335	8,525	8,860	4,482	4,378	8,860
		3.78%	96.22%	100.00%	50.59%	49.41%	100.00%
	Total	3,614	19,187	22,801	15,931	6,870	22,801
TimeOcc1	0	2,131	10,405	12,536	10,784	1,752	12,536
		17.00%	83.00%	100.00%	86.02%	13.98%	100.00%
	1	1,483	8,782	10,265	5,147	5,118	10,265
		14.45%	85.55%	100.00%	50.14%	49.86%	100.00%
	Total	3,614	19,187	22,801	15,931	6,870	22,801
TimeOcc2	0	2,533	13,844	16,377	13,447	2,930	16,377
		15.47%	84.53%	100.00%	82.11%	17.89%	100.00%
	1	1,081	5,343	6,424	2,484	3,940	6,424
		16.83%	83.17%	100.00%	38.67%	61.33%	100.00%
	Total	3,614	19,187	22,801	15,931	6,870	22,801
Time + Pay	0	3,347	11,459	14,806	12,074	2,732	14,806
		22.61%	77.39%	100.00%	81.55%	18.45%	100.00%
	1	267	7,728	7,995	3,857	4,138	7,995
		3.34%	96.66%	100.00%	48.24%	51.76%	100.00%
	Total	3,614	19,187	22,801	15,931	6,870	22,801

In short, it seems feasible to identify who is not a student, but harder to spot who is.

Alternatively, we use education qualification as an indicator of being a student. Panel B – Table 6 illustrates the cross tabulation between our predicted markers and the alternative. In general, by using the education as alternative, we find that the validation of our predicted markers is getting stronger. The best marker is still the time+occupation2 definition. Among 6,424 'potential' students, 61 percent are HE graduates.

We further apply these markers in the wage analysis. We find that our student marker coefficients are positive and statistically significant for the 21-29 age group, but not for the 18-20 age group. This is consistent with our analysis above. Unlike our student markers, the student marker in Census which allows to identify exactly whether he/she is a student at a specific time is negative and statistically significant for the 21-29 age group, but not statistically significant for the younger age group[4].

6. Conclusion

ASHE is a key dataset in the UK, the only one which allows long-term analysis of flows in labour market status and earnings, and hence vitally important in the understanding of low pay and wage progression. Separating out students from non-student workers therefore has considerable value.

This study has tried to create a proxy for 'student working' using the ASHE dataset, and then triangulating with the Census 2011 data which has some of the same people but with an accurate marker for student status. The ASHE-only work suggested that a reasonable indicator for student status could be derived: not precise enough to use as a meaningful variable, but perhaps of sufficient quality to allow researchers to carry out some basic sensitivity tests by separating out the predicted 'students'. Unfortunately, triangulating this with accurate student information on the Census suggested that our preferred method was not notably the 'best'; more importantly, the different measures had radically different impact on regression models, and our 'preferred' proxy may in fact increase bias in the models rather than reducing it.

While this is disappointing, it does illustrate the difficulties of trying to both create and then validate proxy variables using the same data source, even if the validation process is notionally independent of the factors used to derive the proxy. What seemed to be clear and useful from ASHE only turned out, when compared with the additional dataset, to enhance rather than reduce biases.

The problem of identifying students in ASHE remains an unsolved issue. Whilst the ASHE-Census2011 link does allow for some better analysis, this is only point-in-time. However there are ways to address this.

First the Census includes information on highest level of qualification. It may be that this can be used to identify patterns of earnings of past students, to develop a better model which could be re-applied to ASHE. The difficulty with this is assessment: as noted above, modelling and validating the model on the dataset can have unwanted consequences.

Second, the WED team have gained access to HM Revenue and Customs data in 2022, which may include student loans repayments. These are an unambiguous indicator of studentship. These will, of course, only be relevant or students from lower-income families needing student loan support, and so will miss students from higher-income families. On the other hand, as Whittard et al (2022) point out, the need to minimise financial hardship appears a clear driver of student working, and so it is not unreasonable to assume that those with student loans approximate the student workforce.

[4] The details of the wage regressions are not shown here.

Finally, linking ASHE data with information from the Higher Education Statistics Agency (HESA) would directly answer the student status question, as well as providing extremely valuable information on the specific degree subject and place of study. The WED team is currently investigating this possibility with HESA, who have already made some microdata available to researchers. It therefore seems likely that the student issue will be solved, perhaps within the next few years.

REFERENCE

Drew, H., Ritchie, F., & Michail, V. (2016). Understanding apprentice pay: Final report to the Low Pay Commission. https://uwe-repository.worktribe.com/output/908410/understanding-apprentice-pay-final-report

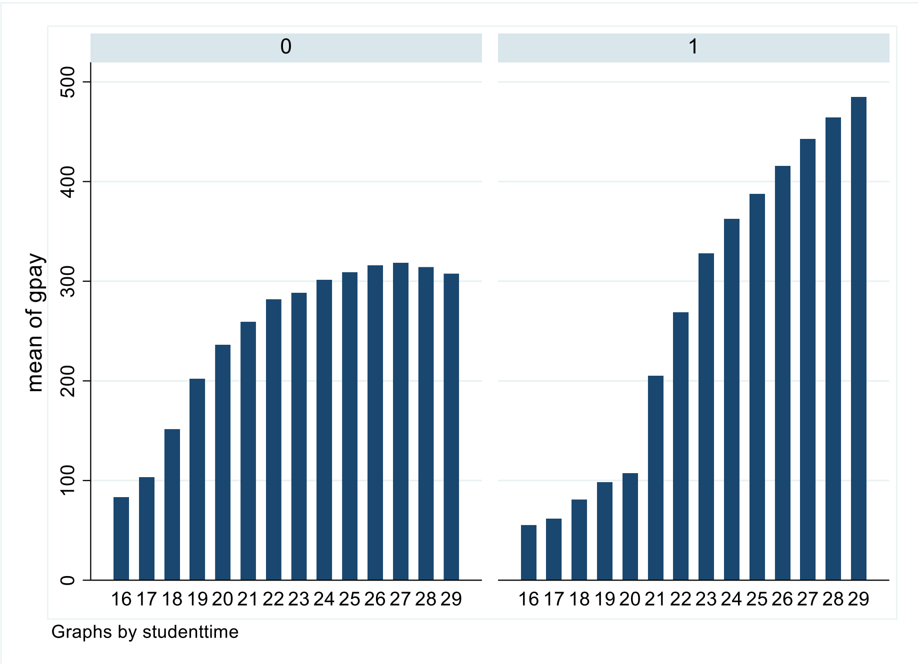
Evans, C., Ritchie, C., Drew, H., & Ritchie, F. (2022). A marriage of convenience: How employers and students working in hospitality view the employment relationship. *Hospitality and Society*, https://doi.org/10.1386/hosp_00055_1. Available from https://uwe-repository.worktribe.com/output/9853358

Forth, J., Phan, V., Ritchie, F., Whittard, D., Stokes, L., Bryson, A., & Singleton, C. (2022). Longitudinal attrition in ASHE. https://www.wagedynamics.com/wp-content/uploads/2022/10/Longitudinal-attrition-in-ASHE-Version-1b.pdf

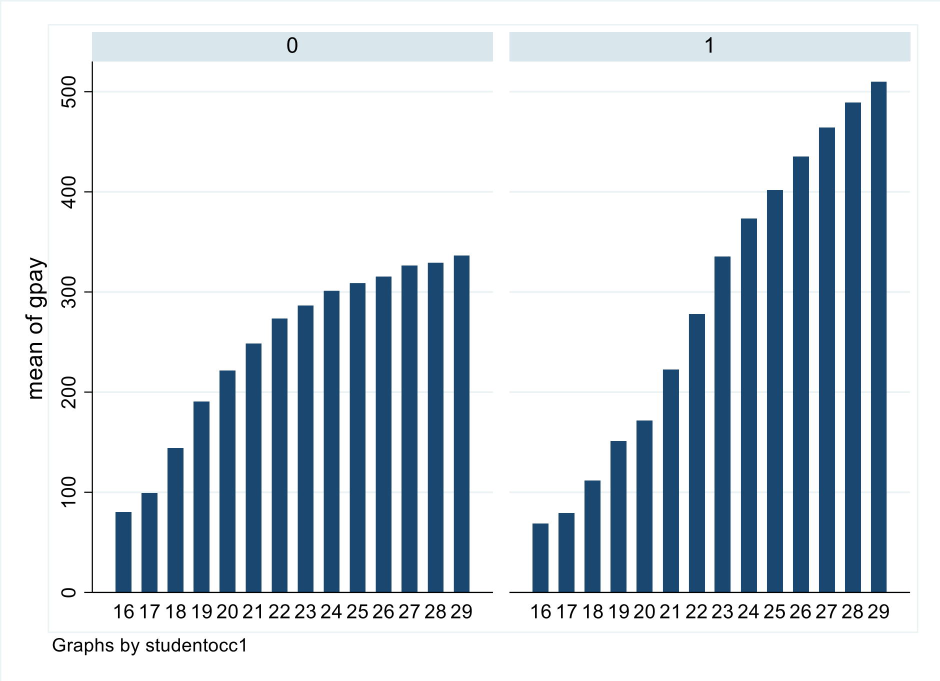
Forth, J., Phan, V., Ritchie, F., Whittard, D., Stokes, L., Bryson, A., & Singleton, C. (2022). User guide Drop 1 of the ASHE – 2011 Census Dataset. http://www.wagedynamics.com/wp-content/uploads/2022/09/ASHE-CEW11-User-Guide-Version-1.2-Drop-1-1.pdf

Whittard, D., Drew, H., & Ritchie, F. (2022). Not just arms and legs: Employer perspectives on student workers. *Journal of Education and Work*, https://doi.org/10.1080/13639080.2022.2126972. Available from https://uwe-repository.worktribe.com/output/9889242

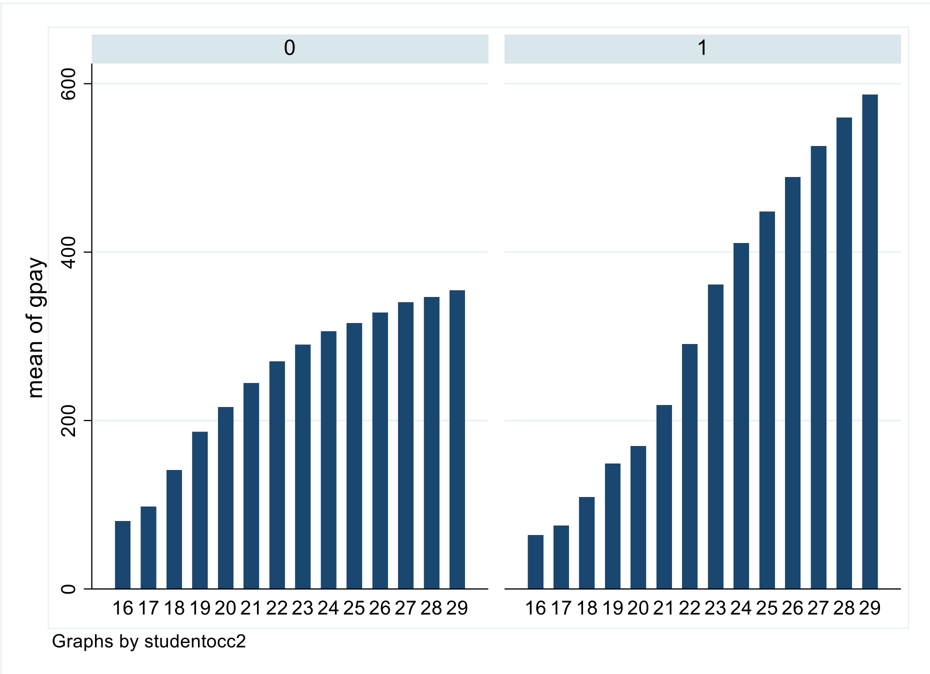
APPENDIX



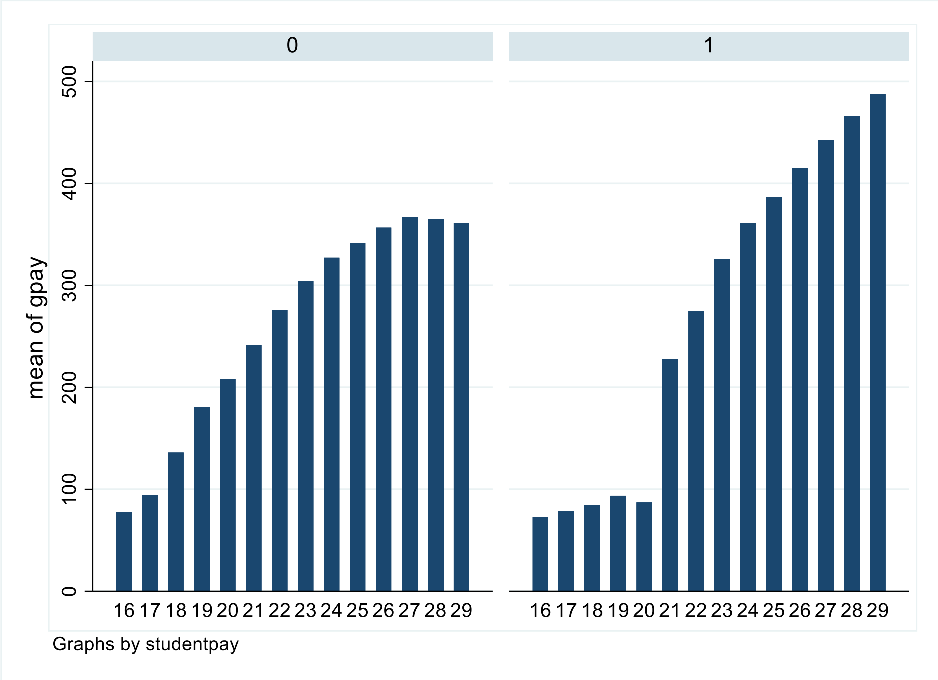
(a) studenttime



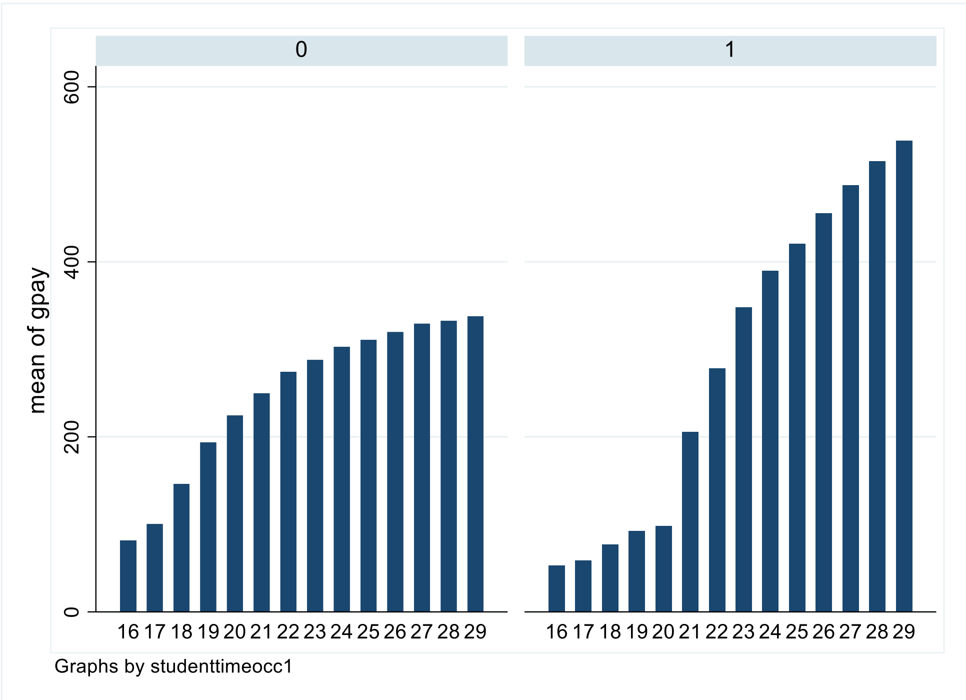
(b) studentocc1



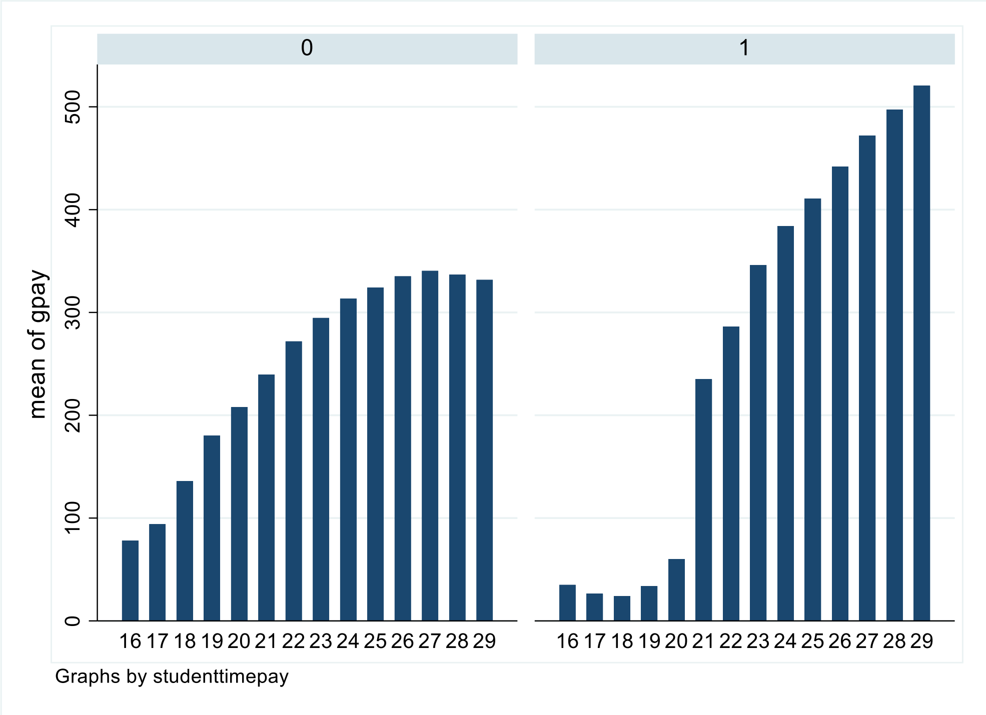
(c) studentocc2



(d) studentpay



(e) studenttimeocc1



(f) studenttimepay

FIGURE 5: AVERAGE GROSS WEEKLY EARNINGS BY AGE BETWEEN NON-STUDENT WORKERS AND WORKING STUDENTS