APPENDIX A: Data Sources

NSLDS

The main data source in this paper is the National Student Loan Data System (NSLDS) which is the primary system used to administer student loan programs. The Department of Education is required to administer the NSLDS by the Higher Education Act of 1965. The NSLDS links information on loans and grants from students, borrowers, lenders, guaranty agencies, schools, and servicers. As of January 2014, the NSLDS contained over 30 billion records on 84,629,538 students and 386,943,660 loans.[1] The data used in this analysis was assembled pursuant to an agreement between the US Treasury and the Department of Education to improve tax administration, to improve education-related tax and fiscal policy, and enhance forecasts and projections of tax and educational policy.

Lenders, guaranty agencies, schools, and servicers are required to report information to the NSLDS within 30 and 120 days of new information arriving. For example, defaults must be reported within 90 days of a loan entering default and changes in enrollment must be reported within 30 days of a change in enrollment status. Updates can be done either electronically or by mail, but today are usually done online. Information in the NSLDS is used to determine eligibility for Title IV programs under rules such as Gainful Employment report standards.

The analysis in this paper is based on a 4 percent random sample of student loan borrowers matched to individual earnings records. The raw loan and demographic data is used to construct a person-by-year individual panel providing information on student characteristics, institutional information, and federal loan information as of the close of each federal fiscal year.

The educational records are sampled as of the end of fiscal years from 1970 to 2013 from transactions records from FSA's operational database. The sample is intended to reflect loan balances and status as of the close of the fiscal year and to reflect characteristics of borrowers and institutions as reported on aid applications and by institutions corresponding to the years in which loans were disbursed. The sample was initially created by the Department of Education's Budget Service Division to be used in budget projections. The sample includes federal direct and federally guaranteed students loans, including both the Federal Family Education Loan Programs and the Direct Loan Program. The sample does not include Perkins loans. Parent PLUS loans are included in the analysis but the outcomes of parent borrowers are excluded from our analysis of the outcomes of borrowers after entering repayment. Private student loans are not included in the analysis sample if they were not made under the FFEL program. The NSLDS

[1] For more information on the NSLDS see The Department of Education (2014).

contains the vast majority of direct student loans, as private student loans not guaranteed by the federal government only account for a tenth of the market, and much less prior to 2005.[2]

Most basic loan information is available from all sample borrowers since 1970 including loan balances, type of loan, the institution the loan was disbursed to (including the type and control of the institution, and the academic level of the borrower), and information regarding whether the student completed the program, and the dates borrowing was initiated, when repayment on each loan began, and whether the loan was in deferment, forbearance, default, or in certain alternative repayment plans. In general, borrowers join the sample in the year when they receive their first loan. Borrower information from the FAFSA is generally only available for loans originated after the 1995 fiscal year. Estimates from these data match up closely to aggregate statistics published by the department of Education, regarding the total volume of existing loans over time, the number of borrowers, and estimated cohort default rates. Repayment is defined as the date at which a borrowers' last loan enters repayment, as some borrowers take out multiple loans for different programs.

The main NSLDS file has been matched to information from the Free Application for Federal Student Aid (FAFSA), in the years those data were available for the first completed FAFSA for each borrower (except for borrowers in 1995, this is generally when their first loan is disbursed). The FAFSA contains detailed information on student demographics and family background. Students are required to fill out the FAFSA in each year that they receive aid or loans. The NSLDS has also been matched to Pell Grant records which contain additional grant applications and receipt. The main NSLDS sample is merged to a panel of administrative tax and earnings records that span the period from 1999-2013.

Administrative Tax and Earnings Records

The main source of earnings records is tax records spanning the period from 1999 to 2013. These records contain data from federal income tax records between 1996-2013, compiled from individual returns and W-2 information returns. Income is measured at the tax unit level; earnings at the individual level.

Loan Volumes: Loan volumes refer to the last loan balance recorded in the NSLDS in each calendar year. Outstanding balances are reported to the NSLDS and updated within 120 days of loans being disbursed. Dollar values are in 2014 dollars unless noted otherwise.

Default: Default rates are defined as an indicator of whether an individual enters into default with a certain number of years since repayment begins. The last recorded repayment and default rates are used for an individual. A loan goes into default if payments are more than 270 days late. Servicers have 90

[2] For more information on private student loans, see Bricker, Brown, Hannon and Pence (2015),

days to report default to the NSLDS after a loan goes into default.

Dependency Status: This variable is constructed from the FAFSA. A students' first recorded dependency status is recorded, from the first FAFSA filed.

Family Income: Family income is obtained from the FAFSA. A students' first recorded family income is used, from the first FAFSA filed.

Active Borrower: Active borrowers are borrowers who have received loan disbursements in the fiscal year determined by the NSLDS.

Poverty: Federal poverty guidelines are used to determine distance for poverty line thresholds as defined by HHS.[3] Earnings and family size are determined by tax returns using the CDW.

Pell Grants: Pell grants awarded are determined from Pell Grant records.

Earnings: Earnings are defined as Medicare wages plus self-employment earnings.

Enrollment: Enrollment is determined by the NSLDS. Schools are required to report enrollment through the Student Status Confirmation Report within 30 days of a change in enrollment status according to Federal Regulation 34 CFR 682.610.[4]

School Types: Are identified using the ownership control type of the first institution at which a student borrowed. School types are defined as public, private, or for-profit and two or four year and by the institution to which the loan was originated.

Entry/Repayment Year: Entry years are assigned based on the fiscal year during which the borrower's first loans were originated. Repayment year is defined by the fiscal year during which the borrower's last loan enters into repayment, and all loans are in repayment.

School Selectivity: Source: Barron's Profiles of American Colleges. 2008. Among 4-year public and private institutions we compress the Barron's Profiles categories into three groups: non-selective (corresponding to Barron's "Non-competitive and Less Competitive"); somewhat selective ("Competitive") and selective ("Very Competitive, Highly Competitive, and Most Competitive).

[3] See HHS Poverty Guidelines for details.
[4] For more on enrollment compliance, see the Enrollment Reporting Guide.

APPENDIX B: Online Data Appendix

The tabulations of NSLDS data underlying most of the charts and tables or otherwise described in the text and the program files that produce the charts and tables are available as a data appendix. These databases summarize the information on borrowers included in the merged ED/Treasury database by institutional type using three temporal concepts: the time of entry (characteristics of new borrowers in the year the borrower first entered the loan system), by fiscal year (characteristics of all borrowers with outstanding loan balances), and by repayment cohort (characteristics of borrowers in the year they entered repayment plus loan and economic outcomes subsequent to entering repayment).

Most figures in the text are derived directly from these databases, and readers may use the same data (and programs) to recreate those figures and to construct alternatives (e.g. means instead of medians; comparisons between alternative or multiple years etc.) These databases also provide a broader range of demographic, institutional, economic, and loan-related variables than directly described in the text. In particular, these data provide detail on:

New Borrowing and Entry into Repayment

The appendix data provide information on the number of borrowers, the amount borrowed, the number entering into borrowing for the first time, and when they entered into repayment. For instance, the data show that the number of borrowers entering repayment increased sharply from 1.4 million in 2007 to 2.4 million in 2011 and 2.9 million in 2013. As a result, the volume of debt entering repayment in 2013 ($89 billion) was almost three times the amount in 2007 ($33 billion).

Educational Outcomes of Borrowers

The NSLDS data include institution-reported indicators of whether a borrower has completed (graduated) from their program of study or withdrawn (i.e. without a degree). While most institutions appear to report completion and withdrawal accurately, the sole purpose of the reporting is to indicate to FSA that a student is no longer enrolled and there is no consequence to differentiating between the two options. Some institutions appear to use them interchangeably or to report only withdrawals. To supplement this measure, we also estimate the fraction of new borrowers who enter repayment within one year of starting borrowing. These borrowers are unlikely to have completed their programs, even at 2-year institutions (Bound, Lovenheim and Turner (2010)). These data show that more than half of borrowers at 2-year for-profit institutions have attended for a year or less, as had about 35 percent of borrowers at 2-year public and 4-year for-profit institutions. In contrast, less than 15 percent of borrowers at 4-year public and private institutions dropped out after one year. Because borrowers may re-enter school, some spells of enrollment are censored in the last several years, increasing the share of short-term enrollments. As a result of the shifting enrollment patterns, many more borrowers have

attended institution types where dropping out is the norm, hence, the fraction of borrowers in recent cohorts who have not completed their programs has increased.

Loan Amounts

The data appendix provides the estimates of the average and median loan balances of borrowers in their first year by entry cohort, in the year that the loan entered repayment (by repayment cohort), and for the overall stock of borrowers in each fiscal year. This also includes the 25th, 75th, and 95th percentiles. In addition, it provides the average and median amount owed by students for undergraduate and graduate loans.

Characteristics of Borrowers

The data appendix provides additional information on the characteristics of borrowers at entry, repayment, and in each fiscal year. These data are drawn from information provided as of the first FAFSA application provided by the student. Figure 4 provides further information on the age and income distribution of borrowers and provides further demographic information.

Labor market outcomes

The data tables by repayment cohort and for borrowers by fiscal year provide estimates of mean and median earnings and income (based on modified AGI defined as AGI plus adjustments) and the fraction not employed or in poverty (based on their income and filing status) for repayment cohorts two years after entering repayment and for the stock of borrowers in repayment.

APPENDIX C: Decomposition Method

To determine the effect of individual characteristics, we use variations on the Oaxaca-Blinder decomposition. The standard Blinder (1973); Oaxaca (1973) framework for decomposition analysis imposes a linear framework. Let Y^j be the default rate in year *j*, and let X^j be a vector of explanatory variables including a constant row of ones in year *j*, and moreover let $\hat{\beta}^{11}$ be the coefficient from the regression $Y^{11} = \hat{\beta}^{11} X^{11} + \varepsilon$.[5] In this case, the analysis is straightforward and the effect of a particular explanatory variable between 2011 and 2000 is measured using the change in the mean of an explanatory variable multiplied by the coefficient from a linear regression:$[\bar{X}^{11} - \bar{X}^{00}]\hat{\beta}^{11}$.

The decomposition is used to determine what default rates would have been in 2011 if the cohort had the same characteristics as individuals who were repaying loans in 2011. The estimated composition effects due to changes in observable characteristics solely reflect the effect of differences in the distribution of characteristics between borrowers in 2000 and 2011 under the following assumptions: (i) simple counterfactual treatment (ii) overlapping support and (iii) conditional independence.[6]

The linear framework is not ideal for decomposing the change in student loan defaults for two reasons. First, the gap in loan defaults lies in the tail of the distribution, where linear estimators tend to perform poorly. Second, there are large gaps between different years in a number of explanatory variables such as the total amount borrowed. This can lead to predicted probabilities below zero or above one in the linear framework. In a nonlinear framework, the change in the mean of an explanatory variable cannot be multiplied by the regression coefficient because for any non-linear function it is not necessarily true that for any $\mathrm{F}(.), \mathrm{E}[\mathrm{Y}] \neq \mathrm{F}(\mathrm{E}[\bar{X}\,\hat{\beta}])$.

Simulating the change in all observables can be done using a logit decomposition, assuming that the dependent variable is of the form $\mathrm{Y_F} = \mathrm{F}(\mathrm{X}\beta) = \frac{e^{\mathrm{X}\hat{\beta}}}{1+e^{\mathrm{X}\hat{\beta}}}$.[7] The procedure is implemented by first estimating the default logit regression $\mathrm{Y} = \mathrm{F}(\hat{\beta}\mathrm{X})$ using the pooled data and then using the predicted $\hat{\beta}$ to simulate the counterfactual predicted default rate using the 2011 explanatory variables. The decomposition can thus be written as

[5] Y^j is an $N^j \times 1$ vector, X^j is an $N^j \times K$ matrix of independent variables and $\hat{\beta}^j$ is a $K \times 1$ vector of coefficients. *K* denotes the number of variables N^j, the number of observations and *j* the year.

[7] The logit model has the desirable property that if a constant term is included, the predicted values of the logit model in year i will be $\bar{Y}^i$.

$$\bar{Y}_F^{11} - \bar{Y}_F^{00} = \left[\sum_{i=1}^{N^{11}} \frac{\mathrm{F}(\hat{\beta}^{11}\mathrm{X}_i{}^{11})}{N^{11}} - \sum_{i=1}^{N^{00}} \frac{\mathrm{F}(\hat{\beta}^{11}\mathrm{X}_i{}^{00})}{N^{00}}\right] + \left[\sum_{i=1}^{N^{00}} \frac{\mathrm{F}(\hat{\beta}^{11}\mathrm{X}_i{}^{00})}{N^{00}} - \sum_{i=1}^{N^{00}} \frac{\mathrm{F}(\hat{\beta}^{00}\mathrm{X}_i{}^{00})}{N^{00}}\right]$$

The first term in brackets is the part of the change in defaults due to changes in the distribution of observables X between 2000 and 2011. The second term in brackets is the part of the change in defaults that is not explained by changes in the distribution of observables. The default logit regression also provides a framework for analyzing the factors that influence the default decision, and whether or not they have changed since the seminal study on student loans by Knapp and Seaks (1992). The above framework allows us to estimate the total change in defaults due to compositional changes in observables; it is also possible to estimate the effect of changes in individual explanatory variables such as the total amount borrowed and earnings following Yun (2004).[8]

Contributions of a single variable can be obtained by weighing the contribution of each variable, constructing weight by evaluating the value of a function using mean characteristics and then linearizing the coefficients and characteristics around the predicted values in each year. The decomposition equation is thus given by

$$\bar{Y}_F^{11} - \bar{Y}_F^{00} =$$
$$\sum_{t=1}^{t=K} W_{\Delta X}^t \left[\sum_{i=1}^{N^{11}} \frac{\mathrm{F}(\hat{\beta}^{11}\mathrm{X}_i{}^{11})}{N^{11}} - \sum_{i=1}^{N^{00}} \frac{\mathrm{F}(\hat{\beta}^{11}\mathrm{X}_i{}^{00})}{N^{00}}\right] + \sum_{t=1}^{t=K} W_{\Delta\beta}^t \left[\sum_{i=1}^{N^{00}} \frac{\mathrm{F}(\hat{\beta}^{11}\mathrm{X}_i{}^{00})}{N^{00}} - \sum_{i=1}^{N^{00}} \frac{\mathrm{F}(\hat{\beta}^{00}\mathrm{X}_i{}^{00})}{N^{00}}\right]$$

Where K is the number of variables and $\mathrm{W}_{\Delta\mathrm{X}}^{\mathrm{t}} = \frac{(\overline{\mathrm{X}_{\mathrm{t}}^{11}} - \overline{\mathrm{X}_{\mathrm{t}}^{00}})\beta_{\mathrm{t}}^{11}}{(\overline{\mathrm{X}^{11}} - \overline{\mathrm{X}^{00}})\beta^{11}}$. Note that the t subscript refers to an individual characteristic.

The aim of the decomposition is to determine what default rates would have been in 2011 if the cohort had the same characteristics as individuals who were repaying loans in 2000. In regard to this point, the estimated composition effects due to changes in observable characteristics solely reflect the effect of differences in the distribution of characteristics between borrowers in 2000 and 2011 if the following assumptions are met: (i) simple counterfactual treatment (ii) overlapping support and (iii) conditional independence. Overlapping support is trivially satisfied in the context of the NSLDS as no single variable predicts whether or not an individual was surveyed in the 2000 cohort or the 2011 cohort. The simple counterfactual assumption, whether or not another counterfactual default structure exists, is also

[8] An alternative is the Fairlie (1999) or Bound, Lovenheim and Turner (2010) counterfactual simulation procedure.

satisfied as there are unlikely to be significant general equilibrium effects between individuals across the two surveys,[9] The conditional independence assumption, often called *unconfoundedness* or *selection on observables* is a somewhat stronger assumption and warrants further discussion.

The conditional independence assumption is satisfied if the errors are independent of belonging to each year of the NSLDS conditional on observables. This is a threat to identification if unobservable determinants of default have changed between 2000 and 2011, or if individuals have selected into borrowing across years based on different unobservable characteristics. While it is impossible to observe unobservables, the results in table 9 provide some evidence that at least the relationship between observables and default has remained similar during the two sample periods. Table 8 indicates that the relationship between observable characteristics remains similar in sign and magnitude between 2000 and 2011. The framework and incentives of student loan programs also remained largely similar in both samples, making it unlikely that students selected into borrowing based on different unobserved characteristics in 2000 and in 2011.[10] Table 9 indicates that, there is no evidence that the relationship between observables and default has changed.

[9] The validity of the simple counterfactual assumption rests on whether or not another counterfactual default structure exists. Other counterfactuals may exist due to general equilibrium effects, for example, Fortin, Lemeiux, and Firpo (2010) use the example of a counterfactual wage structure in which there are no unions in the labor market. This could violate the simple counterfactual assumption as the presence of unions is likely to have general equilibrium effects on wages.

[10] The Bankruptcy Abuse Prevention and Consumer Protection Act of 2005 reformed private student loans, where this study focuses on default in Federal student loans. Wage garnishment rates increased from 10 percent to 15 percent in 2006 as part of the Deficit Reduction Act of 2005, however we focus on loans taken in the first year of enrollment prior to the increase in garnishment amounts. If the garnishment had any effect on the adverse selection of riskier borrowers, this would cause the decomposition results to underestimate the impact of the increase in total borrowing on default rates.