DISCUSSION PAPER SERIES

DP16550

Student Employment and Education: A Meta-Analysis

Katerina Kroupova, Tomas Havranek and Zuzana Irsova

LABOUR ECONOMICS

CEPR

ISSN 0265-8003

Student Employment and Education: A Meta-Analysis

Katerina Kroupova, Tomas Havranek and Zuzana Irsova

Discussion Paper DP16550
Published 14 September 2021
Submitted 13 September 2021

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

Student Employment and Education: A Meta-Analysis

Abstract

Educational outcomes have many determinants, but one that most young people can readily control is choosing whether to work while in school. Sixty-nine studies have estimated the effect, but results vary from large negative to positive estimates. We show that the results are systematically driven by context, publication bias, and treatment of endogeneity. Studies ignoring endogeneity suffer from an upward bias, which is almost fully compensated by publication selection in favor of negative estimates. Net of the biases, the literature suggests a negative but economically inconsequential mean effect. The effect is more negative for high-intensity employment and educational outcomes measured as decisions to dropout, but it is positive in Germany. To derive these results we collect 861 previously reported estimates together with 32 variables reflecting estimation context, use recently developed nonlinear techniques to correct for publication bias, and employ Bayesian and frequentist model averaging to assign a pattern to the heterogeneity in the literature.

JEL Classification: C83, I21, J22

Keywords: student employment, Educational outcomes, meta-analysis, Publication bias, Bayesian model averaging

Katerina Kroupova - kacka.kroupova@gmail.com
Charles University, Prague

Tomas Havranek - tomas.havranek@ies-prague.org
Charles University, Prague and CEPR

Zuzana Irsova - zuzana.irsova@ies-prague.org
Charles University, Prague

Student Employment and Education: A Meta-Analysis*

Katerina Kroupova[a], Tomas Havranek[a,b], Zuzana Irsova[a]

[a]Charles University, Prague
[b]CEPR

September 13, 2021

Abstract

Educational outcomes have many determinants, but one that most young people can readily control is choosing whether to work while in school. Sixty-nine studies have estimated the effect, but results vary from large negative to positive estimates. We show that the results are systematically driven by context, publication bias, and treatment of endogeneity. Studies ignoring endogeneity suffer from an upward bias, which is almost fully compensated by publication selection in favor of negative estimates. Net of the biases, the literature suggests a negative but economically inconsequential mean effect. The effect is more negative for high-intensity employment and educational outcomes measured as decisions to dropout, but it is positive in Germany. To derive these results we collect 861 previously reported estimates together with 32 variables reflecting estimation context, use recently developed nonlinear techniques to correct for publication bias, and employ Bayesian and frequentist model averaging to assign a pattern to the heterogeneity in the literature.

Keywords: Student employment, educational outcomes, meta-analysis, publication bias, Bayesian model averaging

JEL Codes: C83, I21, J22

1 Introduction

A vast literature has examined the beneficial effects of education: Psacharopoulos & Patrinos (2018) collect 1,120 estimates for 139 countries over 65 years to conclude that the global annual return to a year of schooling is 9%, while Benos & Zotou (2014), Xue *et al.* (2021), Cui & Martins (2021), and Huang *et al.* (2009) provide meta-analyses of the nexus between education and economic growth, health, collective spillovers, and individual social capital, respectively. Other studies report strong causal effects of education on crime (Machin *et al.*, 2011), happiness (Cunado & Gracia, 2012), political interest (Milligan *et al.*, 2004), fertility (Basu, 2002), and

*Corresponding author: Zuzana Irsova, zuzana.irsova@ies-prague.org. Data and code are available in an online appendix at meta-analysis.cz/students.

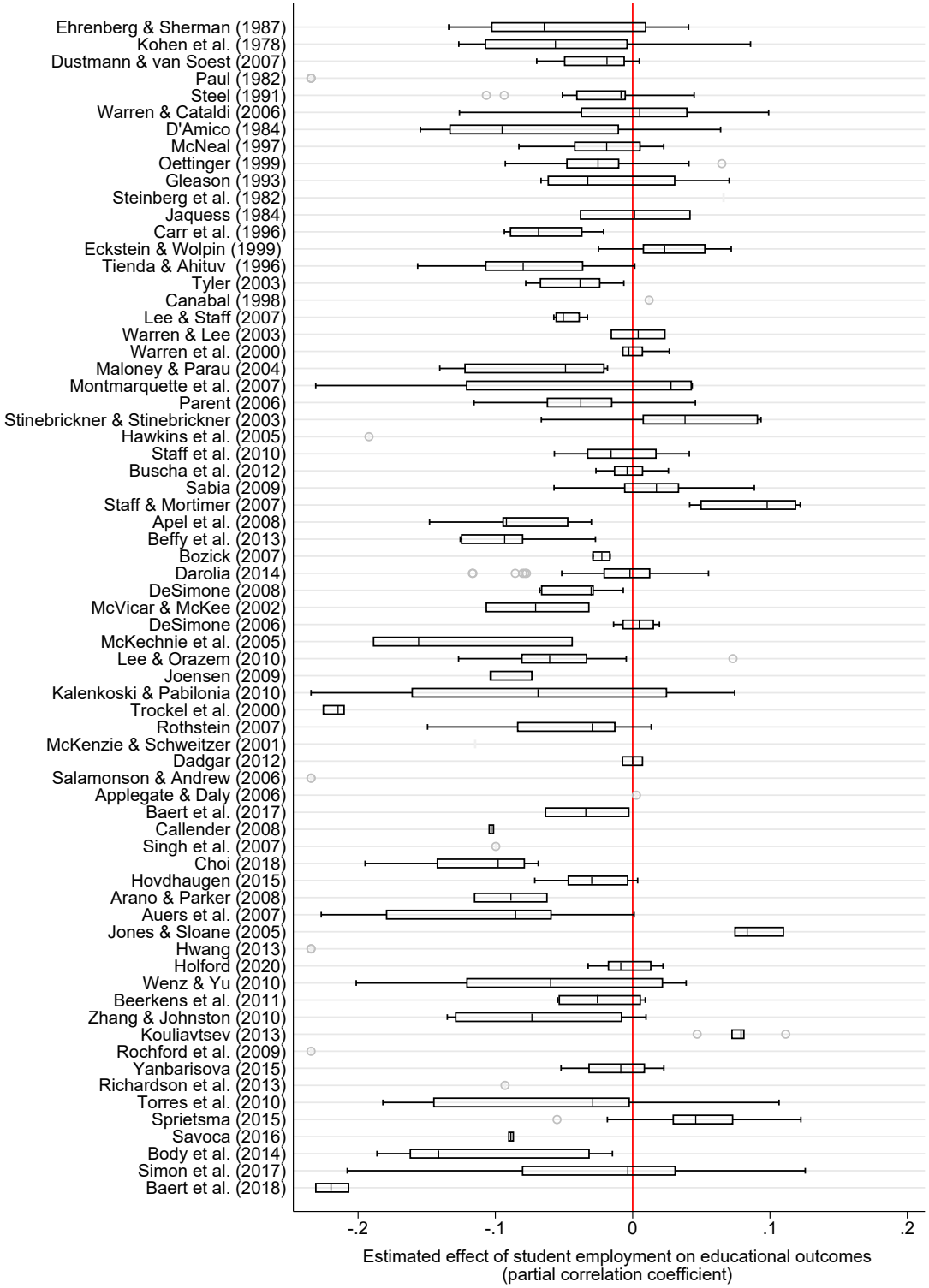
risk aversion (Jung, 2015). But it is hard if not impossible for young people to influence most of the commonly cited determinants of educational outcomes: ability, ethnicity, gender, parental education and affluence. One frequently mentioned determinant, however, is under students' control: employment. True, some young people need to work while in school to sustain their studies. But the ratios of American students employed as of 2019, 45% in college and 19% in high school (BLS, 2020), despite the precipitous decrease since 2000 (when the corresponding ratios stood at 58% and 33%), are too high to be explained by necessity. The effect of employment on educational outcomes, correctly estimated, therefore yields straightforward consequences for students, parents, and policymakers.

Figure 1 shows the main motivation for our paper. Sixty-nine studies have attempted to estimate the effect in question, and collectively they have produced 861 estimates. We recompute these estimates into a comparable metric (partial correlation coefficient) and observe that the results differ greatly across but also within studies: some studies report exclusively negative estimates, a few studies report exclusively positive ones, but most studies report both. The studies in the figure are sorted by data vintage so that the study using the newest data rests on the bottom. No time trend emerges, and the most recent studies are as far from any consensus as the literature was back in the 1980s; the results are all over the place. The lack of consensus makes it difficult to provide clear recommendations for students, parents, and policymakers on this topic, even though excellent narrative surveys of the complex literature are available (most prominently, Neyt *et al.*, 2019). We provide the first quantitative synthesis, a meta-analysis, of this literature, which allows us to isolate the impact of endogeneity and publication bias and to assign a pattern to the heterogeneity apparent in Figure 1.

Our results show that, in most contexts, working while in school does not affect educational outcomes materially. The effect is more negative for high-intensity employment compared to low-intensity employment and for decisions whether to continue with schooling compared to the grades that students receive. Low-intensity employment has no effect on grades, but even the effect of high-intensity employment on dropout probability is small. Regarding cross-country differences, Germany is the only country for which the research literature shows that student employment improves educational outcomes on average. Though for comparability with the rest of the sample we do not consider estimates that address apprenticeships in German Berufsschulen (vocational schools), the long German tradition of effectively combining work and education translates into a corresponding synergy even at the college level (for details on the German system, see, for example, Rözer & van de Werfhorst, 2020). On balance, the 861 estimates reported in the 69 existing studies are consistent with the conclusion that student employment does not hurt education—and, of course, typically has positive influence on other aspects of young people's lives.

On the more technical side, we find an unusual interaction between endogeneity and publication biases, and the fundamental results described above are corrected for both biases as well as other misspecifications. Endogeneity is key here because mostly unobserved characteristics, especially ability, influence both educational outcomes and the decision to work while in

Figure 1: The 861 estimates vary widely both within and across studies



Notes: The figure shows a box plot of partial correlation coefficients (computed from reported coefficients for comparability) reflecting the estimated relationship between student employment and educational outcomes; the studies are sorted by the age of the data they use from oldest to youngest. The length of each box represents the interquartile range (P25-P75), and the line inside the box represents the median. The whiskers represent the smallest and largest estimates within 1.5 times the range between the upper and lower quartiles. Circles denote outliers; the vertical line denotes zero. Extreme outliers are excluded from the figure but included in all statistical tests.

school: able students can combine work and study with good results, displaying both more hours worked and better educational outcomes. If a researcher ignores ability, she wrongly concludes that student employment improves education. But it is also plausible that in some cases the endogeneity bias is negative: for example, students from disadvantaged families can be forced to work in order to sustain their studies, while also showing a propensity for weaker educational outcomes whether or not they work. Researchers have tackled the problem by employing quasi-experimental techniques (matching, instrumental variables, difference-in-differences) or by using a proxy for ability (such as IQ) together with variables reflecting family background (ethnicity, parental education, family affluence). About a half of the estimates are computed while ignoring endogeneity, which makes them obviously suspicious. Instead of omitting these estimates, we use them to identify the mean endogeneity bias in the literature. The bias is positive, suggesting that the ability explanation dominates the family background explanation.

The second major source of bias in the literature is publication selection (Stanley, 2001),[1] which can, in the absence of pre-registered replications (and pre-registration is no panacea for observational research, where authors can inspect data prior to pre-registration), only be addressed by meta-analysis. Researchers write their papers with the intention to publish, and some may consider negative estimates more intuitive and thus publishable compared to positive estimates, especially when the estimates are statistically significant and therefore appear strong and important. Of course, publication selection bias does not equal cheating. An unintuitive result may indicate an issue with the data or the model, and the researcher can often improve the results by running a different specification. The problem is that unintuitive (positive or insignificant) results are easy to spot, while large negative estimates, which might also be due to issues with data or methods, are hard to identify. The asymmetry in the selection rule causes a bias away from zero in most fields of economics (Ioannidis *et al.*, 2017); the bias is natural, inevitable, and it is thus the task of those who take stock of the literature to identify and correct for the bias. We find that the estimates tackling endogeneity, being almost always slightly negative, are free from the bias—a rare finding in economics. In contrast, estimates that ignore endogeneity are plagued by the publication bias, because in the absence of selection they naturally gravitate towards positive and thus less intuitive results.

For the identification of and correction for publication bias we employ the property of econometric techniques used in the literature on the effect of student employment on education: in the absence of publication bias, estimates and standard errors are statistically uncorrelated quantities. The authors of these studies report t-statistics, which only make sense if the ratio of estimates and standard errors can be expected to follow a symmetrical distribution (such as the t-distribution). If estimates and standard errors were correlated, the t-statistics reported in the studies would be meaningless. To explain the identification procedure, it is useful to invoke McCloskey & Ziliak (2019), who liken publication selection to the Lombard effect in

[1]For recent papers on publication selection bias in economics see Blanco-Perez & Brodeur (2020), Brodeur *et al.* (2020), Brodeur *et al.* (2016), Bruns & Ioannidis (2016), Card *et al.* (2018), Christensen & Miguel (2018), DellaVigna *et al.* (2019), Havranek (2015), Imai *et al.* (2020), and Ioannidis *et al.* (2017). Earlier important papers on the topic include Ashenfelter *et al.* (1999), Card & Krueger (1995), Stanley (2005), and Stanley (2008).

psychoacoustics: speakers tend to increase their vocal effort proportionally in response to noise. Similarly some researchers can work harder in response to noisy data in order to obtain statistically significant estimates by altering the estimation technique, control variables, or treatment of outliers. A correlation between estimates and standard errors follows, because larger standard errors (and thus more noise) require larger point estimates to yield statistical significance. Both quantities become similarly correlated if researchers prefer a particular sign of regression estimates since imprecise estimates, in comparison to precise estimates, are more likely to display the "wrong" sign simply by chance. Because in this framework publication bias is a linear function of the standard error, correcting for publication bias involves deriving an estimate conditional on infinite precision, the intercept in a regression of estimates on standard errors.

The linear model of publication bias described in the previous paragraph has two main problems. First, publication bias can form a complex function of the standard error. For example, when the standard error is very small and the t-statistic thus very large, a small change in the standard error is unlikely to influence the probability with which the estimate is published. In contrast, when the t-statistic is slightly above 2, a small increase in the standard error can render the estimate unpublishable in the researcher's eyes. Therefore we also employ recently developed nonlinear techniques for publication bias correction, namely the weighted average of adequately powered estimates (Ioannidis *et al.*, 2017), stem method (Furukawa, 2021), endogenous kink model (Bom & Rachinger, 2019), and selection model (Andrews & Kasy, 2019). Second, the standard error can be endogenous because i) it is itself an estimate, ii) publication selection can work on the standard error in addition to the point estimate, and iii) some aspects of the estimation context can influence both estimates and standard errors. In almost all applications of meta-analysis, standard errors are expected to be given. But unlike in medical research, where meta-analysis was developed, in economics the estimation of standard errors forms an important part of any empirical exercise; the standard error is not exogenous. Our solution is to use the inverse of the number of observations as an instrument for the standard error and additionally employ the new p-uniform* technique recently developed in psychology (van Aert & van Assen, 2021) that does not need the exogeneity assumption.

In the second part of the paper we investigate the sources of heterogeneity in the literature beyond publication and endogeneity biases. We collect 32 aspects that reflect the context in which the estimate was obtained: characteristics of the data (e.g., definition of variables), structural variation (e.g., gender, race, country), estimation method (e.g., matching, instrumental variables), and publication characteristics (e.g., study citations). Regressing these 32 variables on the collected estimates of the effect of student employment on educational outcomes has two problems. First, model uncertainty: we do not know ex ante which of the variables truly belong to the underlying model, but we still want to control for their potential impact on the estimates reported in the literature in order to avoid omitted variable bias. Including all variables in an OLS regression would greatly increase the standard error even for the most important variables. As a solution we choose Bayesian model averaging (for details, see, for example Eicher *et al.*, 2011), which is the natural response to model uncertainty in the Bayesian setting (Steel, 2020).

Bayesian model averaging runs many regressions with different combinations of the 32 explanatory variables and weights them according to model fit and parsimony. Second, collinearity: interpretation of individual partial correlations is difficult. Although collinearity is not large in our dataset, we use the dilution prior by George (2010), which partly addresses the issue. As robustness checks, we use priors according to Fernandez *et al.* (2001) and Ley & Steel (2009); in addition, we use frequentist model averaging with Mallows weights Hansen (2007) using the orthogonalization of covariate space suggested by Amini & Parmeter (2012).

The model averaging analysis confirms the importance of endogeneity and publication biases even after controlling for additional aspects of study design. Studies that ignore endogeneity tend to report more positive estimates, while studies that employ matching, instrumental variables, difference-in-differences, or studies that include a proxy for ability, tend to report more negative estimates. Publication bias affects studies that ignore endogeneity, while studies that control for endogeneity appear to be mostly free of the bias. Other study characteristics that systematically affect the reported effects of student employment on education are the measurement of educational outcomes (average grades vs. decisions to dropout), structure of the data (panel vs. cross-section), employment intensity (high vs. low), country (Germany vs. others), and the use of control variables (motivation, ethnicity). As the bottom line of our analysis, we create a hypothetical study that is derived as a weighted average over all the estimates in our dataset but uses the results of Bayesian model averaging to give more weight to studies that are more credible—so that, for example, little weight is placed on imprecise studies ignoring endogeneity, more weight is placed on highly-cited studies published in top journals, etc. We construct the hypothetical study for several scenarios reflecting different context (e.g., high vs. low-intensity employment). In all scenarios the implied negative effect of student employment on educational outcomes is too small to be important in practice.

2 Data

In this section we describe how we collect data for the meta-analysis. The description requires a brief discussion of how researchers typically measure the effect of student employment on education. More details on measurement follow in Section 4, and an in-depth discussion, which we do not replicate in this paper, is available in Neyt *et al.* (2019). Put simply, the estimates that we collect stem from models that can be reduced into the following regression:

$$Educational\ outcome_{jt} = \beta_0 + \beta_1 Employment_{jt} + \beta_2 Controls_{jt} + \epsilon_{jt}, \tag{1}$$

where $Educational\ outcome_{jt}$ denotes education of student j in time t, $Employment_{jt}$ denotes the student's employment, ϵ_{jt} is the error term, and vector of $Controls_{jt}$ denotes the set of variables controlling for preexisting heterogeneity. The vector contains characteristics of individuals (such as age, race, religious affiliation, past performance, and motivation), family background (such as parents' marital status, educational attainment, number of siblings, and family income), or the specifics of the schooling institution (class size, public vs. private school,

regional unemployment). The coefficient of main interest is β_1, which is what we collect from the studies—together with the standard error and 31 other variables that reflect the context in which the coefficient was produced.

Educational outcomes can be measured in a variety of ways. Some researchers define educational outcomes in terms of study habits, which refer to measures such as class attendance or time spent studying (Schoenhals *et al.*, 1998; Marsh & Kleitman, 2005). Some define them as choices made during the course of studies, for example whether to continue with further education (Steel, 1991). A natural measure of educational outcomes is a test result, and this definition is also the one most commonly used in the literature (DeSimone, 2008; Dustmann & van Soest, 2007). Other researchers focus on educational attainment, which comprises of students' probable and actual achievements (Beffy *et al.*, 2013; Dadgar, 2012). Measuring student employment is only slightly easier. Most studies estimate β_1 in terms of the effect of employment intensity on education, while the rest estimate β_1 in terms of the effect of employment status on education. Researchers using employment status as the response variable simply distinguish between working and non-working students, defining student employment as a dummy variable (see, for example McKenzie & Schweitzer, 2001; McNeal, 1997). In contrast, researchers using employment intensity define the variable either as a continuous (average hours worked per week, such as in Kalenkoski & Pabilonia, 2010) or a categorical variable (defining several categories of work intensity, such as in Torres *et al.*, 2010; Tyler, 2003). The coefficient β_1 thus has different interpretation depending on study design. Even early researchers in this field admit that "*the range of findings may be an artifact of the different operationalisations*" (McNeal, 1997, p. 208).

Narrative surveys of this literature date back to Newman (1942), and all struggle with the differences in the definitions of both variables and, fundamentally, with results as shown in Figure 1 in the Introduction. As Riggert *et al.* (2006, p. 85) put it: "*A critical reading of the empirical literature on student employment could legitimately lead different readers to different conclusions.*" One solution is to review the literature narrowly, focusing only on one definition of the effect (for example, how much an additional hour of work per week changes the grade point average). Such an restrictive approach would, however, eliminate 90% of the results reported in the literature. While we use the restrictive approach as a robustness check, for the main analysis we convert all estimates to a comparable metric, partial correlation coefficient (PCC):

$$PCC(\beta_1)_{is} = \frac{T(\beta_1)_{is}}{\sqrt{T(\beta_1)_{is}^2 + DF(\beta_1)_{is}}}, \tag{2}$$

where $PCC(\beta_1)_{is}$ represents the partial correlation coefficient of i-th estimate reported in study s, $T(\beta_1)_{is}$ denotes the corresponding t-statistic, and $DF(\beta_1)_{is}$ represents the number of degrees of freedom relevant to β_1 from (1). The standard errors of PCCs are calculated as a ratio of PCC to the respective t-statistic $T(\beta_1)_{is}$.

To search for studies reporting the effect of student employment on educational outcomes (we will call them primary studies), we use Google Scholar; for details on the search query and other aspects of literature search, see Figure A1. We examine the abstract of the first 500

studies returned by the query. If the abstract indicates any possibility that the study might contain empirical estimates that we can use, we download the study and inspect it in detail. We follow the guidelines of Havranek *et al.* (2020) for collecting data in meta-analysis; we add the last study on May 21, 2021.

We use three inclusion criteria. First, not to introduce additional heterogeneity into our sample, we exclude two broad definitions of educational outcomes: time spent on study habits and time to obtain a degree. Measures of study habits (such as time spent doing homework or time spent preparing for class, see Marsh & Kleitman, 2005; Manthei & Gilmore, 2005; Schoenhals *et al.*, 1998) represent in our view a process leading to an educational outcome rather than the educational outcome itself. Moreover, these measures are almost always self-reported and, as Applegate & Daly (2006) document, subject to individual over- or under-estimation and hence strong measurement error. Measures of time to obtain a degree (as in Theune, 2015) are affected by trends in study patterns, mostly by the habit of taking gap years or prolonging studies in order to exploit the tax benefits of the student status. Though it is difficult to draw lines, the line has to be drawn somewhere, and we do not consider studies employing the two definitions mentioned above quantitatively comparable with the rest of the literature.

Second, again for the sake of comparability we exclude three definitions of student employment. We do not use studies focusing on student employment in the primary school setting (as in Post & Pong, 2000) since in this context student work is illegal, rare, and mostly limited to a few specific developing countries. Similarly, we discard studies examining the impact of "sandwich work" placement (a year-long integrated period of work experience in students' study program) because such programs are specifically designed to be part of the curriculum with the aim to enhance student academic performance (Jones *et al.*, 2017; Scott-Clayton & Minaya, 2016). Finally, we exclude studies investigating the relationship between summer employment and educational outcomes (Leos-Urbel, 2014, for example) and strictly adhere to research papers focusing on work during school terms.

Third, to be included in the meta-analysis the study must report the standard error or another measure from which the standard error can be reconstructed. We thus exclude several studies that do not report any measure of uncertainty or report only the number of asterisks to represent significance (as in Marsh & Kleitman, 2005; McCoy & Smyth, 2007; Wang *et al.*, 2010, among others). Following Stanley (2001), no study is disqualified on the basis of publication form. Therefore, aside from peer-reviewed journal articles we use working papers, book chapters, and dissertations and control for study quality later in the analysis. The final sample includes 861 estimates collected from 69 studies listed in Table 1.

Before transforming the collected estimates into partial correlation coefficients via (2), we make a number of adjustments to ensure the comparability of these estimates. Several studies, including Bozick (2007) and Warren & Lee (2003), employ logistic regression and report odds ratios. We transform the reported odds ratios (or) into the regression coefficients using the formula $\widehat{or}_{is} = e^{\beta_{is}}$, where β_{is} is our desirable effect estimate from the i-th specification in study s; we follow Oehlert (1992) and define the odds-ratio adjusted standard error as $SE(\widehat{or}_{is}) =$

Table 1: The 69 studies included in the meta-analysis

Apel *et al.* (2008)	Gleason (1993)	Rochford *et al.* (2009)
Applegate & Daly (2006)	Hawkins *et al.* (2005)	Rothstein (2007)
Arano & Parker (2008)	Holford (2020)	Sabia (2009)
Auers *et al.* (2007)	Hovdhaugen (2015)	Salamonson & Andrew (2006)
Baert *et al.* (2017)	Hwang (2013)	Savoca (2016)
Baert *et al.* (2018)	Jaquess (1984)	Simon *et al.* (2017)
Beerkens *et al.* (2011)	Joensen (2009)	Singh *et al.* (2007)
Beffy *et al.* (2013)	Jones & Sloane (2005)	Sprietsma (2015)
Body *et al.* (2014)	Kalenkoski & Pabilonia (2010)	Staff & Mortimer (2007)
Bozick (2007)	Kohen *et al.* (1978)	Staff *et al.* (2010)
Buscha *et al.* (2012)	Kouliavtsev (2013)	Steel (1991)
Callender (2008)	Lee & Orazem (2010)	Steinberg *et al.* (1982)
Canabal (1998)	Lee & Staff (2007)	Stinebrickner & Stinebrickner (2003)
Carr *et al.* (1996)	Maloney & Parau (2004)	Tienda & Ahituv (1996)
Choi (2018)	McKechnie *et al.* (2005)	Torres *et al.* (2010)
Dadgar (2012)	McKenzie & Schweitzer (2001)	Trockel *et al.* (2000)
D'Amico (1984)	McNeal (1997)	Tyler (2003)
Darolia (2014)	McVicar & McKee (2002)	Warren & Cataldi (2006)
DeSimone (2006)	Montmarquette *et al.* (2007)	Warren & Lee (2003)
DeSimone (2008)	Oettinger (1999)	Warren *et al.* (2000)
Dustmann & van Soest (2007)	Parent (2006)	Wenz & Yu (2010)
Eckstein & Wolpin (1999)	Paul (1982)	Yanbarisova (2015)
Ehrenberg & Sherman (1987)	Richardson *et al.* (2013)	Zhang & Johnston (2010)

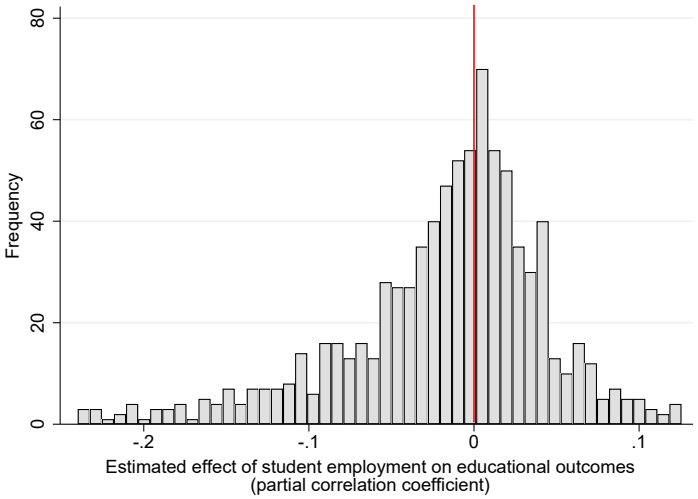
Notes: The dataset, together with R and Stata codes, is available at meta-analysis.cz/students.

$SE(\beta_{is})e^{\beta_{is}}$, where $SE(\beta_{is})$ is the standard error of the original estimate. Similarly, some studies examine a nonlinear effect of student employment on educational outcomes and report an estimate for the quadratic term. Here we linearize the effect to $\beta_{is} = \widehat{\beta}_{lis} + \widehat{\beta}_{qis}\overline{x}_{es}$, with $\widehat{\beta}_{lis}$ being the estimate of the linear term and $\widehat{\beta}_{qis}$ being the estimate of the quadratic term, multiplied by the sample mean of the variable corresponding to student employment $\overline{x}_{es}$ as used in study s. The corresponding standard error is defined as $SE(\beta_{is}) = \sqrt{SE(\widehat{\beta}_{lis})^2 + SE(\widehat{\beta}_{qis})^2\overline{x}_{es}}$.

Furthermore, two studies in our dataset estimate an interaction effect between student employment and other variables (Steel, 1991; Carr *et al.*, 1996). Here we calculate the average marginal effect of student employment on education as $\beta_{is} = \widehat{\beta}_{lis} + \widehat{\beta}_{tis}\overline{x}_{is}$ and approximate the corresponding standard error using the delta method as $SE(\beta_{is}) = \sqrt{SE(\widehat{\beta}_{lis})^2 + SE(\widehat{\beta}_{tis})^2\overline{x}_{is}}$, where $\widehat{\beta}_{tis}$ is the estimate of the included interaction term and $\overline{x}_{is}$ is the the mean value of the variable included in the interaction term. In several instances, we adjust the signs of the reported estimates so that they correctly reflect the direction of the effect (compare the effect for educational outcome defined as students' dropout likelihood of McNeal, 1997, with the effect for outcome defined as the likelihood of completing secondary education, as in Carr *et al.*, 1996). A few extreme outliers appear in the dataset, and we thus winsorize the estimates at the 1% level.

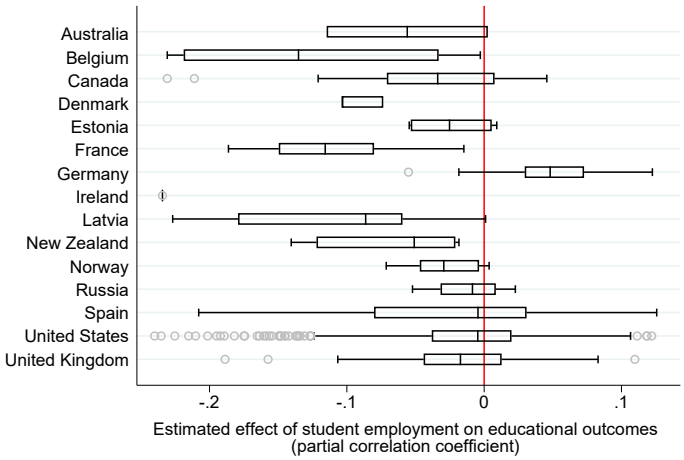
The mean partial correlation coefficient is -0.017, while the median is -0.006. To put these numbers into perspective, consider Doucouliagos (2011), who collects 22,000 partial correlation coefficients produced in economics and creates guidelines for what can be considered zero, small, moderate, and large effects. The boundary between zero and small effects is 0.07, so the bulk

Figure 2: Most common in the literature are zero estimates



Notes: The figure shows the distribution of the partial correlation coefficients reflecting the estimated relationship between student employment and academic achievement. The vertical line represents zero. For ease of exposition, extreme outliers are excluded from the figure but included in all statistical tests.

Figure 3: Estimates vary within and across countries



Notes: The figure shows a box plot of partial correlation coefficients reflecting the estimated relationship between student employment and academic achievement as reported for different countries. The vertical line denotes zero. For ease of exposition, extreme outliers are excluded from the figure but included in all statistical tests.

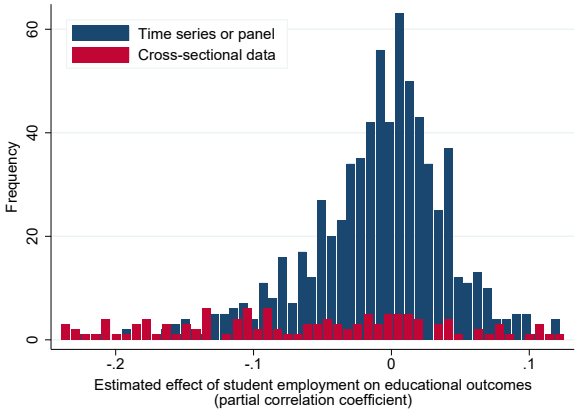
of the literature is consistent with the notion that working while in school has no material effect on educational outcomes. Estimates very close to zero are generally most common in the literature, which is apparent from the histogram in Figure 2. In the absence of publication bias, small-sample bias, and heterogeneity, we would expect the observed distribution of the estimates to be symmetrical. The histogram shows only a slight asymmetry, and 45% of the estimates have the less intuitive positive sign, which is unusual in economics. Next, Figure 3 documents the cross-country variation in our dataset. Two patterns stand out. First, only for one country the average estimate is positive: Germany. While for comparability we exclude estimates derived for German vocational schools (where combination of work and study is the central educational principle), these results suggest the German system is efficient in combining work and study even for other types of schools. Second, the only countries for which the mean partial correlation coefficient is smaller than -0.1 are France and Belgium. While the simple mean is only indicative, it is true that the educational system in both countries shares many common features and differs from the German model in terms of the traditional interaction of study and work (see, for example, Rözer & van de Werfhorst, 2020, for details).

Figure 4 shows six aspects of heterogeneity that are frequently discussed in the primary studies: the dimension of data, students' gender, employment intensity, educational level, control for endogeneity, and differences between the United States and other countries. In panel (a), we see that the distribution of estimates stemming from cross-sectional models is close to uniform, while the distribution of time series or panel estimates is closer to normal and much less likely to deliver estimates below -0.1. Panel (b) shows that students' gender typically does not matter for the results. From panel (c) it is apparent that high-intensity employment, compared to low-intensity employment, is more detrimental to educational outcomes. Panel (d) suggests that while estimates for secondary education are concentrated close to 0, estimates for tertiary education are much more dispersed. From panel (e) we see that most positive estimates are derived from techniques that ignore endogeneity (that is, techniques that do not use matching, instrumental variables, difference-in-differences, nor attempt to use a proxy to control for ability). Panel (f) shows little evidence that estimates for the United States differ from other countries, with the exception of Germany; the figure also shows that the number of estimates for Germany is limited.

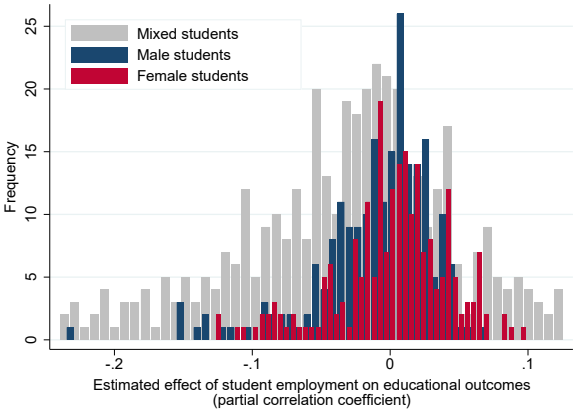
Table 2 displays more detailed comparisons of various subsets of the data. The left-hand part of the table shows unweighted means and the corresponding 95% confidence intervals, while the right-hand part of the table shows means weighted by the inverse of the number of estimates reported in each study. That is, in the left-hand part of the table each estimate has the same weight, while in the right-hand part of the table each study has the same weight. Several patterns stand out on top of those discussed earlier in relation to Figure 4. First, it matters how educational outcomes are measured. The effect of student employment is much more negative when educational outcomes are measured in terms of choices that students can make: typically whether to dropout or whether to apply to college after high school. The corresponding mean partial correlation reaches -0.08 when each study is given the same weight. In comparison, mean

Figure 4: Selected patterns in the data

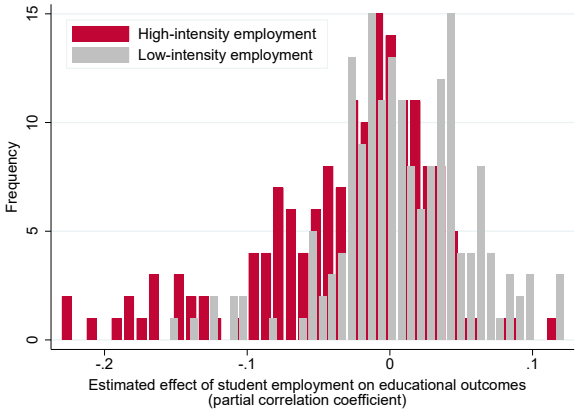
(a) Data dimension



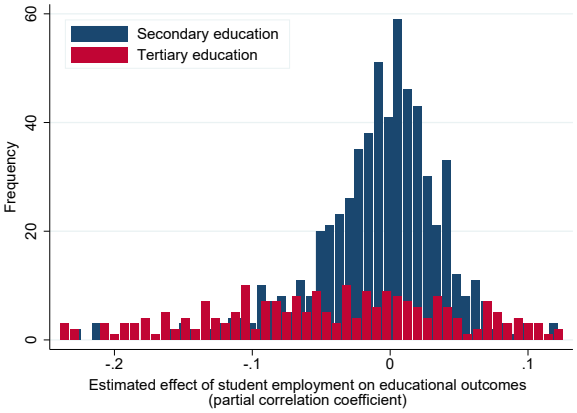
(b) Gender differences



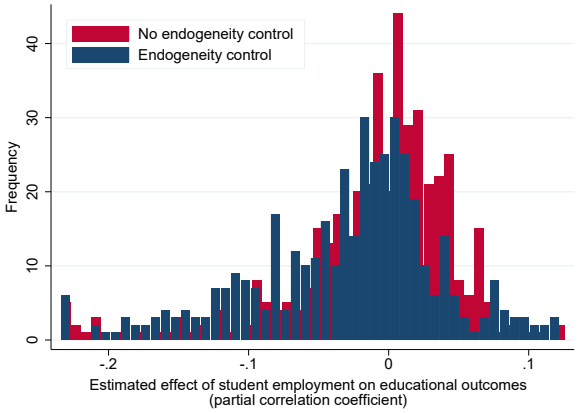
(c) Employment intensity



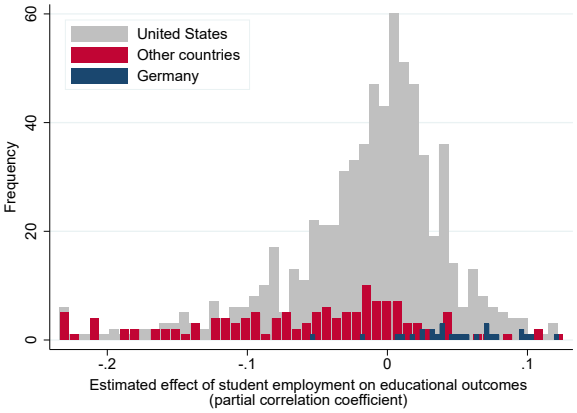
(d) Education level



(e) Method choice



(f) Spatial differences



Notes: The figure depicts, for different subsets of data, histograms of partial correlation coefficients reflecting the estimated relationship between student employment and academic achievement.

Table 2: Summary statistics for different subsets of the literature

		Unweighted			Weighted		
	No. of obs.	Mean	95% conf. int.		Mean	95% conf. int.	
Data characteristics							
Employment: continuous variable	147	-0.030	-0.040	-0.020	-0.040	-0.049	-0.031
Employment: dummy variable	116	-0.022	-0.036	-0.008	-0.039	-0.056	-0.022
Employment: categorical variable	598	-0.014	-0.019	-0.008	-0.057	-0.065	-0.050
Educational outcome: choices	261	-0.039	-0.048	-0.030	-0.080	-0.092	-0.069
Educational outcome: attainment	158	-0.007	-0.016	0.003	-0.024	-0.034	-0.013
Educational outcome: test scores	442	-0.008	-0.014	-0.003	-0.022	-0.029	-0.015
Self-reported education	224	-0.031	-0.041	-0.020	-0.048	-0.060	-0.036
Longitudinal data	729	-0.008	-0.012	-0.004	-0.025	-0.030	-0.020
Cross-sectional data	132	-0.069	-0.086	-0.052	-0.087	-0.105	-0.069
Structural variation							
Male students	218	-0.013	-0.020	-0.007	-0.056	-0.067	-0.044
Female students	222	0.002	-0.003	0.007	-0.023	-0.028	-0.017
Mixed-gender students	421	-0.030	-0.038	-0.022	-0.053	-0.062	-0.045
Caucasian students	33	-0.025	-0.055	0.004	-0.081	-0.124	-0.038
Minority students	46	-0.002	-0.018	0.014	-0.025	-0.042	-0.008
Part-time students	33	0.004	-0.003	0.011	-0.002	-0.010	0.006
Secondary education	621	-0.008	-0.012	-0.004	-0.030	-0.035	-0.025
Tertiary education	240	-0.042	-0.054	-0.030	-0.069	-0.082	-0.057
Low-intensity employment	185	0.013	0.006	0.021	0.014	0.004	0.024
Medium-intensity employment	94	-0.011	-0.023	0.000	-0.035	-0.051	-0.020
High-intensity employment	163	-0.031	-0.041	-0.021	-0.044	-0.055	-0.034
On-campus employment	17	-0.042	-0.095	0.011	-0.063	-0.120	-0.006
Spatial variation							
United States	694	-0.013	-0.018	-0.009	-0.043	-0.049	-0.037
Germany	29	0.052	0.037	0.068	0.052	0.037	0.068
Other countries	138	-0.053	-0.067	-0.039	-0.071	-0.086	-0.055
Estimation methods							
Endogeneity control	425	-0.027	-0.034	-0.021	-0.036	-0.044	-0.029
No endogeneity control	436	-0.008	-0.013	-0.002	-0.075	-0.084	-0.066
OLS method	525	-0.013	-0.020	-0.007	-0.057	-0.066	-0.049
Matching method	29	-0.041	-0.057	-0.025	-0.060	-0.073	-0.047
DID method	44	-0.005	-0.011	0.002	-0.016	-0.026	-0.007
IV method	138	-0.041	-0.051	-0.030	-0.045	-0.055	-0.034
Other method	125	-0.008	-0.017	0.001	-0.024	-0.037	-0.012
Publication characteristics							
Unpublished study	76	-0.004	-0.021	0.014	-0.018	-0.035	-0.002
Published study	785	-0.019	-0.023	-0.014	-0.058	-0.064	-0.052
Published before 1991	40	-0.060	-0.087	-0.034	-0.055	-0.089	-0.022
Published in 1991-2000	103	-0.030	-0.040	-0.020	-0.039	-0.053	-0.025
Published in 2001-2010	453	-0.010	-0.016	-0.004	-0.053	-0.062	-0.045
Published after 2010	265	-0.019	-0.027	-0.011	-0.053	-0.063	-0.042
All estimates	861	-0.017	-0.022	-0.013	-0.051	-0.057	-0.045

Notes: In the left-hand portion of the table each estimate has the same weight. In the right-hand portion of the table each study has the same weight; in other words, there we weight estimates by the inverse of the number of estimates reported per study.

estimates are four times smaller when educational outcomes are measured by test scores or other proxies for educational attainment. Second, the mean estimates for individual techniques for addressing endogeneity differ quite a lot: for example, when each study has the same weight, the mean estimate for matching is -0.06, while only -0.016 for difference-in-differences. Third, the estimates seem to be stable in time, no trend emerges, which is consistent with Figure 1 discussed in the Introduction. But of course conclusions based on group means can be affected by heterogeneity (omitted variable bias) and selection reporting (publication bias), issues to which we turn in the next two sections.

3 Publication and Endogeneity Biases

Publication bias, the tendency to report estimates that are easier to publish, can distort inference drawn from individual studies and literature reviews alike. Ioannidis *et al.* (2017) show that the mean estimate reported in economics is exaggerated twofold because of the bias. Another problem that must be taken into account in the literature on the effect of student employment on educational outcomes is endogeneity bias: unobserved characteristics of students may drive both decisions to work and educational outcomes. In this section we examine the interaction between the two biases. To provide some context, in three introductory paragraphs we describe how both negative and positive estimates have been interpreted in the literature. While theoretical arguments can be made in favor of positive estimates as well, negative estimates are easier to sell, and therefore one expects publication bias against positive (and insignificant) estimates in most contexts. Next, we evaluate publication bias in the entire sample of the estimates we collect, with mixed results. Finally, we separate estimates that ignore endogeneity from estimates that take them into account. While the latter are not affected by publication bias, we find strong publication bias for the former. Estimates ignoring endogeneity are inherently biased upwards, because more able students are more likely to seek a job and at the same time achieve good results in school. But the positive estimates of the effect of student employment on educational outcomes are unintuitive to many researchers, and thus publication selection against such estimates follows and leads to publication bias in studies ignoring endogeneity.

Concerning the interpretation of negative and positive estimates, a theoretical case can be made for substitutability as well as complementarity between student employment and educational outcomes. The *developmental model* (Marsh, 1991) predicts that students' work can contribute to the development of relevant knowledge (Wang *et al.*, 2010; Geel & Backes-Gellner, 2012) and soft skills (including problem-solving, organizational skills, time-management, communication, working under pressure, and presentation skills, see Darolia, 2014) that spill over to the academic setting (Buscha *et al.*, 2012). Stern & Briggs (2001) and Rothstein (2007) argue that an early-age work experience might aid students to ascertain their career goals and motivate them to work harder during their studies. In contrast, the *zero-sum model* predicts that employment crowds out the time which should be devoted to academic activities (Marsh, 1991). Employment does not only reduce the time available for homework and independent study (Choi, 2018; D'Amico, 1984), but can also impair students' involvement in the academic

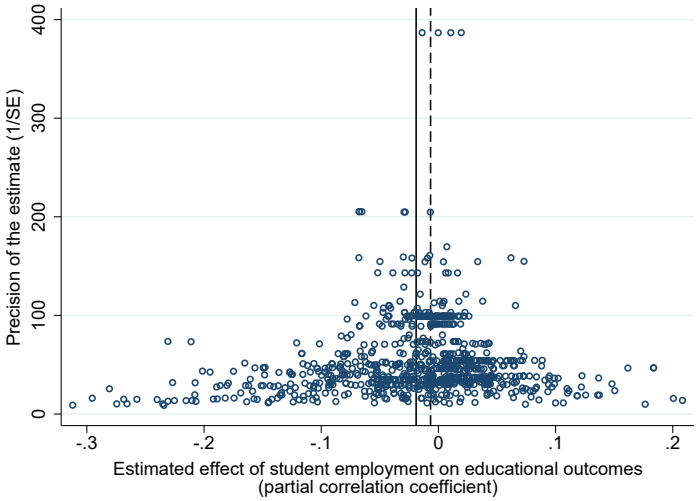
community (undermining their academic commitment, see Darolia, 2014) and produce excessive fatigue, decreasing students' attentiveness (Oettinger, 1999).

The so-called *threshold model* reconciles the theoretical mechanisms behind the aforementioned theories: with increasing working hours, the marginal benefits of student employment decrease and, after surpassing a certain threshold, they begin to crowd out the time crucial for academic success. Some studies, such as Choi (2018), even show that working may be simultaneously a complement and a substitute to academic performance. The *primary-orientation* perspective (Choi, 2018; Lee & Staff, 2007) holds that various socio-psychological factors (including family attitudes towards education, motivation, and educational aspirations) form altogether an individual commitment towards education or work experience (Warren, 2002, for example, argues that educational engagement develops before the decision to work). The investigated effect thus becomes non-significant or less pronounced and does not necessarily have to be causal, as Baert *et al.* (2018) shows. Many researchers also document *background heterogeneity* between the students: the effect varies greatly by ethnic group (D'Amico, 1984), gender (Buscha *et al.*, 2012; Holford, 2020), job type (McNeal, 1997; Sabia, 2009), motivation to work (Wenz & Yu, 2010), job industry (Dadgar, 2012), and educational level (Neyt *et al.*, 2019).

A student of the literature might therefore reject the notion that a dominant theory should drive publication selection bias. Not only are there multiple arguments offering plausible explanations for both positive and negative estimates, but also researchers themselves often acknowledge that there seems to be little consensus on whether student employment hinders or improves academic performance (see, for instance Oettinger, 1999; Sabia, 2009; Tyler, 2003). Nonetheless, it is our overall impression after reading the literature that most researchers believe that the underlying effect is negative, which is also the most intuitive conclusion. For example, Buscha *et al.* (2012, p. 383) admit that "*the view that part-time work has a detrimental effect on educational attainment [...] is increasingly widespread in the last 10 years.*" Similarly, the large authoritative survey by Neyt *et al.* (2019) argues that the most convincing studies report more negative effect estimates compared to less advanced studies.

We begin our investigation of publication bias by employing a visual tool called a funnel plot (Egger *et al.*, 1997). The funnel plot represents a scatter plot in which the estimate's magnitude is depicted on the horizontal axis against a measure of precision (the inverted standard error) on the vertical axis. The most precise estimates should lie close to the true mean effect in the top portion of the graph, with variance increasing at the bottom as precision decreases. Therefore, in the absence of publication bias, the graph should form a symmetrical inverted funnel (Stanley, 2005). In contrast, an asymmetry of the funnel plot indicates the presence of publication bias via preference for positive or negative estimates. (Though the asymmetry can also be caused by small-sample bias or heterogeneity.) The funnel plot presented in Figure 5 roughly forms the predicted inverted funnel shape with a high level of symmetry. Even very imprecise estimates concentrated at the bottom of the figure are reported. Perhaps the left-hand part of the figure is slightly heavier, but overall few funnel plots in economics display so little asymmetry (Ioannidis *et al.*, 2017). The visual test does not indicate publication bias.

Figure 5: Funnel plot shows little publication bias on average



Notes: The solid vertical line represents the mean estimate, the dashed vertical line represents the median estimate. In the absence of publication bias the scatter plot should resemble an inverted funnel symmetrical around the mean. Outliers are excluded from the figure but included in all statistical tests.

Funnel asymmetry can be tested formally by regressing the values on the horizontal axis (PCCs) on the inverted vertical axis values (standard errors) as in Card & Krueger (1995):

$$PCC_{is} = PCC_0 + \gamma SE(PCC_{is}) + \epsilon_{is}, \tag{3}$$

where PCC_{is} are the partial correlation coefficients, $SE(PCC_{is})$ are the corresponding standards errors, and ϵ_{is} represents the error term. We interpret the constant PCC_0 as the true effect corrected for publication bias (that is, conditional on infinite precision) but, as we have noted, later introduce extensions that allow for nonlinearity and endogeneity. Coefficient γ conveys information regarding the existence, direction, and magnitude of publication bias: if we obtain an estimate of γ statistically different from zero, we find evidence for funnel asymmetry, i.e. a non-zero correlation between estimates and their standard errors. To account for potential within-study correlation, we cluster standard errors at the study level. Moreover, we also report wild bootstrap confidence intervals (Roodman *et al.*, 2018).

Panel A and Panel B of Table 3 report the test results for the full sample of 861 partial correlation coefficients in different model specifications. The first column of Panel A represents the benchmark test estimated by ordinary least squares. But the standard error on the right-hand side of the regression can be endogenous for at least three reasons: i) it is itself an estimate, ii) publication selection can work on the standard error (for example, by choosing an alternative clustering approach that yields smaller standard errors) instead of influencing only the point estimate, and iii) some method choices may affect both point estimates and standard errors (for example, the use of instrumental variables, which is supposed to address endogeneity bias in

the point estimate, but also produces larger standard errors). In response, we use the square root of the number of observations as an instrument for the standard error. The instrument is correlated with the standard error by definition but is not estimated, can rarely be artificially increased by the researcher, and in this literature it is mostly unrelated to the chosen estimation technique. In the third column of Panel A we weight each observation by the inverse of the number of estimates reported per study; this way we give each study the same weight. In the last column, following Stanley (2005), we assign more weight to more precise estimates: we weight estimates by the inverse of standard error $1/SE(PCC_{is})$, which has the advantage of addressing the heteroscedasticity inherent to (3).

Table 3: Tests suggest small publication bias overall

Panel A: Linear techniques					
	OLS	IV	Study	Precision	
Standard error	-0.881***	-0.914***	-1.094**	-0.544*	
(Publication bias)	(0.312)	(0.343)	(0.444)	(0.310)	
	[-1.542, -0.245]	[-1.709, -0.233]	[-2.413, -0.126]	[-1.235, 0.161]	
Constant	0.00597	0.00692	0.0136	-0.00299	
(Effect beyond bias)	(0.0123)	(0.0126)	(0.0176)	(0.00673)	
	[-0.0211, 0.0353]	[-0.0220, 0.0370]	[-0.0272, 0.0533]	[-0.0190, 0.0112]	
Observations	861	861	861	861	
Panel B: Between- and within-study variation					
	BE	FE	RE		
Standard error	-1.959***	0.189	-0.405**		
(Publication bias)	(0.358)	(0.573)	(0.200)		
Constant	0.0159	-0.0225	-0.0344***		
(Effect beyond bias)	(0.0145)	(0.0152)	(0.0102)		
Observations	861	861	861		
Panel C: Nonlinear techniques					
	WAAP	Stem method	Kinked model	Selection model	p-uniform*
Effect beyond bias	0.00756	0.00996	-0.0103***	-0.0130***	-0.0293*
	(0.0130)	(0.0265)	(0.00270)	(0.005)	(0.0178)
Observations	861	861	861	861	861

Notes: The table reports, for linear techniques, the results of regression $PCC_{is} = PCC_0 + \gamma SE(PCC_{is}) + \epsilon_{is}$ estimated for the whole sample of 861 estimates (for which the mean estimate equals -0.017). PCC_{is} denotes the partial correlation coefficient of the i-th estimate from the s-th study and $SE(PCC_{is})$ denotes its standard error. The standard errors of the regression parameters are clustered at the study level and shown in parentheses; 95% confidence intervals obtained using wild bootstrap are shown in brackets. Panel A: OLS = ordinary least squares, IV = the inverse of the square root of the number of observations used as an instrument for the standard error, Study = weighted by the inverse of the number of estimates reported per study, Precision = weighted by the inverse of the estimate's standard error. Panel B: BE = study-level between effects, FE = study-level fixed effects, RE = study-level random effects. Panel C: WAAP (weighted average of adequately powered, Ioannidis *et al.*, 2017), stem method (Furukawa, 2021), kinked model (Bom & Rachinger, 2019), selection model (Andrews & Kasy, 2019), p-uniform* (van Aert & van Assen, 2021). ***, **, and * denote statistical significance at the 1%, 5%, and 10% level.

The results of Panel A in Table 3 suggest mild publication bias in favor of negative estimates. While the bias is statistically significant, it is practically unimportant because the corrected mean is essentially zero, close to -0.017 prior to the correction. A similar finding emerges from Panel B, in which we exploit the panel data nature of our dataset. First, we use exclusively

between-study variation and find evidence for bias that is a bit stronger than what we found in Panel A. In contrast, using within-study variation (running a model with study-level fixed effects) gives no evidence of publication bias. It should be noted that fixed effects are generally problematic in meta-analysis because some studies report only a few estimates while other studies report many of them, often as robustness checks. Identification thus rests on studies reporting many estimates, which makes little sense conceptually. Finally, we also employ study-level random effects, which combine within- and between-study variation. The random effects estimation suggests mild publication bias, similarly to the results reported in Panel A.

In Panel C of Table 3 we perform several nonlinear alternatives to the simple test of publication bias discussed earlier. Stanley *et al.* (2010) document cases in which the linear relation between publication bias and the standard error is violated. For example, estimates concentrated at the top of the funnel plot (the highly precise ones) are less likely to be contaminated by publication bias due to their sufficiently small standard errors and statistical significance at the strictest conventional levels. Put in another way, publication bias is more of an issue when the t-statistic is around 2 than when the t-statistic is 10. To overcome the limitations of the linear technique, we first employ the method designed by Ioannidis *et al.* (2017) and compute the weighted average of adequately powered (WAAP) estimates. The WAAP estimator calculates the underlying effect using only estimates with statistical power above 80% and gives us an estimate of 0.008, which is almost identical to the result of the stem-based method (Furukawa, 2021) in the second column of Panel C. The stem-based method exploits the trade-off between bias and variance: when only the most precise studies are used, publication bias is diminished, but variance increases due to the omitted information. Furukawa (2021) presents an algorithm that finds the optimal balance between bias and variance.

The third method reported in Panel C of Table 3 is the endogenous kink method as proposed by Bom & Rachinger (2019). It assumes that more precise estimates are less likely to suffer from publication bias; therefore, it tries to isolate them and use them to compute the average effect. Similarly to Furukawa's method, the kinked model finds the fraction of the most precise estimates endogenously: it obtains the cut-off value by fitting a piecewise linear meta-regression of estimates on their standard errors. The regression consists of two branches, a horizontal branch for the most precise estimates featuring no relation with their standard errors and a negatively-sloped branch mirroring the correlation between standard errors and estimates contaminated by publication bias. The kink, at which the branches meet, signifies the cut-off value. The model gives us an estimate of -0.01. The fourth nonlinear test is the selection model introduced by Andrews & Kasy (2019). They assume that the chance of publishing an estimate is dependent on its statistical significance and that this chance changes only once a certain level of t-statistic is achieved (for example, 1.96). The method uses maximum likelihood to identify the publication probability for different ranges of estimates bounded by critical t-statistic thresholds. Consequently, it calculates how many estimates in these ranges are underrepresented and assigns them more weight. The selection model given us a result almost identical to the kink model: the effect is statistically significant but economically negligible.

Finally, we use a novel technique recently developed in psychology, p-uniform* (van Aert & van Assen, 2021). The technique does not rely on the exogeneity assumption between effect estimates and standard errors embedded in the previous nonlinear tests but uses the distribution of p-values to identify the true effect. The technique uses the statistical principle that the p-values should be uniform at the true effect size: so, to compute the corrected mean effect, it searches for a number that would be most consistent with a uniform distribution of p-values. The result is -0.0293, more negative than the previous estimates but still far from values that could be considered economically important based on the guidelines of Doucouliagos (2011). On balance, in the entire sample we find little evidence for publication bias in either direction, and the corrected mean values implied by various techniques are close to the uncorrected mean of -0.017.

As a robustness check, we test publication bias in a subsample of estimates not transformed into partial correlation coefficients. We use estimates stemming from specifications that measure student employment as the average amount of hours worked per week during the academic year. Moreover, they use the 4-point grade average as the educational outcome. Admittedly, restricting ourselves to estimates that use grade point average as the response variable disqualifies most research papers conducted outside the US. Nonetheless, including other grading schemes would make the economic effect again incomparable. The results of this robustness check are reported in Table B1 in the Appendix. The resulting non-transformed dataset consists of 86 estimates from 16 studies and is characterized by a simple unweighted mean of -0.007 and a median of -0.004. The results of publication bias tests are similar to the baseline case: while we find some statistically significant evidence of downward publication bias, the bias is not important economically as the corrected mean effect is close to zero. But so far we have ignored another source of bias in the primary literature: endogeneity. There are two general sources of potential endogeneity: first, omitted variable and selection bias, and second, reverse causality.

Concerning the first source of endogeneity, students' labor supply decisions are determined by both observable (e.g., family background, gender, ethnicity, etc.) and mostly unobservable characteristics (e.g., ability, motivation, work ethic, time preference, social and peer networks, etc.) that simultaneously influence students' academic performance (Beffy *et al.*, 2013). These characteristics may systematically differ between students who participate in the labor market and students who do not (Rothstein, 2007). The omitted variable bias occurs when researchers fail to include into (1) relevant personal observable characteristics. The selection bias occurs when researchers do not account for students' unobserved characteristics in their methodological approach. When estimating the model (1) with ordinary least squares (OLS), the estimate of β_1 is inconsistent. The OLS estimate of β_1 can be, nevertheless, biased both positively (upward) or negatively (downward). For example, if ability plays the dominant role, OLS estimates will be biased upwards because more able students will work more and also show better results at school. But if family background is important, students from disadvantaged families may be forced to work in order to sustain their studies, while simultaneously show worse study results compared to students from richer families.

Table 4: Publication bias plagues studies that ignore endogeneity

[Block 1] Studies ignoring endogeneity					
Panel A: Linear techniques					
	OLS	IV	Study	Precision	
Standard error	-1.858***	-1.945***	-2.420***	-0.959**	
(Publication bias)	(0.457)	(0.483)	(0.545)	(0.397)	
	[-3.189, -0.866]	[-3.23, -0.921]	[-3.83, -1.17]	[-2.058, -0.24]	
Constant	0.0405**	0.0425**	0.0591***	0.0171	
(Effect beyond bias)	(0.0172)	(0.0176)	(0.0189)	(0.0112)	
	[0.0001, 0.084]	[0.002, 0.086]	[0.0003, 0.105]	[-0.001, 0.067]	
Observations	436	436	436	436	
Panel B: Between- and within-study variation					
	BE	FE	RE		
Standard error	-2.535***	-0.903	-1.310***		
(Publication bias)	(0.489)	(0.794)	(0.255)		
Constant	0.0331	0.0156	-0.0111		
(Effect beyond bias)	(0.0217)	(0.0207)	(0.0146)		
Observations	436	436	436		
Panel C: Nonlinear techniques					
	WAAP	Stem method	Kinked model	Selection model	p-uniform*
Effect beyond bias	·	-0.00959**	0.00187	0.000	·
	(·)	(0.00432)	(0.00386)	(0.00500)	(·)
Observations	436	436	436	436	436

[Block 2] Studies trying to take endogeneity into account					
Panel A: Linear techniques					
	OLS	IV	Study	Precision	
Standard error	-0.311	-0.311	-0.244	-0.449	
(Publication bias)	(0.347)	(0.392)	(0.431)	(0.387)	
	[-1.174, 0.710]	[-1.306, 0.829]	[-1.556, 0.987]	[-1.3, 0.489]	
Constant	-0.0189**	-0.0185*	-0.0218	-0.0151**	
(Effect beyond bias)	(0.00932)	(0.00984)	(0.0159)	(0.00751)	
	[-0.040, -0.0001]	[-0.039, -0.0002]	[-0.057, 0.011]	[-0.034, 0.002]	
Observations	425	425	425	425	
Panel B: Between- and within-study variation					
	BE	FE	RE		
Standard error	-1.334***	1.041***	0.235		
(Publication bias)	(0.457)	(0.373)	(0.287)		
Constant	0.00317	-0.0556***	-0.0431***		
(Effect beyond bias)	(0.0169)	(0.0101)	(0.0126)		
Observations	425	425	425		
Panel C: Nonlinear techniques					
	WAAP	Stem method	Kinked model	Selection model	p-uniform*
Effect beyond bias	-0.0200***	0.00996	-0.0138***	-0.0260***	-0.0319***
	(0.00687)	(0.0294)	(0.00369)	(0.008)	(0.0106)
Observations	425	425	425	425	425

Notes: The table reports, for the linear techniques, the results of regression $PCC_{is} = PCC_0 + \gamma SE(PCC_{is}) + \epsilon_{is}$ estimated for the sample of 436 estimates where the endogeneity of students' decision to work is not controlled for [Block 1] and for the sample of 425 estimates where this endogeneity is controlled for [Block 2]. PCC_{is} denotes the partial correlation coefficient of the i-th estimate from the s-th study, and $SE(PCC_{is})$ denotes its standard error. The standard errors of the regression parameters are clustered at the study level and shown in parentheses; 95% confidence intervals from wild bootstrap clustering are shown in brackets. Panel A: OLS = ordinary least squares, IV = the inverse of the square root of the number of observations used as an instrument for the standard error, Study = weighted by the inverse of the number of estimates reported per study, Precision = weighted by the inverse of the estimate's standard error. Panel B: BE = study-level between effects, FE = study-level fixed effects, RE = study-level random effects. Panel C: WAAP (weighted average of adequately powered, Ioannidis *et al.*, 2017), stem method (Furukawa, 2021), kinked model (Bom & Rachinger, 2019), selection model (Andrews & Kasy, 2019), p-uniform* (van Aert & van Assen, 2021). In Block 1, WAAP and p-uniform* do not converge. ***, **, and * denote statistical significance at the 1%, 5%, and 10% level.

Regarding the second source of endogeneity, both student employment and educational outcomes can be jointly determined (DeSimone, 2006). This occurs when the estimated effect of student work on academic performance partly reflects a causal impact of academic performance on student work. The bias usually permeates cross-sectional studies where researchers do not distinguish between time periods at which the student employment and academic achievement are measured. A number of estimates in our sample is produced without accounting for these sources of endogeneity. They include the results from early studies (utilizing OLS and other elementary estimation methods) that treat student employment as exogenous (Ruhm, 1997) but they also include robustness checks where researchers intentionally show what happens when endogeneity is not accounted for. In any case, if a certain estimation method fails to appropriately control for the pre-existing heterogeneity between students, we cannot conclude that the estimated effect of student employment on educational outcomes is directly attributable to students' employment.

In Table 4 we run tests of publication bias separately for estimates that ignore and try to account for endogeneity, respectively. We say that the study tries to account for endogeneity if it employs a quasi-experimental technique such as matching, instrumental variables, or difference-in-differences, or if it uses a proxy for ability (such as IQ). We identify 425 estimates from 50 studies that conform to this definition; in contrast, 436 estimates from 36 studies ignore endogeneity. The mean partial correlation coefficients, prior to correction for potential publication bias, are similar: -0.027 for estimates controlling for endogeneity, -0.008 for estimates ignoring endogeneity. But Table 4 shows that correction for publication bias paints a different story. If endogeneity is taken into account, little publication bias follows, and the corrected mean partial correlation coefficient is, according to most techniques, close to the uncorrected mean. But for estimates that ignore endogeneity we find strong publication bias and a positive corrected mean according to most specifications. (For two techniques, WAAP and p-uniform*, no results are reported: the former does not identify any study that would have sufficient power, and the latter does not converge.) That is, the results of primary studies differ fundamentally depending on whether endogeneity is taken into account: if not, the results tend to be positive. Because positive estimates are less intuitive, some researchers try different specifications until they obtain a negative coefficient. Publication bias towards negative estimates follows, and the mean reported estimate is negative even if researchers ignore endogeneity.

4 Heterogeneity

In this section we examine why the estimates reported in the literature differ so much. In doing so, we also test the robustness of our results concerning publication bias, because heterogeneity may interact with the bias; indeed, heterogeneity can make the funnel plot asymmetrical even if the literature is free of selective reporting. In the previous section we use several advanced techniques that in important aspects go beyond the funnel asymmetry test, but their results are broadly consistent with the more straightforward linear approach. That is important, because the more complex tests of publication bias cannot be used in the Bayesian model averaging

framework, which is the workhorse of the present section due to substantial model uncertainty, while the linear test can be easily incorporated into Bayesian model averaging. The section is inspired by Tyler (2003, p. 386), who notes in his survey of the earlier literature on the nexus between student employment and education: "*Taken as a whole these studies do not offer consistent lessons about the relationship between school-year work and academic achievement. The reasons for the inconsistencies are likely related to some combination of different data sets, different age students, different dependent variables, and different empirical methods across the studies.*" We collect 32 variables that reflect such differences. We describe these variables, estimate their effect on the results reported in individual studies, and, as the most important outcome of our analysis, in subsection 4.4 we present partial correlation coefficients implied for different contexts by best practice methodology. Our previous results regarding publication and endogeneity biases continue to hold.

4.1 Variables

For ease of exposition we group the variables into four blocks: data characteristics, structural variation, estimation methods, and publication characteristics. Table 5 introduces the definitions of the variables, their mean, standard deviation, and mean weighted by the inverse of the number of estimates reported per study. The correlations between individual variables are not excessive, as shown by Figure B1 in the Appendix; also, all variance-inflation factors are below 10. But, as will be discussed later, we still use the dilution prior for Bayesian model averaging that takes potential collinearity into account.

Table 5: Description and summary statistics of regression variables

Variable	Description	Mean	SD	WM
PCC	The partial correlation coefficient corresponding to the estimated effect of student employment on educational outcomes.	-0.017	0.066	-0.051
Standard error (SE)	The standard error of the PCC.	0.027	0.017	0.034
SE * No endogeneity control	An interaction between the standard error and ignoring endogeneity (proxy for publication bias in studies that ignore endogeneity).	0.013	0.016	0.015
Data characteristics				
Employment: continuous variable	= 1 if student employment is measured by a continuous variable.	0.303	0.460	0.492
Employment: dummy variable	= 1 if student employment is measured by a dummy variable.	0.184	0.387	0.158
Employment: categorical variable	= 1 if student employment is measured by a categorical variable (reference category).	0.513	0.500	0.350
Educational outcome: choices	= 1 if educational outcome is specified as educational decision (e.g. continue next year, enroll at a college).	0.171	0.376	0.168
Educational outcome: attainment	= 1 if educational outcome is specified as educational attainment (e.g. probability of graduation).	0.135	0.342	0.176
Educational outcome: test scores	= 1 if educational outcome is specified as test and exam results (reference category).	0.695	0.461	0.656

Continued on next page

Table 5: Description and summary statistics of regression variables (continued)

Variable	Description	Mean	SD	WM
Self-reported education	= 1 if educational outcome (dependent variable) is self-reported.	0.260	0.439	0.449
Longitudinal data	= 1 if longitudinal data are used to estimate the effect.	0.847	0.360	0.580
Cross-sectional data	= 1 if cross-sectional survey data are used to estimate the effect (reference category).	0.153	0.360	0.420
Data year	The logarithm of the mean year of the data used minus the earliest average year in our data (base = 1967).	3.283	0.516	3.336
Structural variation				
Male students	= 1 if the effect is estimated for male students only.	0.253	0.435	0.106
Female students	= 1 if the effect is estimated for female students only.	0.258	0.438	0.083
Mixed-gender students	= 1 if the effect is estimated for students of all genders (reference category).	0.489	0.500	0.810
Caucasian students	= 1 if the effect is estimated for white students only.	0.038	0.192	0.045
Minority students	= 1 if the effect is estimated for minority students only.	0.053	0.225	0.027
Part-time students	= 1 if the effect is estimated for part-time students only.	0.038	0.192	0.009
Secondary education	= 1 if the effect is estimated for students involved in secondary education.	0.721	0.449	0.464
Tertiary education	= 1 if the effect is estimated for students involved in tertiary education.	0.279	0.449	0.536
Low-intensity employment	= 1 if the effect is estimated for low-intensity workers (fewer than 15 hours per week).	0.215	0.411	0.129
Medium-intensity employment	= 1 if the effect is estimated for medium-intensity workers (15–30 hours per week).	0.109	0.312	0.072
High-intensity employment	= 1 if the effect is estimated for high-intensity workers (more than 30 hours per week).	0.189	0.392	0.142
On-campus employment	= 1 if the effect is estimated for jobs situated on the school premises.	0.020	0.139	0.043
United States	= 1 if the country of analysis is the US.	0.806	0.396	0.638
Germany	= 1 if the country of analysis is Germany.	0.034	0.181	0.014
Other countries	= 1 if the country of analysis is not the US or Germany.	0.049	0.216	0.116
Estimation methods				
OLS method	= 1 if elementary approaches (OLS, logit regression, etc.) are used for estimation.	0.610	0.488	0.674
Matching method	= 1 if the propensity score matching approach is used for estimation.	0.034	0.181	0.029
DID method	= 1 if the difference-in-differences approach is used for estimation.	0.051	0.220	0.011
IV method	= 1 if the instrumental variable approach or simultaneous equation modeling is used for estimation.	0.160	0.367	0.179
Other methods	= 1 if panel methods such as fixed-effects or random-effects are used for estimation (reference category).	0.145	0.352	0.107

Continued on next page

Table 5: Description and summary statistics of regression variables (continued)

Variable	Description	Mean	SD	WM
Endogeneity control	= 1 if the estimation accounts for potential endogeneity (IV approach, difference-in-differences, matching, simultaneous equations, and ability control). The variable is not included in BMA because of collinearity (but its inversion 'No endogeneity control' interacted with the standard error is included).	0.494	0.500	0.618
Number of variables	The logarithm of the number of explanatory variables used in the model in the primary study.	2.568	0.856	2.461
Ability control	= 1 if estimation accounts for students' ability, e.g. SAT scores, prior education, class rank, etc.	0.366	0.482	0.555
Motivation control	= 1 if estimation controls for students' academic motivation.	0.338	0.473	0.237
Parental education control	= 1 if estimation includes variable(s) reflecting parents' educational level.	0.545	0.498	0.445
Age control	= 1 if estimation controls for students' age.	0.462	0.499	0.419
Ethnicity control	= 1 if estimation includes control variables reflecting students' ethnicity.	0.596	0.491	0.453
Publication characteristics				
Impact factor	The Journal Citation Reports impact factor of the journal in which the primary study was published (collected in August 2021).	1.583	1.237	1.573
Citations	The logarithm of the mean number of Google Scholar citations received per year since the study was published (collected in August 2021).	1.695	1.006	1.552
Published study	= 1 if the study was published in a peer-reviewed journal.	0.912	0.284	0.826

Notes: Collected from primary studies. SD = standard deviation, WM = mean weighted by the inverse of the number of estimates reported per study.

Data characteristics. Researchers use various specifications to capture the student employment status. As discussed earlier, most of them utilize student employment as a *continuous* variable, while others create a *categorical* or a *dummy* variable. We identify different continuous measures of student employment intensity in the existing studies. For instance, Carr *et al.* (1996) use total hours worked during a semester to estimate the effect, while D'Amico (1984) relies on the percentage of the school year's weeks with work hours being either above or below 20 hours. Nonetheless, researchers usually measure the intensity of student employment as average hours worked during the interview week (Ruhm, 1997), a typical non-summer week (Sabia, 2009), midterm week (Kalenkoski & Pabilonia, 2010), or during two reference weeks in the academic year (Darolia, 2014). Nevertheless, as explained by Oettinger (1999), imputing the typical or survey week's hours worked to the entire school year might contribute to a significant measurement bias. To correct for the bias, Oettinger (1999) suggests combining the amount of weeks worked during the year and the average weekly hours worked in the resulting student employment measure. In contrast to Oettinger (1999), Ruhm (1997) argues that work hours reported for the week preceding the survey might better reflect the reality than work hours reported for periods preceding the survey by several months, given the time proximity.

Similarly to the continuous variable specification, the categorical specification of student employment intensity can also take various forms (see, for example Gleason, 1993; Torres *et al.*, 2010; Staff *et al.*, 2010). Hovdhaugen (2015), for instance, divides his sample into three bands: 1–19 hours per week, 20–30 hours per week, and more than 30 hours per week. Alternatively, Torres *et al.* (2010) use five work intensity categories and Tyler (2003) uses ten categories, each representing a 5-hour increment. Researchers defining employment as a dummy variable simply distinguish between working and non-working students (see, for example McKenzie & Schweitzer, 2001; McNeal, 1997).

Next, researchers examine the effect of student employment on various educational outcomes including educational *choices*, *test scores*, and *attainment*. Neyt *et al.* (2019) distinguish four classes of educational outcomes: habits, decisions, tests scores, and attainment. Educational engagement refers to students' habits associated with their class preparation and discipline they display in school-related activities. This category comprises measures such as class attendance/absence (Schoenhals *et al.*, 1998), time spent doing homework or devoted to independent study (Marsh & Kleitman, 2005), truancy (Staff *et al.*, 2010), or paying attention during class (Sabia, 2009). Study decisions refer to the choices of dropping out from a course or study program (Warren & Cataldi, 2006), or continuing to higher education (Steel, 1991). The class of test results is the most frequently used in the literature and employs the grade point average (DeSimone, 2008; Gleason, 1993; Sabia, 2009), specific course grades (Kouliavtsev, 2013), test scores (Tyler, 2003), or results of high school final exams (Dustmann & van Soest, 2007). The last category, educational attainment, comprises of students' probable and actual achievements, e.g. the probability of graduation from high school (Beffy *et al.*, 2013) or credits earned during a specific time period (Dadgar, 2012). Applegate & Daly (2006) show that *self-reported* measures are subject to measurement error as they are quite often over- or under-estimated by individuals. Also because of the potential measurement error, we eliminate habit estimates from our sample and try to control for the remaining self-reported educational outcomes by including a corresponding dummy variable.

Another difference in data characteristics involves the dimension of the data: 85% of the estimates exploit *longitudinal datasets* (Apel *et al.*, 2008; Lee & Orazem, 2010; Kalenkoski & Pabilonia, 2010, among others). Longitudinal data allow researchers to control for the time-invariant individual unobserved heterogeneity, and thus help to account for the endogeneity of student employment (Oettinger, 1999). Cross-sectional studies cannot distinguish between time periods at which the student employment and academic achievement are measured. Longitudinal data overcome the issue of mismatched time periods for used variables: student employment observed at time t is used as a regressor for educational outcomes measured at time $t + x$, where x represents several months or years (Warren *et al.*, 2000). Considering the indisputable time order between measuring student employment and educational outcomes, longitudinal data often allow researchers to draw causal inferences between these two variables. Last, we create the variable *data year*, denoting the average year of the data used in the study. We assume that estimates capturing the effect of student employment on educational outcomes can differ across

generations due to varying work and study habits. For instance, Babcock & Marks (2011) show that university students substantially decreased time devoted to study between 1961 and 2003.

Structural variation. The primary studies often differ structurally, reporting estimates for different groups of students in terms of gender, ethnicity, education level, employment intensity, and different countries. We code for many of such differences but eventually use only those that have a sufficient portion of observations in our dataset (that make up at least 3% of the estimates). For example, we explore whether students' gender can drive heterogeneity in the effect of student employment on educational outcomes. Prior research provides some motivation. For instance, Montmarquette *et al.* (2007) find a negative association between student employment and educational outcomes for *males* only. Likewise, Sabia (2009) and Holford (2020) report less negative estimates for *female* students. About a quarter of our data is estimated for male and a quarter for female students separately. About 4% of our data involves estimates for Caucasian students and about 5% of our data involves estimates for minority groups. While Steel (1991) finds the effect for non-Hispanic white students less negative than for the rest of the population, the latter studies, such as Sabia (2009), report rather mixed results.

Part-time students differ from full-time students in their prioritization of work instead of education (DeSimone, 2008): they do not focus on academic pursuits as much as professional ones. Chen & Carroll (2007) report part-time students to be older, married, and more independent, but these students also often hail from challenged backgrounds and have lower rates of persistence. Darolia (2014) claims that there are substantial differences in the effect of employment between part-time and full-time students: while he does find some negative effect in the full-time student sample, he finds none in the part-time sample. The part-time students are exclusive to tertiary education, though, and we also control for the educational level of students with a separate explanatory variable. The existing literature presents opposing views on how the effect of student employment on educational outcomes differs between *secondary education* and higher education students. Bozick (2007) argues that university students compared to high school students enjoy a more flexible study environment (less in-person attendance and a richer choice of classes) and more favorable attitude to education. Thus, one would expect the effect to be less negative for university students. On the other hand, our descriptive statistics presented in Table 2 show a more negative effect for tertiary students instead. Neyt *et al.* (2019) also report more negative effects for university students and explain that tertiary education students might be less successful in combining work and study due to the more challenging content and less structured setting of their studies.

Depending on the intensity of student employment, some studies show that work may be simultaneously a complement and a substitute to academic performance (Choi, 2018). This intensity-dependent perspective (discussed under the threshold model earlier) holds that work has positive consequences on study engagement only up to a certain threshold of hours worked. After exceeding this threshold the effect of student employment on educational outcomes reverses as working hours begin to interfere with academic pursuits (Buscha *et al.*, 2012). The

literature does not agree on the actual threshold at which the effect reverses (Marsh & Kleitman, 2005). While Montmarquette *et al.* (2007) report an inflection point of 15 hours worked per week, Tessema *et al.* (2014) find the threshold at 10 hours worked per week. Whenever possible, we code for different workload intensity: variable *low-intensity employment* applies to estimates capturing the effect for students working up to 15 hours per week and variable *high-intensity employment* captures the weekly intensity above 30 hours per week. Finally, we account for the geographical variation among the primary studies. We have shown some patterns of this variation in Figure 4 panel (f), where Germany stands out. While most primary studies utilize datasets obtained in the *United States* (81% of the collected estimates), we also code for *Germany* separately. The remaining countries include parts of Europe, Australia, Canada, and Russia.

Estimation methods. We codify five dummy variables that reflect estimation methods: *OLS method* which encompasses not only simple ordinary least squares but also other elementary techniques such as linear probability models, *Matching method* representing the propensity score matching approach, *DID method* that stands for the difference-in-differences approach, *IV method* that includes not only instrumental variable approaches but also the simultaneous modeling approach. Considering the varying underlying assumptions of these techniques and the degree to which these estimation methods account for students' unobservable differences, we expect estimation approaches to affect the reported estimates. Indeed, using the same dataset, Stinebrickner & Stinebrickner (2003) employ OLS, fixed-effects, and IV approach to estimate the relationship between student work and academic performance and obtain three fundamentally different estimates.

Ordinary least squares are employed in recent studies mostly as a robustness check because they fail to account for endogeneity. Some studies address endogeneity using the propensity score matching (3% of the dataset) that accounts for observable heterogeneity between working and non-working students (Choi, 2018). The propensity score matching technique pairs working and non-working students based on their similarity in various observable socio-psychological and demographic characteristics composing together the propensity score (Lee & Staff, 2007). Consequently, the effect of student employment on educational outcomes is compared between the matched students. Difference-in-differences (*DID method*) tries to mimic experimental research design while using observational data (Buscha *et al.*, 2012). Combined with the matching model, it can address selection on both observables and unobservables associated with work decisions without the need for instrumental variable and thus serve as a useful tool to obtain the causal effect.

Another approach to obtaining a consistent estimate is the instrumental variable procedure. Many researchers taking advantage of the instrumental variable approach rely on the availability of local labor market conditions, e.g. youth unemployment rate, as the instrumental variable (see Rothstein, 2007; Beffy *et al.*, 2013; Holford, 2020; Lee & Orazem, 2010). Other studies use child labor laws (Tyler, 2003; Apel *et al.*, 2008), the proportion of unearned income (DeSimone, 2006), paternal schooling (DeSimone, 2008), socio-economic status of the family (Simon *et al.*,

2017), amount of financial aid students obtain (Sprietsma, 2015), or the variation in area house prices (Darolia, 2014) as their instrumental variables. Related to the instrumental variable estimation, some researchers rely on the simultaneous equation modeling (Parent, 2006). Similarly to the instrumental variable approach the simultaneous equations model the effect of student work on educational outcomes by estimating a system of linear equations. Nevertheless, instead of relying on the two-stage-least-squares estimator, the model is usually estimated via maximum likelihood estimator (Kalenkoski & Pabilonia, 2010). Another method addressing the endogeneity bias is the dynamic discrete approach explicitly modeling students' decision-making process to work (Eckstein & Wolpin, 1999; Montmarquette *et al.*, 2007). Given the small number of observations using this method (6), we incorporate the technique in the *IV method* dummy. The dynamic discrete approach estimates the likelihood function of participating in the labor market exploiting the finite number of discrete types of students who differ in unobservable characteristics (Eckstein & Wolpin, 1999).

The remaining set of techniques include panel methods. One solution allowing researchers to control for unobserved differences between working and non-working students entails the addition of individual unobserved fixed-effects into (1). By subtracting the individual-specific means from the variable values at each time period, the fixed-effects model allows researchers to control for the time-invariant student-level unobserved characteristics (Darolia, 2014). However, as noted by Apel *et al.* (2008), the fixed-effects model yields unbiased and consistent estimates only under the assumption that unobserved student characteristics determining student work habits and academic performance are constant over time. As explained by Oettinger (1999), this assumption is questionable as students' motivation is likely to fluctuate over time. Typically, students pursuing enrollment at tertiary education institutions increase their academic effort before their high school leaving exams in order to enhance their chances of being accepted to their top-choice universities.

An important aspect of estimation is the potential control for individual characteristics. One such characteristic is students' intrinsic *motivation*. Empirically, Richardson *et al.* (2013) demonstrate that employment is less likely to hamper academic performance if students work because they want to than because they have to. Another important factor researchers control for (if possible) is students' cognitive *ability* (Arano & Parker, 2008; McNeal, 1997; Staff & Mortimer, 2007). We consider this variable to be the strongest form of endogeneity control among the covariates commonly employed by researchers. For example, Oettinger (1999) finds that more able students systematically select different employment schedules than less able students. But students' educational outcomes could be influenced by the economic situation of their parents, and we include a dummy reflecting control for parental education. Carneiro & Heckman (2003) suggest that student educational choices are better explained by family permanent features, such as parents' education levels which directly contribute to family permanent income. Apart from that, students growing up in families with higher education levels are likely to perform better academically as education is more valued in such families (Arano & Parker, 2008). In addition to *parental education*, we include dummy variables for studies controlling for

standard demographic characteristics such as students' *ethnicity* and *age*. Empirically, these factors have been shown to have a substantial impact on the link between student work and academic performance. For instance, Oettinger (1999) finds a negative effect of student employment on their GPA only for students from ethnic minorities. Kohen *et al.* (1978) argue that the negative association is less pronounced for older students who tend to be more mature and committed to their educational and occupational goals.

Publication characteristics. Even though we attempt to control for many aspects of study design that we hope capture the quality of a study, some aspects of quality are hard to codify. Therefore we also include publication characteristics that may reflect quality aspects not reflected by the variables described above. We include a dummy variable indicating whether the study was published in a peer-reviewed journal. Although the quality of peer review differs across journals, peer review is a basic indicator of the reliability of the results (especially once corrected for potential publication bias, which might stem not only from the preferences of the authors, but also editors and referees). To partially account for differences in peer review across journals, we control for the Journal Citation Reports impact factor of the journal, and assume that journals with higher impact factors tend to have stricter peer-review procedures. Finally, we control for the number of per-year citations the study has received. We again assume that, after controlling for publication bias, the number of citations is positively correlated with the quality of the analysis.

4.2 Estimation

Our intention is to find out which variables help explain the heterogeneity in the estimates reported in the literature. One solution is to include all variables into one regression, but the problem is that we do not know ex ante which of the 32 explanatory variables belong into the underlying model. We believe all of them might be important in explaining the heterogeneity, but in practice most likely only a few will prove to be, and including all into one regression would substantially decrease the precision of the entire estimation, complicating inference even for the most important variables. Thus we face substantial model uncertainty, the natural response to which emerges in the Bayesian setting: Bayesian model averaging (BMA, Steel, 2020).

BMA addresses model uncertainty by considering all possible models with different choices of covariates (Raftery, 1995). In essence, BMA estimates a large amount of regressions using different subsets of explanatory variables. Consequently, it constructs a weighted average of all the possible combinations of explanatory variables (Zeugner & Feldkircher, 2009) using posterior model probabilities as weights. Posterior model probabilities arise from Bayes theorem: they are proportional to the product of the integrated likelihood of the model capturing the probability of utilized data considering the model and the prior model probability. This product is then divided by the sum of integrated likelihoods of regression models. While the posterior model probability indicates the goodness-of-fit of the model, the prior model probability refers to researchers' prior beliefs regarding the probability of a model before considering the data (Zeugner, 2011).

Consequently, BMA uses the computed posterior model probabilities to calculate the weighted posterior mean and the weighted posterior variance (or weighted posterior standard deviation) for each included explanatory variable. These two statistics can be compared to the estimate of a regression coefficient and the standard error of the estimated regression parameter in the frequentist setting. The posterior inclusion probability (PIP) of a variable is defined as the sum of the posterior model probabilities (PMP) of models which include this variable. We interpret PIP as the probability that a given variable is a useful predictor of the dependent variable.

When applying BMA we face two computational problems. First, computing integrals included in the integrated likelihood function is demanding (Hoeting *et al.*, 1999). Second, the enormous model space makes the estimation infeasible for a standard personal computer. For instance, with 32 explanatory variables there are be 2^{32} possible regressions, representing a serious computational challenge. One way to overcome this computational obstacle is to apply the Markov chain Monte Carlo method using the Metropolis-Hastings algorithm. Markov chain Monte Carlo diminishes the computational demands of BMA by estimating only models with the highest PMP. As Zeugner (2011) shows, the Metropolis-Hastings algorithm determines these models by comparing a benchmark model with a competing model in terms of their posterior model probabilities. If one model is accepted in favor of the other, a new competing model is selected and compared. If the opposite occurs and the other model is accepted, it becomes a new benchmark model and the procedure is repeated.

Before we proceed with the application of BMA, we specify prior distributions on regression parameters and model probabilities. Given that the amount of prior information on the parameter space available to us is small, we follow Eicher *et al.* (2011) and opt for the unit information prior (UIP). UIP provides approximately the same amount of information as one observation in the dataset. Regarding our prior choice on model space, we do not follow the traditional approach of using the uniform model prior assigning the same probability to each model, irrespective of the number of included control variables. Instead, we follow George (2010) and employ the collinearity adjusted dilution model prior. Unlike to uniform model prior, the dilution model prior relaxes the assumption of zero correlation between explanatory variables. When applying the dilution model prior, the posterior probabilities of models including highly correlated covariates are adequately down-weighted to account for this collinearity (Hasan *et al.*, 2018); because of the large number of variables, the use of this prior is important in meta-analysis even though in our case all variance-inflation factors are below 10. Given the choices described above, BMA estimated with the unit information prior and dilution model prior represents our baseline model. We provide a robustness check following Fernandez *et al.* (2001) and choose the BRIC prior instead of UIP; for the model size we use the beta-binomial random prior advocated by Ley & Steel (2009).

In practice each regression run by BMA has the following form:

$$PCC_{is} = \gamma_0 + \gamma_1 SE(PCC)_{is} * \textit{No endogeneity control}_{is} + \gamma_2 X_i + \epsilon_{is}, \tag{4}$$

where PCC_{is} represents the estimated partial correlation coefficient, X_{is} stands for the ex-

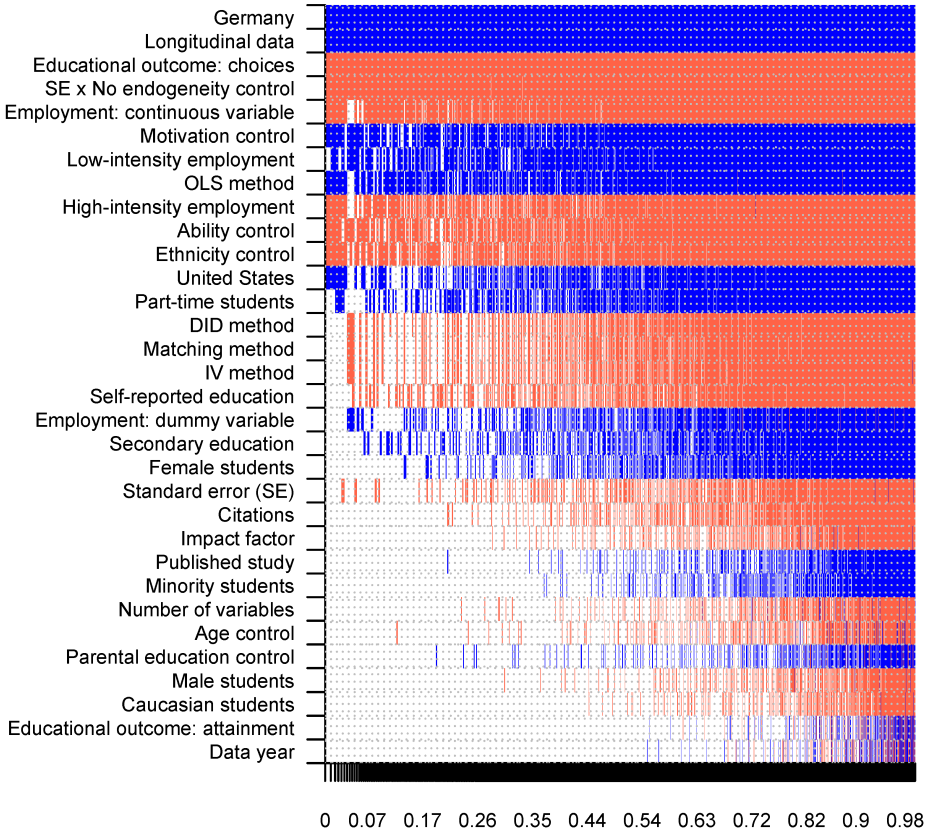
planatory variables including the standard error, γ_1 measures the direction and magnitude of publication bias in the sample of estimates disregarding endogeneity, and ϵ_{is} denotes the error term. The constant γ_0 has no interpretation per se as it reflects the mean effect corrected for publication bias conditional on the covariates. On top of the baseline model we employ frequentist model averaging (FMA). Similarly to BMA, FMA accounts for model uncertainty. Nevertheless, in contrast to BMA, FMA is entirely data-dependent and does not require prior specification (Wang *et al.*, 2009). To implement FMA, we adopt the approach suggested by Hansen (2007). Following his approach we estimate the model averaging estimator that determines the weights by minimizing the Mallows criterion (Amini & Parmeter, 2012). The smaller the Mallows criterion, the smaller the model variance and the better the goodness of fit of the model. The application builds upon Magnus *et al.* (2010) and reduces the model space from 2^{32} to the number of explanatory variables equal to 32, taking advantage of the orthogonalization of the covariate space (Amini & Parmeter, 2012). Steel (2020) provides a detailed overview of frequentist and Bayesian model averaging techniques used in economics.

4.3 Results

The results of our BMA exercise are visualized in Figure 6. The vertical axis lists the explanatory variables in descending order from top to bottom according to their posterior inclusion probability. Hence the most important predictors lie on the top of the plot. The horizontal axis depicts the individual models; the width of each column corresponds to the posterior model probability, so the best models are on the left. White color signifies the exclusion of the particular variable from the model, red color (lighter in grayscale) indicates a negative coefficient for the particular variable, and blue color (darker in grayscale) indicates a positive coefficient. We identify ten variables with PIP above 0.5: *publication bias* interacted with endogeneity bias, *employment: continuous variable*, *educational outcome: choices*, *longitudinal data*, *high-intensity* and *low-intensity employment*, *Germany*, and control variables of *ability*, *motivation*, and *ethnicity*. When interpreting the magnitude of PIP, researchers usually follow Jeffreys (1961). Jeffreys (1961) distinguishes between weak, positive, strong, and decisive effect if the value of the corresponding PIP falls into the interval of 0.5-0.75, 0.75.-0.95, 0.95-0.99, and 0.99-1, respectively, and we follow the convention.

We accompany the graphical output of BMA with quantitative results reported in the left-hand part of Table 6. (In addition, Figure B4 in the Appendix shows posterior coefficient distributions for selected variables.) The numerical results corroborate the conclusions drawn from the plot. In light of Jeffreys's 1961 categorization, variables representing publication bias, education operationalized as a choice, longitudinal data, and German datasets have a decisive effect on the estimated partial correlation coefficient. Furthermore, the results indicate a positive effect for employment defined as a continuous variable; weak effects are identified for the variables representing high-intensity and low-intensity employment, use of OLS, and control variables for motivation, ability, and ethnicity. The interpretation of posterior means from Table 6 correspond to the marginal effects of the characteristic on the calculated PCC.

Figure 6: Model inclusion in Bayesian model averaging



Notes: The figure depicts the results of the benchmark BMA model reported in Table 6. We employ the unit information g-prior (the prior has the same weight as one observation of data) recommended by Eicher *et al.* (2011) and the dilution prior suggested by George (2010), which accounts for collinearity. The explanatory variables are ranked according to their posterior inclusion probabilities from the highest at the top to the lowest at the bottom. The horizontal axis shows the values of cumulative posterior model probability. Blue color (darker in grayscale) = the estimated parameter of the corresponding explanatory variable is positive. Red color (lighter in grayscale) = the estimated parameter of the corresponding explanatory variable is negative. No color = the corresponding explanatory variable is not included in the model. Numerical results are reported in Table 6. All variables are described in Table 5.

For example, the decision to define the educational outcome as a choice means, ceteris paribus, that the calculated PCCs is on average smaller by -0.029 compared to the educational outcome defined as a test score (the category omitted from the regression). The OLS robustness check on the right side of Table 6 and further robustness checks in Table B2 corroborate the findings of baseline BMA. Before we turn to the discussion of the results for the variables in individual categories, it is worth mentioning that our results concerning publication bias hold: studies ignoring endogeneity suffer from publication bias, while studies taking endogeneity into account are free of the bias.

Data characteristics. The results of BMA suggest that defining educational outcomes as *Educational Choice* typically generates more negative PCCs. This finding resonates with Neyt *et al.* (2019), who report that studies operationalizing educational outcome as decisions to

Table 6: Why estimates vary

Response variable: partial correlation coefficient	Bayesian model averaging (baseline model)			OLS (robustness check)		
	P. mean	P. SD	PIP	Coef.	SE	p-value
Intercept	-0.041	NA	1.000	-0.040	0.017	0.019
Standard error (SE)	-0.069	0.171	0.167			
SE * No endogeneity control	-0.874	0.305	0.991	-1.009	0.280	0.000
Data characteristics						
Employment: continuous variable	-0.026	0.014	0.847	-0.027	0.010	0.007
Employment: dummy variable	0.006	0.010	0.275			
Educational outcome: choices	-0.029	0.007	0.996	-0.032	0.011	0.004
Educational outcome: attainment	0.000	0.001	0.009			
Self-reported education	-0.005	0.008	0.321			
Longitudinal data	0.044	0.009	1.000	0.053	0.014	0.000
Data year	0.000	0.000	0.009			
Structural variation						
Male students	0.000	0.001	0.031			
Female students	0.002	0.005	0.190			
Caucasian students	0.000	0.003	0.025			
Minority students	0.001	0.004	0.048			
Part-time students	0.014	0.017	0.464			
Secondary education	0.005	0.009	0.251			
Low-intensity employment	0.016	0.013	0.696	0.016	0.009	0.076
High-intensity employment	-0.015	0.012	0.666	-0.018	0.008	0.018
United States	0.010	0.011	0.482			
Germany	0.067	0.015	1.000	0.056	0.011	0.000
Estimation methods						
OLS method	0.015	0.012	0.676	0.026	0.007	0.000
Matching method	-0.013	0.020	0.341			
DID method	-0.014	0.020	0.369			
IV method	-0.009	0.014	0.338			
Number of variables	0.000	0.001	0.047			
Ability control	-0.014	0.012	0.653	-0.023	0.010	0.021
Motivation control	0.012	0.009	0.722	0.016	0.008	0.055
Parental education control	0.000	0.002	0.041			
Age control	0.000	0.002	0.046			
Ethnicity control	-0.011	0.009	0.639	-0.018	0.008	0.020
Publication characteristics						
Impact factor	0.000	0.001	0.064			
Citations	-0.001	0.002	0.124			
Published study	0.001	0.005	0.050			
Studies	69			69		
Observations	861			861		

Notes: The response variable is the estimate of the effect of student employment on educational outcomes (recomputed to the partial correlation coefficient). SE = standard error, P. mean = posterior mean, P. SD = posterior standard deviation, PIP = posterior inclusion probability. In the left-hand part of the table we employ Bayesian model averaging (BMA) using the unit information g-prior recommended by Eicher *et al.* (2011) and the dilution prior suggested by George (2010). The specification in the right-hand part of the table employs ordinary least squares (OLS) using variables with at least 50% PIP in BMA. The posterior mean in Bayesian model averaging (or alternatively the estimated coefficient in the frequentist model) denotes the marginal effect of a study characteristic on the partial correlation coefficient corresponding to the effect reported in the literature. For a detailed description of all the variables see Table 5.

dropout deliver a consistently more negative relationship compared to studies using other educational outcomes. Intuitively, one can explain the negative relationship via the mechanism of the zero-sum theory. Crowding out of study time translates into poor test performance and exam failures, resulting progressively in a situation in which students prefer to dropout from a certain course or study program (Parent, 2006). Hence, our finding provides support for the notion that the effect of student employment "*grows in cumulative importance*" (Warren *et al.*, 2000, p. 949) and has long-term effects on educational outcomes. Nevertheless, this explanation overlooks students' diverse backgrounds and expectations, mediating the relationship. Eckstein & Wolpin (1999) develop a structural model of high school attendance and show that although student employment increases the probability of dropout, the effect is driven by students' specific characteristics such as their ability, motivation, and preferences concerning leisure.

Another factor negatively influencing the estimated PCCs is whether student *employment* is specified as a *continuous* variable. The finding shows that what primarily matters for the effect of student employment on educational outcomes is the intensity of students' work schedule. The result is consistent with the zero-sum perspective and conclusion cited in multiple studies: working long hours while studying has a detrimental impact on educational outcomes (D'Amico, 1984; Montmarquette *et al.*, 2007; Buscha *et al.*, 2012; Lee & Staff, 2007). For instance, Montmarquette *et al.* (2007, p. 759) show that "*working less than fifteen hours per week is not necessarily detrimental to success in school.*" Beffy *et al.* (2013) confirm this inflection point and show that spending at work more than 16 hours per week has strong negative effect on the graduation probability, whereas working less than 16 hours has a much weaker effect.

Taking the advantage of longitudinal datasets seems to have a substantial impact on explaining the differences in the estimated PCCs. *Longitudinal data* systematically generate more positive estimates of students' employment-education relationship compared to cross-sectional studies. This is in line with prior research demonstrating that studies based on longitudinal data yield less negative (Rothstein 2007; Oettinger 1999) or more positive (Stinebrickner & Stinebrickner, 2003) estimates. The advantages of longitudinal studies over cross-sectional studies are twofold. First, longitudinal data tackle better the endogeneity of the decision to work (Neyt *et al.* 2019; Rothstein 2007). Due to the available time span, longitudinal data mitigate the self-selection bias by differencing out unobserved individual heterogeneity (Oettinger, 1999). Second, longitudinal data overcome the difficulties of drawing causal inferences as work habits are measured before educational outcomes (Moulin *et al.*, 2013). As a result, cross-sectional studies failing to control for time-invariant individual characteristics generate downward-biased estimates. In contrast to the predictive importance of *longitudinal data*, our results indicate that *data year* of the original dataset has no impact on the heterogeneity of PCCs, showing no structural differences among student populations over the years. This result is consistent with the conclusion of Warren & Cataldi (2006), who find little time variation in the relationship between student work and high school dropout between years 1966-1997.

Structural variation. The difference in the estimated effects for *part-time students* fails to manifest itself in our baseline model but is more apparent in the FMA robustness check

presented in Table B2. These FMA results are in line with the findings of Darolia (2014), who suggests that the effect for part-time students is quite small while the effect for full-time students is negative. More importantly, we highlight the importance and direction of the results for variables *low-intensity employment* and *high-intensity employment*. Estimates generated with low work intensities yield systematically more positive effects of student employment on educational outcomes compared to estimates conditional on high work intensity. This finding is intuitive and also in line with Buscha *et al.* (2012), who argue that less intense work involvement is beneficial to study outcomes. The findings are consistent with the threshold perspective, which asserts that student employment has a positive effect on educational outcomes up to a certain amount of working hours, after which the effect reverses.

Regarding cross-country heterogeneity, once again we find that the estimates reported for Germany are substantially more positive than estimates reported for other countries. As we have noted, we specifically exclude estimates pertaining to German vocational schools, which combine work and study by definition, and the corresponding estimates are thus incomparable with the rest of the sample. The estimates using German data in our dataset are relevant to college students, and it is apparent that the long German tradition of effectively combining work and study is not limited to vocational schools but spills over to other parts of the educational system as well.

Estimation methods. We find that the use of OLS typically brings larger estimates of the effect of student employment on education. The result is in line with our previous findings that once endogeneity is not accounted for, the true effect tends to be positive on average. Our findings also suggest that quasi-experimental techniques (matching, instrumental variables, and difference-in-differences) tend to yield more negative estimates compared to other methods, but the corresponding posterior inclusion probabilities for these variables are between 0.3 and 0.4. Among the quasi-experimental techniques the smallest PIP we obtain is for the use of instrumental variables. As noted by Oettinger (1999), it is challenging to find a suitable instrument in the case of this literature. For instance, Baert *et al.* (2017) explain that conditions on the local labor market, often used as an instrument, may affect students' decision to work, e.g. a highly saturated market labor decreases students' chance of finding a job, and hence influence students' educational outcomes. Similarly, Buscha *et al.* (2012) argue that state child labor laws do not have to be necessarily exogenous to educational outcomes as they reflect the general importance of academic attainment in the specific region.

Our BMA results further indicate that accounting for students' *age* and *parental education* in primary studies is not important for explaining the variation in the estimated effect of student employment on educational outcomes. In contrast, controlling for students' *ability* after filtering out publication bias results in more negative estimates. Again, the finding is consistent with the notion that ignoring endogeneity results in spuriously positive estimates of the effect of student employment on education. We observe a similar pattern for *Ethnicity control*. *Motivation control*, on the other hand, seem to influence the estimates in the opposite direction, and the sign of the posterior mean is puzzling; nevertheless, motivation is much more difficult for the

researcher to proxy than ethnicity (race) or ability (IQ). Even so, the importance of including the motivation control has been documented widely. Wenz & Yu (2010) argue that students seeking career-specific skills achieve higher test scores while students seeking general work experience will achieve lower test scores. Stinebrickner & Stinebrickner (2003) argue, for example, that students with low motivation to earn good grades find it more important to engage in term-time employment.

Publication characteristics. In our baseline BMA exercise we fail to find evidence that *published* status and journal quality measured by *impact factor* or in terms of number of *citations* systematically influence the reported estimates. In contrast, the number of *citations* turns out to be important in the FMA robustness check: frequently cited studies yield systematically more negative effect estimates. The following three explanations are plausible: i) researchers cite these studies more often to corroborate their negative findings, ii) researchers refer to studies reporting negative estimates when highlighting the improvements of their studies that yield more positive estimates, iii) research papers yielding negative estimates are of higher methodological quality, and hence are cited more often. Unfortunately, our analysis cannot confirm nor reject any of these explanations. In any case, the marginal effect of the variable in the FMA specification is relatively small.

4.4 Implied Estimate

As the bottom line of our analysis, we use the results of Bayesian model averaging to compute the implied value of the partial correlation coefficient in different contexts (e.g., data for Germany, female students, part-time employment, decisions to dropout, etc.) while correcting for publication, endogeneity, and other biases in the literature. We do so by computing the fitted values from the BMA exercise conditional on the values of individual variables that correspond to best practice in the literature. Of course, best practice is subjective, so we define it only for variables for which there is reasonable consensus in the most recent literature; for other variables we use sample means. Because we want to correct the implied estimate for publication bias, we plug in zero for the standard error. To take endogeneity control into account, we prefer studies that use any of the following approaches: matching, instrumental variables, difference-in-differences, or a proxy for ability. We also prefer if the study controls for motivation, parental education, age, and ethnicity (that is, we plug in "1" for the corresponding dummy variables). Concerning the measure of employment, we prefer if the study uses a continuous variable. We also plug in zero for the dummy variable corresponding to self-reported data, which might entail substantial measurement error. Moreover, we prefer panel data of recent vintage and put more weight on highly-cited studies published in journals with a high impact factor. The other variables are set to their sample means.

Table 7 shows the mean implied estimates for 11 different situations: data for the United States, data for Germany, data for other countries, male students, female students, part-time students, low-intensity employment, high-intensity employment, educational outcomes mea-

Table 7: Best-practice estimates in different contexts

	Mean	95% conf. int.	
USA	-0.039	-0.078	0.001
Germany	0.019	-0.046	0.083
Other countries	-0.048	-0.104	0.008
Male students	-0.039	-0.083	0.005
Female students	-0.037	-0.079	0.005
Part-time students	-0.025	-0.069	0.020
Low-intensity employment	-0.023	-0.068	0.022
High-intensity employment	-0.054	-0.098	-0.010
Educational outcome: choices	-0.062	-0.107	-0.017
Educational outcome: test scores	-0.033	-0.075	0.009
Educational outcome: attainment	-0.033	-0.080	0.014
Overall effect	-0.038	-0.079	0.003

Notes: The table presents the mean partial correlation coefficients implied by the Bayesian model averaging exercise and our definition of best practice for various contexts. That is, we compute fitted values from BMA conditional on selected values of regression variables (for example, 0 for ignoring endogeneity). The confidence intervals are approximate and constructed using OLS with the standard errors clustered at the study level.

sured by decisions to dropout, outcomes measured by test scores, and outcomes measured in a different way. The overall mean is -0.038, and all individual means are negative with the exception of data for Germany. Because the implied estimates are based on the results of Bayesian model averaging, the differences between individual means reflect the discussion presented earlier in this section: most importantly, student employment affects decisions to dropout more than it affects test scores, and the effect of working part-time is generally smaller than the effect of working full-time. But even the largest effects we identify are too small to matter much in practice.

5 Concluding Remarks

We show that the literature examining the impact of student employment on educational outcomes, represented by 861 estimates reported in 69 studies, is consistent with no practically important causal effect. Publication bias interacts with endogeneity bias, and several data and method choices systematically affect the reported estimates as well. After correcting for both biases and controlling for 32 aspects of data, method, and publication characteristics, we derive estimates of the causal effect of employment on education in 11 different contexts (for example, USA vs. Germany, male vs. female students, low- vs. high-intensity employment, and grades vs. decisions to dropout). The effect is statistically insignificant for all but 2 contexts. We find the strongest effect for dropout decisions, but the corresponding partial correlation coefficient is still low: -0.06. Doucouliagos (2011) provides a survey of 22,000 partial correlations computed in economics and notes that 75% of them are larger in the absolute value than 0.07; in his guidelines for interpreting partial correlations values below 0.07 are considered immaterial. In the well-known earlier guidelines for social sciences by Cohen (1988) the threshold for an

effect to be considered at least small is even stricter, 0.1. The boundaries of the 95% confidence intervals for our 11 contexts range from −0.11 (dropout decisions) to 0.08 (Germany), which leads us to rule out anything but weak effects of employment on education.

Three qualifications of our results are in order. First, we work with partial correlation coefficients instead of elasticities, which complicates inference. Unfortunately primary studies almost never report elasticities and use different units and functional forms. It is infeasible to recompute these estimates into a common economic metric, and the partial correlation coefficient thus represents the only choice for comparing the estimates reported in the literature. As a robustness check, we also compute the overall mean effect using estimates that employ the same units and functional form so that they can be directly compared. The implied mean corrected for publication bias is zero, consistent with our other results. Second, our main analysis rests on the assumption that publication bias is a linear function of the (exogenous) standard error. As a robustness check, we employ more complex methods that do not need linearity, exogeneity, or both. In this case the more complex techniques give us similar results, and we thus keep the simple linear specification that allows for straightforward incorporation into Bayesian model averaging. Third, in the analysis of heterogeneity we examine 32 variables, and with so many variables collinearity can complicate the interpretation of individual marginal effects. We show that the collinearity problem in our case is not large and additionally use the dilution prior in Bayesian model averaging, which is designed to minimize the consequences of collinearity.

Despite the lack of prima facie consensus in Figure 1 presented in the Introduction, our analysis shows that the literature taken as a whole and corrected for biases arrives at a clear conclusion: any effects of student employment on educational outcomes are weak at best. That is not to say that the thousands of pages written on this topic has amounted to much ado about nothing. Genuine and useful variation exists in the effect depending on the context under examination; moreover, without the meticulous work of previous researchers, it would be impossible to credibly isolate the effects of both endogeneity and publication biases. After quantitatively examining the literature, we believe it is safe to say with confidence that part-time employment does not hurt students' educational outcomes in any plausible context.

References

van Aert, R. C. & M. van Assen (2021): "Correcting for publication bias in a meta-analysis with the p-uniform* method." *Working paper*, Tilburg University & Utrecht University.

Amini, S. M. & C. F. Parmeter (2012): "Comparison of model averaging techniques: Assessing growth determinants." *Journal of Applied Econometrics* **27(5)**: pp. 870–876.

Andrews, I. & M. Kasy (2019): "Identification of and correction for publication bias." *American Economic Review* **109(8)**: pp. 2766–2794.

Apel, R., S. D. Bushway, R. Paternoster, R. Brame, & G. Sweeten (2008): "Using state child labor laws to identify the causal effect of youth employment on deviant behavior and academic achievement." *Journal of Quantitative Criminology* **24(4)**: pp. 337–362.

Applegate, C. & A. Daly (2006): "The impact of paid work on the academic performance of students: A case study from the University of Canberra." *Australian Journal of Education* **50(2)**: pp. 155–166.

Arano, K. & C. Parker (2008): "How Does Employment Affect Academic Performance Among College Students?" *Journal of Economic Insight* **34(2)**: pp. 65–82.

Ashenfelter, O., C. Harmon, & H. Oosterbeek

(1999): "A review of estimates of the schooling/earnings relationship, with tests for publication bias." *Labour Economics* **6(4)**: pp. 453–470.

Auers, D., T. Rostoks, & K. Smith (2007): "Flipping burgers or flipping pages? Student employment and academic attainment in post-Soviet Latvia." *Communist and Post-Communist Studies* **40(4)**: pp. 477–491.

Babcock, P. & M. Marks (2011): "The Falling Time Cost of College: Evidence from Half a Century of Time Use Data." *The Review of Economics and Statistics* **93(2)**: pp. 468–478.

Baert, S., I. Marx, B. Neyt, E. V. Belle, & J. V. Casteren (2018): "Student employment and academic performance: an empirical exploration of the primary orientation theory." *Applied Economics Letters* **25(8)**: pp. 547–552.

Baert, S., B. Neyt, E. Omey, & D. Verhaest (2017): "Student Work, Educational Achievement, and Later Employment: A Dynamic Approach." *IZA Discussion Papers 11127*, Institute of Labor Economics (IZA), Bonn.

Basu, A. M. (2002): "Why does Education Lead to Lower Fertility? A Critical Review of Some of the Possibilities." *World Development* **30(10)**: pp. 1779–1790.

Beerkens, M., E. Magi, & L. Lill (2011): "University studies as a side job: Causes and consequences of massive student employment in Estonia." *Higher Education* **61(6)**: pp. 679–692.

Beffy, M., D. Fougere, & A. Maurel (2013): "The Effect of College Employment on Graduation: Evidence from France." *CEPR Discussion Papers 9565*, C.E.P.R. Discussion Papers.

Benos, N. & S. Zotou (2014): "Education and Economic Growth: A Meta-Regression Analysis." *World Development* **64(C)**: pp. 669–689.

Blanco-Perez, C. & A. Brodeur (2020): "Publication Bias and Editorial Statement on Negative Findings." *Economic Journal* **130(629)**: pp. 1226–1247.

BLS (2020): "Bureau of Labor Statistics news resealeses." *Various years, retrieved from https://www.bls.gov/bls/newsrels.htm*.

Body, K. M.-D., L. Bonnal, & J.-F. Giret (2014): "Does student employment really impact academic achievement? The case of France." *Applied Economics* **46(25)**: pp. 3061–3073.

Bom, P. R. D. & H. Rachinger (2019): "A kinked meta-regression model for publication bias correction." *Research Synthesis Methods* **10(4)**: pp. 497–514.

Bozick, R. (2007): "Making it through the first year of college: The role of students' economic resources, employment, and living arrangements." *Sociology of Education* **80(3)**: pp. 261–285.

Brodeur, A., N. Cook, & A. Heyes (2020): "Methods Matter: P-Hacking and Causal Inference in Economics." *American Economic Review* **110(11)**: pp. 3634–3660.

Brodeur, A., M. Le, M. Sangnier, & Y. Zylberberg (2016): "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics* **8(1)**: pp. 1–32.

Bruns, S. B. & J. P. A. Ioannidis (2016): "p-Curve and p-Hacking in Observational Research." *PloS ONE* **11(2)**: p. e0149144.

Buscha, F., A. Maurel, L. Page, & S. Speckesser (2012): "The effect of employment while in high school on educational attainment: A conditional difference-in-differences approach." *Oxford Bulletin of Economics and Statistics* **74(3)**: pp. 380–396.

Callender, C. (2008): "The impact of term-time employment on higher education students' academic attainment and achievement." *Journal of Education Policy* **23(4)**: pp. 359–377.

Canabal, M. E. (1998): "College student degree of participation in the labor force: Determinants and relationship to school performance." *College Student Journal* **32(4)**: pp. 597–605.

Card, D., J. Kluve, & A. Weber (2018): "What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations." *Journal of the European Economic Association* **16(3)**: pp. 894–931.

Card, D. & A. B. Krueger (1995): "Time-series minimum-wage studies: A meta-analysis." *The American Economic Review* **85(2)**: pp. 238–243.

Carneiro, P. M. & J. J. Heckman (2003): "Human capital policy." *IZA Discussion Papers 821*, Institute of Labor Economics (IZA), Bonn.

Carr, R. V., J. D. Wright, & C. J. Brody (1996): "Effects of high school work experience a decade later: Evidence from the National Longitudinal Survey." *Sociology of Education* **69(1)**: pp. 66–81.

Chen, X. & C. Carroll (2007): "Part-time undergraduates in postsecondary education: 2003–04." *Technical report*, Washington, DC: US Department of Education.

Choi, Y. (2018): "Student employment and persistence: Evidence of effect heterogeneity of student employment on college dropout." *Research in Higher Education* **59(1)**: pp. 88–107.

Christensen, G. & E. Miguel (2018): "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* **56(3)**: pp. 920–980.

Cohen, J. (1988): *Statistical Power Analysis in the Behavioral Sciences*. Hillsdale: Erlbaum, 2nd edition.

Cui, Y. & P. S. Martins (2021): "What Drives Social Returns to Education? A Meta-Analysis." *IZA Discussion Papers 14332*, Institute of Labor Economics (IZA).

Cunado, J. & F. Gracia (2012): "Does Education Affect Happiness? Evidence for Spain." *Social In-*

dicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement **108(1)**: pp. 185–196.

Dadgar, M. (2012): "The academic consequences of employment for students enrolled in community college." *CCRC Working Paper 46*, Community College Research Center.

D'Amico, R. (1984): "Does employment during high school impair academic progress?" *Sociology of Education* **57(3)**: pp. 152–164.

Darolia, R. (2014): "Working (and studying) day and night: Heterogeneous effects of working on the academic performance of full-time and part-time students." *Economics of Education Review* **38(C)**: pp. 38–50.

DellaVigna, S., D. Pope, & E. Vivalt (2019): "Predict science to improve science." *Science* **366(6464)**: pp. 428–429.

DeSimone, J. S. (2006): "Academic Performance and Part-Time Employment among High School Seniors." *The BE Journal of Economic Analysis & Policy* **6(1)**: pp. 1–36.

DeSimone, J. S. (2008): "The Impact of Employment during School on College Student Academic Performance." *NBER Working Papers 14006*, National Bureau of Economic Research, Inc.

Doucouliagos, H. (2011): "How large is large? Preliminary and relative guidelines for interpreting partial correlations in economics." *Working Papers 5/2011*, Deakin University.

Dustmann, C. & A. van Soest (2007): "Part-time work, school success and school leaving." *Empirical Economics* **32(2)**: pp. 277–299.

Eckstein, Z. & K. I. Wolpin (1999): "Why youths drop out of high school: The impact of preferences, opportunities, and abilities." *Econometrica* **67(6)**: pp. 1295–1339.

Egger, M., G. D. Smith, M. Schneider, & C. Minder (1997): "Bias in meta-analysis detected by a simple, graphical test." *BMJ* **315(7109)**: pp. 629–634.

Ehrenberg, R. G. & D. R. Sherman (1987): "Employment While in College, Academic Achievement, and Postcollege Outcomes: A Summary of Results." *Journal of Human Resources* **22(1)**: pp. 1–23.

Eicher, T. S., C. Papageorgiou, & A. E. Raftery (2011): "Default priors and predictive performance in Bayesian model averaging, with application to growth determinants." *Journal of Applied Econometrics* **26(1)**: pp. 30–55.

Fernandez, C., E. Ley, & M. F. J. Steel (2001): "Benchmark priors for Bayesian Model Averaging." *Journal of Econometrics* **100(2)**: pp. 381–427.

Furukawa, C. (2021): "Publication bias under aggregation frictions: Theory, evidence, and a new correction method." *Working paper*, MIT.

Geel, R. & U. Backes-Gellner (2012): "Earning while learning: When and how student employment is beneficial." *Labour* **26(3)**: pp. 313–340.

George, E. I. (2010): "Dilution priors: Compensating for model space redundancy." In "IMS Collections Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown," volume 6, p. 158–165. Institute of Mathematical Statistics.

Gleason, P. M. (1993): "College Student Employment, Academic Progress, and Postcollege Labor Market Success." *Journal of Student Financial Aid* **23(2)**: pp. 5–14.

Hansen, B. (2007): "Least Squares Model Averaging." *Econometrica* **75(4)**: pp. 1175–1189.

Hasan, I., R. Horvath, & J. Mares (2018): "What Type of Finance Matters for Growth? Bayesian Model Averaging Evidence." *World Bank Economic Review* **32(2)**: pp. 383–409.

Havranek, T. (2015): "Measuring intertemporal substitution: The importance of method choices and selective reporting." *Journal of the European Economic Association* **13(6)**: pp. 1180–1204.

Havranek, T., T. D. Stanley, H. Doucouliagos, P. Bom, J. Geyer-Klingeberg, I. Iwasaki, W. R. Reed, K. Rost, & R. C. M. van Aert (2020): "Reporting Guidelines for Meta-Analysis in Economics." *Journal of Economic Surveys* **34(3)**: pp. 469–475.

Hawkins, C. A., M. L. Smith, R. C. Hawkins, II, & D. Grant (2005): "The relationships among hours employed, perceived work interference, and grades as reported by undergraduate social work students." *Journal of Social Work Education* **41(1)**: pp. 13–27.

Hoeting, J. A., D. Madigan, A. E. Raftery, & C. T. Volinsky (1999): "Bayesian model averaging: A tutorial." *Statistical Science* **14(4)**: pp. 382–401.

Holford, A. (2020): "Youth employment, academic performance and labour market outcomes: Production functions and policy effects." *Labour Economics* **63(C)**. Art. 101806.

Hovdhaugen, E. (2015): "Working while studying: The impact of term-time employment on dropout rates." *Journal of Education and Work* **28(6)**: pp. 631–651.

Huang, J., H. Maassen van den Brink, & W. Groot (2009): "A meta-analysis of the effect of education on social capital." *Economics of Education Review* **28(4)**: pp. 454–464.

Hwang, J.-K. (2013): "Employment and student performance in Principles of Economics." *International Review of Economics Education* **13(C)**: pp. 26–30.

Imai, T., T. A. Rutter, & C. F. Camerer (2020): "Meta-Analysis of Present-Bias Estimation Using Convex Time Budgets." *Economic Journal* **ueaa115(forthcoming)**.

Ioannidis, J. P., T. D. Stanley, & H. Doucouliagos (2017): "The Power of Bias in Economics Research." *Economic Journal* **127(605)**: pp. F236–F265.

Jaquess, S. N. (1984): *The influence of part-time employment and study habits and attitudes on academic performance of high school juniors.* Doctoral thesis, The University of Oklahoma. 152 pages.

Jeffreys, H. (1961): *Theory of Probability.* Oxford Classic Texts in the Physical Sciences. Oxford: Oxford University Press, third edition.

Joensen, J. S. (2009): *Academic and labor market success: The impact of student employment, abilities, and preferences.* Doctoral thesis, Stockholm School of Economics. 71 pages.

Jones, C. M., J. P. Green, & H. E. Higson (2017): "Do work placements improve final year academic performance or do high-calibre students choose to do work placements?" *Studies in Higher Education* **42(6)**: pp. 976–992.

Jones, R. & P. Sloane (2005): "Students and term-time employment." *A report for the economic research unit*, Welsh Economy Labour Markets Evaluation Research Centre, Department of Economics.

Jung, S. (2015): "Does education affect risk aversion? Evidence from the British education reform." *Applied Economics* **47(28)**: pp. 2924–2938.

Kalenkoski, C. & S. Pabilonia (2010): "Parental transfers, student achievement, and the labor supply of college students." *Journal of Population Economics* **23(2)**: pp. 469–496.

Kohen, A. I., G. Nestel, & C. Karmas (1978): "Factors affecting individual persistence rates in undergraduate college programs." *American Educational Research Journal* **15(2)**: pp. 233–252.

Kouliavtsev, M. (2013): "The Impact of employment and extracurricular involvement on undergraduates' performance in a business statistics course." *Journal of Economics and Economic Education Research* **14(3)**: pp. 53–66.

Lee, C. & P. F. Orazem (2010): "High school employment, school performance, and college entry." *Economics of Education Review* **29(1)**: pp. 29–39.

Lee, J. C. & J. Staff (2007): "When work matters: The varying impact of work intensity on high school dropout." *Sociology of Education* **80(2)**: pp. 158–178.

Leos-Urbel, J. (2014): "What is a summer job worth? The impact of summer youth employment on academic outcomes." *Journal of Policy Analysis and Management* **33(4)**: pp. 891–911.

Ley, E. & M. F. Steel (2009): "On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression." *Applied Econometrics* **24(C)**: pp. 651–674.

Machin, S., O. Marie, & S. Vujić (2011): "The Crime Reducing Effect of Education." *Economic Journal* **121(552)**: pp. 463–484.

Magnus, J. R., O. Powell, & P. Prufer (2010): "A comparison of two model averaging techniques with an application to growth empirics." *Journal of Econometrics* **154(2)**: pp. 139–153.

Maloney, T. & A. Parau (2004): "The effects of in-school and in-tertiary employment on academic achievement and labour market transitions: Evidence from the Christchurch Health and Development Study." *Report to the Labour Market Policy Group*, New Zealand Department of Labour.

Manthei, R. J. & A. Gilmore (2005): "The effect of paid employment on university students' lives." *Education + Training* **47(3)**: pp. 202–215.

Marsh, H. W. (1991): "Employment during high school: Character building or a subversion of academic goals?" *Sociology of Education* pp. 172–189.

Marsh, H. W. & S. Kleitman (2005): "Consequences of employment during high school: Character building, subversion of academic goals, or a threshold?" *American Educational Research Journal* **42(2)**: pp. 331–369.

McCloskey, D. N. & S. T. Ziliak (2019): "What Quantitative Methods Should We Teach to Graduate Students? A Comment on Swann's Is Precise Econometrics an Illusion?" *The Journal of Economic Education* **50(4)**: pp. 356–361.

McCoy, S. & E. Smyth (2007): "So much to do, so little time: part-time employment among secondary students in Ireland." *Work, Employment and Society* **21(2)**: pp. 227–246.

McKechnie, J., K. Dunleavy, & S. Hobbs (2005): "Student employment and its educational impact: a Scottish study." *Scottish Educational Review* **37(1)**: pp. 58–67.

McKenzie, K. & R. Schweitzer (2001): "Who succeeds at university? Factors predicting academic performance in first year Australian university students." *Higher Education Research & Development* **20(1)**: pp. 21–33.

McNeal, R. B. (1997): "Are students being pulled out of high school? The effect of adolescent employment on dropping out." *Sociology of Education* **70(3)**: pp. 206–220.

McVicar, D. & B. McKee (2002): "Part-time work during post-compulsory education and examination performance: Help or hindrance?" *Scottish Journal of Political Economy* **49(4)**: pp. 393–406.

Milligan, K., E. Moretti, & P. Oreopoulos (2004): "Does education improve citizenship? evidence from the united states and the united kingdom." *Journal of Public Economics* **88(9-10)**: pp. 1667–1695.

Montmarquette, C., N. Viennot-Briot, & M. Dagenais (2007): "Dropout, School Performance, and Working while in School." *The Review of Economics and Statistics* **89(4)**: pp. 752–760.

Moulin, S., P. Doray, B. Laplante, & M. C. Street (2013): "Work intensity and non-completion of university: Longitudinal approach and causal inference." *Journal of Education and Work* **26(3)**: pp. 333–356.

Newman, S. C. (1942): *Employment problems of college students.* Studies in Education. Washington, DC: American Council on Public Affairs. 158 pages.

Neyt, B., E. Omey, D. Verhaest, & S. Baert (2019): "Does Student Work Really Affect Educational Outcomes? A Review Of The Literature." *Journal of Economic Surveys* **33(3)**: pp. 896–921.

Oehlert, G. W. (1992): "A note on the delta method." *The American Statistician* **46(1)**: pp. 27–29.

Oettinger, G. S. (1999): "Does High School Employment Affect High School Academic Performance?" *ILR Review* **53(1)**: pp. 136–151.

Parent, D. (2006): "Work while in high school in Canada: Its labour market and educational attainment effects." *Canadian Journal of Economics/Revue Canadienne d'Economique* **39(4)**: pp. 1125–1150.

Paul, H. (1982): "The impact of outside employment on student achievement in macroeconomic principles." *The Journal of Economic Education* **13(2)**: pp. 51–56.

Post, D. & S.-l. Pong (2000): "Employment during middle school: The effects on academic achievement in the US and abroad." *Educational Evaluation and Policy Analysis* **22(3)**: pp. 273–298.

Psacharopoulos, G. & H. A. Patrinos (2018): "Returns to investment in education: a decennial review of the global literature." *Education Economics* **26(5)**: pp. 445–458.

Raftery, A. E. (1995): "Bayesian model selection in social research." *Sociological Methodology* **25(C)**: pp. 111–163.

Richardson, J. J., S. Kemp, S. Malinen, & S. A. Haultain (2013): "The academic achievement of students in a New Zealand university: Does it pay to work?" *Journal of Further and Higher Education* **37(6)**: pp. 864–882.

Riggert, S. C., M. Boyle, J. M. Petrosko, D. Ash, & C. Rude-Parkins (2006): "Student employment and higher education: Empiricism and contradiction." *Review of Educational Research* **76(1)**: pp. 63–92.

Rochford, C., M. Connolly, & J. Drennan (2009): "Paid part-time employment and academic performance of undergraduate nursing students." *Nurse Education Today* **29(6)**: pp. 601–606.

Roodman, D., J. G. MacKinnon, M. O. Nielsen, & M. D. Webb (2018): "Fast and wild: Bootstrap inference in Stata using boottest." *Queen's Economics Department Working Paper 1406*, Department of Economics, Queen's University, Canada: Kingston.

Rothstein, D. S. (2007): "High School Employment and Youths' Academic Achievement." *Journal of Human Resources* **42(1)**: pp. 194–213.

Ruhm, C. J. (1997): "Is High School Employment Consumption or Investment?" *Journal of Labor Economics* **15(4)**: pp. 735–776.

Rözer, J. & H. G. van de Werfhorst (2020): "Three Worlds of Vocational Education: Specialized and General Craftsmanship in France, Germany, and The Netherlands." *European Sociological Review* **36(5)**: pp. 780–797.

Sabia, J. J. (2009): "School-year employment and academic performance of young adolescents." *Economics of Education Review* **28(2)**: pp. 268–276.

Salamonson, Y. & S. Andrew (2006): "Academic performance in nursing students: Influence of part-time employment, age and ethnicity." *Journal of Advanced Nursing* **55(3)**: pp. 342–349.

Savoca, M. (2016): *Campus employment as a high-impact practice: Relationship to academic success and persistence of first-generation college students.* Phd thesis, Colorado State University. 126 pages.

Schoenhals, M., M. Tienda, & B. Schneider (1998): "The educational and personal consequences of adolescent employment." *Social Forces* **77(2)**: pp. 723–761.

Scott-Clayton, J. & V. Minaya (2016): "Should student employment be subsidized? Conditional counterfactuals and the outcomes of work-study participation." *Economics of Education Review* **52(C)**: pp. 1–18.

Simon, H., J. M. C. Diaz, & J. L. C. Costa (2017): "Analysis of university student employment and its impact on academic performance." *Electronic Journal of Research in Educational Psychology* **15(2)**: pp. 281–306.

Singh, K., M. Chang, & S. Dika (2007): "Effects of part-time work on school achievement during high school." *The Journal of Educational Research* **101(1)**: pp. 12–23.

Sprietsma, M. (2015): "Student employment: Advantage or handicap for academic achievement?" *ZEW Discussion Papers 085/2015*, ZEW - Leibniz Centre for European Economic Research.

Staff, J. & J. T. Mortimer (2007): "Educational and work strategies from adolescence to early adulthood: Consequences for educational attainment." *Social Forces* **85(3)**: pp. 1169–1194.

Staff, J., J. E. Schulenberg, & J. G. Bachman (2010): "Adolescent work intensity, school performance, and academic engagement." *Sociology of Education* **83(3)**: pp. 183–200.

Stanley, T. D. (2001): "Wheat from Chaff: Meta-analysis as Quantitative Literature Review." *Journal of Economic Perspectives* **15(3)**: pp. 131–150.

Stanley, T. D. (2005): "Beyond publication bias." *Journal of Economic Surveys* **19(3)**: pp. 309–345.

Stanley, T. D. (2008): "Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection." *Oxford Bulletin of Economics and Statistics* **70(1)**: pp. 103–127.

Stanley, T. D., S. B. Jarrell, & H. Doucouliagos (2010): "Could It Be Better to Discard 90% of the

Data? A Statistical Paradox." *The American Statistician* **64(1)**: pp. 70–77.

Steel, L. (1991): "Early work experience among white and non-white youths: Implications for subsequent enrollment and employment." *Youth & Society* **22(4)**: pp. 419–447.

Steel, M. F. J. (2020): "Model Averaging and its Use in Economics." *Journal of Economic Literature* **58(3)**: pp. 644–719.

Steinberg, L. D., E. Greenberger, L. Garduque, & S. McAuliffe (1982): "High school students in the labor force: Some costs and benefits to schooling and learning." *Educational Evaluation and Policy Analysis* **4(3)**: pp. 363–372.

Stern, D. & D. Briggs (2001): "Does paid employment help or hinder performance in secondary school? Insights from US high school students." *Journal of Education and Work* **14(3)**: pp. 355–372.

Stinebrickner, R. & T. R. Stinebrickner (2003): "Working during school and academic performance." *Journal of Labor Economics* **21(2)**: pp. 473–491.

Tessema, M. T., K. J. Ready, & M. Astani (2014): "Does part-time job affect college students' satisfaction and academic performance (GPA)? The case of a mid-sized public university." *International Journal of Business Administration* **5(2)**: pp. 50–59.

Theune, K. (2015): "The working status of students and time to degree at German universities." *Higher Education* **70(4)**: pp. 725–752.

Tienda, M. & A. Ahituv (1996): "Ethnic Differences in School Departure: Does Youth Employment Promote or Undermine Educational Attainment?" In G. L. Mangum & S. L. Mangum (editors), "Of Heart and Mind: Social Policy Essays in Honor of Sar A. Levitan," chapter 4, pp. 93–110. Kalamazoo, Michigan: Upjohn Institute for Employment Research.

Torres, V., J. P. Gross, & A. Dadashova (2010): "Traditional-age students becoming at-risk: Does working threaten college students' academic success?" *Journal of College Student Retention: Research, Theory & Practice* **12(1)**: pp. 51–68.

Trockel, M. T., M. D. Barnes, & D. L. Egget (2000): "Health-related variables and academic performance among first-year college students: Implications for sleep and other behaviors." *Journal of American College Health* **49(3)**: pp. 125–131.

Tyler, J. H. (2003): "Using State Child Labor Laws to Identify the Effect of School-Year Work on High School Achievement." *Journal of Labor Economics* **21(2)**: pp. 353–380.

Wang, H., M. Kong, W. Shan, & S. K. Vong (2010): "The effects of doing part-time jobs on college student academic performance and social life in a Chinese society." *Journal of Education and Work* **23(1)**: pp. 79–94.

Wang, H., X. Zhang, & G. Zou (2009): "Frequentist model averaging estimation: A review." *Journal of Systems Science and Complexity* **22(4)**: pp. 732–748.

Warren, J. R. (2002): "Reconsidering the relationship between student employment and academic outcomes: A new theory and better data." *Youth & Society* **33(3)**: pp. 366–393.

Warren, J. R. & E. F. Cataldi (2006): "A historical perspective on high school students' paid employment and its association with high school dropout." *Sociological Forum* **21(1)**: pp. 113–143.

Warren, J. R. & J. C. Lee (2003): "The impact of adolescent employment on high school dropout: Differences by individual and labor-market characteristics." *Social Science Research* **32(1)**: pp. 98–128.

Warren, J. R., P. C. LePore, & R. D. Mare (2000): "Employment during high school: Consequences for students' grades in academic courses." *American Educational Research Journal* **37(4)**: pp. 943–969.

Wenz, M. & W.-C. Yu (2010): "Term-time employment and the academic performance of undergraduates." *Journal of Education Finance* **35(4)**: pp. 358–373.

Xue, X., M. Cheng, & W. Zhang (2021): "Does Education Really Improve Health? A Meta?Analysis." *Journal of Economic Surveys* **35(1)**: pp. 71–105.

Yanbarisova, D. (2015): "The effects of student employment on academic performance in Tatarstan higher education institutions." *Russian Education & Society* **57(6)**: pp. 459–482.

Zeugner, S. (2011): "Bayesian model averaging with BMS." *Tutorial to the R-package BMS 1e30*, European Center for Advanced Research in Economics.

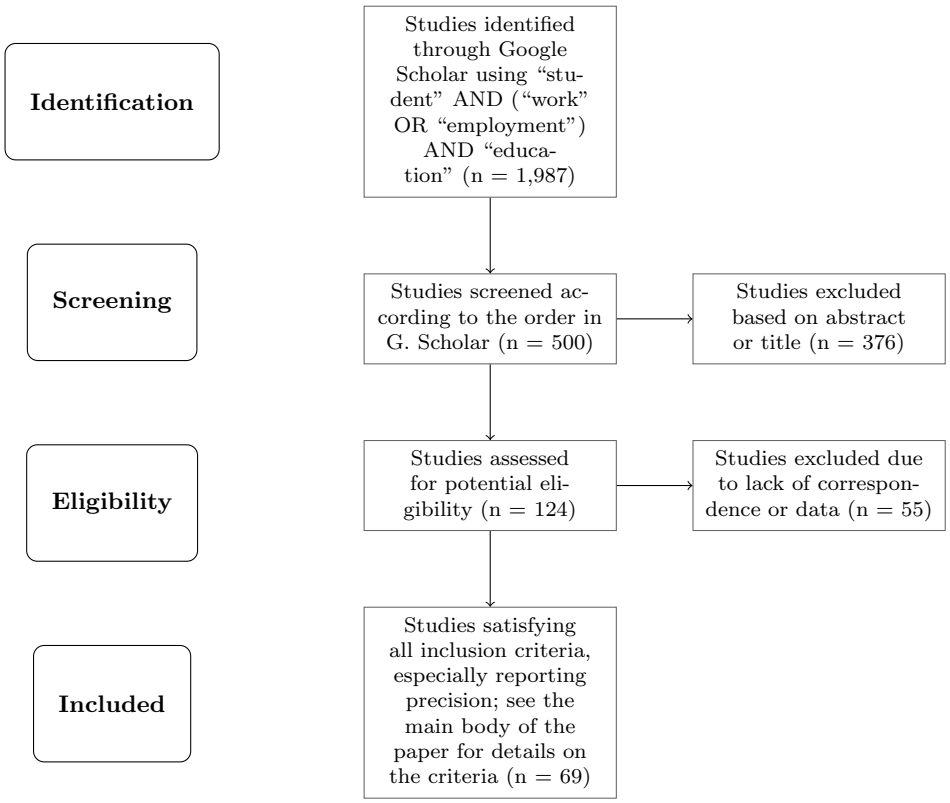
Zeugner, S. & M. Feldkircher (2009): "Benchmark priors revisited: On adaptive shrinkage and the supermodel effect in Bayesian model averaging." *IMF Working Paper 9/202*, International Monetary Fund.

Zhang, G. & C. Johnston (2010): "Employment and the academic performance of undergraduate business students." *Southwestern Business Administration Journal* **10(2)**: pp. 54–83.

Appendices

A Details of Literature Search (for online publication)

Figure A1: PRISMA flow diagram



Notes: Preferred reporting items for systematic reviews and meta-analyses (PRISMA) is an evidence-based set of items for reporting in systematic reviews and meta-analyses. More details on PRISMA and reporting standards of meta-analysis in general are provided by Havranek *et al.* (2020).

B Additional Statistics, Robustness Checks, and BMA Diagnostics (for online publication)

Figure B1: Correlations between regression variables

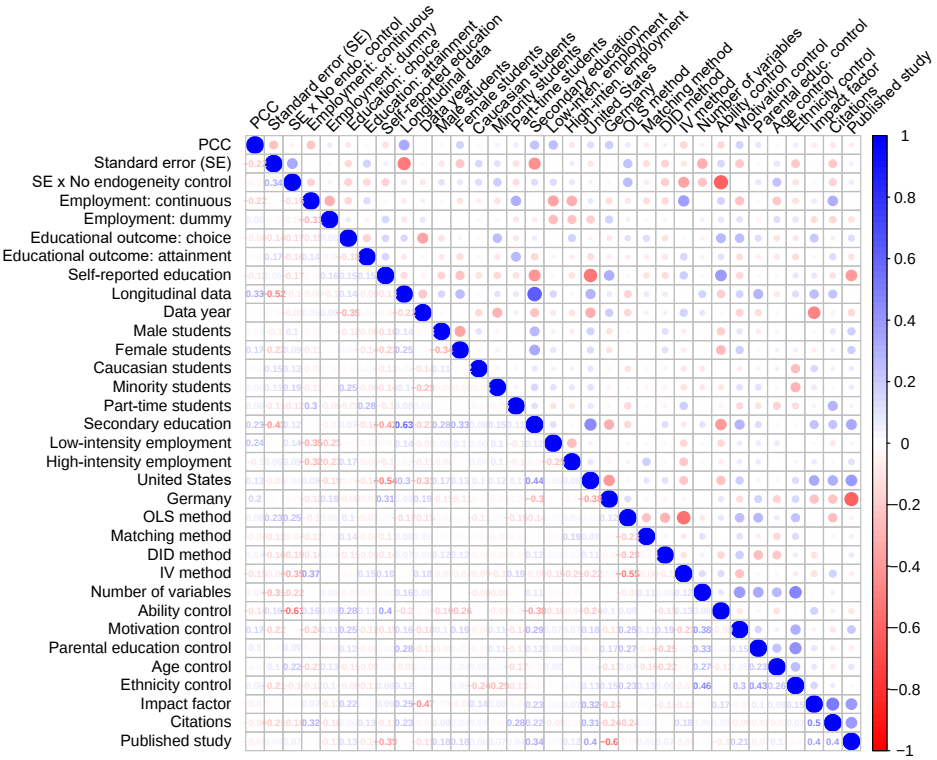


Table B1: Tests of publication bias (homogeneous estimates)

Panel A: Linear techniques				
	OLS	IV	Study	Precision
Standard error	-1.104	0.228	-1.03	-0.968
(Publication bias)	(0.830)	(0.643)	(0.846)	(0.620)
	[-2.215, 1.205]	[-2.179, 2.102]		[-2.062, 0.811]
Constant	0.00740	-0.00991	0.0154	0.00108
(Effect beyond bias)	(0.00671)	(0.0107)	(0.0108)	(0.00369)
	[-0.008, 0.029]	[-0.015, 0.043]		[-0.011, 0.011]
Observations	86	86	86	86
Panel B: Between- and within-study variation				
	BE	FE	RE	
Standard error	-2.025***	-1.196***	-1.373***	
(Publication bias)	(0.374)	(0.258)	(0.218)	
Constant	0.0110	0.00859	0.00345	
(Effect beyond bias)	(0.0128)	(0.00558)	(0.0106)	
Observations	86	86	86	
Panel C: Nonlinear techniques				
	WAAP	Stem-based	Endogenous kink	Selection model	p-uniform*
Effect beyond bias	-0.00242***	-0.002	-0.00211***	-0.001	-0.007
	(0.000241)	(0.00123)	(0.000223)	(0.003)	(0.00791)
Observations	86	86	86	86	86

Notes: See notes to Table 3.

Table B2: Why estimates vary (robustness checks)

Response variable: partial correlation coefficient	Bayesian model averaging (robustness check)			Frequentist model averaging (robustness check)		
	P. mean	P. SD	PIP	Coef.	SE	p-value
Intercept	-0.043	NA	1.000	-0.030	0.026	0.245
Standard error (SE)	-0.067	0.170	0.160	-0.300	0.187	0.119
SE * No endogeneity control	-0.862	0.305	0.989	-0.952	0.231	0.000
Data characteristics						
Employment: continuous variable	-0.026	0.014	0.851	-0.025	0.008	0.001
Employment: dummy variable	0.005	0.010	0.264	0.010	0.008	0.196
Educational outcome: choices	-0.029	0.007	0.995	-0.031	0.007	0.000
Educational outcome: attainment	0.000	0.001	0.009	0.004	0.007	0.562
Self-reported education	-0.005	0.008	0.321	-0.004	0.007	0.597
Longitudinal data	0.044	0.010	0.999	0.034	0.009	0.000
Data year	0.000	0.000	0.009	0.002	0.005	0.734
Structural variation						
Male students	0.000	0.001	0.029	-0.001	0.006	0.907
Female students	0.002	0.005	0.168	0.009	0.006	0.134
Caucasian students	0.000	0.003	0.021	-0.012	0.012	0.317
Minority students	0.001	0.004	0.046	0.008	0.011	0.452
Part-time students	0.013	0.017	0.440	0.027	0.012	0.020
Secondary education	0.005	0.010	0.267	0.002	0.008	0.803
Low-intensity employment	0.016	0.013	0.678	0.019	0.007	0.008
High-intensity employment	-0.015	0.012	0.664	-0.015	0.008	0.043
United States	0.009	0.011	0.462	0.021	0.008	0.008
Germany	0.067	0.015	1.000	0.078	0.016	0.000
Estimation methods						
OLS method	0.016	0.011	0.715	-0.001	0.008	0.917
Matching method	-0.011	0.019	0.300	-0.039	0.014	0.005
DID method	-0.012	0.019	0.321	-0.041	0.013	0.001
IV method	-0.008	0.013	0.296	-0.023	0.009	0.012
Number of variables	0.000	0.001	0.040	-0.003	0.003	0.382
Ability control	-0.014	0.012	0.636	-0.013	0.008	0.090
Motivation control	0.012	0.009	0.682	0.017	0.006	0.005
Parental education control	0.000	0.002	0.036	0.008	0.006	0.160
Age control	0.000	0.002	0.038	-0.006	0.005	0.227
Ethnicity control	-0.010	0.009	0.630	-0.011	0.007	0.111
Publication characteristics						
Impact factor	0.000	0.001	0.060	-0.001	0.003	0.767
Citations	-0.001	0.002	0.106	-0.006	0.003	0.024
Published study	0.001	0.004	0.042	0.017	0.011	0.132
Studies	69			69		
Observations	861			861		

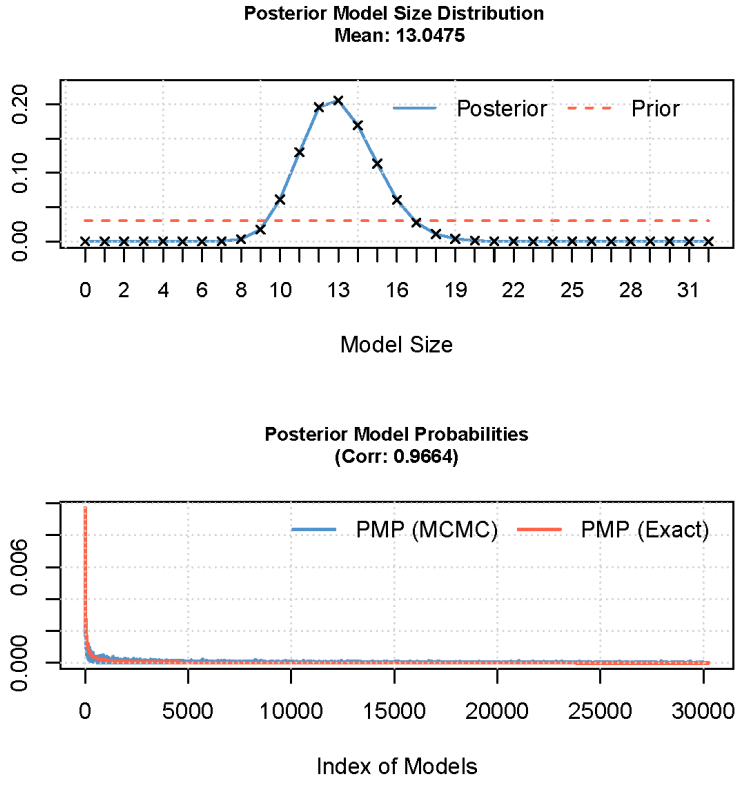
Notes: Response variable is the estimate of the effect of student employment on educational outcomes (reflected by a partial correlation coefficient). SE = standard error, P. mean = posterior mean, P. SD = posterior standard deviation, PIP = posterior inclusion probability. In the first specification from the left we employ Bayesian model averaging (BMA) using BRIC g-prior suggested by Fernandez *et al.* (2001) and the beta-binomial model prior according to Ley & Steel (2009). The specification on the right employs frequentist model averaging by applying Mallows weights Hansen (2007) using orthogonalization of the covariate space suggested by Amini & Parmeter (2012) to reduce the number of estimated models. The posterior mean in Bayesian model averaging (or alternatively the estimated coefficient in frequentist model averaging) denotes the marginal effect of a study characteristic on the partial correlation coefficient of the effect reported in the literature. For detailed description of all the variables see Table 5.

Table B3: Diagnostics of the baseline BMA estimation

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
13.0475	$3 \cdot 10^5$	$1 \cdot 10^5$	1.036119 mins	85,272
Modelspace	*Visited*	*Topmodels*	*Corr PMP*	*No. obs.*
$4.3 \cdot 10^9$	0.20%	100%	0.9664	861
Model prior	*g-prior*	*Shrinkage-stats*		
Random-dilution / 16	UIP	Av = 0.9988		

Notes: In the baseline model we employ the unit information g-prior recommended by Eicher *et al.* (2011) (the prior provides the same amount of information as one observation from the data) and the dilution prior suggested by George (2010), which accounts for collinearity.

Figure B2: Model size and convergence of the baseline BMA estimation



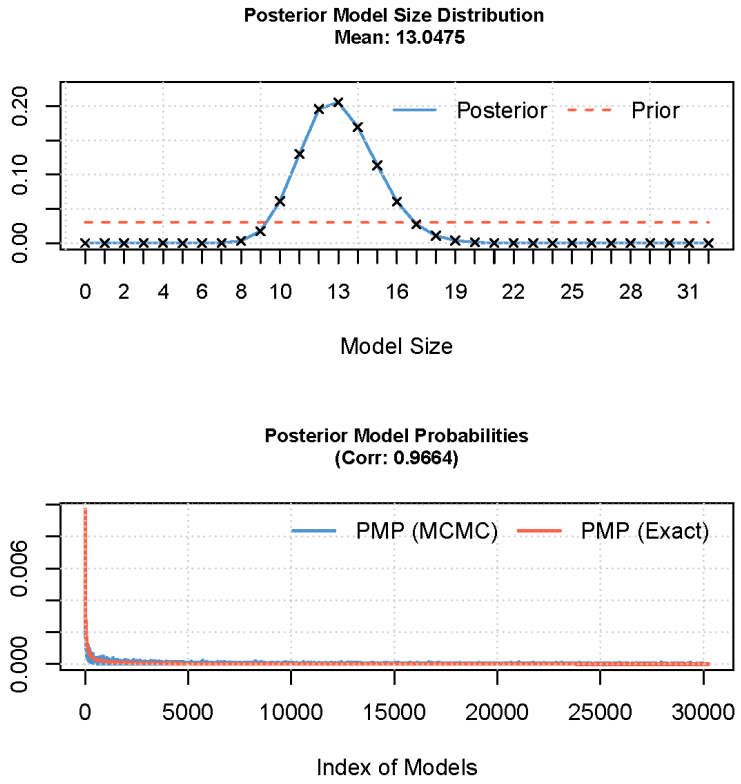
Notes: The figure depicts the posterior model size distribution and the posterior model probabilities of the BMA estimation reported in Table 6.

Table B4: Diagnostics of the BMA estimation (BRIC and random priors)

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
12.6514	$3 \cdot 10^5$	$1 \cdot 10^5$	58.25228 secs	82702
Modelspace	*Visited*	*Topmodels*	*Corr PMP*	*No. obs.*
$4.3 \cdot 10^9$	0.19%	100%	0.9678	861
Model prior	*g-prior*	*Shrinkage-stats*		
Random / 16	BRIC	Av=0.999		

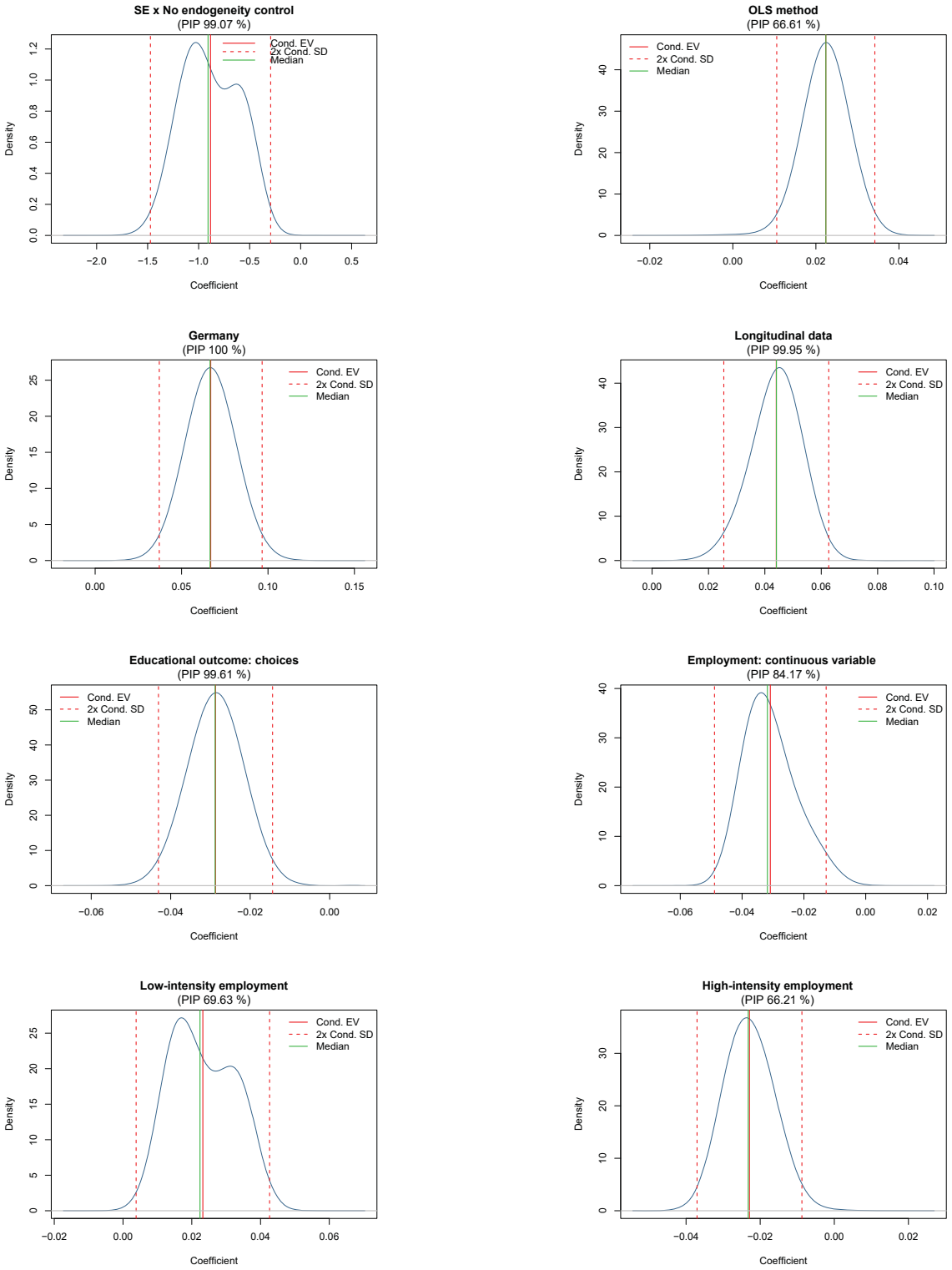
Notes: The specification uses a BRIC g-prior suggested by Fernandez *et al.* (2001) and the beta-binomial model prior according to Ley & Steel (2009).

Figure B3: Model size and convergence of the BMA estimation (BRIC and random priors)



Notes: The figure depicts the posterior model size distribution and the posterior model probabilities of the BMA estimation reported in Table B2.

Figure B4: Posterior coefficient distributions for selected variables



Notes: The figure depicts the posterior coefficient distributions of the regression coefficients corresponding to selected variables in the baseline BMA estimation. For instance, we see that the coefficient corresponding to *educational outcome: choices* is negative in all models irrespective of other variables being included or ignored.