B Web Appendix to "Student Employment and Education: A Meta-Analysis"

B.1 Theories on the Nexus Between Student Work and Education

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). Nonethe-less, 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.*"

B.2 Details on Data Transformation and Summary Statistics

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 $SE(\widehat{or}_{is}) = SE(\beta_{is})e^{\beta_{is}}$, 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 consider an interaction of 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 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.

Table B1 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 drop out 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 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.

Table B1: 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.

B.3 Details on Publication Bias Correction

For the identification of and correction for publication bias we employ the property of econometric techniques used to estimate the effect of 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 implies that 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 compare publication selection to the Lombard effect in 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. In this framework publication bias is a linear function of the standard error. Hence, 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 Panel C of Table 2 we show nonlinear alternatives to the simple test of publication bias. 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 2 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.

B.4 Details on Variables Explaining Heterogeneity

Data characteristics. Researchers use various specifications to capture 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 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 B1 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,

without proper controls for student ability and other variables, 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 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.

Students' educational outcomes could also 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.

B.5 Technical Details on Bayesian Model Averaging

When applying BMA we face two computational problems. First, computing the 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 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 the 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:

$$PCC_{is} = \gamma_0 + \gamma_1 SE(PCC)_{is} * \textit{No endogeneity control}_{is} + \gamma_2 X_i + \epsilon_{is}, \tag{A1}$$

where PCC_{is} represents the estimated partial correlation coefficient, X_{is} stands for the explanatory 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.

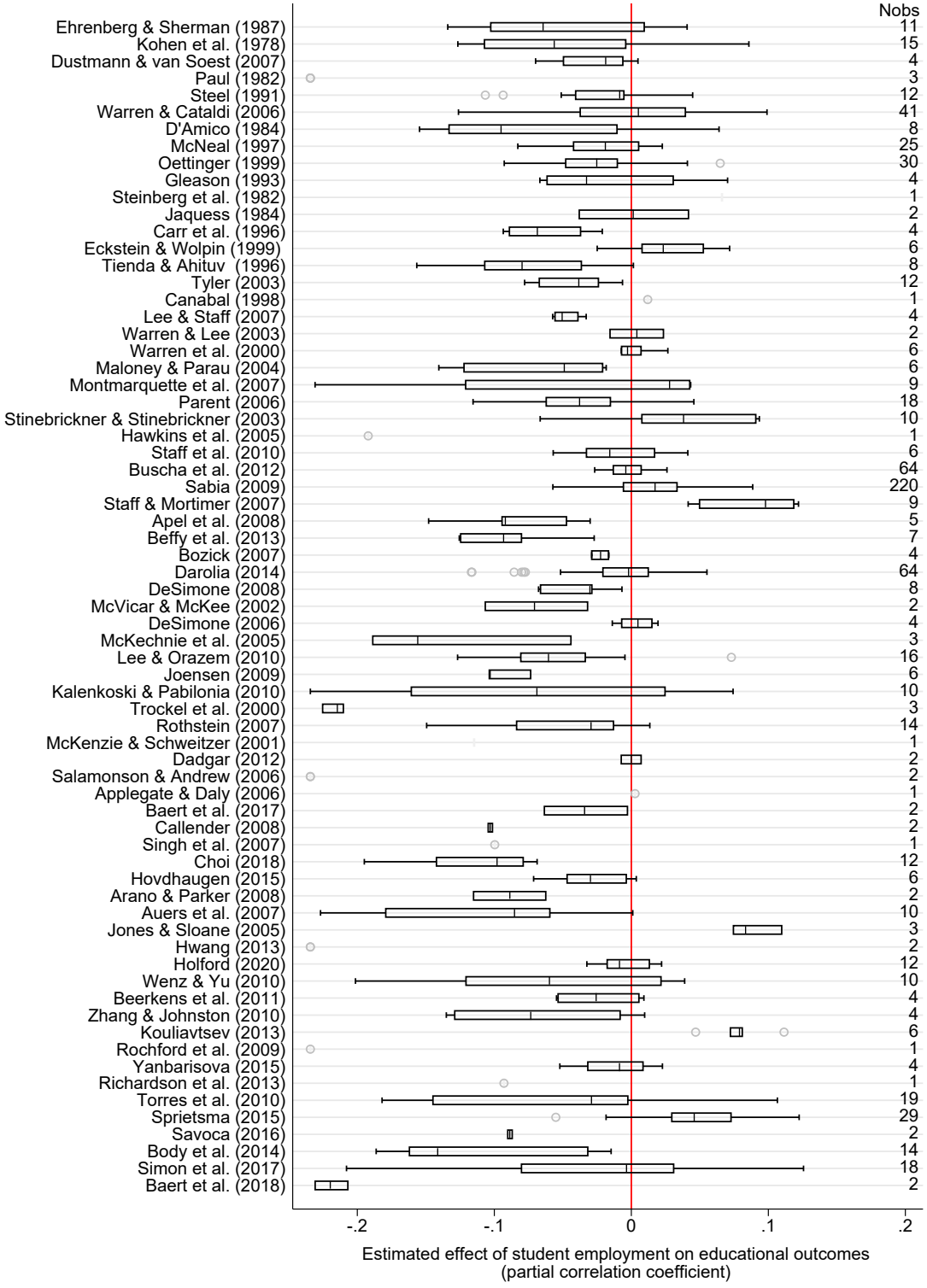
B.6 Additional Tables and Figures

Table B2: Publication bias tests, expanded sample

Panel A: Linear techniques					
	OLS	IV	Study	Precision	
Standard error	-0.892***	-0.924***	-1.102**	-0.560*	
(Publication bias)	(0.306)	(0.337)	(0.440)	(0.301)	
	[-1.589, -0.264]	[-1.714, -0.247]	[-2.337, -0.196]	[-1.188, 0.087]	
Constant	0.00602	0.00699	0.0136	-0.00287	
(Effect beyond bias)	(0.0123)	(0.0126)	(0.0176)	(0.00675)	
	[-0.022, 0.0377]	[-0.022, 0.037]	[-0.031, 0.052]	[-0.020, 0.011]	
Observations	872	872	872	872	
Panel B: Between- and within-study variation					
	BE	FE	RE		
Standard error	-1.926***	0.193	-0.411**		
(Publication bias)	(0.353)	(0.572)	(0.200)		
Constant	0.0160	-0.0230	-0.0336***		
(Effect beyond bias)	(0.0142)	(0.0153)	(0.0101)		
Observations	872	872	872		
Panel C: Nonlinear techniques					
	WAAP	Stem method	Kinked model	Selection model	p-uniform*
Effect beyond bias	0.00756	0.0100	-0.0101***	-0.0120***	-0.0292*
	(0.0136)	(0.0266)	(0.00268)	(0.005)	(0.0169)
Observations	872	872	872	872	872

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 872 estimates (for which the mean estimate equals −0.018) from the expanded sample of 73 studies (we have included four more studies not identified by our baseline search: Bekova, 2021; Kocsis, 2021; Logan *et al.*, 2016; Titus, 2010). 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.

Figure B1: Boxplot of estimates sorted by data year



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.

Table B3: Publication bias tests, sample without US estimates

Panel A: Linear techniques				
	OLS	IV	Study	Precision
Standard error	-2.016***	-2.201***	-2.634***	-1.325*
(Publication bias)	(0.691)	(0.738)	(0.728)	(0.701)
	[-3.633, -0.454]	[-3.824, -0.464]	[-4.235, -0.716]	[-3.272, -0.036]]
Constant	0.0250	0.0307	0.0475	0.00442
(Effect beyond bias)	(0.0268)	(0.0275)	(0.0338)	(0.0164)
	[-0.029, 0.090]	[-0.024, 0.106]	[-0.026, 0.129]	[-0.026, 0.057]
Observations	167	167	167	167

Panel B: Between- and within-study variation			
	BE	FE	RE
Standard error	-2.539***	2.357	-2.404***
(Publication bias)	(0.570)	(7.222)	(0.600)
Constant	0.0291	-0.105	0.0250
(Effect beyond bias)	(0.0245)	(0.215)	(0.0234)
Observations	167	167	167

Panel C: Nonlinear techniques					
	WAAP	Stem method	Kinked model	Selection model	p-uniform*
Effect beyond bias	·	-0.058**	-0.038***	-0.013***	-0.038
	·	(0.028)	(0.013)	(0.012)	(0.027)
Observations	167	167	167	167	167

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 167 non-US estimates (for which the mean estimate equals -0.065). 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.

Table B4: Tests of publication bias, hours worked and 4.0 GPA scale

[Block 1] All studies with homogenous 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

[Block 2] Studies trying to take endogeneity into account

Panel A: Linear techniques

	OLS	IV	Study	Precision
Standard error	-1.038	0.955	-1.043	-0.875
(Publication bias)	(0.900)	(0.922)	(0.968)	(0.684)
	[-2.212, 1.264]	[-2.169, 2.001]	[-12.66, 2.337]	[-2.040, 1.064]
Constant	0.0143	-0.0215	0.0351*	0.00116
(Effect beyond bias)	(0.0125)	(0.0209)	(0.0190)	(0.00507)
	[-0.006, 0.0720]	[-0.014, 0.046]	[-0.013, 0.082]	[-0.0141, 0.0149]
Observations	50	50	50	50

Panel B: Between- and within-study variation

	BE	FE	RE
Standard error	-1.741***	-1.204***	-1.150***
(Publication bias)	(0.434)	(0.348)	(0.288)
Constant	0.0177	0.0173*	0.0144
(Effect beyond bias)	(0.0154)	(0.0100)	(0.0114)
Observations	50	50	50

Panel C: Nonlinear techniques

	WAAP	Stem-based	Endogenous kink	Selection model	p-uniform*
Effect beyond bias	-0.00231***	-0.002*	-0.00204***	-0.006	-0.00670*
	(0.000288)	(0.00114)	(0.000255)	(0.004)	(0.00389)
Observations	50	50	50	50	50

Notes: See notes to Table 2. Here we only include studies that use the 4.0 GPA scale to measure educational outcomes and hours worked per week to measure student work.

Table B5: Tests of publication bias, hours worked and dropout rate

[Block 1] All studies with homogenous estimates

Panel A: Linear techniques

	OLS	IV	Study	Precision
Standard error	-0.376***	-0.821*	-0.416**	-1.013***
(Publication bias)	(0.113)	(0.473)	(0.170)	(0.363)
	[-1.772, 0.884]	.	[-1.741, -0.0893]	[-2.969, -0.363]
Constant	-0.0258*	-0.00969	-0.0293***	-0.00269**
(Effect beyond bias)	(0.0150)	(0.0207)	(0.0106)	(0.00112)
	[-0.0932, 0.0005]	.	[-0.0528, -0.0094]	[-0.0133, -0.0005]
Observations	22	22	22	22

Panel B: Between- and within-study variation

	BE	FE	RE
Standard error	-0.715*	0.595**	-0.150
(Publication bias)	(0.264)	(0.271)	(0.225)
Constant	-0.0151	-0.0610***	-0.0313**
(Effect beyond bias)	(0.0127)	(0.0125)	(0.0152)
Observations	22	22	22

Panel C: Nonlinear techniques

	WAAP	Stem-based	Endogenous kink	Selection model	p-uniform*
Effect beyond bias	-0.00176***	0.002	-0.000789	-0.004	-0.0105
	(0.000607)	(0.00152)	(0.000544)	(0.010)	(0.0345)
Observations	22	22	22	22	22

[Block 2] Studies trying to take endogeneity into account

Panel A: Linear techniques

	OLS	IV	Study	Precision
Standard error	-0.349***	-0.811	-0.385**	-0.984***
(Publication bias)	(0.126)	(0.512)	(0.168)	(0.363)
	[-1.464, 0.615]	.	[-1.603, -0.0744]	[-3.032, -0.347]
Constant	-0.0293*	-0.0109	-0.0333***	-0.00398
(Effect beyond bias)	(0.0170)	(0.0241)	(0.0121)	(0.00261)
	[-0.094, 0.00220]	.	[-0.0625, -0.00946]	[-0.0189, 0.000142]
Observations	20	20	20	20

Panel B: Between- and within-study variation

	BE	FE	RE
Standard error	-0.702*	0.595*	-0.306
(Publication bias)	(0.260)	(0.290)	(0.210)
Constant	-0.0162	-0.0669***	-0.0305**
(Effect beyond bias)	(0.0125)	(0.0145)	(0.0148)
Observations	20	20	20

Panel C: Nonlinear techniques

	WAAP	Stem-based	Endogenous kink	Selection model	p-uniform*
Effect beyond bias	-0.00122	-0.00375	0.0008	-0.0110	-0.0121
	(0.00122)	(0.00446)	(0.001)	(0.0100)	(0.0143)
Observations	20	20	20	20	20

Notes: See notes to Table 2. Here we only include studies that use the dropout rate to measure educational outcomes and hours worked per week to measure student work.

Table B6: Tests of publication bias, employment dummy and dropout rate

[Block 1] All studies with homogenous estimates

Panel A: Linear techniques

	OLS	IV	Study	Precision
Standard error	-0.0212	-0.0385	-0.0855***	-0.0320
(Publication bias)	(0.0496)	(0.0968)	(0.0314)	(0.170)
	[-0.559, 0.774]	.	[-0.157, 0.295]	[-0.395, 0.633]
Constant	-0.160*	-0.150	-0.0932	-0.154**
(Effect beyond bias)	(0.0837)	(0.104)	(0.118)	(0.0696)
	[-1.113, 1.477]	.	[-0.320, 0.144]	[-0.391, -0.020]
Observations	37	37	37	37

Panel B: Between- and within-study variation

	BE	FE	RE
Standard error	0.224	-0.412***	-0.227*
(Publication bias)	(0.230)	(0.144)	(0.125)
Constant	-0.515	0.0655	-0.0486
(Effect beyond bias)	(0.304)	(0.115)	(0.264)
Observations	37	37	37

Panel C: Nonlinear techniques

	WAAP	Stem-based	Endogenous kink	Selection model	p-uniform*
Effect beyond bias	-0.129***	-0.0644	0.0211	-0.305***	-0.220
	(0.0418)	(0.118)	(0.0272)	(0.064)	(0.631)
Observations	37	37	37	37	37

[Block 2] Studies trying to take endogeneity into account

Panel A: Linear techniques

	OLS	IV	Study	Precision
Standard error	-0.0287	-0.0423	-0.0946***	-0.00390
(Publication bias)	(0.0539)	(0.0946)	(0.0302)	(0.165)
	[-0.531, 0.808]	.	[-0.158, 0.251]	[-0.362, 0.738]
Constant	-0.132	-0.124	-0.0590	-0.146**
(Effect beyond bias)	(0.0912)	(0.103)	(0.118)	(0.0676)
	[-0.218, 1.483]	.	[-0.283, 0.184]	[-0.362, -0.0168]
Observations	36	36	36	36

Panel B: Between- and within-study variation

	BE	FE	RE
Standard error	0.119*	-0.412***	-0.0287
(Publication bias)	(0.0100)	(0.144)	(0.0836)
Constant	-0.224**	0.0933	-0.132
(Effect beyond bias)	(0.0152)	(0.117)	(0.102)
Observations	36	36	36

Panel C: Nonlinear techniques

	WAAP	Stem-based	Endogenous kink	Selection model	p-uniform*
Effect beyond bias	-0.129***	-0.0642	0.0211	-0.284***	-0.0643
	(0.0418)	(0.115)	(0.0272)	(0.062)	(0.727)
Observations	36	36	36	36	36

Notes: See notes to Table 2. Here we only include studies that use the dropout rate to measure educational outcomes and a 0/1 indicator to measure student work.

Figure B2: Correlations between regression variables

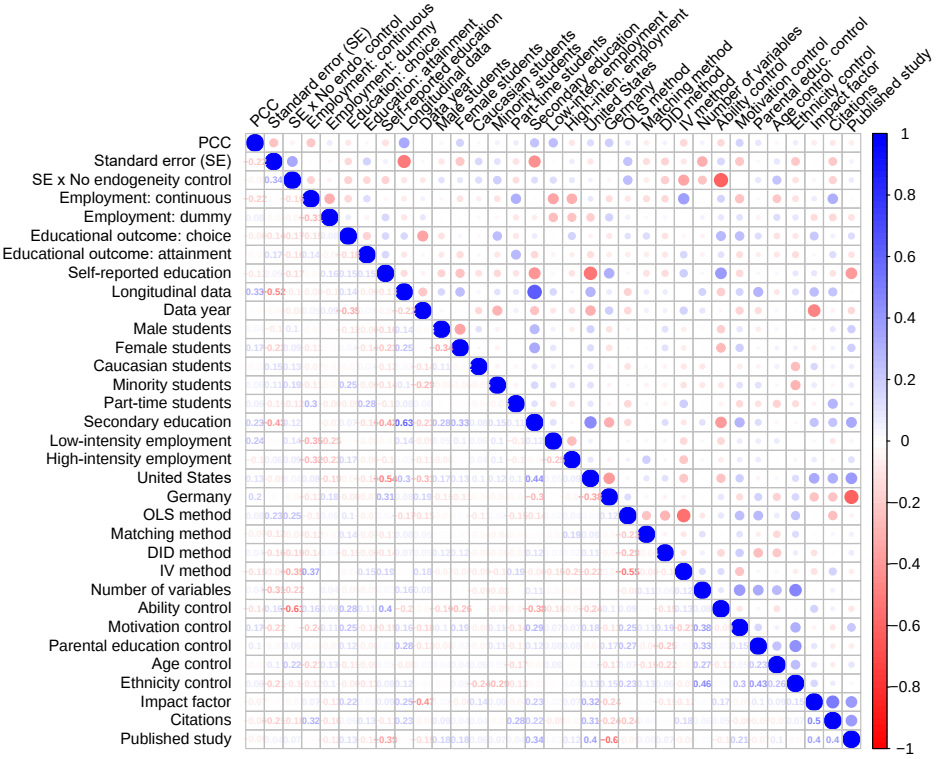
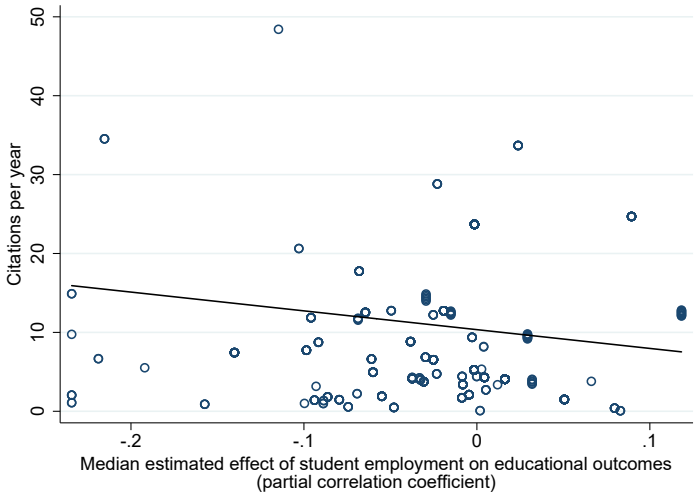


Figure B3: Estimate size and citations received



Notes: The figure shows citations per year against the median effect reported in each study. Extreme outliers are excluded from the figure but included in all statistical tests.

Table B7: 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		

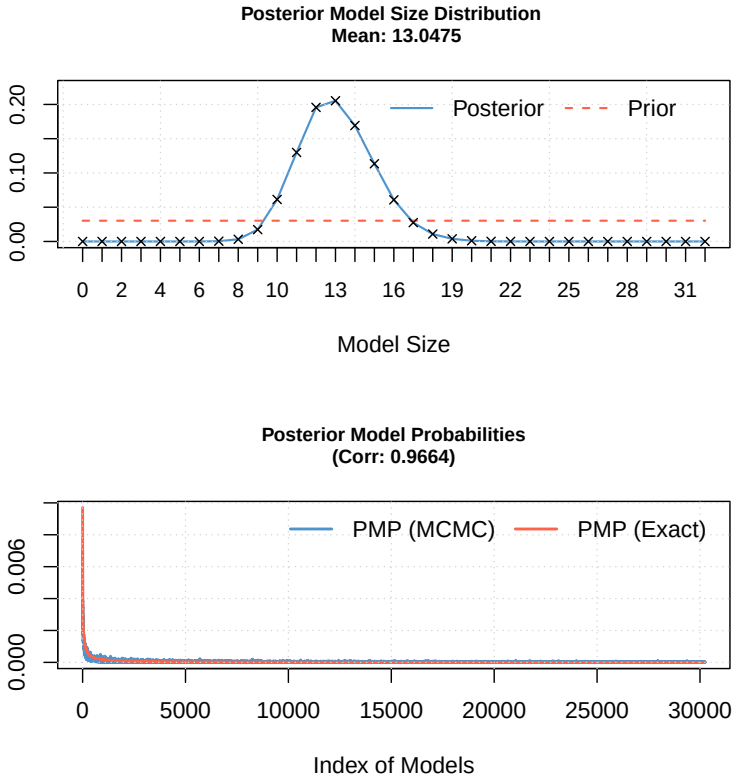
Notes: Response variable is the estimate of the effect of student employment on educational outcomes (reflected by a partial correlation coefficient). SE = standard error, P. mean = posterior mean, P. SD = posterior standard deviation, PIP = posterior inclusion probability. In the first specification 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 4.

Table B8: 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 B4: Model size and convergence of the baseline BMA estimation



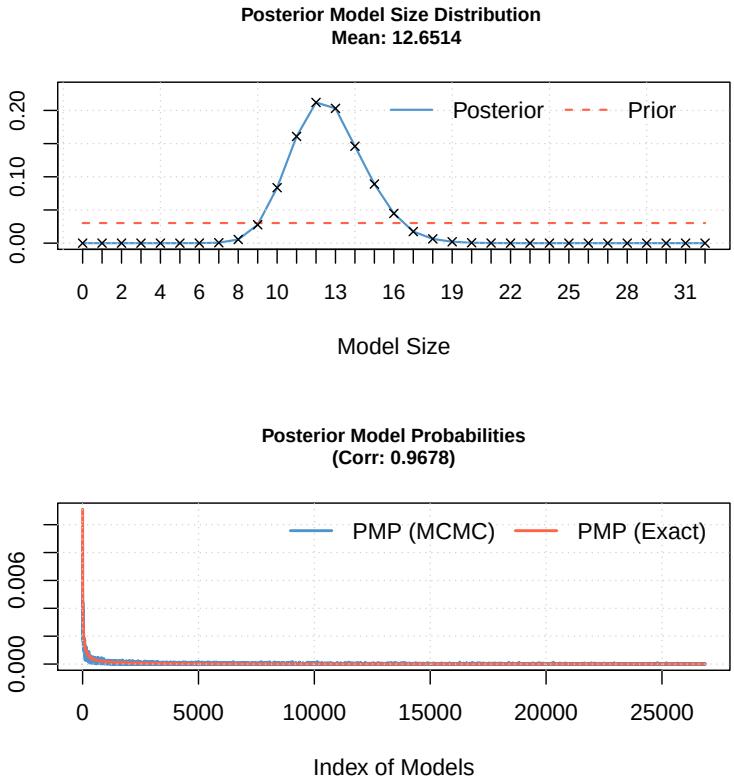
Notes: The figure depicts the posterior model size distribution and the posterior model probabilities of the BMA estimation reported in Table 5.

Table B9: 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		

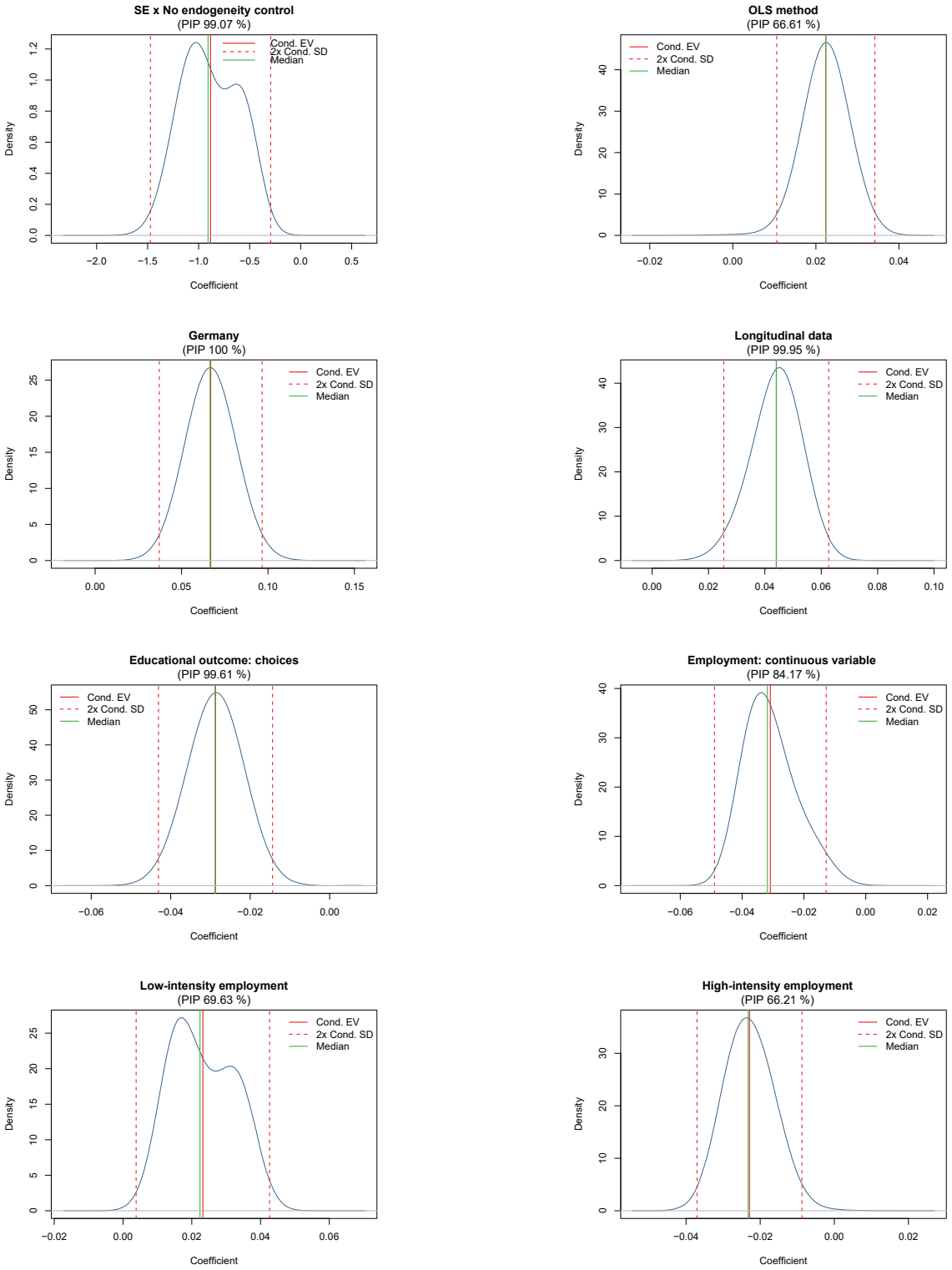
Notes: The specification uses a BRIC g-prior suggested by Fernandez *et al.* (2001) and the beta-binomial model prior according to Ley & Steel (2009).

Figure B5: 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 B7.

Figure B6: 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.

Table B10: Explaining heterogeneity; publication characteristics excluded

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	-0.044	NA	1.000	-0.018	0.024	0.450
Standard error (SE)	-0.055	0.153	0.141	-0.226	0.182	0.216
SE * No endogeneity control	-0.931	0.294	0.995	-0.997	0.230	0.000
Data characteristics						
Employment: continuous variable	-0.027	0.013	0.879	-0.030	0.008	0.000
Employment: dummy variable	0.005	0.009	0.249	0.009	0.008	0.225
Educational outcome: choices	-0.029	0.007	0.997	-0.033	0.007	0.000
Educational outcome: attainment	0.000	0.001	0.014	0.000	0.006	0.987
Self-reported education	-0.005	0.008	0.325	-0.006	0.007	0.381
Longitudinal data	0.044	0.009	0.999	0.029	0.009	0.001
Data year	0.000	0.001	0.013	0.001	0.005	0.790
Structural variation						
Male students	0.000	0.002	0.038	-0.001	0.006	0.931
Female students	0.002	0.005	0.199	0.010	0.006	0.102
Caucasian students	0.000	0.003	0.032	-0.015	0.012	0.204
Minority students	0.001	0.004	0.052	0.006	0.011	0.604
Part-time students	0.014	0.017	0.464	0.025	0.011	0.025
Secondary education	0.005	0.009	0.245	0.004	0.008	0.644
Low-intensity employment	0.015	0.013	0.678	0.018	0.007	0.011
High-intensity employment	-0.016	0.012	0.710	-0.016	0.008	0.034
On-campus employment United States	0.007	0.010	0.397	0.018	0.008	0.018
Germany	0.066	0.015	1.000	0.072	0.015	0.000
Estimation methods						
OLS method	0.021	0.012	0.793	0.002	0.008	0.823
Matching method	-0.009	0.017	0.233	-0.036	0.014	0.010
DID method	-0.009	0.017	0.260	-0.039	0.013	0.002
IV method	-0.006	0.011	0.224	-0.022	0.009	0.015
Number of variables	0.000	0.001	0.052	-0.003	0.003	0.365
Ability control	-0.015	0.011	0.722	-0.014	0.007	0.055
Motivation control	0.014	0.009	0.784	0.020	0.006	0.001
Parental education control	0.000	0.002	0.054	0.010	0.006	0.071
Age control	0.000	0.002	0.051	-0.006	0.005	0.257
Ethnicity control	-0.011	0.009	0.698	-0.015	0.007	0.020
Studies	69			69		
Observations	861			861		

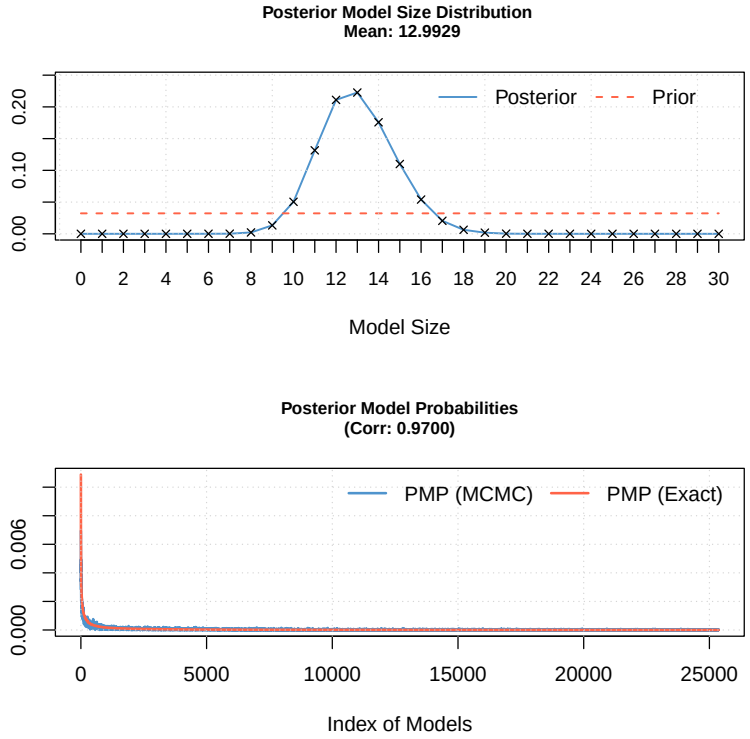
Notes: Response variable is the estimate of the effect of student employment on educational outcomes (reflected by a partial correlation coefficient). SE = standard error, P. mean = posterior mean, P. SD = posterior standard deviation, PIP = posterior inclusion probability. In the first specification 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 4.

Table B11: Diagnostics of BMA estimation; publication characteristics excluded

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
12.9929	$3 \cdot 10^5$	$1 \cdot 10^5$	2.850391 mins	88,312
Modelspace	*Visited*	*Topmodels*	*Corr PMP*	*No. obs.*
$1.1 \cdot 10^9$	0.82%	100%	0.9700	861
Model prior	*g-prior*	*Shrinkage-stats*		
Random-dilution / 15	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 B7: Model size and convergence of BMA estimation; publication characteristics excluded



Notes: The figure depicts the posterior model size distribution and the posterior model probabilities of the BMA estimation reported in Table B10.