Opatrny, Matej; Havranek, Tomas; Irsova, Zuzana; Scasny, Milan

Working Paper

Publication Bias and Model Uncertainty in Measuring the Effect of Class Size on Achievement

Publication Bias and Model Uncertainty in Measuring the Effect of Class Size on Achievement*

Matej Opatrny[a], Tomas Havranek[a,b], Zuzana Irsova[a,c] , Milan Scasny[a]

[a]Charles University, Prague

[b]Centre for Economic Policy Research, London

[c]Anglo-American University, Prague

May 24, 2023

Abstract

Class size reduction mandates are frequent and invariably justified by studies reporting positive effects on student achievement. Yet other studies report no effects, and the literature as a whole awaits correction for potential publication bias. Moreover, if identification drives results systematically, the relevance of individual studies will vary. We build a sample of 1,767 estimates collected from 62 studies and for each estimate codify 42 factors reflecting estimation context. We employ recently developed nonlinear techniques for publication bias correction and Bayesian model averaging techniques that address model uncertainty. The results suggest publication bias among studies featured in top five economics journals, but not elsewhere. The implied class size effect is zero for all identification approaches except Tennessee's Student/Teacher Achievement Ratio project. The effect remains zero for disadvantaged students and across subjects, school types, and countries.

Keywords: Class size, student learning, meta-analysis, publication bias, Bayesian model averaging

JEL Codes: C83, H52, I21

*An online appendix with data and code is available at `meta-analysis.cz/class`. Corresponding author: Zuzana Irsova, `zuzana.irsova@ies-prague.org`.

1 Introduction

Since 2010, at least 17 jurisdictions have mandated or incentivized class size reductions in countries including Australia, Canada, Finland, France, Germany, India, Israel, New Zealand, Norway, Portugal, South Korea, Spain, the United Kingdom, and the United States (Table B1 in Appendix B). Prior to 2010, at least 24 US states had started to mandate or incentivize reductions (Whitehurst & Chingos, 2011). The policy is universally popular amongst parents and teachers. According to one survey, 90% of American teachers believe that smaller classes can "strongly" or "very strongly" improve student learning (Scholastic, 2012). Aside from robust intuition, reduction mandates claim justification in empirical evidence. For example, the legislation mandating significant class size reductions in New York City starting in September 2023 includes the following rationale:

> *Studies have shown that students learn faster and perform better in smaller classes.*
> (New York State Senate, 2022)

We show that the claim is inconsistent with the bulk of empirical evidence. The implied class size effect is close to zero across methods, students, schools, and jurisdictions. Even disadvantaged students benefit little from class size reductions—there is no systematic evidence suggesting otherwise. Different identification approaches, in general, do not bring systematically different results. Yet the prevailing public impression, expressed in Wikipedia entries, ChatGPT replies, and legislative justifications, is that empirical research shows benefits of reductions, at least for some students. The impression is to a large extent driven by two influential, high-quality studies: Angrist & Lavy (1999) and Krueger (1999). Together, they have attracted more than 5,000 citations in Google Scholar. But the two studies are not corroborated by the rest of the literature, including recent contributions by Angrist *et al.* (2017) and Angrist *et al.* (2019). We document that the zero finding is a robust feature of current data and methods.

Our main contribution is twofold. First, we take into account publication bias. Meta-analyses of the class size effect are not rare: indeed, one was conducted by the founding father of the method, Gene Glass, soon after he coined the term "meta-analysis" (Glass & Smith, 1979). But no meta-analysis has attempted to correct the literature for publication bias or p-hacking, although such selective reporting in economics routinely exaggerates typical reported estimates

by a factor of 2 or more (Ioannidis *et al.*, 2017). We use recently developed techniques for publication bias and p-hacking correction. Second, we explicitly address model uncertainty both in meta-analysis and the underlying literature. Existing meta-analyses either give equal weight to each estimate (Hanushek, 1997, 1999) or each study (Mishel & Rothstein, 2002; Krueger, 2003), assign weights proportional to reported precision (Hedges & Stock, 1983; Greenwald *et al.*, 1996; Nye *et al.*, 2002), or restrict their analysis to a handful of estimates they deem particularly reliable (10 studies in the case of Filges *et al.*, 2018). We collect 42 factors that capture estimation context and, using Bayesian and frequentist model averaging, connect them to differences in reported results.

Publication bias, stemming from the preference of editors, referees, and authors for intuitive and significant results, is particularly threatening in class size research. Intuition provides a clear prediction: smaller classes should improve student learning or, at the very least, not be detrimental. Doucouliagos & Stanley (2013) show that fields with a strong underlying intuition tend to suffer more from the bias. The debate concerning class size effects has been heated and sometimes personal (Mishel & Rothstein, 2002). Several high-quality recent papers document the extent of the publication bias problem in economics, often in areas with fewer ex ante reasons to expect bias (Andrews & Kasy, 2019; Blanco-Perez & Brodeur, 2020; Brown *et al.*, 2023; Card *et al.*, 2018; DellaVigna & Linos, 2022; Elliott *et al.*, 2022; Imai *et al.*, 2021; Iwasaki, 2022; Neisser, 2021; Stanley *et al.*, 2021; Ugur *et al.*, 2020; Xue *et al.*, 2020). It is therefore all the more remarkable that we find little publication bias in the class size literature. The overall research record in the field is surprisingly undistorted. The significant exception is studies published in top five economics journals, where optimistic results concerning class size reductions are published too often, even holding identification approaches constant.

Publication bias is sometimes distinguished from p-hacking. In this narrower definition, publication bias denotes the decision (editors', referees', or authors') to publish or suppress the results, which are individually unbiased. P-hacking, then, denotes the intentional or unintentional effort of authors to produce desirable results, typically those that are intuitive and statistically significant. Under p-hacking, even individual estimates can be biased. Both phenomena give rise to a correlation between estimates and standard errors, which should otherwise be zero. But each phenomenon has a different solution. For example, selection models, long used

in meta-analysis to correct for publication bias, assume that estimates are individually unbiased (Mathur, 2022)—these models compute the relative publication probability of significant and insignificant results and then re-weight the estimates (Hedges, 1992; Andrews & Kasy, 2019). In addition, these models are weighted by inverse variance, which creates a bias if standard errors are underestimated due to p-hacking. Unfortunately, publication bias and p-hacking are observationally equivalent in applied meta-analysis. For the sake of parsimony, we use the term "publication bias" in place of "publication bias and/or p-hacking," reserving the term p-hacking for when it is necessary to distinguish it from publication bias.

Novel meta-analysis techniques can accommodate some forms of p-hacking. Irsova *et al.* (2023) develop the meta-analysis instrumental variable estimator (MAIVE), which builds on funnel plot models in the tradition of Egger *et al.* (1997), Stanley (2005), and Stanley (2008). Classical funnel plot techniques seek to recover the estimate conditional on maximum precision. That is, these models allow for p-hacking on point estimates. McCloskey & Ziliak (2019) provide a useful analogy to the Lombard effect in psychoacoustics: speakers increase their vocal effort in response to noise. In a similar vein, researchers can respond to noise in their data (imprecision) by more effort (search over specifications) in order to produce large point estimates and reach statistical significance. But standard errors are assumed to be given to the researcher and cannot be manipulated, consciously or unconsciously. The assumption is unlikely to hold in observational research. The corresponding analogy is Taylor's law in ecology: variance decreases with a smaller mean (originally describing population density for various species, Taylor, 1961). Some researchers may be tempted, for example, to use less conservative standard errors when their estimates are small. By exploiting the statistical relationship between the standard error and sample size, Irsova *et al.* (2023) show in simulations that using the latter as an instrument for the former addresses most forms of p-hacking as well as method heterogeneity that can produce correlation between estimates and standard errors in the absence of selection.

The class size research as a whole is unbiased. The finding, which is rare in economics, is supported by the rigorously founded selection model due to Andrews & Kasy (2019), the simplified selection model (p-uniform*) due to van Aert & van Assen (2021), the endogenous kink model due to Bom & Rachinger (2019), the weighted average of adequately powered estimates (WAAP) model due to Ioannidis *et al.* (2017), the stem-based model due to Furukawa (2021),

the instrumental MAIVE estimator due to Irsova *et al.* (2023), as well as classical funnel-based meta-regression techniques with different weights and study-level fixed effects (Stanley, 2005; Stanley & Doucouliagos, 2014). In contrast, we find evidence of publication bias among studies published in top five journals. The bias is not strong, but suffices to shrink the implied high-published effect of class size reductions to an economically and statistically insignificant value: the largest corrected effect across all the techniques for top five journals corresponds to a 0.035 standard-deviation increase in test scores after a class size reduction of 10 students, about a tenth of the largest estimate reported by Krueger (1999).

The unconditional meta-analysis mean can be misleading if different identification approaches lead to systematically different results. Because variation in class size is generally far from random, the choice of an identification approach matters in principle. Empirical studies use five main approaches: i) ordinary least squares with controls, ii) student or class fixed effects (e.g., Chingos, 2012; Lindahl, 2005), iii) instrumental variables with, for example, enrollment or population used as instruments for class size (Borland *et al.*, 2005; Hoxby, 2000), iv) regression discontinuity design using jurisdiction-level limits on class size (Angrist *et al.*, 2017; Urquiola & Verhoogen, 2009), and v) experiments (Krueger, 1999; Shin & Raudenbush, 2011). The first approach is unlikely to succeed in recovering the causal estimate, and researchers typically use OLS only to show what happens if they ignore endogeneity. The class size literature has been an important laboratory of the credibility revolution in empirical economics: the canonical application of regression discontinuity design is due to Angrist & Lavy (1999), and the large-scale Tennessee's Student/Teacher Achievement Ratio (STAR) experiment (Krueger, 1999) helped propel the drive in economics towards randomized controlled trials.

Aside from analyzing these five groups of studies separately, we also take into account the broader issue of model uncertainty in estimation. Researchers make numerous data and method choices at various stages: we collect 42 factors that reflect the context in which researchers obtain their estimates. We then connect these 42 factors to the observed differences in reported class size effects. As the baseline technique, we employ Bayesian model averaging (Steel, 2020), which constitutes the natural response to model uncertainty in the Bayesian framework. To account for collinearity we use the dilution prior due to George (2010). We also report the results of frequentist model averaging with Mallows' weights (Hansen, 2007) using the orthogonalization

of covariate space due to Amini & Parmeter (2012). As the bottom line of our analysis, we use the Bayesian model averaging results to construct a hypothetical ideal study and compute implied estimates of the class size effect for various estimation contexts.

The results suggest little systematic dependence of reported effects on estimation design. Among the five basic identification approaches, four deliver class size effects robustly close to zero. The only exception is the STAR experiment, where even after correction for potential publication bias and other issues we find a mean effect almost of the size reported in the influential study by Krueger (1999). One possible interpretation is that the STAR experiment data are qualitatively superior to other studies and so the corresponding evidence is the only reliable one. But the rest of the literature includes high-quality studies with eminently plausible identification approaches, especially when regression discontinuity is used, and covers many countries and types of schools. After dozens of attempts, the literature has been unable to replicate the results of the STAR experiment—which are, as we document, not driven by publication bias. The most convincing explanation is that randomization failed in the STAR experiment, and we briefly comment on that issue in the Conclusion.

The remainder of the paper is structured as follows. Section 2 describes the dataset of class size effects. Section 3 investigates publication bias. Section 4 examines model uncertainty. Section 5 concludes the paper. Appendix A gives details on how we select studies for inclusion in the meta-analysis. Appendix B provides additional details on the data set and robustness checks (eventually for online publication). The web appendix at at `meta-analysis.cz/class` features data and codes for R and Stata.

2 Data

To search for studies reporting empirical estimates of the effect of class size on student achievement, we use Google Scholar because of its universal coverage and ability to inspect the full text of studies, not only the title, abstract, and keywords. Appendix A reports details on our search strategy. We read the abstracts of the first 500 studies identified by the Google Scholar query and download those that show any promise of containing estimates of the class size effect. There are 216 such studies, and we record their references. Next, we go through the 100 studies most frequently cited among the 216 ones identified in the previous stage. This addi-

tional step, which is intended to capture important studies potentially omitted by the Google Scholar search, yields additional 26 papers that may provide estimates of the class size effect. Next, we skim the full text of the 242 prospective studies. The ones that could be included in meta-analysis are listed in Table 1.

Table 1: Studies included in the meta-analysis

Akerhielm (1995)	Etim *et al.* (2020)	Li & Konstantopoulos (2017)
Angrist & Lavy (1999)	Francis & Barnett (2019)	Lindahl (2005)
Angrist *et al.* (2017)	Fredriksson *et al.* (2013)	McKee *et al.* (2015)
Angrist *et al.* (2019)	Gerritsen *et al.* (2017)	Milesi & Gamoran (2006)
Arias & Walker (2004)	Gottfried (2014)	Nandrup (2016)
Asadullah (2005)	Heinesen (2010)	Rivkin *et al.* (2005)
Babcock & Betts (2009)	Hojo & Oshio (2012)	Sandy & Duncan (2010)
Bandiera *et al.* (2010)	Hojo (2013)	Shen & Konstantopoulos (2017)
Becker & Powers (2001)	Hojo & Senoh (2019)	Shen & Konstantopoulos (2021)
Bonesronning (2003)	Hoxby (2000)	Shen & Konstantopoulos (2022)
Boozer & Rouse (2001)	Jakubowski & Sakowski (2006)	Shin & Raudenbush (2011)
Borland *et al.* (2005)	Jepsen & Rivkin (2009)	Sims (2008)
Bosworth (2014)	Kara *et al.* (2021)	Sims (2009)
Bressoux *et al.* (2009)	Kedagni *et al.* (2021)	Suryadarma *et al.* (2006)
Browning & Heinesen (2007)	Kennedy & Siegfried (1997)	Urquiola (2006)
Bruhwiler & Blatchford (2011)	Kokkelenberg *et al.* (2008)	Urquiola & Verhoogen (2009)
Chetty *et al.* (2011)	Konstantopoulos & Shen (2016)	Vaag Iversen & Bonesronning (2013)
Chingos (2012)	Krueger (1999)	Woessmann (2005b)
Cho *et al.* (2012)	Leuven *et al.* (2008)	Woessmann & West (2006)
Dobbelsteen *et al.* (2002)	Leuven & Ronning (2016)	Woessmann (2005a)
Engin-Demir (2009)	Levin (2001)	

Notes: Details on the literature search, which was terminated on February 1, 2023, are shown in Appendix A. The dataset, together with R and Stata codes, is available at meta-analysis.cz/class.

We impose three inclusion criteria. First, the study must report an estimated relationship between test scores (not other measures of performance) and a continuous measure of class size (not a dummy variable for a "small class"). Second, the study must report standard errors or other statistics from which standard errors can be computed. Third, the study must report the standard deviations of test scores so that we can convert all estimates to a common metric. For the common metric we choose the change in the percentage points of the standard deviations of test scores corresponding to an increase in class size by one student. That is, an estimate of -1 in our dataset means that a class size reduction by 10 students is associated with an improvement in test scores by 0.1 standard deviations. In total, 46 studies comply with the three aforementioned inclusion criteria. For a robustness check, we also include additional 16 studies that comply with the first two but not the third criterion; in that case we recompute the reported effects to partial correlation coefficients. Because treatment and control class sizes vary

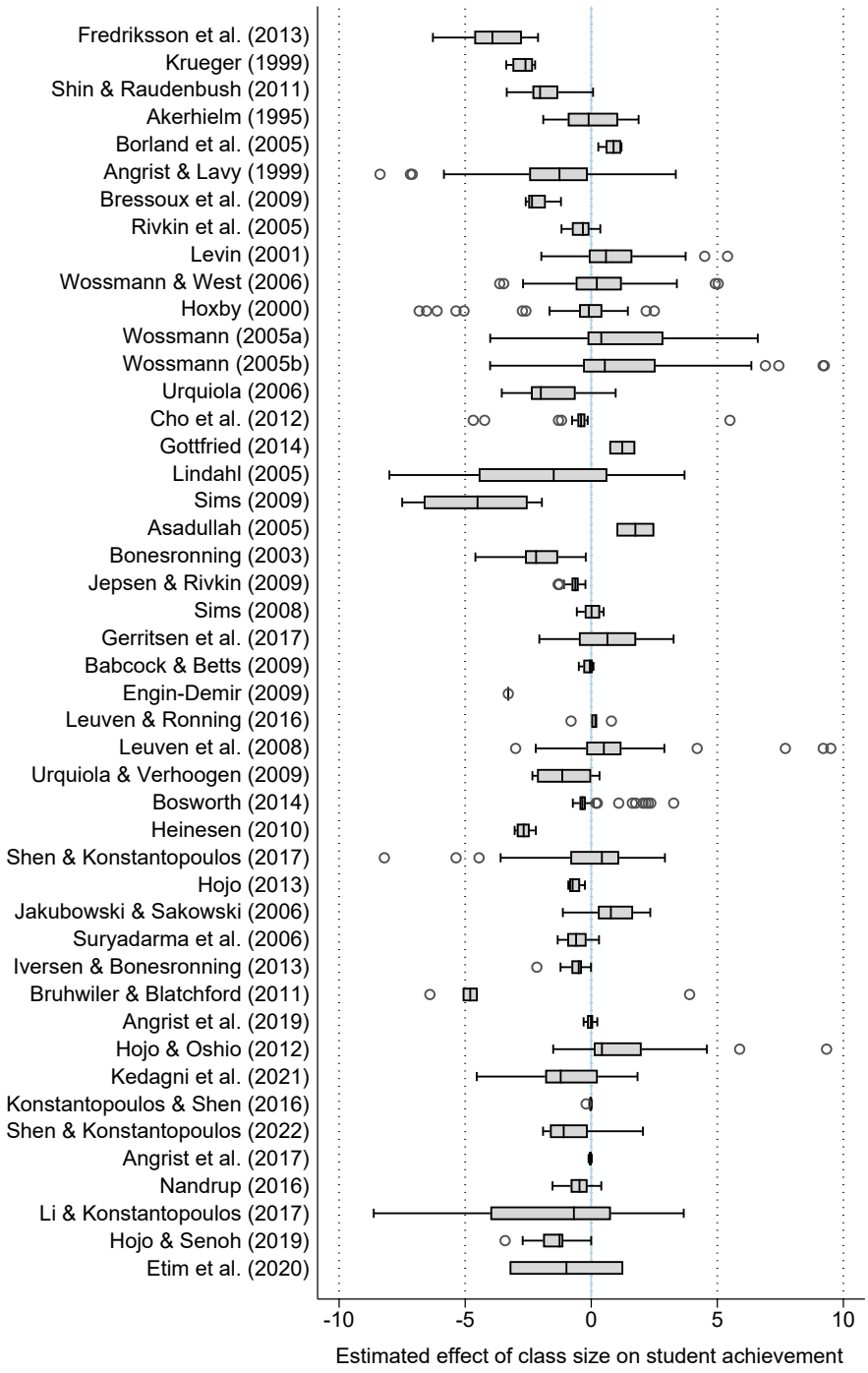
across studies and especially across countries (see Figure B1 and Figure B2 in Appendix B), for another robustness check we recompute the effects to represent a change in the percentage points of the standard deviations of test scores corresponding to an increase in class size by one standard deviation (Table B3 in Appendix B). The robustness checks provide results qualitatively similar to those of our main analysis.

In total, we gather 1,767 estimates of the class size effect reported in 62 primary studies. For each estimate we collect the standard error and 42 factors that reflect the context in which the estimate is obtained: subjects tested; the characteristics of students, schools, and jurisdictions; estimation characteristics; and publication characteristics. Despite recent advances in large language models, the data collection process for meta-analysis cannot be automated or delegated to research assistants. So, two of the co-authors of this paper collected the required tens of thousands of data points by hand after reading the 62 primary studies in detail. Then they compared their datasets and corrected typos and other mistakes. The final clean dataset, together with codes in R and Stata, is available in an online appendix at meta-analysis.cz/class.

Figure 1 shows the box plot of studies satisfying all three inclusion criteria. The studies are sorted by the age of the data from oldest to youngest. Three observations stand out. First, there is no apparent time trend in the reported estimates. Studies using recent data do not seem to report results systematically different from older studies. Second, within-study variation in results is large and often larger than variation in mean results across studies. This second observation highlights the importance of collecting all estimates from the literature, not just one representative estimate per study. Third, with a few exceptions, the central estimates of individual studies tend to cluster around negative values close to zero. Note that effects smaller than -1 in absolute value are relatively small in economic terms because they imply less than a 0.1 standard-deviation improvement in test scores following a class size reduction by 10 students. An analogous box plot of countries instead of studies (Figure B3 in Appendix B) gives a similar intuition concerning the prevalence of small effects.

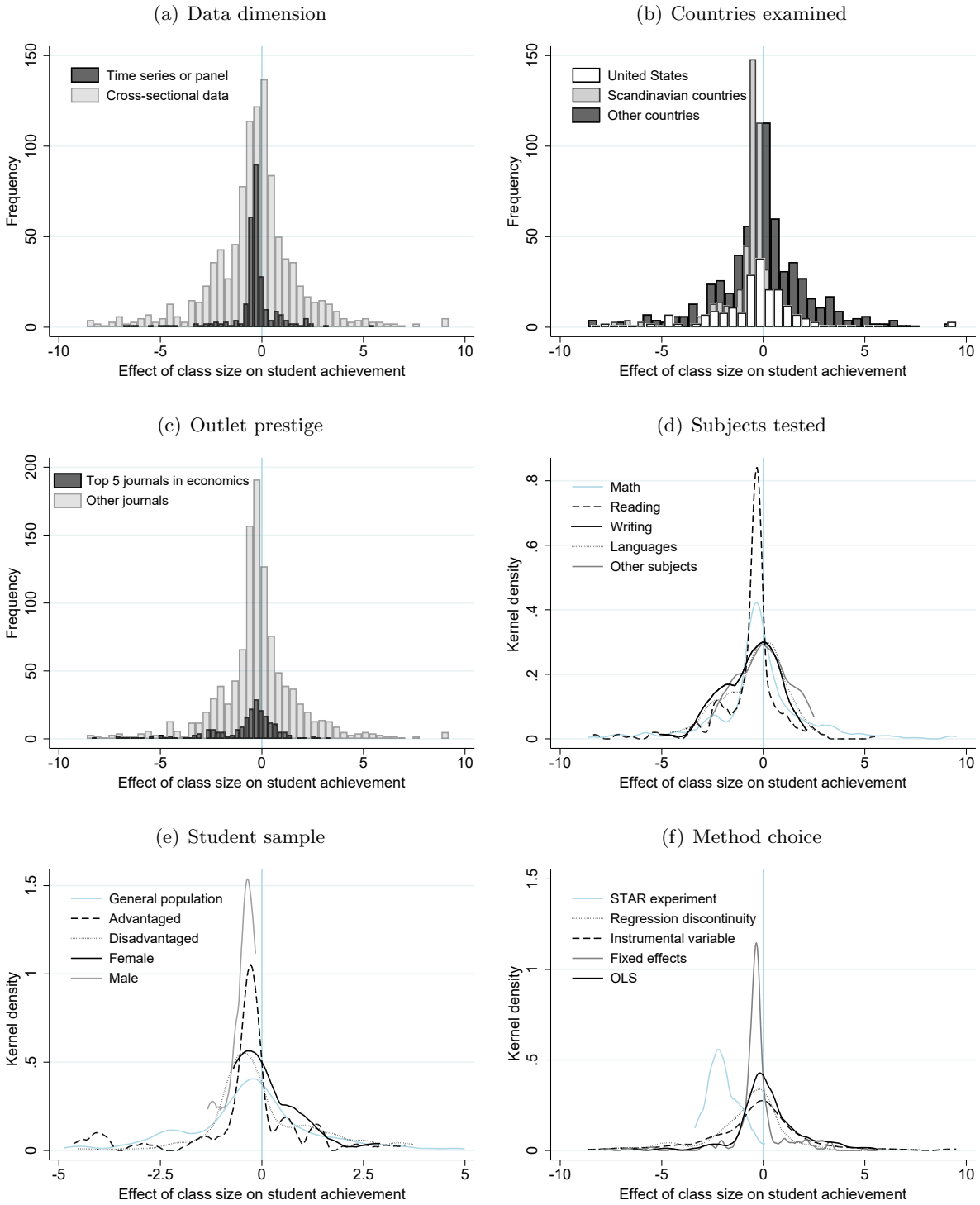
Figure 2 provides a bird's-eye view of the potential sources of systematic heterogeneity in the literature. Nevertheless, little heterogeneity is apparent at first sight. The one difference that stands out is the substantially larger negative effect reported in studies focusing on the STAR experiment compared to all other identification approaches. Regression discontinuity,

Figure 1: Estimates vary widely within and across studies, often cover zero



Notes: The figure shows a box plot of the estimated effects of class size on achievement. The effects are normalized to represent a change in the percentage points of the standard deviations of test scores corresponding to an increase in class size by one student. That is, an estimate of -1 means that a class size reduction by 10 students is associated with an improvement in test scores by 0.1 standard deviations. The studies are sorted by the age of the data 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. Extreme outliers are excluded from the figure for ease of exposition but included in all statistical tests.

Figure 2: Little prima facie systematic heterogeneity



Notes: The figure depicts, for different subsets of data, histograms of the estimated effects of class size on achievement. The effects are normalized to represent a change in the percentage points of the standard deviations of test scores corresponding to an increase in class size by one student. That is, an estimate of -1 means that a class size reduction by 10 students is associated with an improvement in test scores by 0.1 standard deviations. Extreme outliers are excluded from the figure for ease of exposition but included in all statistical tests.

instrumental variables, fixed effects, and OLS give usually zero or very mildly negative results. (Recall that a negative coefficient here means a negative effect of class size on student achievement, and therefore a positive effect of class size reduction policy.) Zero or tiny effects are also reported on average for individual subjects (math, reading, languages, and others), students (advantages, disadvantaged, female), and countries (United States, Scandinavia, other countries). The top five journals in economics most commonly publish results close to zero, similarly to other journals, but the top journals feature a larger proportion of substantially negative results. Panel data yield, on average, similar results to cross-sectional data, but results based on the latter are often quite widely dispersed on both sides of zero.

More detailed numerical information on the differences in the reported class size effects are available in Table 2. The left-hand part of the table provides simple unweighted summary statistics: each estimate has the same weight. In the right-hand part of the table, estimates are weighted by the inverse of the number of estimates reported per study—so that each study has the same weight. The appropriateness of various weights has been a subject of controversy in literature surveys on the class size effect. Hanushek (1997) gives each estimate the same weight, while Krueger (2003) gives each study the same weight. Even with a very different dataset, we confirm the observation of Krueger (2003) that giving each study the same weight results in more substantial estimates of the class size effect. Nevertheless, a different weighting scheme is traditionally used in meta-analysis: inverse variance weights (Greenwald *et al.*, 1996; Hedges & Stock, 1983), which maximize the efficiency of the resulting meta-analysis estimate. Inverse variance weights are not shown in Table 2 but are used later in our analysis. If employed in Table 2, inverse-variance weights would push all means very close to zero.

Similarly to Figure 2, Table 2 provides little evidence of systematic heterogeneity in the literature. The mean estimate is -0.36 (-0.65 when each study is given the same weight), which implies an economically small effect. When data from the STAR experiment are used, the primary study is likely to report estimates around -2, a relatively large effect that could justify some policies of class size reductions (Krueger, 1999). But other identification strategies, and all other contexts of data and estimation, show much smaller effects—perhaps with the exception of the kindergarten grade, but there we only have 20 estimates collected from the literature. The table reveals a substantial difference between the mean estimate reported by

Table 2: Summary statistics for subsets of the literature

		Unweighted			Weighted		
	Observations	Mean	95% conf. int.		Mean	95% conf. int.	
Subjects tested							
Math	765	-0.25	-0.45	-0.05	-0.70	-0.87	-0.53
Reading	301	-0.72	-0.92	-0.52	-0.91	-1.11	-0.71
Writing	46	-0.61	-1.03	-0.20	-0.78	-1.11	-0.45
Languages	144	-0.46	-0.71	-0.22	-1.21	-1.47	-0.94
Other subjects	114	-0.18	-0.62	0.26	-0.21	-0.86	0.45
Class and student characteristics							
Kindergarten	20	-1.65	-2.15	-1.15	-1.18	-1.82	-0.53
Primary school	769	-0.88	-1.03	-0.73	-0.91	-1.05	-0.77
Secondary school	571	0.34	0.12	0.57	-0.45	-0.70	-0.20
Female students	23	0.12	-0.26	0.50	0.38	0.11	0.65
Male students	15	-0.50	-0.69	-0.31	-0.94	-1.14	-0.75
Minority students	46	0.01	-0.68	0.71	-0.83	-1.90	0.25
Disadvantaged students	127	-0.35	-0.83	0.13	-0.41	-0.85	0.03
Advantaged students	84	-0.95	-1.50	-0.39	-0.99	-1.58	-0.40
General population students	1,032	-0.33	-0.48	-0.19	-0.66	-0.81	-0.52
Data characteristics							
Longitudinal data	270	-0.39	-0.54	-0.24	-0.93	-1.12	-0.74
Cross-sectional data	1,076	-0.36	-0.51	-0.20	-0.61	-0.76	-0.46
United States	495	-0.68	-0.81	-0.55	-0.72	-0.88	-0.57
Scandinavian countries	214	-0.37	-0.77	0.04	-1.28	-1.60	-0.95
Other countries	641	-0.12	-0.33	0.10	-0.37	-0.58	-0.16
Estimation characteristics							
STAR experiment	56	-1.99	-2.19	-1.78	-2.29	-2.49	-2.09
Regression discontinuity	133	-0.78	-1.18	-0.37	-1.23	-1.54	-0.91
Instrumental variable	574	-0.39	-0.65	-0.13	-0.42	-0.67	-0.16
Fixed effects	354	-0.34	-0.48	-0.19	-0.39	-0.57	-0.21
Endogeneity control attempted	1,117	-0.50	-0.65	-0.35	-0.68	-0.83	-0.53
Endogeneity ignored	233	0.29	0.08	0.50	-0.56	-0.84	-0.28
Publication characteristics							
Top 5 journals in economics	218	-0.91	-1.16	-0.66	-1.65	-1.90	-1.40
Other journals	1,132	-0.26	-0.41	-0.11	-0.50	-0.65	-0.36
All estimates	1,350	-0.36	-0.49	-0.23	-0.65	-0.78	-0.52

Notes: The table shows subsample-specific means for estimated effects of class size on achievement. The effects are normalized to represent a change in the percentage points of the standard deviations of test scores corresponding to an increase in class size by one student. That is, an estimate of -1 means that a class size reduction by 10 students is associated with an improvement in test scores by 0.1 standard deviations. 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. For the definition of subsamples see Table 6.

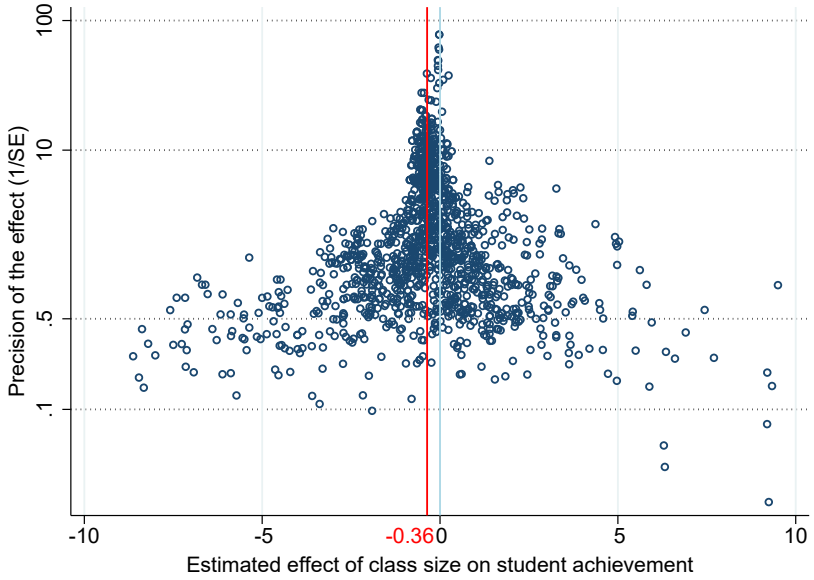
the top five economics journals and other journals. The mean for the top journals is three times the mean for other journals, and the difference holds for both weighting schemes. One possible explanation for this difference is publication bias.

3 Publication Bias

The phenomenon most commonly associated with publication bias is a correlation between estimates and standard errors. The lack of any correlation is the expected consequence of all econometric techniques used to estimate the class size effect: otherwise the reported t-statistic would be meaningless. The correlation arises when researchers (or editors or referees) preferentially publish results that have the intuitive sign and are statistically significant, even by chance. If the point estimate happens to be large enough to offset the standard error, researchers obtain statistical significance and can publish the result more easily. The larger the standard error, the larger the point estimate has to be. A related p-hacking mechanism is the Lombard effect described in the Introduction: given a lot of imprecision, authors may be tempted to try many different specifications until they get a point estimate large enough to produce a t-statistic above 1.96. A visual tool related to this intuition and commonly used in meta-analysis is the so-called funnel plot (Egger *et al.*, 1997; Stanley, 2005): a scatter plot of estimates (on the horizontal axis) on their precision (inverse of the standard error, on the vertical axis). An asymmetry of the funnel plot indicates a correlation between estimates and standard errors.

Figure 3 shows the funnel plot for the class size literature. We observe the theoretically predicted funnel shape: the most precise estimates at the top are concentrated close to each other, while the least precise ones are widely dispersed. Remarkably, the funnel is symmetrical, which is rare in economics (Ioannidis *et al.*, 2017). The apparent absence of publication bias is surprising given the strong intuition in favor of negative effects (that is, effects supporting the notion that larger classes hurt achievement) and the potential need to justify class size reduction policies popular with teachers and parents. The finding testifies to the honesty of researchers in the field. In any case, the most precise estimates are very close to zero. Many meta-analysis techniques are based on the idea that the top of the funnel is the most informative part of the literature, and therefore they try to estimate the mean reported coefficient conditional on max-

Figure 3: Funnel plot shows no publication bias on average



Notes: The estimated effects reported in individual studies are normalized to represent a change in the percentage points of the standard deviations of test scores corresponding to an increase in class size by one student. That is, an estimate of -1 means that a class size reduction by 10 students is associated with an improvement in test scores by 0.1 standard deviations. In the absence of publication bias the scatter plot should resemble an inverted funnel symmetrical around the mean. Extreme outliers are excluded from the figure for ease of exposition but included in all statistical tests. The vertical line represents the mean estimate (-0.36).

imum precision. As we will soon see, this common meta-analysis approach can be problematic in economics if p-hacking makes reported standard errors too small and if correlation between estimates and standard errors comes from different sources than publication bias (for example, heterogeneity).

Table 3 shows meta-analysis tests of publication bias and the corresponding estimates of the underlying effect corrected for the bias. The first block of the table focuses on all estimates in the literature, the second block focuses on estimates published in the top five economics journals. In Appendix B, Table B2 presents the results of tests for other subsets of the literature (STAR experiment, regression discontinuity, instrumental variables, OLS), and Table B3 considers other definitions of the effect size (partial correlation coefficients, effects recomputed to represent standard-deviation changes in class size). These robustness checks yield results on publication bias similar to the baseline analysis in the first block of Table 3; regarding the mean class size effect beyond potential bias, the result is always zero with the exception of the STAR experiment.

Each block has two panels. Panel A reports the results of a simple linear regression of estimates on standard errors, though with different flavors. Panel B reports the results of more complex nonlinear models, two of which are selection models independent of the funnel plot.

The basic formalization of the funnel plot intuition, a simple OLS regression, is shown in the first column. Usually called the "Egger regression" after Egger *et al.* (1997), it was actually first used by Card & Krueger (1995). The next specification adds study fixed effects to filter out unobserved study-level heterogeneity. (Observed heterogeneity, both within- and between-study, will be addressed in the next section.) The third specification uses the meta-analysis instrumental variable estimator (MAIVE) due to Irsova *et al.* (2023). If standard errors are p-hacked in a mechanism analogous to the Taylor's law discussed in the Introduction, for example by using inappropriate clustering, the top of the funnel can be a biased estimate of the underlying mean effect. Also, some method choices can jointly influence estimates and their standard errors, rendering the canonical publication bias test unreliable. The straightforward solution is to use the inverse of the square root of sample size as an instrument for the reported standard error. Sample size is related to the standard error by definition, and it is difficult to exaggerate via p-hacking. To the extent that sample size does not drive the selection of methods that, in turn, systematically influence both estimates and standard errors, sample size is a valid instrument. Note that in most applications of MAIVE in this paper the instrument is weak, and we thus report weak-instrument-robust confidence intervals due to Andrews (2018). The next two specifications in Panel A use alternative weights: proportional to the inverse of the number of estimates reported per study (Krueger, 2003) and to the inverse variance of the estimates (Hedges & Stock, 1983).

Panel B of Table 3 shows the results of five nonlinear models that correct the reported mean effect for potential publication bias. The first model is the weighted average of adequately powered estimates (WAAP) developed by Ioannidis *et al.* (2017). This model is based on the funnel plot, discards estimates with retrospective power below 80%, and computes an inverse-variance-weighted mean of the remaining estimates. The next model, stem-based technique due to Furukawa (2021), extends the previous one by endogenously determining what proportion of the most informative estimates to use. The proportion is determined by exploiting the trade-off between bias and variance: it is inefficient to discard estimates (variance increases), but

Table 3: No publication bias outside top five journals

Block 1: All estimates					
Panel A: Linear	OLS	FE	IV	Study	Precision
Publication bias	-0.088	-0.107	-0.576	0.111	-0.488
(*standard error*)	(0.243)	(0.228)	(0.515)	(0.288)	(0.308)
	[-0.793, 0.448]		[-6.557, .432]	[-0.486, 0.849]	[-1.208, 0.216]
			{-1.585, 0.433}		
Effect beyond bias	-0.260	-0.238	0.316	-0.785***	-0.095
(*constant*)	(0.177)	(0.269)	(0.602)	(0.277)	(0.063)
	[-0.607, 0.104]		[-0.813, 4.361]	[-1.422, -0.191]	[-0.383, 1.103]
First-stage robust F-stat			7.311		
Panel B: Nonlinear	WAAP	Stem	Kink	p-uniform*	Selection
Publication bias			-0.488***		P = 0.591
			(0.083)		(0.169)
Effect beyond bias	-0.046**	-0.165	-0.095***	-0.640***	-0.258***
	(0.020)	(0.156)	(0.010)	(0.200)	(0.057)
Observations	1,350	1,350	1,350	1,350	1,350
Block 2: Top five journals					
Panel A: Linear	OLS	FE	IV	Study	Precision
Publication bias	-1.430***	-1.385	-0.713	-1.822***	-1.010***
(*standard error*)	(0.500)	(0.753)	(0.935)	(0.347)	(0.380)
	[-2.808, -0.042]		[NA]	[-2.457, .1896]	[-5.686, -0.331]
			{-2.546, 1.120}		
Effect beyond bias	0.210	0.175	-0.351	-0.129	-0.0471
(*constant*)	(0.284)	(0.589)	(0.558)	(0.476)	(0.0520)
	[-2.632, 1.097]		[NA]	[-2.014, .7471]	[-0.146, 0.623]
First-stage robust F-stat			3.826		
Panel B: Nonlinear	WAAP	Stem	Kink	p-uniform*	Selection
Publication bias			-1.01***		P = 0.231
			(0.202)		(0.201)
Effect beyond bias	-0.153	-0.082	-0.047	-0.105	-0.140
	(0.096)	(0.232)	(0.054)	(0.294)	(0.129)
Observations	218	218	218	218	218

Notes: Panel A reports the results of a linear regression: $e_{ij} = e_0 + \beta \cdot SE(e_{ij}) + \epsilon_{ij}$, where e_{ij} denotes the i-th class size effect estimated in the j-th study, and $SE(e_{ij})$ denotes the standard error. The class size effects are normalized to represent a change in the percentage points of the standard deviations of test scores corresponding to an increase in class size by one student. That is, an estimate of -1 means that a class size reduction by 10 students is associated with an improvement in test scores by 0.1 standard deviations. FE: study-level fixed effects. IV: reported standard errors are instrumented by the inverse of the square root of sample size. Study: estimates are weighted by the inverse of the number of estimates reported per study. Precision: estimates are weighted by their inverse variance. In Panel B, WAAP denotes the weighted average of adequately powered estimates (Ioannidis *et al.*, 2017), Stem denotes the stem-based technique (Furukawa, 2021), Kink denotes the endogenous kink model (Bom & Rachinger, 2019), p-uniform* denotes the technique due to van Aert & van Assen (2021), and Selection denotes the technique due to Andrews & Kasy (2019). In the selection model, P denotes the probability that estimates insignificant at the 5% level are published relative to the probability that significant estimates are published. Standard errors, clustered at the study level, are reported in parentheses. In square brackets we report 95% confidence intervals from wild bootstrap (Roodman *et al.*, 2018). For IV, in curly brackets we report the two-step weak-instrument-robust 95% confidence interval based on Andrews (2018) and Sun (2018).* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

imprecise estimates are more likely to be selectively reported (publication bias increases). The stem-based technique minimizes the sum of bias and variance. The third nonlinear model, the endogenous kink technique due to Bom & Rachinger (2019), is based on the Egger regression but adds a constant segment for highly statistically significant estimates, when it probably does not matter for publication bias if the standard error changes. The fourth model, p-uniform* (van Aert & van Assen, 2021), is a simplified selection model based on the statistical principle that p-values should be uniformly distributed at the mean underlying effect size. Finally, the rigorously founded selection model by Andrews & Kasy (2019) computes the probability that each estimate within a particular significance bracket is published, and weights the reported estimates by the inverse of that probability.

Table 3 shows little evidence of publication bias when all studies, irrespective of publication outlet, are considered. (The results would remain very similar if from Block 1 we removed studies published in the top five journals.) The tests in Panel A corroborate the intuition of the funnel plot: no correlation appears between estimates and the corresponding standard errors. The mean class size effect (the constant in the regression) corrected for potential publication bias is therefore similar to the simple mean presented earlier. Inverse-variance weights, common in meta-analysis and used in the last specification of Panel A, yield a smaller mean effect than alternative weighting schemes. Models in Panel B serve as more reliable estimators of the underlying mean effect because they do not assume (as Panel A techniques do) that publication bias is a linear function of the standard error. As noted by Andrews & Kasy (2019) and Stanley & Doucouliagos (2014), the linearity assumption is unlikely to hold in meta-analysis. But in the case of the entire class size literature, nonlinear methods give results similar to those of linear methods: mean effects corrected for publication bias are close to simple means. Any potential publication bias in the literature taken as a whole appears weak.

The story is different for studies published in the top five economics journals (Block 2). Here we find a substantial correlation between estimates and standard errors irrespective of the weight used. The correlation is marginally insignificant at the 5% level in the fixed effects specification, and imprecisely estimated in the instrumental variable specification, where the instrument is weak. Crucially, however, all nonlinear correction techniques point to corrected mean estimates very close to zero, shrinking the simple mean substantially, and thus providing

Table 4: Specification test for the Andrews & Kasy (2019) model

	All estimates	Top 5 journals	Regression discontinuity
Correlation	0.513 [0.351, 0.499]	0.407 [0.072, 0.556]	0.769 [0.642, 0.895]
Observations	1,350	218	133
	Instrumental variable	Fixed effects	Endogeneity ignored
Correlation	0.521 [0.330, 0.571]	0.412 [0.227, 0.638]	0.506 [0.350, 0.691]
Observations	574	354	233

Notes: Following Kranz & Putz (2022), the table shows, for various subsets of the literature, the correlation coefficient between the logarithm of the absolute value of the estimated class size effect and the logarithm of the corresponding standard error, weighted by the inverse publication probability estimated by the Andrews & Kasy (2019) model. If the assumptions of the model hold, the correlation is zero. Not enough estimates are available to conduct this test separately for the STAR experiment. Bootstrapped 95% confidence intervals in parentheses.

Table 5: Tests due to Elliott *et al.* (2022)

	20 bins	15 bins	10 bins	5 bins
Test for non-increasingness	0.227	0.619	0.594	0.847
Test for monotonicity and bounds	0	0.023	0.001	0.570
Observations ($p <= 0.15$)	621	621	621	621
Total observations	1,350	1,350	1,350	1,350

Notes: The table shows p-values for each test; the null hypothesis is no p-hacking. The techniques rely on the conditional chi-squared test of Cox & Shi (2023). The first technique is a histogram-based test for non-increasingness of the p-curve, the second technique is a histogram-based test for 2-monotonicity and bounds on the p-curve and the first two derivatives. Both tests feature cluster-robust variance estimators. To work well, these models require a large sample (Havranek *et al.*, 2023), so they cannot be applied to the subsample of studies published in top five journals.

evidence for bias. For example, the selection model by Andrews & Kasy (2019) finds that estimates statistically significant at the 5% level are more than four times more likely to be published than statistically insignificant estimates. Why does publication bias appear only in the top economics journals? One possible explanation is that the peer-review process in the top journals places more weight on solid identification, and the canonical examples of early well-identified studies are Krueger (1999) on the STAR experiment and Angrist & Lavy (1999) on regression discontinuity, both showing substantial effects of class size reductions. Subsequent positive results, confirming the two influential studies, could have been easier to publish high than contradictory evidence.

As we have noted, most meta-analysis techniques are based on strong assumptions. The main one is the lack of correlation between estimates and standard errors in the absence of publication bias. The model of Andrews & Kasy (2019) allows an indirect test of this assumption (Kranz & Putz, 2022). If all assumptions of the selection model hold, estimates and standard errors (re-weighted by the inverse publication probability computed by the selection model) should be uncorrelated. Table 4 shows the results for various subsets of data: we always find a substantial correlation. The finding indicates that the zero correlation assumption is tenuous. We address the issue using three strategies. First, we estimate the MAIVE model due to Irsova *et al.* (2023). The results are broadly in line with the baseline techniques, but the instrument is weak. Second, we use the p-uniform* technique developed by van Aert & van Assen (2021), which does not require the orthogonality assumption. The results of p-uniform* are very similar to other techniques. Third, we employ the novel tests due to Elliott *et al.* (2022). The first technique is a histogram-based test for non-increasingness of the p-curve, the second technique is a histogram-based test for 2-monotonicity and bounds on the p-curve and the first two derivatives. Neither test relies on the orthogonality assumption. We obtain some evidence of publication bias using the second test, but not the first one. All in all, there is no strong evidence of publication bias outside the top five journals.

4 Model Uncertainty

The correlation between estimates and standard errors, attributed in the previous section to publication bias, can also arise due to heterogeneity in the class size literature. One task of the

present section, therefore, is to make sure our conclusions regarding publication bias and the mean underlying effect survive an explicit treatment of heterogeneity. In this task we face the twin problem of model uncertainty: one at the level of individual class size studies, the other at the level of meta-analysis. Regarding the former, the literature lacks a clear consensus on how a "best practice" class size study should look, and individual studies differ in dozens aspects, big and small. Regarding the latter, it is unclear which of the dozens of characteristics potentially reflecting heterogeneity should be added to the final meta-analysis model. Our intention is to address the former source of model uncertainty by systematically tracing the differences in results to differences in the data and methodology used by the primary studies. In the process we also address model uncertainty in meta-analysis by using Bayesian and frequentist model averaging. As the bottom line we provide estimates of the class size effect, corrected for potential publication bias and misspecifications, in various context.

We collect 42 aspects that reflect the context in which the corresponding estimates are obtained. The resulting variables are described and summarized in Table 6. For ease of exposition, we divide the variables into five groups: subjects tested (math, reading, writing, others), class and student characteristics (kindergarten, primary school, secondary school, class size, female students, minority students, disadvantaged students, advantaged students), data characteristics (cross section, panel, data year, countries), estimation characteristics (STAR experiment, regression discontinuity, instrumental variables, fixed effects, OLS, controls for students, teachers, and schools), and publication characteristics (top five journals, citations, publication year, journal impact factor).

Table 6: Description and summary statistics of variables reflecting heterogeneity

Variable	Description	Mean	SD	WM
Class size effect	Estimated effect of class size on student achievement; normalized to represent a change in the percentage points of the standard deviations of test scores corresponding to an increase in class size by one student (response variable).	-0.36	2.44	-0.65
Standard error (SE)	Standard error of the estimated class size effect.	1.18	1.76	1.20
SE * Top journal	Interaction of the standard error and a dummy that equals one for top journal publication.	0.13	0.39	0.11
Subjects tested				
Test in math	= 1 if the test subject is mathematics.	0.57	0.50	0.53
Test in reading	= 1 if the test subject is reading.	0.22	0.42	0.20
Test in writing	= 1 if the test subject is writing.	0.03	0.18	0.02
Test in languages	= 1 if the test subject is languages.	0.11	0.31	0.17

Continued on next page

Table 6: Description and summary statistics of variables reflecting heterogeneity (continued)

Variable	Description	Mean	SD	WM
Test in other subjects	= 1 if the test subject is other than mathematics, reading, writing or languages (reference category for test subjects).	0.08	0.28	0.17
Class and student characteristics				
Kindergarten	= 1 if the estimate corresponds to the kindergarten grade.	0.01	0.12	0.03
Primary school	= 1 if the estimate corresponds to grades 1–5.	0.57	0.50	0.58
Secondary school	= 1 if the estimate corresponds to grades 6–12 (reference category for grade type).	0.42	0.49	0.42
Class size	The logarithm of the average class size used for the estimation minus sample minimum of class size in the literature.	2.46	0.37	2.47
Female students	= 1 if the effect is estimated for female students only.	0.02	0.13	0.01
Minority students	= 1 if the effect is estimated for minority students only.	0.03	0.18	0.01
Disadvantaged students	= 1 if the effect is estimated only for disadvantaged students (students from low-income families, incomplete families, with low-educated parents, with low-experienced or low-educated teachers, low-performing students, or students with learning disabilities).	0.09	0.29	0.04
Advantaged students	= 1 if the effect is estimated only for advantaged students (students from complete families, with high-educated parents, high-experienced or high-educated teachers, high-performing or gifted students).	0.06	0.24	0.02
General population students	= 1 if the effect is estimated for students representing the general population.	0.76	0.42	0.90
Data characteristics				
Cross-sectional data	= 1 if cross-sectional data are used.	0.80	0.40	0.85
Longitudinal data	= 1 if panel data are used (reference category for data dimension).	0.20	0.40	0.13
Data year	The logarithm of the average year of the time period of the data used to estimate the class size effect minus sample minimum in the literature.	2.79	0.45	2.79
Country: United States	= 1 if the estimate uses data from the United States.	0.37	0.48	0.33
Country: Scandinavian	= 1 if the estimate uses data from Scandinavia (Denmark, Finland, Norway, Sweden).	0.16	0.37	0.18
Country: other	= 1 if the country for which the effect is estimated is other than the United States or Scandinavian countries (reference category for country variables).	0.47	0.50	0.49
Estimation characteristics				
Method: STAR experiment	= 1 if the STAR experiment from Tennessee is used to identify the effect.	0.04	0.20	0.04
Method: RDD	= 1 if regression discontinuity design is used to identify the effect.	0.10	0.30	0.15
Method: IV	= 1 if the instrumental variable approach is used to identify the effect.	0.43	0.49	0.35
Method: FE	= 1 if student fixed-effects are included in the model (or the value-added model is used) to estimate the effect.	0.26	0.44	0.21
Method: OLS	= 1 if the method used to estimate the effect does not explicitly account for endogeneity (reference category for the method variables).	0.17	0.38	0.24

Continued on next page

Table 6: Description and summary statistics of variables reflecting heterogeneity (continued)

Variable	Description	Mean	SD	WM
Number of variables	The logarithm of the number of explanatory variables used in the primary study.	2.33	0.89	2.48
Control: student's gender	= 1 if a control for the gender of students is included.	0.66	0.47	0.67
Control: student's age	= 1 if a control for the age of students is included.	0.47	0.50	0.33
Control: student's ethnicity	= 1 if a control for ethnicity, nationality, or immigration-related status of a student is included.	0.36	0.48	0.33
Control: household income	= 1 if a control for the household income of students' family is included.	0.33	0.47	0.36
Control: parental education	= 1 if a control for the education of students' parents is included.	0.55	0.50	0.47
Control: family status	= 1 if a control for family status (married, cohabiting, same-sex, divorced, or single parent) is included.	0.15	0.36	0.18
Control: peers' ability	= 1 if a control for in-class peer ability is included (e.g. IQ scores of classmates).	0.27	0.44	0.25
Control: teacher's experience	= 1 if a control for the teacher's experience is included.	0.61	0.49	0.51
Control: teacher's gender	= 1 if a control for the teacher's gender is included.	0.46	0.50	0.30
Control: teacher's education	= 1 if a control for the teacher's education is included.	0.40	0.49	0.38
Control: school size	= 1 if a control for the school size (number of the first-year enrollees or the total number of students) is included.	0.16	0.37	0.11
Control: rural population	= 1 if a control for the proportion of people living in rural area within the school district is included.	0.21	0.41	0.13
Publication characteristics				
Top journal	= 1 if the study was published in a top five journal in economics.	0.16	0.37	0.13
Citations	The logarithm of the number of per-year citations received since the study first appeared in Google Scholar.	2.58	1.29	2.33
Publication year	The logarithm of the year when the first draft of the study appeared in Google Scholar minus the sample minimum in the literature.	7.15	1.80	7.44
Impact factor	The discounted recursive RePEc impact factor of the outlet.	0.83	1.26	0.83

Notes: SD = standard deviation, WM = mean weighted by the inverse of the number of estimates reported per study. OLS = ordinary least squares.

The complexity of the literature and the consequently large number of aspects in which individual studies and estimates differ gives rise to model uncertainty in meta-analysis. The natural solution to such model uncertainty in the Bayesian setting is Bayesian model averaging (BMA, Steel, 2020). BMA runs many regressions with the estimated class size effect on the left-hand side and various subsets of the variables introduced in Table 6 on the right-hand side. It then weights the individual regression models by goodness of fit and parsimony. In the implementation of BMA we use the unit information g-prior recommended by Eicher *et al.* (2011) and the dilution model prior due to George (2010). The dilution prior is important because it

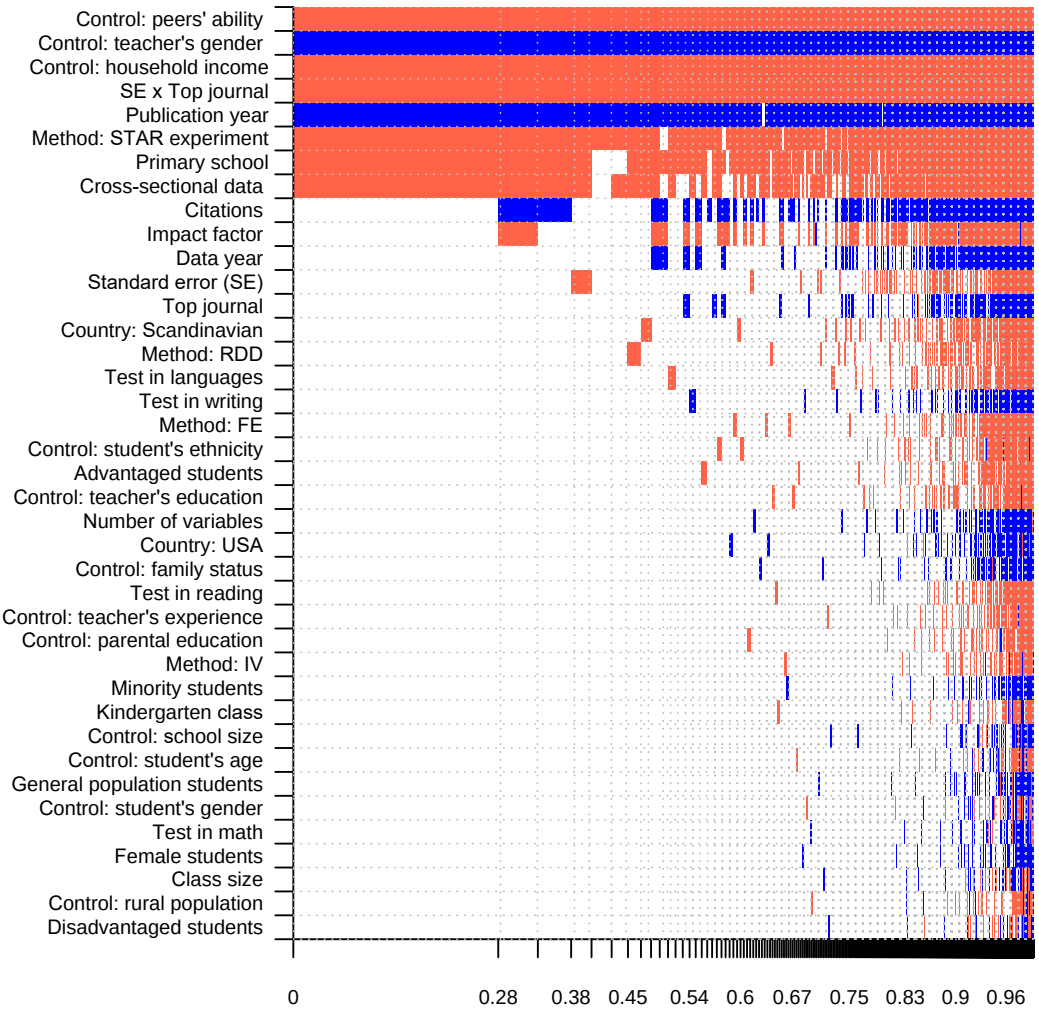
addresses potential collinearity in meta-regression: models that feature a small determinant of the correlation matrix get a smaller weight. Because the choice of priors is inherently subjective, we use two robustness checks. First, a BMA variant with BRIC g-prior based on Fernandez *et al.* (2001) and the beta-binominal model prior according to Ley & Steel (2009). Second, frequentist model averaging with Mallows' weights (Hansen, 2007) using the orthogonalization of the covariate space suggested by Amini & Parmeter (2012). The results of the robustness checks are broadly in line with our baseline results are available in Figure B5 and Table B5 in Appendix B.

The outcome of Bayesian model averaging is depicted graphically in Figure 4. The explanatory variables are ranked according to their posterior inclusion probabilities (analogous to statistical significance in the frequentist sense) from the highest at the top to the lowest at the bottom. The horizontal axis measures cumulative posterior model probability: the BMA weight of individual models; the best models are shown on the left. Blue color (darker in grayscale) means that the estimated parameter of the corresponding explanatory variable is positive. Red color (lighter in grayscale) means the estimated parameter of the corresponding explanatory variable is negative. The figure makes it clear that most of the 42 variables that we collect do not help explain the systematic differences in reported class size effect. Only 8 variables are robustly important, and their corresponding regression coefficients have the same sign irrespective of other variables being included or ignored.

More details on the baseline BMA estimation are available in Table 7 and Figure 5. Table 7 reports the numerical results of BMA together with a simple OLS check: the robustness check only includes variables with posterior inclusion probability above 0.5. Figure 5 shows posterior coefficient distributions for the 8 important variables. The BMA results corroborate the previous findings regarding publication bias: on average in the literature there is no correlation between estimates and standard errors, even if we control for various aspects of data and methodology. The correlation, however, remains strong for studies published in the top five journals, indicating publication bias there. Once again, we also find that the STAR experiment yields results systematically different from those of other identification approaches. Estimates obtained using regression discontinuity, instrumental variables, fixed effects, and OLS are on average close to each other. We also find that class size effects tend to be somewhat stronger in primary schools

compared to secondary schools, that panel data bring weaker class size effects than cross-sectional data, that newer studies yield on average weaker effects, and that some characteristics of students, peers, and teachers can matter for the class size effect.

Figure 4: Model inclusion in Bayesian model averaging



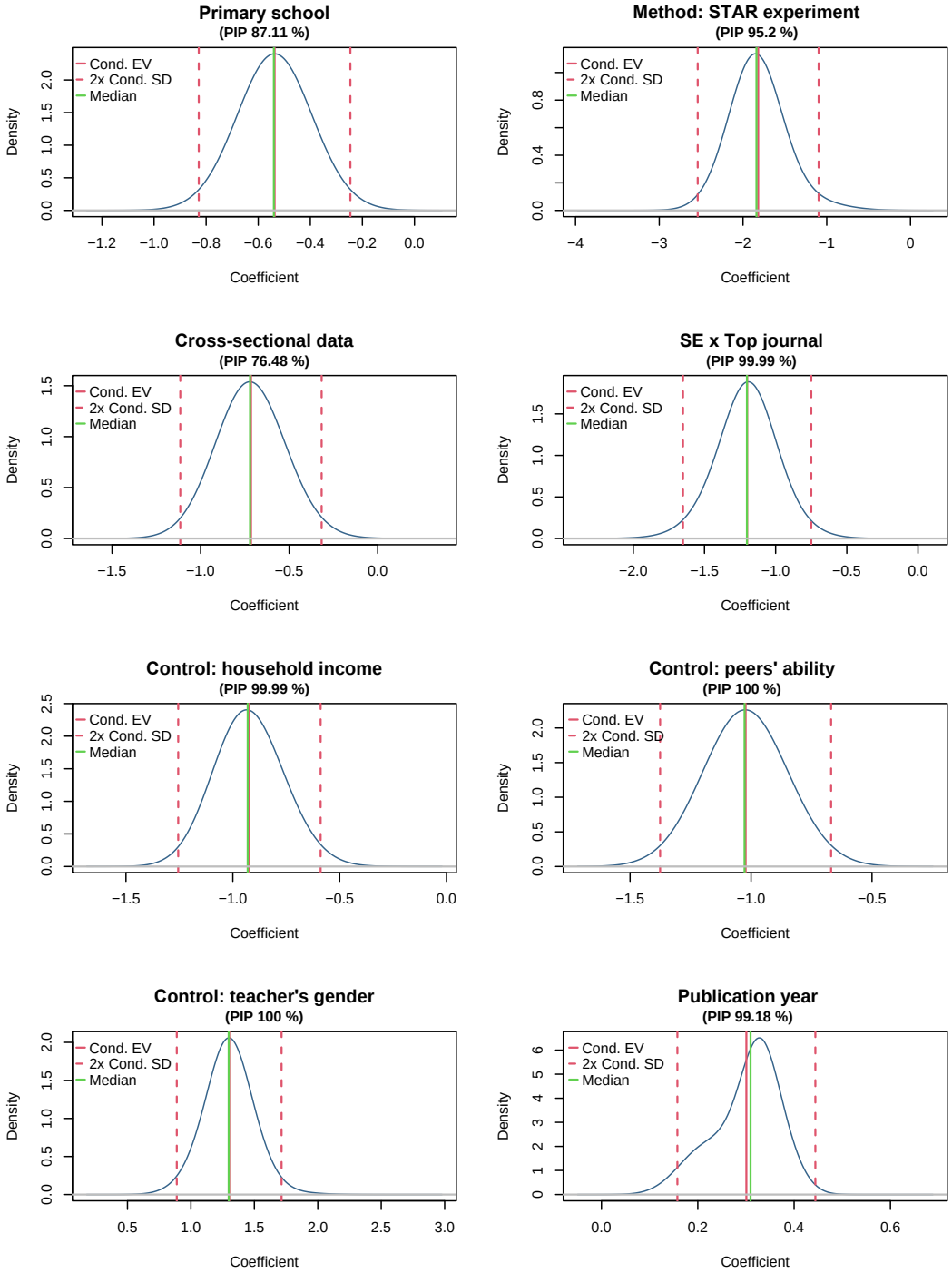
Notes: The figure depicts the results of the benchmark BMA model reported in Table 7. We employ the unit information g-prior (the prior has the same weight as one observation of data) recommended by Eicher *et al.* (2011) and the dilution prior suggested by George (2010), which accounts for collinearity. 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 measures 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 7. All variables are described in Table 6.

Table 7: Why estimates of the class size effect vary

Response variable: reported class size effect	Bayesian model averaging (baseline model)			OLS (robustness check)		
	P. mean	P. SD	PIP	Coef.	SE	p-value
Constant	-2.18	NA	1.00	-1.57	0.53	0.00
Standard error (SE)	-0.01	0.02	0.08			
SE * Top journal	-1.20	0.23	1.00	-1.19	0.26	0.00
Subjects tested						
Test in math	0.00	0.01	0.01			
Test in reading	0.00	0.03	0.02			
Test in writing	0.02	0.13	0.03			
Test in languages	-0.01	0.08	0.04			
Class and student characteristics						
Kindergarten	0.00	0.06	0.01			
Primary school	-0.47	0.23	0.87	-0.54	0.24	0.03
Class size	0.00	0.02	0.01			
Female students	0.00	0.04	0.01			
Minority students	0.00	0.04	0.01			
Disadvantaged students	0.00	0.02	0.01			
Advantaged students	-0.01	0.08	0.03			
General population students	0.00	0.02	0.01			
Data characteristics						
Cross-sectional data	-0.55	0.35	0.76	-0.76	0.29	0.01
Data year	0.10	0.28	0.13			
Country: United States	0.01	0.06	0.02			
Country: Scandinavian	-0.02	0.11	0.06			
Estimation characteristics						
Method: STAR experiment	-1.73	0.52	0.95	-1.82	0.27	0.00
Method: RDD	-0.02	0.11	0.05			
Method: IV	0.00	0.02	0.01			
Method: FE	-0.01	0.06	0.03			
Number of variables	0.00	0.03	0.02			
Control: student's gender	0.00	0.02	0.01			
Control: student's age	0.00	0.02	0.01			
Control: student's ethnicity	-0.01	0.06	0.03			
Control: household income	-0.92	0.17	1.00	-0.97	0.32	0.00
Control: parental education	0.00	0.03	0.01			
Control: family status	0.00	0.05	0.02			
Control: peers' ability	-1.02	0.18	1.00	-1.05	0.30	0.00
Control: teacher's experience	0.00	0.04	0.01			
Control: teacher's gender	1.30	0.21	1.00	1.30	0.29	0.00
Control: teacher's education	-0.01	0.08	0.03			
Control: school size	0.00	0.03	0.01			
Control: rural population	0.00	0.02	0.01			
Publication characteristics						
Top journal	0.07	0.31	0.07			
Citations	0.16	0.24	0.38			
Publication year	0.30	0.08	0.99	0.33	0.03	0.00
Impact factor	-0.11	0.21	0.27			
Observations	1,350			1,350		

Notes: The response variable is the estimate of the effect of class size on achievement normalized to represent a change in the percentage points of the standard deviations of test scores corresponding to an increase in class size by one student. 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 g-prior and model prior recommended by Eicher *et al.* (2011) and additionally 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 effect reported in the literature. For a detailed description of all the variables see Table 6; for details on the BMA procedure see Table B4 and Figure B4.

Figure 5: 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, the coefficient corresponding to *STAR experiment* is negative and substantially far from zero in all models irrespective of other variables being included or ignored.

So, what is the best guess concerning the class size effect in various contexts, based on the entire literature and after correction for potential publication bias and misspecifications? Table 8 gives the answer. The implied effects are computed as fitted values from the Bayesian model averaging exercise. The overall mean, in the first row, is conditional on the following choices: no publication bias even in the top five journals (which means we plug in zero for the standard error), preference for panel data (cross-sectional data = 0), preference for new data (data year set to sample maximum), preference for attempted endogeneity control (OLS = 0), preference for including student, peer, and teacher controls, preference for studies that are published recently, are highly cited, and are featured in an outlet with a high impact factor. All other variables are set to their sample means. The resulting estimate is virtually zero, albeit with a wide credible interval. In the next rows we keep the subjective definition of "best practice" described above and only change the relevant part—for example, in the second row, "STAR experiment", we set the STAR experiment variable equal to one and other identification variables equal to zero.

The results suggest negligible effects of class size on student achievement in all contexts except the STAR experiment. The lack of a systematic, replicable effect seems to be a robust feature of the literature, independent of the specific meta-analysis approach. We would obtain similar conclusions if we focused the entire analysis on studies published in the top five journals, and hence likely avoided studies not subjected to high-quality peer review. The zero effect is eminently unintuitive, as shown by the expectations of teachers cited in the Introduction. How can smaller classes *not* help achievement? One possible explanation is that teachers may not change their teaching practices when the size of the class changes (Ehrenberg *et al.*, 2001; Hattie, 2005). On a more aggregated level, smaller classes require many more teachers. With a sudden reduction in the average class size, it might be difficult for principals to hire additional teachers of the required quality. In consequence, a smaller proportion of students will end up with really good teachers. This negative side effect should, in principle, disappear over time: smaller classes increase the attractiveness of the profession and may motivate young people to become teachers. On the other hand, as noted by Hanushek (1999), long-run time series for the US show a decrease in the mean class size but not a corresponding improvement in test scores.

Table 8: Implied effects of class size on achievement in different contexts

	Mean	95% cred. int.	
Overall best practice	-0.083	-3.186	3.020
Method: STAR experiment	-1.738	-5.941	2.465
Method: Regression discontinuity	-0.029	-3.129	3.071
Method: Instrumental variable	-0.007	-3.182	3.168
Method: Fixed effects	-0.015	-2.956	2.925
Method: Ordinary least squares	-0.005	-3.087	3.076
Kindergarten	0.180	-2.901	3.262
Primary school	-0.284	-3.974	3.406
Secondary school	0.184	-2.857	3.224
Country: United States	-0.075	-3.270	3.121
Country: Scandinavian	-0.106	-2.951	2.740
Country: other	-0.082	-3.389	3.226
Test in math	-0.081	-3.400	3.237
Test in reading	-0.085	-3.172	3.003
Test in writing	-0.061	-3.192	3.069
Test in languages	-0.095	-3.334	3.143
Test in other subject	-0.082	-3.202	3.039
Advantaged students	-0.094	-3.417	3.229
Disadvantaged students	-0.082	-3.205	3.041

Notes: The table uses benchmark BMA results to compute the class size effect conditional on selected aspects of data, methodology, and publication (see text for details). That is, the table attempts to answer the question what the class size effect would be in different contexts if the literature was free of publication bias even in the top journals and all studies used the same identification strategy. The class size effects are normalized to represent a change in the percentage points of the standard deviations of test scores corresponding to an increase in class size by one student. That is, an estimate of -1 means that a class size reduction by 10 students is associated with an improvement in test scores by 0.1 standard deviations.

5 Conclusion

We use recently developed techniques to account for publication bias and model uncertainty in the literature estimating the effect of class size on student achievement. Remarkably, despite the strong intuition favoring a positive effect and the polarization within the research literature, we find minimal evidence of publication bias, except in studies published in the top five economics journals. The bias there is relatively mild and might have been caused by the incentive to replicate the positive results of the STAR experiment, which was viewed as the gold standard in the literature for much of the 1990s and 2000s. Studies employing various identification approaches—such as student and class fixed effects, instrumental variables, or regression discontinuity—and considering different student types (including disadvantaged ones), subjects, schools, and jurisdictions typically do not yield systematically different results. These studies collectively suggest a class size effect that is economically indistinguishable from zero.

The caveat to the conclusion provided above is that we find an economically significant effect of class size reduction in studies examining the large and expensive STAR experiment conducted

in Tennessee in the 1980s. We show that the positive results are a robust feature of the STAR experiment data, not an artifact created by publication bias or methodological approaches in analyzing the data. But the STAR experiment also universally fails to replicate. While we cannot rule out the possibility that the experiment is the only relevant piece of evidence and the rest of the literature is misspecified, that possibility seems unlikely given the quality of many of the other studies, especially those focusing on regression discontinuity. The plausible explanation is that something went wrong with the STAR experiment.

Several researchers have discussed the potential problems in the experiment (Ehrenberg *et al.*, 2001; Hanushek, 1999; Jepsen, 2015), which is described in detail by Mosteller (1995). The experiment was nominally randomized, but it is hard to verify whether students and teachers really were assigned randomly, as the decision on the assignment was in the hands of the principals. The schools had to actively register for the program, and only a fraction of eligible schools did so. It is hard to imagine that at least some principals would not assign teachers to smaller classes strategically, perhaps as a reward for previous good performance. It is also hard to imagine that some parents did not lobby principals to place their kids into smaller classes, and that the lobbying would always fail. As a consequence, the experiment could feature smaller classes with systematically better teachers and advantaged students. But even if the initial randomization was perfect, substantial and uncontrolled flows of students between treatment and control classes appeared in subsequent years, rendering interpretation problematic. Hence we do not elevate the STAR experiment above other studies and assign it the same weight as the regression discontinuity and instrumental variable approaches. Doing so yields a negligible implied effect of class size on student performance.

The bottom line is that the available empirical evidence, taken as a whole, shows no effect of class size in any commonly examined context. That conclusion does not necessarily mean that class size reduction does not help students, at least some of them. Perhaps all studies share a common misspecification and thus the literature has failed to identify an underlying positive effect. But any benefits would have to be massive to justify the immense costs of class size reductions (Rivkin *et al.*, 2005), and it is doubtful whether such large benefits will ever be identified. Until then, class size reduction remains an evidence-based policy in search of evidence.

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.

Akerhielm, K. (1995): "Does class size matter?" *Economics of Education Review* **14(3)**: pp. 229–241.

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. (2018): "Valid two-step identification-robust confidence sets for GMM." *The Review of Economics and Statistics* **100(2)**: pp. 337–348.

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

Angrist, J. D., E. Battistin, & D. Vuri (2017): "In a small moment: Class size and moral hazard in the Italian Mezzogiorno." *American Economic Journal: Applied Economics* **9(4)**: pp. 216–249.

Angrist, J. D. & V. Lavy (1999): "Using Maimonides' rule to estimate the effect of class size on scholastic achievement." *The Quarterly Journal of Economics* **114(2)**: pp. 533–575.

Angrist, J. D., V. Lavy, J. Leder-Luis, & A. Shany (2019): "Maimonides' rule redux." *American Economic Review: Insights* **1(3)**: pp. 309–24.

Argaw, B. & P. Puhani (2018): "Does class size matter for school tracking outcomes after elementary school? Quasi-experimental evidence using administrative panel data from Germany." *Economics of Education Review* **65(C)**: pp. 48–57.

Arias, J. J. & D. M. Walker (2004): "Additional evidence on the relationship between class size and student performance." *The Journal of Economic Education* **35(4)**: pp. 311–329.

Asadullah, M. N. (2005): "The effect of class size on student achievement: Evidence from Bangladesh." *Applied Economics Letters* **12(4)**: pp. 217–221.

Babcock, P. & J. R. Betts (2009): "Reduced-class distinctions: Effort, ability, and the education production function." *Journal of Urban Economics* **65(3)**: pp. 314–322.

Bandiera, O., V. Larcinese, & I. Rasul (2010): "Heterogeneous class size effects: New evidence from a panel of university students." *The Economic Journal* **120(549)**: pp. 1365–1398.

Becker, W. E. & J. R. Powers (2001): "Student performance, attrition, and class size given missing student data." *Economics of Education Review* **20(4)**: pp. 377–388.

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

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

Bonesronning, H. (2003): "Class size effects on student achievement in Norway: Patterns and explanations." *Southern Economic Journal* **69(4)**: pp. 952–965.

Boozer, M. & C. Rouse (2001): "Intraschool variation in class size: Patterns and implications." *Journal of Urban Economics* **50(1)**: pp. 163–189.

Borland, M. V., R. M. Howsen, & M. W. Trawick (2005): "An investigation of the effect of class size on student academic achievement." *Education Economics* **13(1)**: pp. 73–83.

Bosworth, R. (2014): "Class size, class composition, and the distribution of student achievement." *Education Economics* **22(2)**: pp. 141–165.

Bressoux, P., F. Kramarz, & C. Prost (2009): "Teachers' training, class size and students' outcomes: Learning from administrative forecasting mistakes." *The Economic Journal* **119(536)**: pp. 540–561.

Bressoux, P., L. Lima, & C. Monseur (2019): "Reducing the number of pupils in French first-grade classes: Is there evidence of contemporaneous and carryover effects?" *International Journal of Educational Research* **96(C)**: pp. 136–145.

Brown, A. L., T. Imai, F. Vieider, & C. Camerer (2023): "Meta-Analysis of Empirical Estimates of Loss-Aversion." *Journal of Economic Literature* **(forthcoming)**.

Browning, M. & E. Heinesen (2007): "Class size, teacher hours and educational attainment." *Scandinavian Journal of Economics* **109(2)**: pp. 415–438.

Bruhwiler, C. & P. Blatchford (2011): "Effects of class size and adaptive teaching competency on classroom processes and academic outcome." *Learning and Instruction* **21(1)**: pp. 95–108.

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.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, & D. Yagan (2011): "How does your kindergarten classroom affect your earnings? Evidence from project STAR." *The Quarterly Journal of Economics* **126(4)**: pp. 1593–1660.

Chingos, M. M. (2012): "The impact of a universal class-size reduction policy: Evidence from Florida's statewide mandate." *Economics of Education Review* **31(5)**: pp. 543–562.

Cho, H., P. Glewwe, & M. Whitler (2012): "Do

reductions in class size raise students' test scores? Evidence from population variation in Minnesota's elementary schools." *Economics of Education Review* **31(3)**: pp. 77–95.

Cox, G. & X. Shi (2023): "Simple adaptive size-exact testing for full-vector and subvector inference in moment inequality models." *The Review of Economic Studies* **90(1)**: pp. 201–228.

DellaVigna, S. & E. Linos (2022): "RCTs to Scale: Comprehensive Evidence From Two Nudge Units." *Econometrica* **90(1)**: pp. 81–116.

Dobbelsteen, S., J. Levin, & H. Oosterbeek (2002): "The causal effect of class size on scholastic achievement: Distinguishing the pure class size effect from the effect of changes in class composition." *Oxford Bulletin of Economics and statistics* **64(1)**: pp. 17–38.

Doucouliagos, C. & T. D. Stanley (2013): "Are all economic facts greatly exaggerated? Theory competition and selectivity." *Journal of Economic Surveys* **27(2)**: pp. 316–339.

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. J. Brewer, A. Gamoran, & J. D. Willms (2001): "Class size and student achievement." *Psychological Science in the Public Interest* **2(1)**: pp. 1–30.

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.

Elliott, G., N. Kudrin, & K. Wuthrich (2022): "Detecting p-hacking." *Econometrica* **90(2)**: pp. 887–906.

Elpais (2020): "Madrid will delay start of school term for some students and bring down class sizes." *Online article as of Aug 25, 2020*, Ediciones El Pais.

Engin-Demir, C. (2009): "Factors influencing the academic achievement of the Turkish urban poor." *International Journal of Educational Development* **29(1)**: pp. 17–29.

Etim, J. S., A. S. Etim, & Z. D. Blizard (2020): "Class size and school performance: An analysis of elementary and middle schools." *International Journal on Studies in Education* **2(2)**: pp. 66–77.

Evain, F. (2022): "Class size in primary education: the decline continues due to the impact of reducing the last grade of pre-primary classes." *Note d'information 22/08*, Ministere de l'Education nationale et de la Jeunesse, DEPP, France.

Fadulu, L. (2022): "Class Sizes Set to Shrink in New York City Schools, but at What Cost?" *NYT article as of Jun 3, 2022*, The New York Times.

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

Filges, T., C. S. Sonne-Schmidt, & B. C. V. Nielsen (2018): "Small class sizes for improving student achievement in primary and secondary schools: A systematic review." *Campbell Systematic Reviews* **14(1)**: pp. 1–107.

FNAE (2019): "Average group sizes in basic education in Finland below the OECD average." *Online article as of Aug 8, 2019*, Finnish National Agency for Education, Finland.

Francis, J. & W. S. Barnett (2019): "Relating preschool class size to classroom quality and student achievement." *Early Childhood Research Quarterly* **49(4Q)**: pp. 49–58.

Fredriksson, P., B. Ockert, & H. Oosterbeek (2013): "Long-term effects of class size." *The Quarterly Journal of Economics* **128(1)**: pp. 249–285.

Furukawa, C. (2021): "Publication bias under aggregation frictions: From communication model to new correction method." *MIT working paper*, Massachusetts Institute of Technology, Cambridge, MA.

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.

Gerritsen, S., E. Plug, & D. Webbink (2017): "Teacher quality and student achievement: Evidence from a sample of Dutch twins." *Journal of Applied Econometrics* **32(3)**: pp. 643–660.

Glass, G. V. & M. L. Smith (1979): "Meta-analysis of research on class size and achievement." *Educational Evaluation and Policy Analysis* **1(1)**: pp. 2–16.

Gottfried, M. A. (2014): "Peer effects in urban schools: Assessing the impact of classroom composition on student achievement." *Educational Policy* **28(5)**: pp. 607–647.

Greenwald, R., L. V. Hedges, & R. D. Laine (1996): "The effect of school resources on student achievement." *Review of Educational Research* **66(3)**: pp. 361–396.

Han, J. & K. Ryu (2017): "Effects of class size reduction in upper grades: Evidence from Seoul, Korea." *Economics of Education Review* **60(C)**: pp. 68–85.

Hansen, B. E. (2007): "Least squares model averaging." *Econometrica* **75(4)**: pp. 1175–1189.

Hanushek, E. A. (1997): "Assessing the effects of school resources on student performance: An update." *Educational Evaluation and Policy Analysis* **19(2)**: pp. 141–164.

Hanushek, E. A. (1999): "Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects." *Educational Evaluation and Policy Analysis* **21(2)**: pp. 143–163.

Hattie, J. A. C. (2005): "The paradox of reducing class size and improved learning outcomes." *Inter-*

national Journal of Educational Research **43(6)**: pp. 387–425.

Havranek, T., Z. Irsova, L. Laslopova, & O. Zeynalova (2023): "Publication and Attenuation Biases in Measuring Skill Substitution." *The Review of Economics and Statistics* **forthcoming**.

Havranek, T., T. Stanley, H. Doucouliagos, P. Bom, J. Geyer-Klingeberg, I. Iwasaki, W. R. Reed, K. Rost, & R. Van Aert (2020): "Reporting guidelines for meta-analysis in economics." *Journal of Economic Surveys* **34(3)**: pp. 469–475.

Hedges, L. V. (1992): "Modeling Publication Selection Effects in Meta-Analysis." *Statistical Science* **72(2)**: pp. 246–255.

Hedges, L. V. & W. Stock (1983): "The effects of class size: An examination of rival hypotheses." *American Educational Research Journal* **20(1)**: pp. 63–85.

Heinesen, E. (2010): "Estimating class-size effects using within-school variation in subject-specific classes." *The Economic Journal* **120(545)**: pp. 737–760.

Hojo, M. (2013): "Class-size effects in Japanese schools: A spline regression approach." *Economics Letters* **120(3)**: pp. 583–587.

Hojo, M. & T. Oshio (2012): "What factors determine student performance in East Asia? New evidence from the 2007 trends in international mathematics and science study." *Asian Economic Journal* **26(4)**: pp. 333–357.

Hojo, M. & W. Senoh (2019): "Do the disadvantaged benefit more from small classes? Evidence from a large-scale survey in Japan." *Japan and the World Economy* **52(C)**: p. 100965.

Hoxby, C. M. (2000): "The effects of class size on student achievement: New evidence from population variation." *The Quarterly Journal of Economics* **115(4)**: pp. 1239–1285.

Imai, T., T. A. Rutter, & C. F. Camerer (2021): "Meta-Analysis of Present-Bias Estimation Using Convex Time Budgets." *The Economic Journal* **131(636)**: pp. 1788–1814.

Ioannidis, J. P., T. Stanley, H. Doucouliagos *et al.* (2017): "The power of bias in economics research." *The Economic Journal* **127(605)**: pp. 236–265.

Irsova, Z., P. R. D. Bom, T. Havranek, & H. Rachinger (2023): "Spurious Precision in Meta-Analysis." *MetaArXiv 3qp2w*, Center for Open Science.

Iwasaki, I. (2022): "The finance-growth nexus in Latin America and the Caribbean: A meta-analytic perspective." *World Development* **149(C)**: p. 105692.

Jakubowski, M. & P. Sakowski (2006): "Quasi-experimental estimates of class size effect in primary schools in Poland." *International Journal of Educational Research* **45(3)**: pp. 202–215.

Jepsen, C. (2015): "Class size: Does it matter for student achievement?" *IZA World of Labor* **2015(190)**: pp. 1–10.

Jepsen, C. & S. Rivkin (2009): "Class size reduction and student achievement the potential tradeoff between teacher quality and class size." *Journal of Human Resources* **44(1)**: pp. 223–250.

JP (2015): "Ministers approve Bennett reform to reduce class sizes." *Online article as of Nov 16, 2015*, The Jerusalem Post.

Kara, E., M. Tonin, & M. Vlassopoulos (2021): "Class size effects in higher education: Differences across STEM and non-STEM fields." *Economics of Education Review* **82(C)**: p. 102104.

Kedagni, D., K. Krishna, R. Megalokonomou, & Y. Zhao (2021): "Does class size matter? How, and at what cost?" *European Economic Review* **133(C)**: p. 103664.

Kennedy, P. E. & J. J. Siegfried (1997): "Class size and achievement in introductory economics: Evidence from the TUCE III data." *Economics of Education Review* **16(4)**: pp. 385–394.

Kokkelenberg, E. C., M. Dillon, & S. M. Christy (2008): "The effects of class size on student grades at a public university." *Economics of Education Review* **27(2)**: pp. 221–233.

Konstantopoulos, S. & T. Shen (2016): "Class size effects on mathematics achievement in Cyprus: Evidence from TIMSS." *Educational Research and Evaluation* **22(1-2)**: pp. 86–109.

Koreaherald (2016): "Korea to cut class size of high schools." *Online article as of Apr 25, 2016*, The Korea Herald.

Kranz, S. & P. Putz (2022): "Methods matter: p-hacking and publication bias in causal analysis in economics: Comment." *American Economic Review* **112(9)**: pp. 3124–3136.

Krueger, A. B. (1999): "Experimental estimates of education production functions." *The Quarterly Journal of Economics* **114(2)**: pp. 497–532.

Krueger, A. B. (2003): "Economic considerations and class size." *The Economic Journal* **113(485)**: pp. F34–F63.

Leuven, E., H. Oosterbeek, & M. Rønning (2008): "Quasi-experimental estimates of the effect of class size on achievement in Norway." *The Scandinavian Journal of Economics* **110(4)**: pp. 663–693.

Leuven, E. & M. Ronning (2016): "Classroom grade composition and pupil achievement." *The Economic Journal* **126(593)**: pp. 1164–1192.

Levin, J. (2001): "For whom the redundant counts: A quartile regression analysis of family influence on scholastic achievement." *Empirical Economics* **26(1)**: pp. 221–246.

Ley, E. & M. F. Steel (2009): "On the effect of prior assumptions in Bayesian model averaging with applications to growth regression." *Journal of Applied Econometrics* **24(4)**: pp. 651–674.

Li, W. & S. Konstantopoulos (2017): "Does class-size reduction close the achievement gap? Evidence from TIMSS 2011." *School Effectiveness and School Improvement* **28(2)**: pp. 292–313.

Lindahl, M. (2005): "Home versus school learning: A new approach to estimating the effect of class size on achievement." *Scandinavian Journal of Economics* **107(2)**: pp. 375–394.

Mathur, M. B. (2022): "Sensitivity analysis for p-hacking in meta-analyses." *Working paper*, Quantitative Sciences Unit and Department of Pediatrics, Stanford University.

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.

McKee, G., K. R. Sims, & S. G. Rivkin (2015): "Disruption, learning, and the heterogeneous benefits of smaller classes." *Empirical Economics* **48(3)**: pp. 1267–1286.

MECF (2012): "Quality Criteria in Basic Education (In Finnish: Perusopetuksen laatukriteerit. Perusopetuksen, perusopetuksen aamu- ja iltapaivatoiminnan seka koulun kerhotoiminnan laatukriteerit)." *Publications of the Ministry of Education and Culture 2012/29*, Ministry of Education and Culture, Finland.

MERN (2019): "Compulsory education facts and learning outcomes: The Norwegian Education Mirror, 2019." *Education Mirror 2019 online publicaton*, The Norwegian Directorate for Education and Training, Ministry of Education and Research, Norway.

Milesi, C. & A. Gamoran (2006): "Effects of class size and instruction on kindergarten achievement." *Educational Evaluation and Policy Analysis* **28(4)**: pp. 287–313.

Mishel, L. & R. Rothstein (editors) (2002): *The class size debate*. Economic Policy Institute, Washington, DC.

Mosteller, F. (1995): "The Tennessee study of class size in the early school grades." *The Future of Children* **5(2)**: pp. 113–127.

Nandrup, A. B. (2016): "Do class size effects differ across grades?" *Education Economics* **24(1)**: pp. 83–95.

NASUWT (2023): "Class Sizes." *SNCT Handbook as of May 2023*, NASUWT, National Association of Schoolmasters Union of Women Teachers, United Kingdom and SNCT, Scottish Negotiating Committee for Teachers, Scotland UK.

Neisser, C. (2021): "The Elasticity of Taxable Income: A Meta-Regression Analysis." *Economic Journal* **131(640)**: pp. 3365–3391.

New York State Senate (2022): "NY State Senate Bill S9460." *Bill text*, New York State Senate, United States.

NIE (2020): "Not more than 30 students in a class, says New Education Policy." *Online article as of July 30, 2020*, The New Indian Express.

NSWG (2023): "History of New South Wales government schools." *Online article as of Mar 6, 2023*, New South Wales Government, Australia.

NSWTF (2019): "How many students should be in my class?" *Online source from Nov 10, Section: News*, New South Wales Teachers Federation, Australia.

Nye, B., L. V. Hedges, & S. Konstantopoulos (2002): "Do low-achieving students benefit more from small classes? Evidence from the Tennessee class size experiment." *Educational Evaluation and Policy Analysis* **24(3)**: pp. 201–217.

OECD (2020): "Educational Policy Outlook: Portugal." *Educational Policy Outlook 7/2020*, OECD.

OME (2019): "Class size engagement guide." *Online memo as of Jan 2019*, Ontario Ministry of Education, Canada.

QME (2023): "Path 5 - Reduce the number of students per class in elementary school." *Online article as of May 11, 2023*, Quebec Ministry of Education and Ministry of Higher Education, Canada.

Rivkin, S. G., E. A. Hanushek, & J. F. Kain (2005): "Teachers, schools, and academic achievement." *Econometrica* **73(2)**: pp. 417–458.

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.

Rubio, S. (2022): "Class Size Reduction Bill." *Senator's announcement as of Apr 20, 2022*, Senate District 22, US Senate, United States.

Sandy, J. & K. Duncan (2010): "Examining the achievement test score gap between urban and suburban students." *Education Economics* **18(3)**: pp. 297–315.

Scholastic (2012): "Primary Sources 2012: America's Teachers on the Teaching Profession." *Report*, Scholastic Inc. and the Bill and Melinda Gates Foundation, New York, NY: Scholastic.

Shen, T. & S. Konstantopoulos (2017): "Class size effects on reading achievement in Europe: Evidence from PIRLS." *Studies in Educational Evaluation* **53(C)**: pp. 98–114.

Shen, T. & S. Konstantopoulos (2021): "Estimating causal effects of class size in secondary education: Evidence from TIMSS." *Research Papers in Education* **36(5)**: pp. 507–541.

Shen, T. & S. Konstantopoulos (2022): "Are class size and teacher characteristics associated with cognitive outcomes in early grades?" *School Effectiveness and School Improvement* **33(3)**: pp. 333–359.

Shin, Y. & S. W. Raudenbush (2011): "The causal effect of class size on academic achievement: Multivariate instrumental variable estimators with data

missing at random." *Journal of Educational and Behavioral Statistics* **36(2)**: pp. 154–185.

Sims, D. (2008): "A strategic response to class size reduction: Combination classes and student achievement in California." *Journal of Policy Analysis and Management* **27(3)**: pp. 457–478.

Sims, D. P. (2009): "Crowding Peter to educate Paul: Lessons from a class size reduction externality." *Economics of Education Review* **28(4)**: pp. 465–473.

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. & H. Doucouliagos (2014): "Meta-regression approximations to reduce publication selection bias." *Research Synthesis Methods* **5(1)**: pp. 60–78.

Stanley, T. D., H. Doucouliagos, J. P. A. Ioannidis, & E. C. Carter (2021): "Detecting publication selection bias through excess statistical significance." *Research Synthesis Methods* **12(6)**: pp. 776–795.

Steel, M. F. (2020): "Model averaging and its use in economics." *Journal of Economic Literature* **58(3)**: pp. 644–719.

Sun, L. (2018): "Implementing valid two-step identification-robust confidence sets for linear instrumental-variables models." *Stata Journal* **18(4)**: pp. 803–825.

Suryadarma, D., A. Suryahadi, S. Sumarto, & F. H. Rogers (2006): "Improving student performance in public primary schools in developing countries: Evidence from Indonesia." *Education Economics* **14(4)**: pp. 401–429.

Taylor, L. R. (1961): "Aggregation, variance and the mean." *Nature* **189(4766)**: pp. 732–735.

Tinetti, H. J. (2023): "Smaller class sizes to improve teaching and learning outcomes." *The official government website as of April 17, 2023*, Ministry of Education, New Zealand Government.

Ugur, M., S. Awaworyi Churchill, & H. Luong (2020): "What do we know about R&D spillovers and productivity? Meta-analysis evidence on heterogeneity and statistical power." *Research Policy* **49**: p. 103866.

Urquiola, M. (2006): "Identifying class size effects in developing countries: Evidence from rural Bolivia." *The Review of Economics and Statistics* **88(1)**: pp. 171–177.

Urquiola, M. & E. Verhoogen (2009): "Class-size caps, sorting, and the regression-discontinuity design." *American Economic Review* **99(1)**: pp. 179–215.

Vaag Iversen, J. M. & H. Bonesronning (2013): "Disadvantaged students in the early grades: Will smaller classes help them?" *Education Economics* **21(4)**: pp. 305–324.

VARC (2023): "An Integrated Qualitative and Quantitative Evaluation of the SAGE Project." *SAGE project web*, Value-Added Research Center at Wisconsin Center for Education Research at the School of Education, University of Wisconsin-Madison, United States.

WDPI (2016): "Achievement Gap Reduction (AGR) Program." *Strategy Resource Guide*, Wisconsin Department of Public Instruction, United States.

Whitehurst, G. J. & M. M. Chingos (2011): "Class Size: What Research Says and What it Means for State Policy." *Report 5/2011*, Brown Center on Education Policy at Brookings, Washington, DC.

Woessmann, L. (2005a): "Educational production in East Asia: The impact of family background and schooling policies on student performance." *German Economic Review* **6(3)**: pp. 331–353.

Woessmann, L. (2005b): "Educational production in Europe." *Economic Policy* **20(43)**: pp. 446–504.

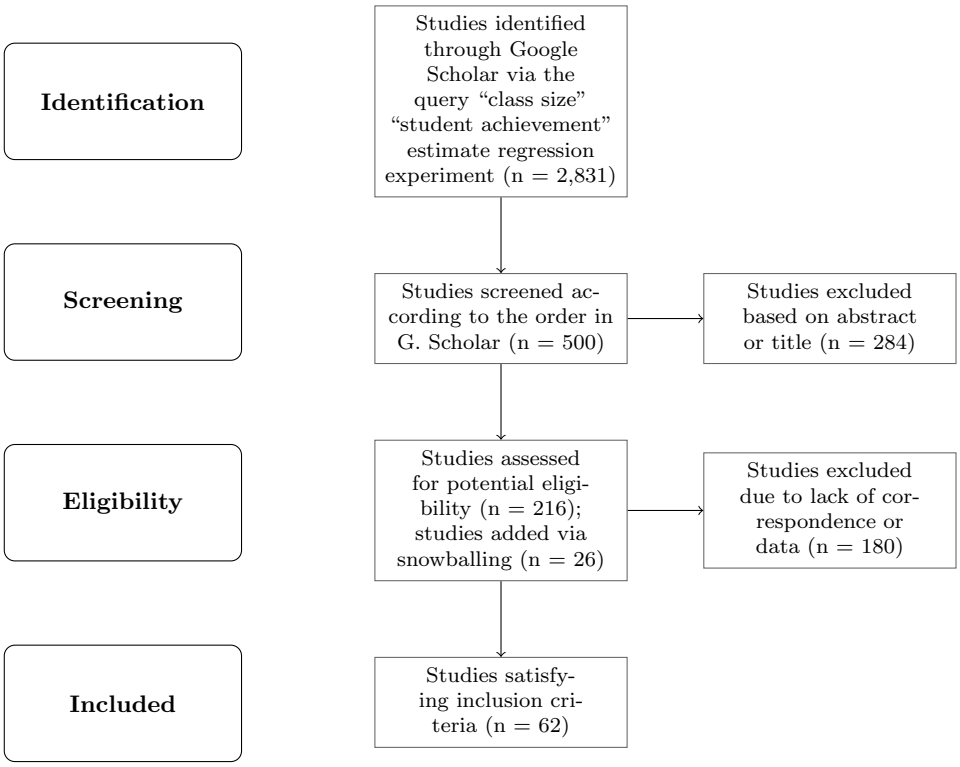
Woessmann, L. & M. West (2006): "Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS." *European Economic Review* **50(3)**: pp. 695–736.

Xue, X., W. R. Reed, & A. Menclova (2020): "Social capital and health: a meta-analysis." *Journal of Health Economics* **72(C)**: p. 102317.

Appendices

A Details of Literature Search

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). Snowballing: we download the references of the potentially eligible studies identified in step "Screening" and inspect the 100 studies most commonly cited among the 216 studies. If, based on the title and abstract, these commonly cited studies show any promise of containing empirical estimates of the class size effect, we add them to the set of potentially eligible studies. Snowballing yields 26 additional studies. Inclusion criteria: 1) the study must report an estimated empirical relationship between test scores (not, for example, total years of schooling) and class size (not, for example, a dummy variable for a "small class"); 2) the study must report standard errors or other statistics from which standard errors can be computed; 3) the study must report the standard deviations of test scores so that we can convert all estimates to a common metric. (Note that, in the robustness check focused on partial correlation coefficients, we also include studies that violate criterion 3.) The literature search was terminated on February 1, 2023. The dataset, together with R and Stata codes, is available at meta-analysis.cz/class.

B Additional Material (for online publication)

Table B1: Jurisdictions with class size reductions since 2010

Jurisdiction	Year	Description	Source
Australia (New South Wales)	2016	Agreement between the government and teachers union to have on average 20 students in grade K, 22 in grade 1, 24 in grade 2, 30 in grades 3–10, 24 in grades 11–12.	NSWTF (2019) NSWG (2023)
Canada (Ontario)	2012	Ontario regulation 132/12, average of 26 students to 2 instructors in kindergarten, cap of 23 in grades K–3, average of 24.5 students in grades 4–8, and average of 22 in grades 9–12.	OME (2019)
Canada (Quebec)	2011	Focus on underprivileged areas with cap at 20 students for grades 3–4, and at 24 students for grades 5–6.	QME (2023)
Finland	2010	In the quality criteria of 2010 and 2012 by the Finnish Ministry	FNAE (2019)
	2012	for Education and Culture recommendation on class size of 20–25 pupils for grades 1–6.	MECF (2012)
France	2017	In 2017 addendum to the Education Law, underprivileged areas cap set at 12 students in grade 1, in 2018 extended to grade 2 (it followed previous experiment in 2002-2003 where in underprivileged areas, cap was 10 students in grade 1). In 2020, in all areas the class size cap at 24 students for grades K–3.	Bressoux *et al.* (2019) Evain (2022)
Germany (Hesse)	2011	In 2011 Hessian Education state law amendment, class sizes reduced for primary schools to a maximum of 25 students.	Argaw & Puhani (2018)
India	2021	In 2021 the Indian New Education policy reducing the student-to-teacher ratio to 25:1 for primary schools and 30:1 for upper primary schools by 2022.	NIE (2020)
Israel	2015	In 2015, the Israeli government approved of a plan to cap grades 1–2 to no more than 34 students per class by 2020.	JP (2015)
New Zealand	2023	In 2023, the NZ Minister of Education announced that student-to-teacher ratios for grades 4–8 will be reduced from 29:1 to 28:1 by 2025.	Tinetti (2023)
Norway	2017	In 2017 Norwegian parliament voted on upper limit on student-	MERN (2019)
	2019	to-teacher ratio 16:1 in grades 1–4 and 21:1 in grades 5–10. In 2019 these ratios were reduced to 15:1 in grades 1–4 and 20:1 in grades 5–10.	
Portugal	2017	In 2017, the Portuguese government announced a class size reduction policy of average class size of 20 students in primary schools and 26 students in secondary schools by 2021.	OECD (2020)
South Korea	2015	The Ministry of Education announced plan to reduce average class size from 30 to 24 and student/teacher ratio from 16.6 to 13.3 till 2022.	Han & Ryu (2017) Koreaherald (2016)
Spain (Madrid)	2020	Cap of 20 students for grades 1–3 due to covid-19 pandemics regionally.	Elpais (2020)
United Kingdom (Scotland)	2010	Cap reduced to 25 students for grade 1 and composite age classes, and to 30 to 33 students for other grades.	NASUWT (2023)
United States (California)	2013	In 2013, class size cap of 24 students in grades K–3, in 2022 Senate	Rubio (2022)
	2022	Bill 1431 lowered student-to-teacher ratio to 20:1 in grades K–3.	
United States (New York City)	2022	In 2022 NY state senate bill S9460, class size cap of 20 students in grades K–3, of 23 students in grades 4–8, and of 25 students per class in high school.	Fadulu (2022)
United States (Wisconsin)	2015	In the follow-up of the SAGE program in 2015 (Wisconsin Acts 53 and 71, Achievement Gap Reduction program), participating schools have to reduce student-to-teacher ratio to max 18:1 or 30:2.	WDPI (2016) VARC (2023)

Notes: The table gives examples on class size reduction mandates or recommendations (accompanied by additional government funding) in various regions since 2010. The list is not exhaustive. Prior to 2010, at least 24 US states started to mandate or incentivize reductions (Whitehurst & Chingos, 2011).

Figure B1: Distribution of class size

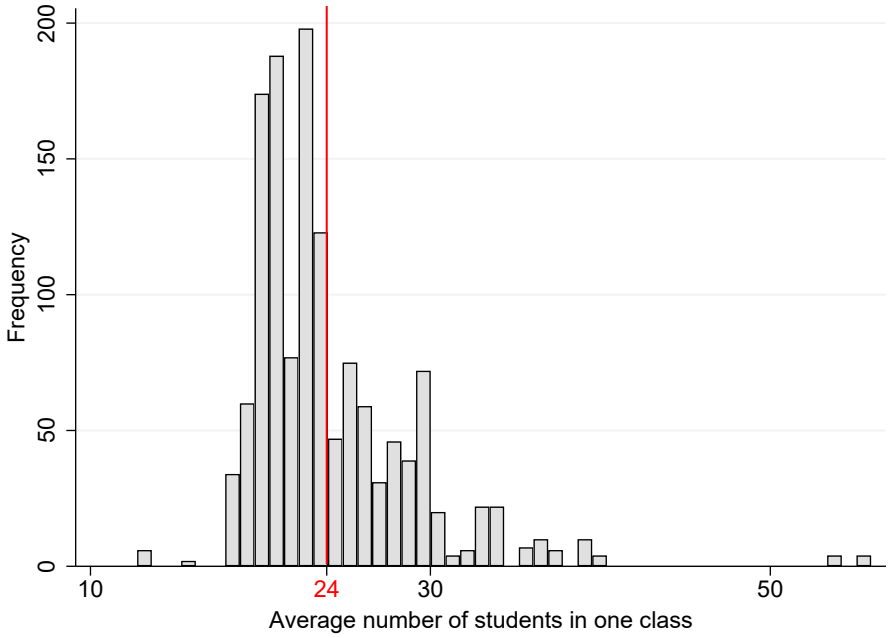
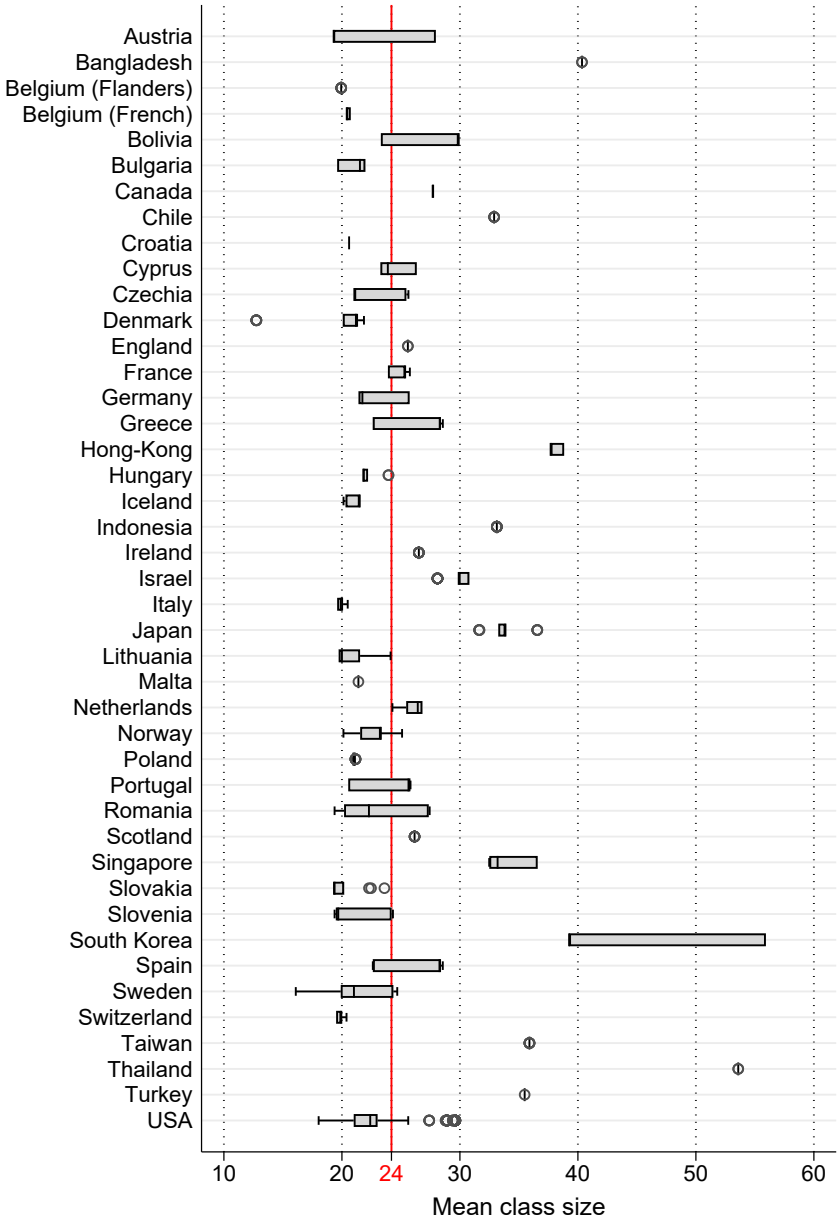
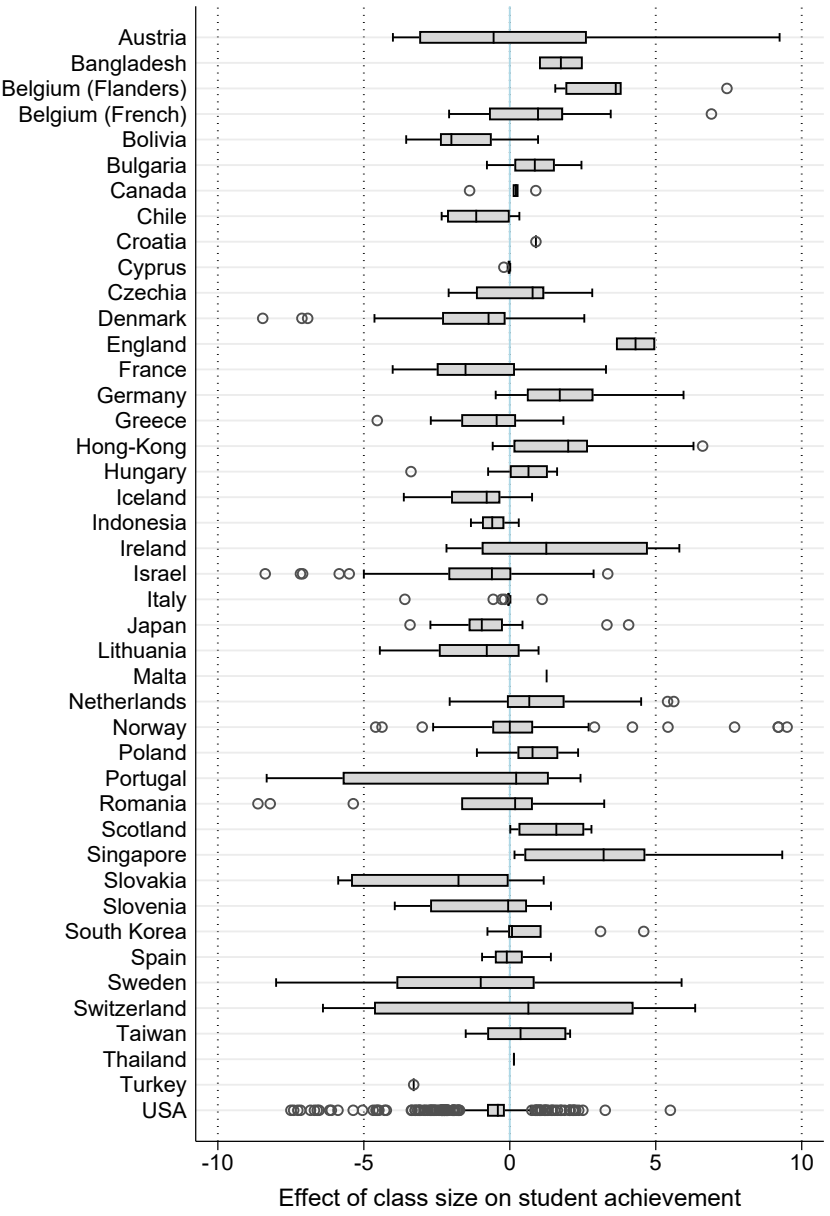


Figure B2: Class size differences within and across countries



Notes: 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 solid vertical line denotes the overall mean (24).

Figure B3: Estimated class size effects vary within and across countries



Notes: The figure shows a box plot of the estimated effects of class size on achievement. The effects are normalized to represent a change in the percentage points of the standard deviations of test scores corresponding to an increase in class size by one student. That is, an estimate of -1 means that a class size reduction by 10 students is associated with an improvement in test scores by 0.1 standard deviations. 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. Extreme outliers are excluded from the figure for ease of exposition but included in all statistical tests.

Table B2: Publication bias tests for subsets of data

Block 1: STAR experiment					
Panel A: Linear	OLS	FE	IV	Study	Precision
Publication bias	0.658	-0.0691	2.172**	1.436***	0.332
(*standard error*)	(0.627)	(0.220)	(1.085)	(0.248)	(0.277)
	[0.038, 4.281]		[NA]	[1.347, 1.552]	[0.044, 3.797]
			{0.045, 4.299}		
Effect beyond bias	-2.407***	-1.943**	-3.376***	-3.138***	-2.207***
(*constant*)	(0.532)	(0.140)	(0.790)	(0.378)	(0.321)
	[-4.609, -1.893]		[-6.276, 2.279]	[-3.508, -2.732]	[-4.622, -1.901]
First-stage robust F-stat			268.598		
Panel B: Nonlinear	WAAP	Stem	Kink	p-uniform*	Selection
Publication bias			0.332		P = 0.512
			(0.66)		(0.309)
Effect beyond bias	-2.058***	-1.840***	-2.207***	-2.092***	-2.120***
	(0.106)	(0.287)	(0.388)	(0.112)	(0.168)
Observations	56	56	56	56	56
Block 2: Regression discontinuity					
Panel A: Linear	OLS	FE	IV	Study	Precision
Publication bias	-0.137	0.186	-1.819**	-0.266	-0.814**
(*standard error*)	(0.285)	(0.293)	(0.852)	(0.193)	(0.353)
	[-1.888, 0.954]		[NA]	[-2.211, 1.376]	[-1.647, .04534]
			{-3.488,-0.150}		
Effect beyond bias	-0.610**	-1.001**	1.434	-0.914***	-0.0286***
(*constant*)	(0.247)	(0.356)	(1.086)	(0.341)	(0.00776)
	[-1.271, -0.211]		[NA]	[-1.910, -0.201]	[-.1925, .08528]
First-stage robust F-stat			19.647		
Panel B: Nonlinear	WAAP	Stem	Kink	p-uniform*	Selection
Publication bias			-0.814***		P = 0.637
			(0.137)		(0.410)
Effect beyond bias	-0.037***	-0.077	-0.029***	-0.029***	-0.034***
	(0.002)	(0.100)	(0.009)	(-0.007)	(-0.004)
Observations	133	133	133	133	133

Continue on next page

Table B2: Publication bias tests for subsets of data—cont.

Block 3: Instrumental variable					
Panel A: Linear	OLS	FE	IV	Study	Precision
Publication bias (*standard error*)	-0.0773 (0.271) [-0.899, 0.535]	-0.0975 (0.241)	0.156 (1.555) [NA] {-2.892, 3.204}	0.297 (0.367) [-0.552, 1.155]	-0.244 (0.229) [-0.866, 0.267]
Effect beyond bias (*constant*)	-0.241 (0.337) [-1.003, 0.466]	-0.202 (0.468)	-0.695 (2.921) [NA]	-0.919* (0.494) [-2.004, 0.097]	-0.007 (0.030) [-0.344, 0.523]
First-stage robust F-stat			0.388		
Panel B: Nonlinear	WAAP	Stem	Kink	p-uniform*	Selection
Publication bias			-0.244*** (0.083)		P = 0.502 (0.310)
Effect beyond bias	NA (NA)	-0.052 (0.307)	-0.007 (0.021)	-0.432* (0.222)	-0.139 (0.113)
Observations	574	574	574	574	574
Block 4: Fixed effects					
Panel A: Linear	OLS	FE	IV	Study	Precision
Publication bias (*standard error*)	-0.209 (0.497) [-1.426, 0.779]	-0.335 (0.694)	-0.118 (0.573) [-5.686, 2.689] {[-1.241, 1.004}	0.278 (0.363) [-1.596, 0.820]	-0.121 (0.360) [-1.042, 0.877]
Effect beyond bias (*constant*)	-0.249 (0.153) [-0.805, 0.657]	-0.197 (0.290)	-0.287 (0.205) [-1.125, 0.899]	-0.626* (0.348) [-1.414, 0.171]	-0.300*** (0.048) [-0.470, -0.169]
First-stage robust F-stat			197.022		
Panel B: Nonlinear	WAAP	Stem	Kink	p-uniform*	Selection
Publication bias			-0.121 (0.245)		P = 0.377 (-0.179)
Effect beyond bias	-0.309*** (0.025)	-0.401 (0.285)	-0.300*** (0.028)	-0.176 (0.281)	-0.357*** (-0.038)
Observations	354	354	354	354	354

Continue on next page

Table B2: Publication bias tests for subsets of data—cont.

Block 5: Ordinary least squares					
Panel A: Linear	OLS	FE	IV	Study	Precision
Publication bias	0.009	1.494**	-0.427	-0.848***	0.489
(*standard error*)	(0.448)	(0.667)	(0.811)	(0.183)	(0.493)
	[-0.812, 1.495]		[-2.001, 1.585]	[-1.136, 0.075]	[-0.624, 1.630]
			{-2.016, 1.161}		
Effect beyond bias	0.286	-0.552	0.532	0.222	-0.056
(*constant*)	(0.237)	(0.377)	(0.491)	(0.314)	(0.0522)
	[-0.310, 0.824]		[-0.481, 1.799]	[-0.424, 0.849]	[-0.614, 1.133]
First-stage robust F-stat			9.288		
Panel B: Nonlinear	WAAP	Stem	Kink	p-uniform*	Selection
Publication bias			0.489**		P = 0.148
			(0.230)		(0.072)
Effect beyond bias	NA	-0.416	-0.056***	0.203	-0.315**
	(NA)	(0.299)	(0.021)	(0.197)	(0.155)
Observations	233	233	233	233	233

Notes: Panel A reports the results of a linear regression: $\mathrm{e}_{ij} = \mathrm{e}_0 + \beta \cdot SE(\mathrm{e}_{ij}) + \epsilon_{ij}$, where e_{ij} denotes the i-th class size effect estimated in the j-th study, and $SE(\mathrm{e}_{ij})$ denotes the standard error. The class size effects are normalized to represent a change in the percentage points of the standard deviations of test scores corresponding to an increase in class size by one student. That is, an estimate of -1 means that a class size reduction by 10 students is associated with an improvement in test scores by 0.1 standard deviations. FE: study-level fixed effects. IV: reported standard errors are instrumented by the inverse of the square root of sample size. Study: estimates are weighted by the inverse of the number of estimates reported per study. Precision: estimates are weighted by their inverse variance. In Panel B, WAAP denotes the weighted average of adequately powered estimates (Ioannidis *et al.*, 2017), Stem denotes the stem-based technique (Furukawa, 2021), Kink denotes the endogenous kink model (Bom & Rachinger, 2019), p-uniform* denotes the technique due to van Aert & van Assen (2021), and Selection denotes the technique due to Andrews & Kasy (2019). In the selection model, P denotes the probability that estimates insignificant at the 5% level are published relative to the probability that significant estimates are published. Standard errors, clustered at the study level, are reported in parentheses. In square brackets we report 95% confidence intervals from wild bootstrap (Roodman *et al.*, 2018). For IV, in curly brackets we report the two-step weak-instrument-robust 95% confidence interval based on Andrews (2018) and Sun (2018).* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B3: Publication bias tests with different definitions of the class size effect

Block 1: Effect of a one-standard-deviation increase in class size					
Panel A: Linear	OLS	FE	IV	Study	Precision
Publication bias	-0.126	-0.119	-0.378	-0.103	-0.488
(*standard error*)	(0.256)	(0.222)	(0.571)	(0.171)	(0.308)
	[-0.813, 0.414]		[NA]	[-0.461, 0.297]	[-1.208, 0.216]
			{-1.496, 0.740}		
Effect beyond bias	-0.0565	-0.0584	0.00584	-0.128***	-0.0947
(*constant*)	(0.0405)	(0.0549)	(0.145)	(0.0399)	(0.0625)
	[-0.134, 0.037]		[NA]	[-0.210, -0.047]	[-0.383, 1.103]
First-stage robust F-stat			3.896		
Panel B: Nonlinear	WAAP	Stem	Kink	p-uniform*	Selection
Publication bias			-0.64***		P = 0.746
			(0.084)		(0.264)
Effect beyond bias	-0.005***	-0.011	-0.013***	-0.110	-0.052***
	(0.001)	(0.036)	(0.002)	(0.072)	(0.019)
Observations	1,350	1,350	1,350	1,350	1,350
Block 2: Effects recomputed to partial correlation coefficients					
Panel A: Linear	OLS	FE	IV	Study	Precision
Publication bias	0.0708	-0.0202	0.0687	-0.572	0.0984
(*standard error*)	(0.295)	(0.316)	(0.297)	(0.363)	(0.401)
	[-0.837, 0.975]		[-0.906, 0.825]	[-1.389, 0.284]	[-0.744, 0.884]
			{-0.514, 0.651}		
Effect beyond bias	-0.00507	-0.00365	-0.00504	-0.00411	-0.00682***
(*constant*)	(0.00447)	(0.00495)	(0.00448)	(0.00566)	(0.00189)
	[-0.015, 0.005]		[-0.014, 0.005]	[-0.016, 0.008]	[-0.024, -0.003]
First-stage robust F-stat			22.01		
Panel B: Nonlinear	WAAP	Stem	Kink	p-uniform*	Selection
Publication bias			0.098		P = 0.512
			(0.119)		(0.088)
Effect beyond bias	-0.005***	-0.003	-0.007***	NA	-0.004***
	(0.001)	(0.008)	(0.001)	(NA)	(0.001)
Observations	1,767	1,767	1,767	1,767	1,767

Notes: Panel A reports the results of a linear regression: $e_{ij} = e_0 + \beta \cdot SE(e_{ij}) + \epsilon_{ij}$, where e_{ij} denotes the i-th class size effect estimated in the j-th study, and $SE(e_{ij})$ denotes the standard error. In Block 1, class size effects are normalized to represent a change in the percentage points of the standard deviations of test scores corresponding to an increase in class size by one standard deviation. In Block 2, class size effects are normalized to represent partial correlation coefficients between class size and student achievement. FE: study-level fixed effects. IV: reported standard errors are instrumented by the inverse of the square root of sample size. Study: estimates are weighted by the inverse of the number of estimates reported per study. Precision: estimates are weighted by their inverse variance. In Panel B, WAAP denotes the weighted average of adequately powered estimates (Ioannidis *et al.*, 2017), Stem denotes the stem-based technique (Furukawa, 2021), Kink denotes the endogenous kink model (Bom & Rachinger, 2019), p-uniform* denotes the technique due to van Aert & van Assen (2021), and Selection denotes the technique due to Andrews & Kasy (2019). In the selection model, P denotes the probability that estimates insignificant at the 5% level are published relative to the probability that significant estimates are published. Standard errors, clustered at the study level, are reported in parentheses. In square brackets we report 95% confidence intervals from wild bootstrap (Roodman *et al.*, 2018). For IV, in curly brackets we report the two-step weak-instrument-robust 95% confidence interval based on Andrews (2018) and Sun (2018).* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B4: Diagnostics of the benchmark BMA estimation (UIP and dilution priors)

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
9.1611	$3 \cdot 10^5$	$1 \cdot 10^5$	1.25 mins	28,454
Model space	*Visited*	*Top models*	*Corr PMP*	*No. obs.*
$5.5 \cdot 10^{11}$	0.0005%	100%	0.9997	1,350
Model prior	g-*prior*	*Shrinkage-stats*		
Random/19.5	UIP	Av = 0.9993		

Notes: We employ the combination of unit information prior recommended by (Eicher *et al.*, 2011) and dilution prior suggested by George (2010), which accounts for collinearity.

Figure B4: Benchmark BMA model size and convergence (UIP and dilution priors)

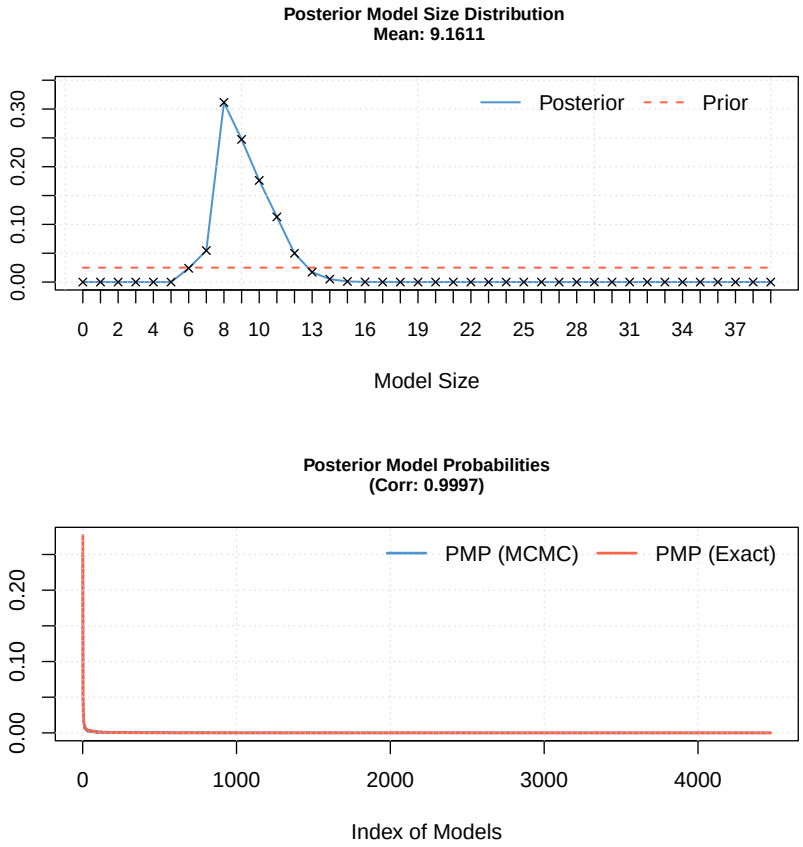
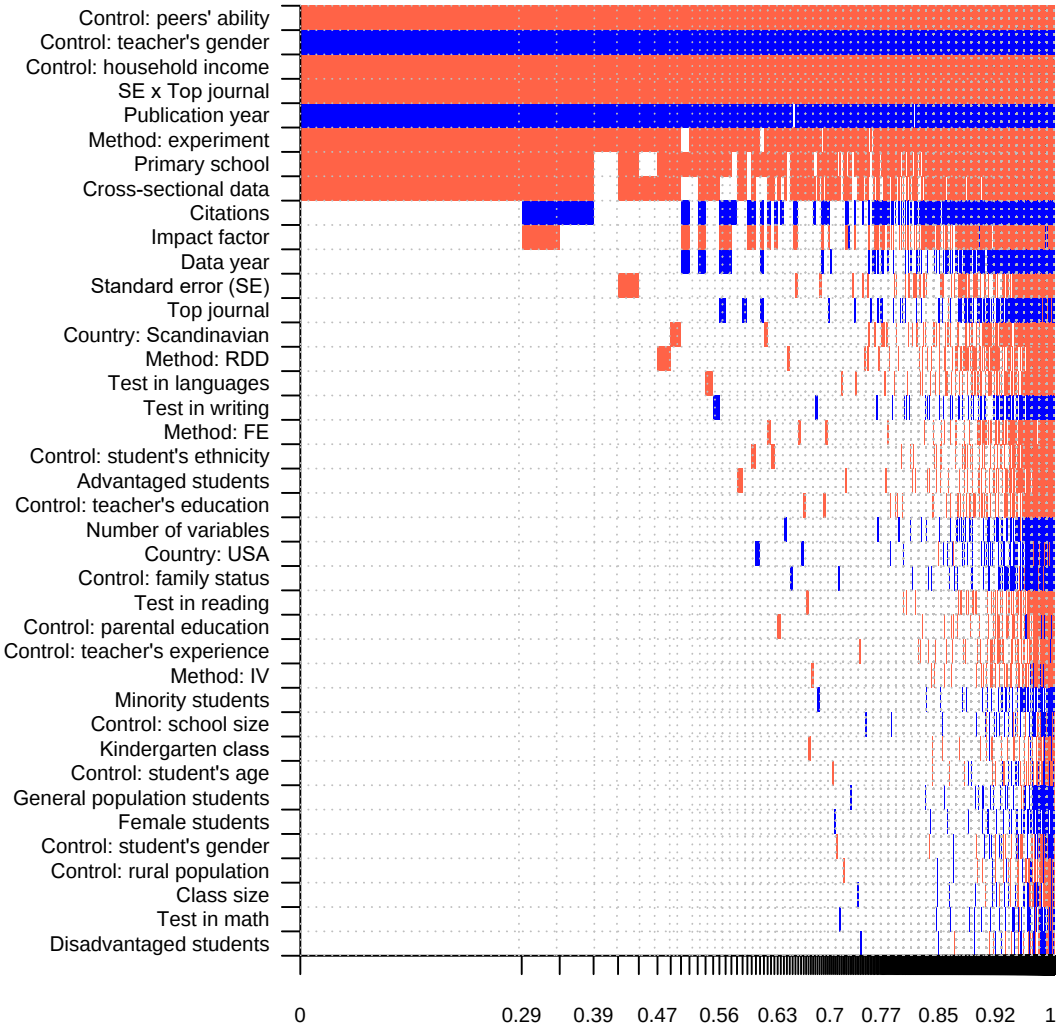


Table B5: Why estimates of the class size effect vary (robustness checks)

Response variable: class size effect	Bayesian model averaging BRIC and random priors			Frequentist model averaging		
	P. mean	P. SD	PIP	Coef.	SE	p-value
Constant	-2.16	NA	1.00	-2.34	1.77	0.19
Standard error (SE)	-0.01	0.02	0.08	0.00	0.06	1.00
SE * Top journal	-1.20	0.22	1.00	-1.05	0.34	0.00
Subjects tested						
Test in math	0.00	0.01	0.01	0.00	0.27	1.00
Test in reading	0.00	0.03	0.01	0.00	0.59	1.00
Test in writing	0.02	0.13	0.03	0.00	0.07	1.00
Test in languages	-0.01	0.08	0.04	0.00	0.39	1.00
Class characteristics						
Kindergarten	0.00	0.06	0.01	0.00	0.01	1.00
Primary school	-0.46	0.23	0.86	-0.36	0.32	0.26
Class size	0.00	0.02	0.01	0.00	0.34	1.00
Female students	0.00	0.04	0.01	0.00	0.16	1.00
Minority students	0.00	0.04	0.01	0.00	0.30	1.00
Disadvantaged students	0.00	0.02	0.01	0.00	0.04	1.00
Advantaged students	-0.01	0.08	0.02	0.00	0.32	1.00
General population students	0.00	0.02	0.01	0.00	0.04	1.00
Data characteristics						
Cross-sectional data	-0.55	0.35	0.76	-0.44	0.39	0.26
Data year	0.09	0.27	0.11	0.14	0.53	0.79
Country: United States	0.01	0.06	0.02	0.00	0.09	1.00
Country: Scandinavian	-0.02	0.11	0.06	0.00	0.61	1.00
Estimation technique						
Method: STAR experiment	-1.74	0.51	0.96	-1.52	0.55	0.01
Method: RDD	-0.02	0.11	0.05	0.00	0.62	1.00
Method: IV	0.00	0.02	0.01	0.00	0.46	1.00
Method: FE	-0.01	0.06	0.03	0.00	0.72	1.00
Number of variables	0.00	0.03	0.02	0.00	0.43	1.00
Control: student's gender	0.00	0.02	0.01	0.00	0.02	1.00
Control: student's age	0.00	0.02	0.01	0.00	0.06	1.00
Control: student's ethnicity	-0.01	0.05	0.03	0.00	0.23	1.00
Control: household income	-0.92	0.17	1.00	-0.85	0.23	0.00
Control: parental education	0.00	0.03	0.01	0.00	0.60	1.00
Control: family status	0.00	0.05	0.02	0.00	0.24	1.00
Control: peers' ability	-1.02	0.18	1.00	-0.97	0.25	0.00
Control: teacher's experience	0.00	0.04	0.01	0.00	0.07	1.00
Control: teacher's gender	1.31	0.21	1.00	1.20	0.64	0.06
Control: teacher's education	-0.01	0.07	0.02	0.00	0.69	1.00
Control: school size	0.00	0.03	0.01	0.00	0.13	1.00
Control: rural population	0.00	0.02	0.01	0.00	0.13	1.00
Publication characteristics						
Top journal	0.07	0.29	0.06	0.00	1.40	1.00
Citations	0.15	0.24	0.35	0.23	0.35	0.52
Publication year	0.30	0.08	0.99	0.26	0.13	0.04
Impact factor	-0.10	0.20	0.24	-0.17	0.42	0.69
Observations	1,350			1,350		

Notes: The response variable is the estimate of the effect of size class on student achievement. The class size effects are normalized to represent a change in the percentage points of the standard deviations of test scores corresponding to an increase in class size by one student. 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 effect size reported in the literature. For detailed description of all the variables see Table 6.

Figure B5: Model inclusion in BMA (BRIC and random priors)



Notes: On the vertical axis 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 a corresponding explanatory variable is positive. Red color (lighter in grayscale) = the estimated parameter of a corresponding explanatory variable is negative. No color = the corresponding explanatory variable is not included in the model. Numerical results are reported in Table B5. All variables are described in Table 6.

Table B6: Diagnostics of the BMA estimation (BRIC and random priors)

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
9.0522	$3 \cdot 10^5$	$1 \cdot 10^5$	1.18mins	26,918
Model space	*Visited*	*Top models*	*Corr PMP*	*No. obs.*
$5.5 \cdot 10^{11}$	0.0005%	100%	0.9995	1,350
Model prior	*g-prior*	*Shrinkage-stats*		
Random/19.5	BRIC	Av = 0.9993		

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 B6: BMA model size and convergence (BRIC and random priors)