IZA Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16853

What Works? Interventions Aimed At Reducing Student Dropout in Higher Education

Kristoffer Ibsen
Michael Rosholm

MARCH 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16853

What Works? Interventions Aimed At Reducing Student Dropout in Higher Education

Kristoffer Ibsen
Aarhus University

Michael Rosholm
Aarhus University and IZA

MARCH 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

What Works? Interventions Aimed At Reducing Student Dropout in Higher Education*

High student dropout rates are a continuing concern within higher education and have received considerable attention for decades. Despite numerous proposed interventions, it remains unclear which interventions are effective, and what the pivotal active ingredients are. This paper systematically synthesizes the effects of different interventions on dropout and graduation rates in higher education. Screening around 7250 abstracts of potentially relevant studies, we found 38 (quasi-)experimental studies that met our pre-specified inclusion criteria. We describe the studies in various dimensions and conduct a narrative systematic review as well as a meta-analysis. Our findings indicate a general tendency towards positive effects. Interventions manipulating peer group composition, particularly those adjusting for gender and ability, show the most promising outcomes. We discuss the economic implications of effect sizes and show that they are considerable. This paper contributes to understanding effective strategies for addressing student dropout, providing valuable insights for higher education institutions and policymakers.

JEL Classification: I23, I24, I29

Keywords: student dropout, intervention, higher education, meta-analysis, educational economics, efficiency

Corresponding author:
Michael Rosholm
Department of Economics and Business Economics
Aarhus University
Fuglesangs Allé 4
DK-8210 Aarhus V
Denmark

E-mail: rom@econ.au.dk

* We gratefully acknowledge helpful comments from Katrine Bønneland Tølbøll, Erik Ø. Sørensen, Elena Mattena, and Eric Bettinger. Declarations of interest: none.

Reducing student dropout rates remains a persistent challenge in higher education and has received considerable attention for more than half a century (Tinto, 1975). Despite recent improvements in graduation rates, primarily attributed to grade inflation (Denning et al., 2022), the issue of college dropout persists.[1] In the United States, between 2013 and 2017, a concerning 25% of students who initially enrolled in bachelor's programs did not return for the second year,[2] while 40% leave college without ever obtaining a degree (Kirp, 2019). Similar trends are evident across 22 OECD countries, where only 67% of students graduate within six years (OECD, 2019). Consequently, numerous interventions have been attempted to address this pressing issue with varying success, and a unified understanding of how to solve the problem continues to be elusive.

While attending college offers some long-term benefits (Zimmerman, 2014; Lovenheim and Smith, 2023), students who successfully complete their college education tend to experience more favorable outcomes especially in terms of economic returns (Card, 1999; Ost, Pan and Webber, 2018). This is in particular the case for students from disadvantaged backgrounds (Oreopoulos, 2021). Completing college not only elevates individual outcomes but also contributes to society at large, where higher educational attainment is associated with increased earnings, expanded employment prospects, and higher tax contributions. Moreover, in terms of distributional impacts, an increase in college completion rates could reduce poverty rates and income inequality (Hershbein, Kearney and Pardue, 2020). Additionally, an upward trajectory in educational attainment could also yield positive effects on more extensive economic measures including both the level and growth of GDP (Sianesi and van Reenen, 2003; Barro, 2013) as there is an increasing demand for high-skilled workers in the labor market due to technological change (Autor, 2014). Hence, it is quite evident that there is a strong economic case for interventions that can increase graduation rates.[3]

From an institutional perspective, student dropout is typically undesirable as it represents a misallocation of resources towards students who ultimately do not complete their education. Concerns arise when dropout is driven by factors such as credit market limitations, scheduling challenges, or social difficulties, as these issues may hinder students from graduating. However, it is important to acknowledge that there are scenarios in which dropout may not be inherently negative or avoidable. These instances include when students change career goals, encounter personal tragedies, or experience an educational mismatch.[4] In some cases, individuals with entrepreneurial aspirations or opportunities

[1] We use (four-year) college and university interchangeably.

[2] "2019, Digest of Education Statistics 2018, Table 326.30," National Center for Education Statistics (NCES), $https://nces.ed.gov/programs/digest/d19/tables/dt19_326.30.asp$ (accessed December 22, 2020).

[3] To the best of our knowledge, there are no existing estimates of the economic costs associated with dropout in higher education. However, as stated, there exists a substantial literature on the economic implications of higher educational attainment, showing nearly uniformly positive effects.

[4] These instances may have negative societal consequences if dropouts do not pursue similar educational paths

in specialized fields may opt to discontinue their formal education when the returns on human capital accumulation no longer outweigh the costs of further investment in continuing education. Nevertheless, institutions should prioritize implementing comprehensive interventions to reduce cases of avoidable dropout. However, there exists limited knowledge about what interventions to deploy and whether they indeed will work across various institutional contexts.

In this paper, we synthesize the numerous interventions that have been attempted in college and examine their effectiveness in reducing college dropout rates. We do so by conducting a comprehensive systematic review and meta-analysis of interventions that are either targeted at mitigating dropout rates or enhancing graduation rates. We include interventions evaluated through experimental or quasi-experimental designs in our analysis. Our perspective in this study is that of a dean, a department head, or a head of studies who grapple with severe dropout challenges and consider implementing some sort of intervention to address the issue.[5] Hence, we disregard interventions based on financial incentives (for a review of impacts of such types of interventions, see e.g. Nguyen, Kramer and Evans (2018) or Dynarski, Page and Scott-Clayton (2022)) and more broad types of programs that are often implemented at the national, state, or overall university level. We also disregard interventions designed to alter centralized admission policies and/or enrollment processes (for a survey of the effectiveness of such interventions, see e.g. French and Oreopoulos (2017)). Our focus is on interventions that can be feasibly implemented at the study, departmental, or faculty level.

A systematic review involves a meticulous synthesis of all available research on interventions aimed at reducing dropout rates. This could help clarify where consensus exists and where discrepancies lie. By combining the qualitative synthesis of the systematic review with the quantitative insights provided by the meta-analysis, we can gain a more robust understanding of the effectiveness of various interventions and the elements that could reduce dropout rates. This knowledge could then be used to develop better strategies for reducing dropout rates in the future.

Across various disciplines, there exist hundreds of thousands of papers utilizing meta-analyses as a research tool (Gurevitch et al., 2018). In recent decades, the application of meta-analyses in economics research has experienced substantial growth and has become increasingly common (Havránek et al., 2020).[6] The primary objective of meta-analyses is to establish evidence-based practices and to resolve seemingly contradictory research outcomes. The advantage of the meta-analysis is that it offers a combined effect size across

afterwards.

[5]The present study was in fact commissioned by the dean of Aarhus BSS with the aim of identifying a promising intervention to test and subsequently implement.

[6]See a list of relevant publications on RePec at: https://ideas.repec.org/k/metaana.html (last accessed: October 10, 2023).

all included studies and goes beyond the particularities of individual studies (Borenstein et al., 2021). Even if every single study has insignificant impacts, the overall effect may still be statistically significant. Moreover, by including a set of variables describing aspects of the different interventions and the environment in which they were implemented, we are able to investigate the relative effect sizes of various types of interventions in different contexts and evaluated using different methods. The main disadvantage of the meta-analysis is that it excludes studies for which it was not possible to extract or calculate a comparable effect size. Optimally, meta-analyses can, if performed properly, point to important policy conclusions and direct future research efforts.

We systematically searched 9 databases, identifying 13,992 studies (7,235 after removing duplicates) published during the period 2000-2021[7]. Through a thorough screening and selection process, we narrowed down our sample to 38 studies, resulting in 56 estimated effects. From these, we were able to calculate 49 comparable effect sizes for utilization in our meta-analysis.

We first conducted a narrative review of all studies and all 56 estimates. This gives an impression of which types of interventions tend to have positive impacts. We complemented the narrative review with a comprehensive meta-analysis of the 49 estimates for which we were able to obtain comparable effect size estimates.

We find an average effect size on first-year persistence rates of 0.08 ($p = 0.01$) and of 0.13 ($p = 0.02$) on six-year graduation rates. We find quite convincing evidence that interventions involving some element of peer group manipulation, in the sense of gender composition or tracking in tutorial groups, have the largest and most consistently positive impacts. We also find a tendency for certain specific courses (e.g. 'University 101') and feedback and support(e.g. student-faculty mentoring programs) to have positive impacts. We do not find any dramatic differences across different estimation methods, geographical regions, or a broad study assessment measure. Neither do we find evidence of outliers or publication bias. We demonstrate in a back-of-the-envelope calculation that the estimated effect sizes are economically important in the sense that they have the potential to generate a lot of additional income at our own institution. We also argue that such effect sizes are likely to be economically important at the societal level.

The remainder of the paper is organized as follows: In the next two subsections, we first discuss the determinants of university dropout and the current evidence on interventions as summarized in various literature reviews. In section II, we discuss methodological issues, and in section III we present a few descriptive statistics of the included studies. Section IV contains the narrative systematic review, while section V presents the meta-analytical results, and section VI discusses results, implications, points to potential future research,

[7]Our last search was July 12, 2021.

and presents study limitations.

A. Determinants of University Dropout

From a theoretical economic perspective, a forward looking student would choose to enroll in university based on a calculation of expected costs (such as tuition fees and foregone earnings while in university) and returns from doing so (see e.g. Becker (1962)). Aina et al. (2018) conduct a review of the economics of university dropout and present a recent theoretical model. In this assessment, an individual decides to stop in the educational system when the productivity obtained through human capital accumulation no longer exceeds the costs of investing in continuing education, irrespective of whether the education is completed or not. This decision naturally depends on a number of characteristics of the potential student, including cognitive ability, social and personality characteristics (e.g. extroversion and conscientiousness), etc. However, the student does not have perfect information, and once enrolled, new information on the cognitive demands, workload, the social environment, events in the family etc. will arrive continuously and influence the decision of whether to continue or to drop out from the study. Similarly, Stinebrickner and Stinebrickner 2012; 2014) point specifically to the importance of learning about ones academic ability and grade performance for the decision to drop out.

The student integration model formulated by the sociologist Vincent Tinto (e.g. Tinto (1975, 1993)) shares many of the same considerations but is more specific when it comes to describing which factors may lead an individual to dropout from college/university and therefore indirectly points to which mechanisms or factors successful interventions could target. He points to the dropout decision as being more closely related to information arriving after enrollment in a dynamic decision process. In particular, students' social and academic integration are seen as important factors. He points specifically to their participation in study groups and their interactions with fellow students and academic staff as being important determinants of dropout.

Both Aina et al. (2018) and Tinto (1975, 1993) adopt the perspective of human capital accumulation and posit that education increases an individual's productivity by equipping them with valuable skills and knowledge. Under this framework, dropping out of an educational institution results in a high cost because the individual misses out on accumulating the remainder of those skills and knowledge, which would have boosted their productivity and, consequently, their earnings potential. Importantly, the human capital acquired is not lost. In contrast, the signalling hypothesis (Spence, 1973) posits that education primarily serves as a 'signal' to employers about an individual's inherent abilities, such as intelligence, perseverance, and commitment. Under this framework, obtaining a degree or diploma does not necessarily increase one's productivity but rather acts as a signal of one's potential productivity. From this viewpoint, dropping out incurs a high cost because the individual fails to send a strong positive signal to potential employers. Without the 'signal' of a completed degree, employers might infer that the dropout lacks the desired qualities, even if this is not the case. The individual might be at a disadvantage in the labor market compared to degree-holders, resulting in potentially lower wages or reduced employment opportunities.

In this context, it is essential to recognize that various factors contribute to the phenomenon of university dropout. Aina et al. (2018), Behr et al. (2020), and Dawson, Kearney and Sullivan (2020) survey the empirical literature on university dropout determinants and altogether identify five groups of determinants: Students' characteristics, abilities and behavior (academic as well as non-academic skills and traits); parental background and family networks; characteristics of the tertiary education system and its institutions; the financial aspects (grants, tuition etc.); and the situation in the labor market. In particular, they note the importance of the students' high school GPA, social skills, and time spent working. Among factors related to the university - which is of particular interest to this review - Aina et al. (2018) mention human and financial resources, admission criteria (including not only GPA but also factors such as the motivation and attitudes towards studying), interaction with peers and university faculty, and for women also the presence of other women in class/study groups. Behr et al. (2020) mention the potential usefulness of data mining techniques for developing early warning systems regarding students at risk of dropping out.

Robbins et al. (2004) examine the association between psychosocial and study-skill factors and retention rates in a meta-analysis. They find that factors such as academic goals (commitment to graduating, preference for long-term goals, degree expectations, desire to finish college, valuing education), academic self-efficacy (academic self-efficacy, academic self-worth, academic self-confidence, course self-efficacy, degree task and college self-efficacy), and academic-related skills (time-management skills, study skills and habits, leadership skills, problem-solving and coping strategies, communication skills) are significantly positively related to retention (ES=0.34, 0.36, 0.37, respectively). However, it is not evident that these associations can be given a causal interpretation. This study nevertheless points to the importance of individual goals, beliefs, and perceived abilities for staying in college. Note that these are all, to varying extent, malleable traits.

From these studies, an important insight emerges; the causes of student dropout are heterogeneous and complex. Hence, any intervention addressing these issues would have to be able to address and/or modify several of these underlying factors. Dawson, Kearney and Sullivan (2020) point to the fact that interventions until quite recently have tended to focus on two aspects; academic under-preparedness and financial aspects (tuition and other costs of studying), while more recent approaches focus on more comprehensive approaches designed to address several challenges at the same time. They mention elements such as mentoring, coaching, case management, financial assistance, academic advice.

B. Current Evidence on Interventions to Prevent Dropout

To the best of our knowledge, there is no other systematic review of the range of interventions that seek to reduce university dropout through locally applicable interventions. A few other reviews have covered subsets of the interventions, but none of them have, to our knowledge, conducted a systematic literature search of the type of interventions that may be implemented locally. Many of the reviews focus on community colleges and four-year colleges together, while our focus is exclusively on studies covering only four-year colleges and/or universities. Moreover, there is a large number of studies reviewing evidence regarding the determinants of university dropout and completion, as discussed above. Below, we summarize the findings of several previous reviews that have focused on specific subsets of interventions similar to those considered in our study.

In an early systematic review, Patton et al. (2006) note that "...no other topic in the published literature has garnered the attention of higher education researchers and administrators more than student persistence." They go on by stating that one of their most important results is the dearth of evidence to support the efficacy claims regarding a wide range of campus-based retention initiatives.

In their review, Andrade (2008) examined a wide range of interventions/ingredients collectively referred to as 'learning communities.' These interventions included mentoring, integrated course contents, academic skills training, and more. The review revealed an overall tendency towards positive effects. However, due to several limitations such as a limited number of relevant studies, selective program participation, and program heterogeneity, it was challenging to pinpoint which specific ingredients were crucial for producing positive outcomes.

Valentine et al. (2011) conduct a systematic review and meta-analysis of college retention programs, but remarks that even the best included studies are methodologically suspect.

Permzadian and Credé (2016) review literature on the effectiveness of first-year seminars, which are defined as courses specifically designed to equip new students with the knowledge, skills, and abilities necessary for overcoming the challenges during the first year in college. They find no significant effect for four-year colleges, neither for GPA (ES=0.01) nor on first-year retention rates (ES=0.10). However, only very few of their included studies are based on randomized trials or other designs allowing a strict causal interpretation.

Sneyers and Witte (2018) conduct a meta-analysis of three types of interventions; academic probation, student-faculty mentoring programmes, and need-based grants. They find, based on 25 (quasi-) experimental studies,[8] that student-faculty mentoring significantly increases retention and graduation rates (ES 0.15 and 0.10, respectively). Academic probation reduces retention rates (ES -0.17) but do not affect graduation rates, while grants have small but significant positive effects on entry, retention, and graduation rates (ES= .5 for all three outcomes).

Dawson, Kearney and Sullivan (2020) focus on eight specific interventions aimed at improving college retention and completion rates, which have all been rigorously evaluated in randomized trials. Some of them target low income first generation students, while others have a broader target group. Five are implemented at community colleges, while three are at 4 year college or university. They all share the feature that they take a comprehensive approach to addressing barriers to completion. Interventions in universities or 4 year colleges all show very small and insignificant impacts on first year retention rates (ES -0.01 - 0.03), while a few of those implemented at community colleges do show significant positive effects, although only one has a significant effect size above 0.2.

In a very broad and comprehensive recent review of non-financial barriers to college access and success, Dynarski et al. (2023) discuss barriers as well as interventions aimed at improving access to as well as completion of college education. They look mostly at the U.S. context and focus on community colleges as well as universities. They find that many broad-based remedial education courses do not work well, while courses taking a more comprehensive approach to preparing and introducing students to the academic expectations and requirements of college show more promise. They find overall positive impacts of intensive in-person counseling as well as multifaceted programs providing students with support across multiple domains that may challenge persistence in college. Finally, they cite evidence that students gain from being taught by high-quality permanent staff members, especially if these staff match students' own demographics in terms of ethnicity, gender, and race. Overall, they tend to find negative impacts of online learning and interventions.

Feygin et al. (2022) review the literature on interventions aimed at improving various outcomes for college students in two- and four-year colleges. They find evidence that interventions taking a holistic approach that combine several components are conducive to better outcomes, and the same holds for interventions incorporating proactive communication and sustained follow-up. Finally, offering non-academic support such as mental health support, child care etc. improves students outcomes.

In recent years, a number of meta-analyses and systematic reviews on specific types of interventions and/or with an otherwise relatively narrow focus have been conducted. In a systematic review of interventions aimed at reducing non-completion, Delnoij et al. (2020) conclude that coaching, remedial teaching, and peer mentoring are promising programs.

[8] In fact, they also include cohort studies as well as unpublished reports.

However, for some reason, their analysis is based on only 10 studies. It is not clear to us whether this relatively low number is due to the number of databases searched, the inclusion criteria or something else. Foster and Francis (2020) conduct a systematic review of a group of interventions they commonly refer to as 'data analytics', which are essentially predictive analytics. Bradford, Beier and Oswald (2021) conduct a meta-analysis of summer STEM bridge programs and find that such programs significantly increase first year retention rates (OR=1.747).

Based on this review of the existing literature in the field, a number of important conclusions emerge:

- Despite the importance of the subject and a relatively large number of studies, existing evidence regarding interventions aimed at reducing dropout rates in universities is still surprisingly weak.
- We do know quite a bit about factors associated with higher dropout risk, and we do know that some of these factors are in fact malleable.
- The existing evidence regarding interventions aimed at reducing dropout mostly show small and insignificant effects.
- Still, student-faculty mentoring programs tend to show significantly positive effects.

I. Methodology

We pre-registered a protocol for the systematic review and meta-analysis with OSF prior to conducting the literature search.[9] Our review complies with the PRISMA[10] guidelines for systematic reviews and meta-analyses as well as the reporting guidelines for meta-analyses in economics (Havránek et al., 2020).

A. Inclusion Criteria

We include studies that present estimations based on evidence from randomized controlled trials (RCTs) or from one of four quasi-experimental designs: instrumental variables (IV), difference in differences (DiD), regression discontinuity (RD) or propensity score matching (PSM). These methods were selected for their ability to control for confounding variables, which is essential for drawing accurate conclusions about causal relationships. While RCTs are considered the gold standard for causal inference, as they involve randomly assigning participants to treatment and control groups, the specified quasi-experimental

[9] See https://doi.org/10.17605/OSF.IO/ZE9D5
[10] Preferred Reporting Items for Systematic Reviews and Meta-Analyses, Page et al. (2021).

designs also provide strong evidence for causality, as they attempt to replicate the conditions of an RCT using observational data. However, there are certain problems associated with the use of PSM in particular, such as the potential for unobserved confounding and the reliance on strong assumptions. On the other hand, PSM can be useful in situations where random assignment is not feasible or ethical. To be included in our analysis, PSM studies were required to include a minimum set of basic characteristics, most importantly gender, high school GPA and some measure of socioeconomic status (SES).

To ensure the relevance and comparability of the studies included in our review, we limit our study population in four ways. First, we exclude studies that focus on interventions at community colleges[11] and instead limit our review to studies that aim to reduce the dropout rate or increase the graduation rate at universities and four-year colleges. The rationale for this criterion is grounded in the significant differences between community college students and university students, particularly in terms of their socioeconomic backgrounds and performance in admissions tests. In particular, students enrolling in community colleges in the United States, where the main literature on community college students origins, are more likely to be underrepresented minorities and economically disadvantaged (Lovenheim and Smith, 2023). At the same time, community college students have traditionally persisted and graduated at relatively lower rates than university students (Denning et al., 2022; Mountjoy, 2022). Although most students who enroll in community colleges initially have the goal of earning a bachelor's degree, the reality is that very few of them actually achieve this goal (Rosenbaum and Rosenbaum, 2013).

Second, we exclude studies that investigate interventions including part-time students. Another important aspect in the exclusion of studies based on community colleges is that students are much more likely to attend part-time than in four-year colleges or universities, resulting in negative effects on credits accumulated (Darolia, 2014). Part-time students may face different challenges and drop out for different reasons than full-time students, as they tend to be older, more financially constrained, and have other competing obligations (e.g. in terms of family). Therefore, part-time students generally tend to have significantly lower completion rates compared to their full-time counterparts (Lovenheim and Smith, 2023).

Third, we exclude studies examining interventions that are part of distance or online programs at universities as well as interventions aimed towards PhD students. Distance or online programs may attract a different type of student, while interventions aimed towards PhD students may not be relevant to the broader population of undergraduate students.

Fourth, we limit the review to interventions conducted at universities located in OECD

[11] We use the terms community colleges and two-year colleges interchangeably when referring to institutions that award associate's degrees as their highest credential.

countries. This ensures that the results stem from relatively comparable settings in terms of economic development and educational systems.

To obtain all the eligible studies, we performed a systematic search in the electronic bibliographic databases of Australian Education Index, Business Source Complete, British Education Index, EconLit, Education Database, Education Research Complete, ERIC, Scopus, and Web of Science. The search terms followed the PICO[12] structure and are shown in Appendix Table A2. We limited the searches to studies published in the period from January 2000 to July, 2021. The cutoff of the twenty-first century ensures that there is a relevance of the findings to current conditions and practices at the universities. In addition, the searches were limited to peer-reviewed publications in English. This restriction means that unpublished studies, such as working papers, dissertations, or conference papers, were not included. Particularly, we want to ensure that the included studies have been through a process of scientific quality assurance. However, this criterion may lead to publication bias if studies with significant results have a higher likelihood of being published. This bias is also known as the "file drawer" problem as studies with null results tend to be filed away. In reality, papers may not even be written up if they contain null results. We address the issue of publication bias in Section IV.B.

After performing the systematic search based on the chosen search terms, the selection process consisted of three stages. First, we retrieved titles and abstracts of the potentially relevant studies and independently screened them to identify studies of relevance to our research question and met our specified inclusion criteria. Second, we independently read the full-text articles of the identified studies to confirm their eligibility for inclusion. For the first two stages, we assessed the interrater reliability by calculating the kappa index (κ). Any disagreements over the eligibility of particular studies were resolved through discussion and consensus. In the third stage, we reviewed the reference lists of the identified literature (backward searching) and newer publications citing identified literature (forward searching) to identify any relevant articles that may have been missed in the first two stages. However, we did not identify any additional studies that met our inclusion criteria in this stage.

Finally, the included studies have to be independent investigations. In the case that two studies report results from the same intervention with an almost identical participant pool, only one of the studies was included; the one with the most comprehensive results.

B. Data Construction

After identifying eligible studies, we assembled the data set by extracting estimation results and selected characteristics of the included studies. To extract data consistently,

[12]P(opulation/problem), I(ntervention/exposure), C(control/comparison), and O(utcome/results). See Appendix Table A1.

we used a pre-specified data extraction sheet designed to capture all information from each study that we assessed ex ante to be relevant. The data extraction sheet included categories such as authors, title, publication year, country, study participants and controls, intervention contents, design and methodology, subgroups, outcomes and times of measurement as well as estimated effects. In the following, we define the non-trivial parts of the extracted data in detail.

'Intervention contents' captures information about the type of intervention that was studied. The included studies involved a diversity of intervention types and contents. Overall, we identified ten different types of interventions, which - for analytical purposes - we grouped further into five crude categories: *feedback and support* cover various types of mentoring; *peer group* covers interventions on study group composition or living & learning communities; *academic skills upgrade* include e.g. specific introductory courses on, say, math; *institutional settings and requirements* are mainly interventions on changing admission policies or academic probation; *specific courses* are a broad category covering courses of the type 'university 101', leadership courses etc.

'Design and methodology' includes information about sample sizes and evaluation design. Our inclusion criteria limited the evaluation designs to RCTS, RD designs, DiD designs, IV methods, and various matching methods, mostly PSM.

The primary outcome measures used are first-year persistence (the reverse of first-year dropout) and graduation within six years. These are binary outcomes. Based on these, we extracted the resulting estimates. To the extent that impacts on first year dropout rates were reported, we reversed the sign of the impact, thus converting it to an impact on first-year persistence.

C. *Standardized Effect Sizes*

To be able to aggregate and compare the effects of the included studies, and to compare to other studies, we standardize the estimated effects into one common metric. We use a standardized effect size called Glass' Δ, which takes the treatment effect and divides it by the standard deviation in the control/comparison group. The advantage of using Glass' Δ is that this effect size is not sensitive to heterogeneous effects; if the treatment affects the variance of the outcome, and, hence, the standard deviation in the treatment group, then measures dividing by the pooled standard error will be sensitive to heterogeneous effects. Where the information available did not allow the calculation of Glass Δ, however, we instead calculated Cohen's d, where the treatment effect is divided by the standard deviation in the pooled sample of treatments and controls/comparisons. Whenever the necessary data to convert the estimates into standardized effect sizes were missing and could potentially be retrieved, we requested the missing information from the study authors.

D. Study Assessment Checklist

To assess the stringency and richness of information in the included studies, we developed a checklist inspired by the National Institutes of Health Quality Assessment Tool for Controlled Intervention Studies (see Appendix Table A3).[13] We independently rated each of the studies and resolved disagreements through discussion. Specifically, the checklist consists of 11 assessment categories. As a study cannot be both a randomised trial and a controlled trial, each study can obtain a score of maximum 10. We group studies with a score of ≥ 7 or above as 'high', studies with a score between ≥ 4 and ≤ 6 as 'medium', and studies with a score a ≤ 3 or below as 'low'. Note that a study — due to the inclusion criteria — must have a score of at least 1 to be eligible for the review.

E. Meta-analysis

We supplement the systematic review with a statistical analysis of the results - a meta-analysis. Meta-analysis is a statistical method used to combine and analyze data from multiple independent studies to draw more generalizable conclusions. The motivation for doing so is that the meta-analysis may reveal patterns and show effects that are not clearly revealed otherwise. For example, a collection of insignificant results all pointing in the same direction may become significant in the meta-analysis. The meta-analysis also weighs each result by its statistical certainty - the lower the standard error on the estimate, the larger the weight. While the systematic review strives to minimize subjectivity and bias through the use of predefined criteria and transparent methods, it is important to acknowledge that a certain degree of judgment in assigning weight to studies is inevitable. In addition, a meta-analysis allows us to calculate the overall effect size based on all the included studies, as well as its uncertainty. Finally, meta-analysis naturally leads to conducting meta-regressions where we can investigate potential moderators and mediators of positive (or negative) effects.

Obviously, there are some drawbacks to meta-analyses when compared to the systematic review. First, it is often not possible to calculate effect sizes that can be used in a meta-analysis for all studies included in the systematic review; some information needed may simply not be present, and contact to authors is often tedious and at worst unsuccessful. Second, meta-analysis may hide more subtle patterns in the data that may be more clearly

[13]A similar rating system was used in Mikkelsen and Rosholm (2018). We specifically do not refer to it as a quality appraisal in this study, since the rating system was developed for a specific medical context. For example, preregistration has not been a tradition within many social science disciplines until very recently, and much of the information on compliance is simply not available in many of the observational studies based on administrative data.

revealed in the systematic review. Therefore, in this paper, we utilize both approaches to present a comprehensive and clear overview of the available evidence.

Both approaches, however, carry the potential for oversimplification, as aggregating findings from diverse studies into a single metric or conclusion can oversimplify complex phenomena into overly simplified representations. Simultaneously, there is a risk of losing valuable contextual insights, as these methods may, at times, omit or diminish the rich context in which the original studies were conducted. These insights may be essential to accurately interpret the results. On the other hand, systematic reviews and meta-analyses rigorously follow predefined criteria for inclusion, analysis, and interpretation. These criteria underscore the importance of minimizing biases and facilitating replication by other researchers, thereby upholding transparency in the process. Additionally, these methods frequently consolidate findings from various contexts, leading to a more comprehensive understanding of a phenomenon and potentially bolstering the generalizability of the results.

For the meta-analysis, we start by considering the model

$$\hat{\theta}_j = \theta_j + \epsilon_j \tag{1}$$

where $\hat{\theta}_j$ is the jth estimate of the true unknown effect size, θ_j, in the data set of K independent studies ($j = 1, ..., K$), and ϵ_j is the sampling error. It is assumed that $\epsilon_j \sim N(0, \sigma_j^2)$ and that the sampling variance σ_j^2 is known. When carrying out a meta-analysis, our aim is to develop a statistical model that allows us to integrate the reported estimates to identify the true effect size. A crucial aspect to consider is whether to adopt a fixed-effects[14] or random-effects approach, which involves distinct assumptions regarding how to model the effect sizes.

The fixed-effects approach assumes that all studies share the same true effect size, i.e. $\theta_j = \theta_{j'} = \theta$ for $j \neq j'$. While observed effect sizes may differ across studies due to sampling error, their true effect sizes are constant and fixed. In the absence of sampling error, if we were to compute the effect size of each study, all true effect sizes would thus be identical. We can obtain the fixed-effects estimate of θ_j using a weighted average of all studies:

$$\bar{\theta}_j = \frac{\sum_{j=1}^{K} w_j \theta_j}{\sum_{j=1}^{K} w_j} \tag{2}$$

where the estimated weights are given by the inverse variance, $w_j = \frac{1}{\sigma_j^2}$, so that estimates with lower variance (and thus higher precision) are assigned larger weights.

[14]Also referred to as a common-effect approach

Unlike the fixed-effects approach, the random-effects approach presupposes that there is not a single true effect size, but rather a distribution of true effect sizes across the studies. Thus, it assumes that effect sizes are different, i.e. $\theta_j \neq \theta_{j'}$ for $j \neq j'$ and that the included studies represent a random sample from a larger population of studies. In addition to the sampling error, this results in an additional error source since the true effect size θ_j is a component of a broader distribution of true effect sizes with mean μ, meaning that $\theta_j = \mu + \zeta_j$ where ζ_j is the sampling error from this distribution. It is assumed that $\zeta_j \sim N(0, \tau^2)$ where τ^2 represents the between-study heterogeneity, which is beyond the mere sampling variability. Specifically, the random-effects model thus assumes that

$$\hat{\theta}_j = \theta_j + \epsilon_j = \mu + \zeta_j + \epsilon_j \tag{3}$$

We can obtain the random-effects estimate from using the weighted average of all studies as in (2), but with adjusted weights given by $w_j = \frac{1}{\sigma_j^2 + \hat{\tau}^2}$, where $\hat{\tau}^2$ is an estimate of τ^2. From τ^2, we are able to quantify the between-study heterogeneity using the I^2 statistic (Higgins and Thompson, 2002). The I^2 statistic is defined as the proportion of total variation in study estimates that is due to heterogeneity between studies.[15] Note that various methods are available to estimate τ^2. We use the restricted maximum likelihood estimator, which is generally recommended based on simulated scenarios (Viechtbauer, 2005; Langan et al., 2019). However, we find similar results for four out of five other methods that we employ as seen in Appendix Table A5.

It is important to note that when $\tau^2 = 0$, the random-effects model reduces to the fixed-effects model. In real-world applications, it is often unrealistic to assume complete homogeneity among studies included in a meta-analysis. Maybe the type of treatment, the duration of the treatment, or the characteristics of the study population vary across the studies. Since we expected a significant level of between-study heterogeneity for exactly these reasons, we employed a random-effects model to pool the effect sizes.

After conducting the initial meta-analysis, we take a step further by performing meta-regressions to try to identify some sources of heterogeneity in effect sizes. This allows us to examine how certain characteristics of the studies and interventions that were included in the meta-analysis may moderate the effect sizes that we observed. Regarding study characteristics, we incorporate variables such as the geographic region where the study was conducted (differentiating between North America and Europe), the score obtained on the study assessment checklist, and the empirical strategy employed. In terms of intervention characteristics, we consider the type of intervention utilized. By analyzing these factors

[15]The calculation of I2 is formally determined by: $I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + s^2} \times 100\%$ where $s^2 = \frac{\sum w_j (K-1)}{(\sum w_j)^2 - \sum w_j^2}$ where K is the number of studies and $w_j = \frac{1}{\sigma_j^2}$

using meta-regressions, we can discern how they might impact the effectiveness of interventions in diverse contexts. Specifically, we estimate a random-effects meta-regression model expressed as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_i + \epsilon_i \tag{4}$$

where Y_i represents the effect size reported in study i, $X_{1i}, \ldots, X_{ki}$ are study-level covariates, $\beta_0, \ldots, \beta_k$ are the regression coefficients, u_i is the between-study variance assumed to follow a normal distribution with mean 0 and variance τ^2, and ϵ_i is the within-study error.

Meta-analysis serves as a powerful tool, aiding researchers in synthesizing findings from various studies and deriving a more accurate estimation of the true effect size. However, the validity of this integrated analysis relies heavily on the assumption that the incorporated studies are statistically independent. This implies that the outcomes of individual studies should not bear any influence on each other. Any breach in this independence can result in skewed or erroneous estimates of the true effect size. For example, if a pair of studies utilize the same data set or participants, the outcomes of one might inadvertently influence those of the other.

Since some of our included studies report multiple estimates, typically from conducting different interventions with the same cohort of students, we employ cluster-robust variance estimation. This approach accommodates correlations of estimates within each study, and thus, it effectively manages inherent dependencies that might occur when multiple effect sizes originate from a single study (Hedges, Tipton and Johnson, 2010). As an alternative approach to address the issue of statistically dependent estimates within each study, we also considered a multi-level approach with three levels, but we find that there is essentially no heterogeneity for the second level (i.e. within studies). Therefore, the results are similar to the standard approach, which consists of two levels.[16]

To further refine the precision of our pooled effect, we incorporate Knapp-Hartung adjustments (Knapp and Hartung, 2003), which especially is appropriate in situations with heterogeneity and a small number of studies (IntHout, Ioannidis and Borm, 2014). This procedure enhances the sensitivity of the confidence interval for the pooled effect size, better reflecting the true underlying variability in the effect sizes across different studies. Specifically, the Knapp-Hartung adjustment takes into account the potential for overestimation of between-study variance that can occur in standard random-effects meta-analysis, leading to more conservative and reliable confidence intervals.

Our results are robust to not applying either the cluster-robust variance estimation or

[16]Results are available upon request.

the Knapp-Hartung adjustments, as evidenced in Appendix Table A6.

Publication Selection Bias. — Publication bias, or more broadly, selective reporting, refers to the situation where studies selected for a scientific review systematically diverge from all relevant available studies. Specifically, publication bias is recognized in the context of meta-analysis as the correlation between the probability of a study being published and the statistical significance of its results. Studies may not be accepted for publication with non-significant findings.

In the scenario where we are missing some studies in our meta-analysis, the implications for our results depend on whether these missing studies are a random sample of all pertinent studies, or if there's a systematic discrepancy. If the unaccounted-for studies serve as a random sample of all relevant studies for our research query, our meta-analytic results, although slightly less precise, will generally remain valid. However, if the missing studies systematically diverge from the known studies - for instance, if smaller studies with non-significant outcomes are consciously not published - it can induce bias in our meta-analytic results. The presence of systematic bias in meta-analyses may have significant implications for policy decisions or interventions. However, it can be argued that meta-analysis offers a partial remedy to the problem since researchers have the opportunity to identify and estimate the impact of bias by examining the information within the distribution of effect sizes from the studies at hand.

One of these methods is the use of funnel plots, which are scatter plots of effect sizes from individual studies ($\hat{\theta}_{ij}$) against a measure of each study's size or precision (i.e. the inverse of the estimates' standard errors: $1/SE_{ij}$) (Light et al., 1984). Asymmetry in these plots can indicate the possibility of publication bias, but it is important to recognize that factors other than publication bias can also contribute to the observed asymmetry. These factors may include the presence of a moderator variable that is correlated with the study effect, or the existence of significant heterogeneity between studies. The FAT (Funnel Asymmetry Test) offers a more systematic methodology for assessing asymmetry in funnel plots (Stanley and Doucouliagos, 2012). Specifically, the test uses a simple regression model to check the presence of bias:

$$\hat{\theta}_j = \beta_0 + \beta_1 SE_j + \epsilon_j \tag{5}$$

where $\hat{\theta}_j$ again is the estimated coefficient of interest (here, the effect size on first year persistence or six year graduation), and SE_j is the associated standard errors of the effect size. The regression is based on the presumption that the reported effect, *ceteris paribus*, is positively correlated with its standard error when publication bias is present.[17] Consequently, the FAT (Funnel Asymmetry Test) infers the presence of publication bias by testing $H_0 : \beta_1 = 0$ (Egger et al., 1997; Stanley and Doucouliagos, 2012) and is considered as a test of whether the funnel graph is asymmetric. In contrast, the PET (Precision Effect Test) tests $H_0 : \beta_0 = 0$ in Equation 5, and thus, whether a true empirical effect is present, independent of any potential distortion caused by publication bias. The intercept serves as a selection-corrected effect-size estimate since the intercept represents an extrapolated effect size with zero SE and hence perfect precision (Vevea, Coburn and Sutton, 2019). However, PET is known to be biased when there is heterogeneity across studies. To account for this issue, the PEESE (Precision Effect Estimate with Standard Errors) step is implemented if PET is significant (Stanley and Doucouliagos, 1997). PEESE regresses effect sizes on the square of standard errors as seen in Equation 6, unlike PET which regresses on standard errors. While both PEESE and PET methods result in biased intercepts, there is a notable distinction in the degree of bias exhibited. PEESE, despite its inherent bias, demonstrates a relatively smaller extent of distortion in the intercept estimation when compared to PET.

$$\hat{\theta}_j = \alpha_0 + \alpha_1 SE_j^2 + \epsilon_j \tag{6}$$

It is important to note that the variance of ϵ_j in this regression will differ across estimates. Consequently, it is commonly recommended to employ weighted least squares (WLS), where the inverse of the variance of the study's estimate $(1/SE_j^2)$ is utilized as the weight (Stanley and Doucouliagos, 2012).

While these methods are helpful tools, they are not infallible, and their results should be interpreted with caution. It is well-known that they suffer from low power and have difficulty in incorporating heterogeneity. Additionally, it is important to keep in mind that they do not confirm the cause of the asymmetry. For instance, in the case of RCTs, researchers might deliberately plan their studies to attain a predetermined level of statistical power. When anticipating a small effect size, a larger sample size may be chosen to ensure the study's ability to detect the effect, resulting in a smaller standard error. Conversely, when expecting a large effect size, a smaller sample size may suffice, leading to a larger standard error. This "endogenous precision" can introduce an asymmetry that is not attributable to publication bias but rather stems from the study's inherent design choices. The presence of "endogenous precision," particularly in RCT contexts, underscores the necessity of approaching funnel plot results and statistical methodologies informed by its principles with caution and considering alternative explanations for observed asymmetries.

[17] In contrast, in accordance with the conventional t-test and random sampling theory, estimates and their standard errors should be independent of each other.

II. Descriptive Statistics

Figure 1 shows a PRISMA flow diagram of the search results. The literature search yielded 13992 potentially relevant publications (7235 after the removal of 6757 duplicates). These 7235 studies were screened against title and abstract, which resulted in the exclusion of 7058 studies.

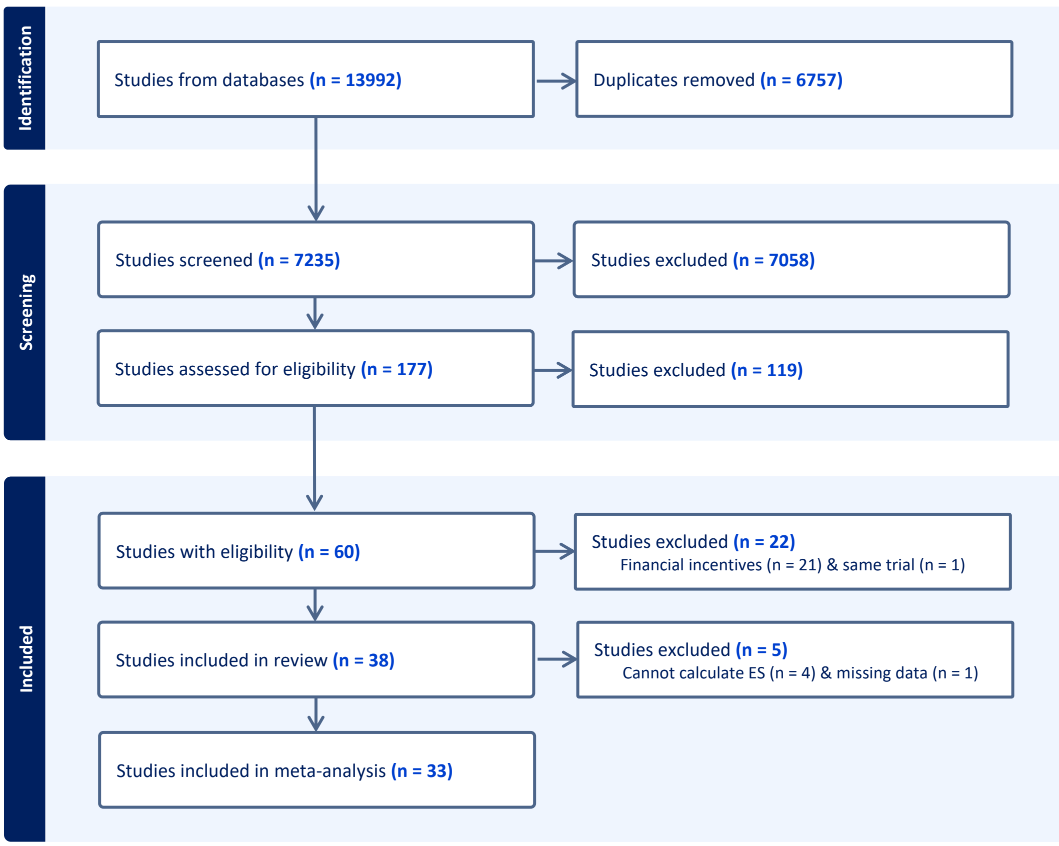


Figure 1. : PRISMA Flow Diagram

The remaining 177 studies were then assessed for full text eligibility. 119 articles were excluded in this second stage, leaving 60 studies. As we were mainly interested in interventions that could be implemented at the local (department, faculty or university) level, we finally decided to exclude 21 studies focusing on financial incentives, which are typically implemented at state or national levels. This left 39 studies to be included in the review. Two of these reported results from the same trial, so we only included the one with the most comprehensive set of results, thus leaving us with a final of 38 studies in the systematic review.

The interrater reliability was 0.55 in the first stage, while it was 0.49 in the second stage. Agreement about inclusions and exclusions was 97 % in the first stage, and 75 % in the second stage. A lot of studies were not included because they did not separate between community college students and university students. Thus, while the interrater reliability is not impressive, most disagreements were caused by the difficulty of assessing the possibility of extracting results from college students separately from community college students.

Other reasons for disagreement were whether or not interventions were exclusively conducted during the first year, that the chosen evaluation design did not meet the inclusion criteria (e.g. a PSM without the pre-specified minimum set of covariates), and that no relevant outcomes were reported.

Table 1 presents the descriptive statistics of the sample. Since some studies reported more than one result (i.e. both first-year persistence and six-year graduation rates, or results for several mutually exclusive groups), the table contains a column that shows descriptive statistics by study and one that shows descriptive statistics by estimates. Table A2 in the Appendix contains a more detailed description of each study.

As already mentioned, for some studies, the information did not allow the calculation of standardized effect sizes. For some studies there was not an estimate of the effect, we could only assess its sign and significance, for instance because the treatment was not dichotomous (e.g. a class size as in Bettinger and Long (2018)). In those cases, the study was only included in the narrative systematic review. These studies were Bettinger and Long (2018), Oosterbeek and van Ewijk (2014), Soria and Werner (2018), and Zilvinskis, Borden and Severtis (2020).

For the remaining studies with insufficient data for calculating effect sizes, we contacted the authors to request the missing information. We contacted the authors of eight studies in total and obtained information from six of them. The two studies where we were unable to obtain information were also not included in the meta-analysis. These were Chiteng Kot (2014) and Casey et al. (2018) (for the latter, only information for the impact on six-year graduation rate was missing). Hence, we were able to calculate standardized effect sizes for 49 out of 56 of the impact estimates included in the narrative review. The characteristics of the estimates included in the meta-analysis are shown in the third column of table 1.

27 studies (42 estimated effects) were from North American universities, while 11 (14) were from European universities. In terms of evaluation method, 10 studies (15 effects) were based on data from randomized trials, while the remaining studies used observational methods - 6 (13) used regression discontinuity designs, 2 (2) used difference-in-differences designs, 3 (5) used instrumental variables techniques, and 17 (21) were based on some type of (propensity score) matching strategy.

42 estimates were first year persistence/dropout, while 14 were six-year graduation rates.

Table 1—: Descriptive statistics

	Studies	**Results**	**Meta-analysis**
Number	38	56	49
Type of intervention			
Feedback and support	12	15	14
Peer groups	8	12	11
Academic skills upgrade	6	12	12
Institutional setting and requirements	7	10	6
Specific courses	5	7	6
Region			
North America	27	42	36
Europe	11	14	13
Experimental method			
RCT	10	15	14
RDD	6	13	12
DID	2	2	1
IV	3	5	3
Matching estimator	17	21	19
Outcome type			
First year persistence		42	37
Six year graduation		14	12
Check list index			
7-10	13	25	23
4-6	16	22	20
1-3	9	9	6
Effect			
Positive		24	22
Insignificant		28	24
Negative		4	3

The included studies involved a diversity of intervention types and contents. Overall, we identified ten different types of interventions, which - for analytical purposes - we grouped further into five crude categories:

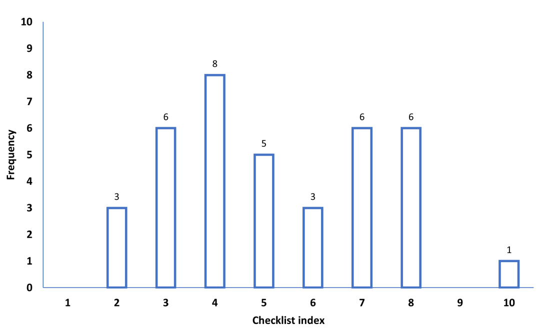
Twelve studies (15 estimates) were of the type *feedback and support* (e.g. various types of mentoring); Aulck et al. (2016), Bergey et al. (2019), Campbell and Campbell (2007), Chiteng Kot (2014), Clark and Cundiff (2011), Culver and Bowman (2020), Deighton et al. (2019), Dennehy and Dasgupta (2017), Dobronyi, Oreopoulos and Petronijevic (2019), Ellis and Gershenson (2020), Plak et al. (2020), and Stegers-Jager, Cohen-Schotanus and Themmen (2013).

Eight studies (12 estimates) involved some kind of *peer group* component (e.g. study group composition or living & learning communities); Booij, Leuven and Oosterbeek (2017), Booth, Cardona-Sosa and Nolen (2018), Johnson et al. (2020), Masserini and Bini (2021), Nix, Roberts and Hughes (2016), Nosaka and Novak (2014), Oosterbeek and van Ewijk (2014), and Reynolds (2020).

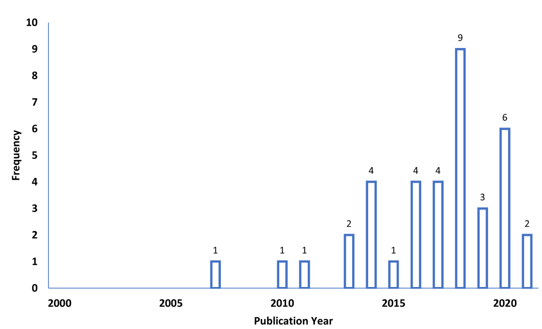
Six studies (12 estimates) had elements of *academic skills upgrade* (e.g. specific introductory courses on, say, math); Attewell and Monaghan (2016), Boatman and Long (2018), De Paola and Scoppa (2014), Duchini (2017), Findley-Van Nostrand and Pollenz (2017), and Wachen, Pretlow and Dixon (2018).

Seven studies (10 estimates) implemented changes in *institutional settings and requirements* (e.g. changing admission policies or academic probation); Bettinger and Long (2018), Carrieri, D'Amato and Zotti (2015), Casey et al. (2018), Lindo, Sanders and Oreopoulos (2010), Novak, Paguyo and Siller (2016), Tafreschi and Thiemann (2016), and Zilvinskis, Borden and Severtis (2020).

Finally, five studies (7 estimates) contained *specific courses* (e.g. 'university 101', leadership courses etc.); Bowman and Culver (2018), Bowman and Holmes (2018), Keller and Lacy (2013), Song et al. (2018), and Soria and Werner (2018).



(a) Studies by Checklist Index



(b) Studies by Publication Year

Table A2 in the online appendix provides the study assessment checklist values for each of the included studies. Figure 2a shows the checklist index. The average index value is 5.6, and the median is 5.5 (28 at 5 or below, 28 at 6 or above). Table A2 reveals that low scores are primarily due to lack of pre-registration, lack of sufficient control variables, missing description of compliance rates, missing (precise) definitions of the outcome variable, etc.

Figure 2b describes the studies by year of publication. The figure reveals that no studies from 2000 until 2006 were included, and only five studies were published before 2014. This is in part due to lack of studies, and in part due to exclusion of studies due to not meeting the inclusion criteria.

III. Systematic Review

Table 1 shows that of the 56 results found in the 38 included studies, 24 (43%) had a significantly positive effect, 28 had no significant impact, while 4 had a significantly negative impact. Hence, we find a clear tendency to positive over negative effects, although half of the studies produce effects that are not significantly different from zero.

If we look at effects by region, the results are identical; both in North America and Europe, 43% have significantly positive effects and 7% have significantly negative impacts. Split by estimation method, 4 (33%) of the RCTs had significantly positive effects (and none negative), 1(8%) of those based on regression discontinuity designs were significantly positive (3 negative), while those based on matching type estimators had 17 (74%) significantly positive effects (none negative). The two remaining methods were used in too few studies to meaningfully infer anything.

For the estimates on first-year persistence rates, 17 (40%) had positive effects (4 negative), while for impact estimates on six-year graduation rates, 7 (50%) were significantly positive (none negative).

Of the studies rated 7-10 on the study assessment checklist, 6 (24%) had significantly positive effect (2 negative), for those rated 4-6, 13 (59%) were significantly positive (2 negative), while for the scoring 3 or below, 5 (56%) were significantly positive (none negative).

Finally, if we split the estimates by the type of intervention, those involving *feedback and support* had 7 (47%) significantly positive estimates (none negative), interventions containing *peer group* elements had 7 (58%) positive and none negative, those involving *academic skills upgrade* had 3 (25%) significantly positive estimates (none negative), interventions regarding *institutional settings and requirements* had 2 (20%) significantly positive and 3 negative, and *specific courses* had 5 (71%) significantly positive (none negative).

Summing up this counting exercise, we find a tendency that RCTs and studies based on RDD designs tend to produce fewer positive effects, while matching estimators in particular are likely to have positive effects. Studies with a higher checklist rating are slightly less likely to be positive, while interventions with *peer groups* or *specific courses* are most likely to have positive effects, and interventions involving changes in *institutional settings and requirements* are least likely to have positive effects.

We now take a closer look at a few of those studies that reveal significantly positive and relatively large effects. First, we look at interventions containing *peer group* elements. Two studies stand out; those by Booij, Leuven and Oosterbeek (2017) and Booth, Cardona-Sosa and Nolen (2018). They both used RCT methods, found relatively large effects, and scored relatively high on the checklist index. Booij, Leuven and Oosterbeek (2017) use data on undergraduate students in economics from a Dutch university and randomize the composition of tutorial groups. They find that three-way tracking by ability increases first-year persistence rates for those with low and medium ability by 12 percentage points, relative to an average persistence rate of 60 percent.

Booth, Cardona-Sosa and Nolen (2018) randomly assign first-year students to all-female, all-male or coed classrooms for one hour per week. They find that women assigned to all-female classrooms have significantly lower dropout rates than women in coed classrooms. The remaining studies in this category (Johnson et al. (2020); Nix, Roberts and Hughes (2016); Nosaka and Novak (2014)) are U.S. based and are in the 'Living and Learning Community' or 'Learning Community' category. They all use some variant of matching technique, and they all score quite low on the checklist index.

The other category that stands out with relatively many significantly positive effects is *specific courses*. These are a mixture of intervention types (honors programs, service-learning and academic leadership courses) and no particular intervention stands out. They are all conducted in the U.S, use either matching or IV techniques and all score relatively low on the checklist index.

Among the remaining studies with significant positive effects, we should also mention the studies by Deighton et al. (2019) and Dennehy and Dasgupta (2017) which also generate relatively large effect sizes, score well on the checklist index, and are based on RCT evaluations. Deighton et al. (2019) study a supportive text messaging intervention employed during potentially stressful periods of the academic year aimed at students with low emotional intelligence. Students were (quasi-) randomly assigned to either a treatment or a control group, and those in the treatment group had 17 percentage points lower dropout rates than individuals in the control group.

Dennehy and Dasgupta (2017) conduct a randomized experiment of a peer mentoring intervention on women's retention in engineering. They find that dropout rates are reduced by 13 percentage points, when women are are assigned to female peer mentors.

Based on this narrative review, we would conclude that the most convincing evidence shows that especially interventions involving some element of systematic and selective peer group interaction, either by gender or by ability grouping, shows promising effects.

IV. Meta-analysis

Utilizing a random-effects model in our meta-analysis, we systematically estimate the overall effect of the interventions on first-year persistence and six-year graduation. Descriptive statistics for the 49 effect sizes included in the meta-analysis are shown in the third column of Table 1. Table 2 reports the results, which are also depicted in forest plots in Figures 3 and 4 for first-year persistence and six-year graduation, respectively. The pooled effect size for first-year persistence is 0.076, which is statistically significantly different from 0 at the 5% significance level, and implies that the interventions on average significantly improved first-year persistence. Similarly, we find a significant positive effect size of 0.125 for the included studies that consider six-year graduation as an outcome.

Table 2—: Effect sizes

	First-year persistence	Six-year graduation
Estimate	0.075	0.124
	(0.028)	(0.044)
p-value	0.012	0.022
τ^2	0.018	0.009
I^2	94.43%	87.37%
Prediction interval	[-0.204, 0.353]	[-0.116, 0.363]
Observations	37	12
Studies	31	9

While the estimated effect sizes vary from -0.42 to 0.39 for first-year persistence, the effect sizes vary slightly less for graduation (from -0.21 to 0.32). The prediction intervals also demonstrate considerable fluctuation. Although the upper boundaries of the prediction intervals appear to be the same for both outcomes, the lower boundary is marginally more negative for the results of first-year persistence. The fact that these prediction intervals encompass negative values suggests that the possibility of negative intervention effects cannot be disregarded in future studies. From the I^2 statistics, we further observe that 94% of the total variability in effects sizes for first-year persistence and 87% for six-year graduation is due to between-observation heterogeneity rather than sampling variance. Both of these numbers are fairly large. Such considerable heterogeneity indicates that there is quite strong evidence against a single true effect size as argued by a fixed-effects approach and implies that the effect of the interventions varied across the different studies.

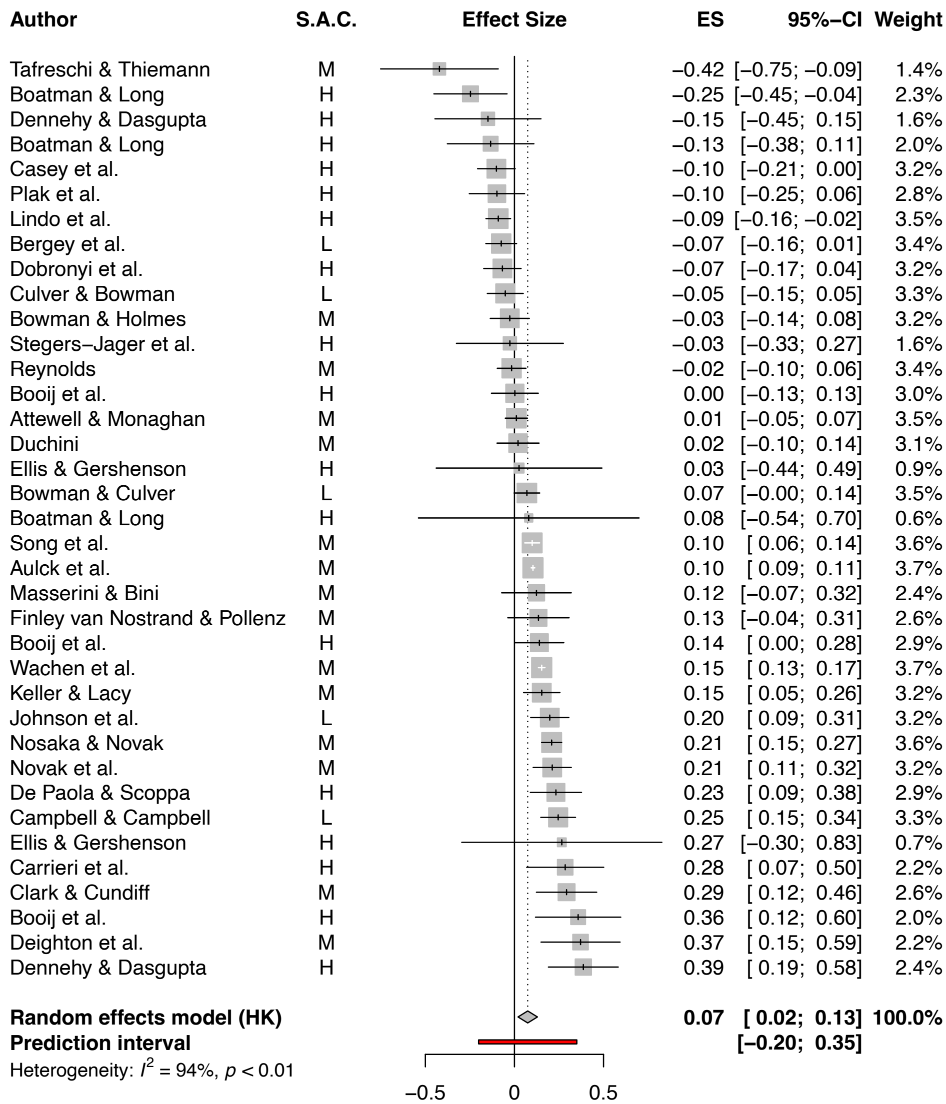


Figure 3. : Forest plot: first-year persistence

This high heterogeneity underscores the need for careful interpretation of our findings, acknowledging the possibility that the observed effects may not be universally applicable across all settings and populations. It also provides a strong rationale for searching for moderators and mediators of effect sizes in meta-regression analyses.

Table 3 presents the results from meta-regressions conducted to analyze subgroup effects. Meta-regressions were only conducted for the outcome 'first-year persistence', as

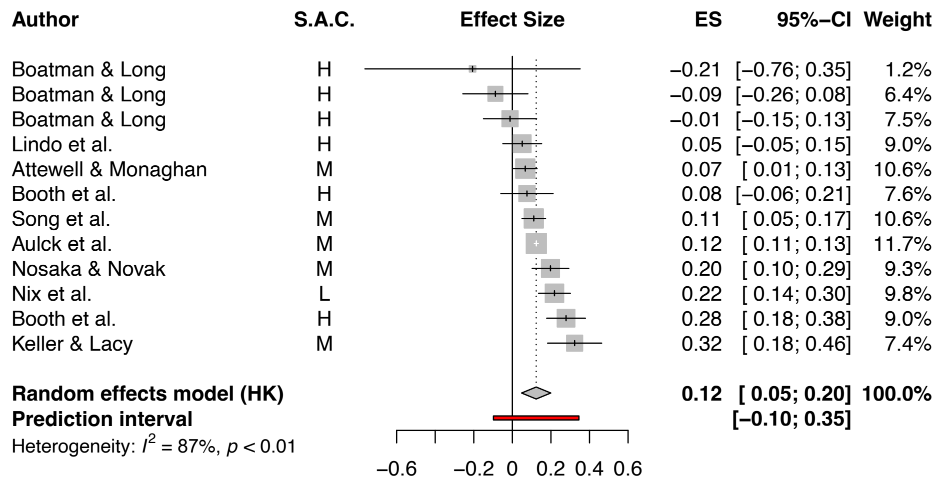


Figure 4. : Forest plot: six-year graduation

the number of estimates for six-year graduation was too small for subgroup analyses. We conducted separate meta-regressions for four different potential moderators: intervention types, study assessment checklist, empirical methods, and the study setting.

In terms of intervention types, the five different categories show estimated effect sizes between -0.01 and 0.13. The only significantly positive effect size arises for interventions in the *peer groups* category (ES=0.13). *Feedback and support* and *specific courses* tend to have positive effects, but none of them are statistically significant. For *academic skills upgrade* and *institutional settings and requirements* effect sizes are very small. The Study Assessment Checklist analysis reveals that studies assessed 'medium' on this scale have significant effects (ES=0.11), while studies with lower or higher scores reveal somewhat smaller and statistically insignificant effects.

When looking at the statistical method involved, we had to group RD, DID, and IV designs into one category. We found that studies using 'Matching' methods found positive effects (ES=0.13), while neither 'RCT' or 'Others' showed any significant effects. This is of course a little worrying, as we would, all else equal, have more faith in studies based on randomization or some quasi-random identification strategy than we would in matching strategies.[18]

Finally, we found statistically significant effects only for studies conducted in North America (ES=0.10) but not for Europe. This would appear largely to be caused by statis-

[18]Note that we did put some minimum requirements—described in the section on inclusion criteria—on matching studies for them to be included in the review.

Table 3—: Meta-regressions / subgroup analyses

Subgroup	Number	**Estimate**	**SE**	**P-value**	τ^2	I^2	N_m	N_s
Intervention	Feedback and support	0.09	(0.04)	0.07	0.03	90.87%	13	11
	Peer groups	0.13	(0.04)	0.02	0.01	76.22%	7	5
	Academic skills upgrade	0.04	(0.07)	0.52	0.02	87.93%	8	6
	inst. settings etc.	-0.01	(0.13)	0.96	0.06	93.46%	5	5
	Specific courses	0.08	(0.04)	0.11	0.00	49.99%	4	4
S.A.C.	1-3	0.08	(0.06)	0.29	0.02	89.05%	5	5
	4-6	0.11	(0.04)	0.02	0.01	94.41%	15	15
	7-10	0.04	(0.05)	0.42	0.03	80.56%	17	11
Methods	RCT	0.08	(0.05)	0.12	0.03	79.91%	12	8
	Others	-0.03	(0.07)	0.65	0.03	85.08%	10	8
	Matching	0.13	(0.03)	0.00	0.01	92.67%	15	15
Continent	Europe	0.10	(0.07)	0.16	0.03	77.90%	11	9
	North America	0.07	(0.03)	0.04	0.02	95.46%	26	22

Note: N_m refers to number of effect sizes, and N_s refers to number of studies.

tical power issues, as the estimated effect size for Europe is actually larger than for North America.

A. *Robustness Check*

We conduct a leave-one-out meta-analysis to examine whether some individual studies have an unduly large impact on the estimated effect sizes. The effect sizes from the leave-one-out meta-analysis are shown in Figures 5 and 6. It is clear that the overall effects size is very robust to leaving out any single study, as the effects sizes vary from 0.07 to 0.08 and 0.11 to 0.14, respectively, and its statistical significance does not change either. The effect is thus robust to the exclusion of single studies.

B. *Publication Bias*

As argued in Section I.E, the funnel plot serves as a useful first step for detecting publication bias. Figure 7 shows funnel plots for first-year persistence in the left panel and six-year graduation in the right panel. We find no clear indication of publication bias, as the funnel plots appear fairly symmetric. Note that the y-axis in funnel plots is inverted (meaning that "higher" values on the y-axis represent lower standard errors). Appendix Figure A1 show contour plots, which includes significance levels to the funnel plots.

Table 4 presents the results for publication bias from the FAT (Egger's regression test) and PET Equation 5 in columns 1 and 3 as well as the PEESE Equation 6 in columns 2 and 4. First, the FAT test reveals no significant publication bias, neither for first-year persistence, nor for six-year graduation. Second, the PET and PEESE tests indicate

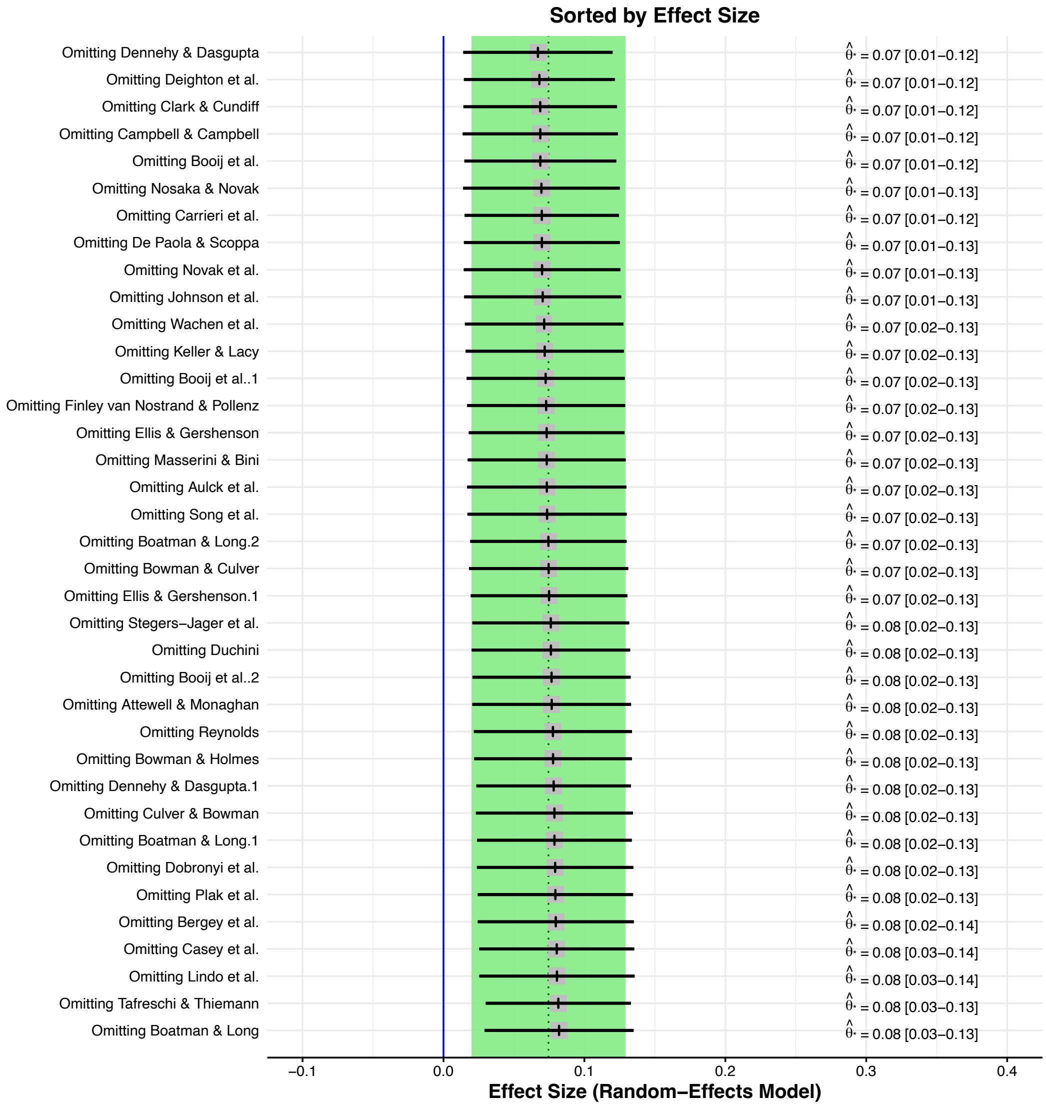


Figure 5. : Leave-One-Out: first-year persistence

that there are indeed significant effects present, even when a correction for (insignificant) publication bias is performed. In sum, we find no evidence of publication bias and hence also no indication that the significant effect sizes we find are generated by publication bias.

V. Conclusion and Discussion

In the present review and meta-analysis, we evaluated what types of interventions can reduce the likelihood of student dropout in higher education. In the systematic review,

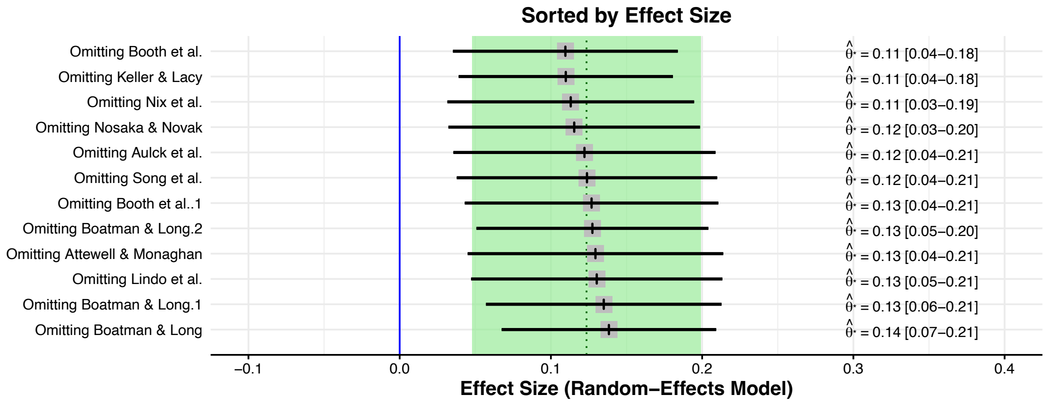
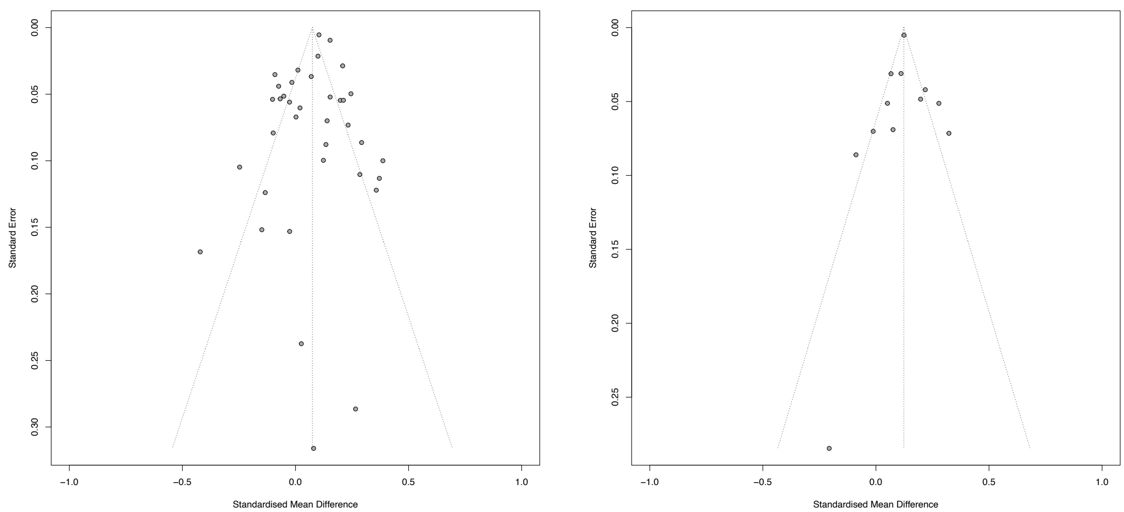


Figure 6. : Leave-One-Out: six-year graduation



(a) First-year persistence

(b) Six-year graduation

Figure 7. : Funnel plots

we argue that the most convincing evidence shows that especially interventions involving some element of systematic and selective peer group interaction, either by gender or by ability grouping, shows promising effects. We also tend to find that studies based on PSM type estimation methods produce more significantly positive effects and that RCTs tend to produce fewer significantly positive effects, although some of those studies that produce the largest effects are in fact based on RCTs. Finally, we find a tendency that studies that

Table 4—: Tests for publication bias

		First-year persistence		Six year graduation	
		(1)	(2)	(3)	(4)
SE of estimate	β_1	-0.681	-2.812	-0.063	-3.186
		(0.503)	(3.788)	(0.709)	(6.020)
Constant	β_0	0.116	0.108	0.125	0.125
		(0.013)	(0.011)	(0.012)	(0.010)
FAT (H_0: $\beta_1 = 0$)	p-value	0.184	0.463	0.930	0.608
PET / PEESE (H_0: $\beta_0 = 0$)	p-value	0.000	0.000	0.000	0.000
Observations		37	37	12	12
Number of studies		31	31	12	12

score highest on the study assessment checklist produce fewer significantly positive effects.

In the meta-analysis, we split the analysis by whether the dependent variable was first-year persistence or six-year graduation. We found significant positive effects in both cases (ES=0.08 and 0.13, respectively). When looking at meta-regressions, we found significantly positive, and larger than average, effect sizes for interventions involving manipulating peer groups by e.g. gender or ability. We found a tendency for interventions involving *feedback and support* as well as *specific courses* to produce positive effects, although none of these were statistically significant. We found that the meta-analysis was robust to leaving out single studies, and we found no evidence of publication bias. While previous reviews overall tended to conclude that interventions involving mentoring and student-faculty interactions showed promising effects, we do find a tendency to the same in this study (those intervention are labeled *feedback and support* in our study), but we find clearly the strongest evidence for studies manipulating peer groups in a systematic way. This result is present in both the narrative systematic review as well as the meta-analysis, hence, we are quite confident that this is the type of intervention we would recommend for further study.[19]

We now briefly discuss the economic importance of the estimated effect sizes and the potential gains from interventions relative to the costs of interventions. We do so primarily in the context of our own department where we have a good grasp of the economics involved. Interventions involving peer groups showed significantly positive effects, while two types of interventions tended to have positive effects; those involving various types

[19] Indeed, together with a few colleagues, we are currently conducting a randomized trial at our own department, where we manipulate the composition of study groups by gender and ability.

of *specific courses* and those involving *feedback and support*. Peer group interventions - to the extent that they just re-allocate students in teams or study groups - are likely almost zero-cost interventions, while *specific courses* as well as certain types of *feedback and support* are also perceived to be relatively low cost (per student). From the perspective of the department, the cost of dropout is the lost subsidy (or tuition) from the government. At our department, this corresponds to around DKK 170,000 per students dropping out during the first semester, assuming, perhaps optimistically, that they would otherwise all have continued to complete a Master's degree.[20]

As an example, the study of tracking study groups by ability by Booij, Leuven and Oosterbeek (2017) found effect sizes of 0.36 standard deviations for the third of the students with lowest GPA. With a conservatively specified dropout rate of 50% in the lowest third and 25% in the medium ability group, this corresponds to a 18 percentage point reduction in the dropout rate for the third of the student intake with the lowest GPA and a 6% reduction for the medium group. Hence, the potential gains from such an intervention are large from the perspective of a single department. With an annual intake of, say, 1000 students, this would amount to additional permanent funding of (170,000 x 0.18 x 1,000 x 1/3) + (170,000 x 0.06 x 1,000 x 1/3) = DKK 13,600,000 per year. However, a reduced drop-out rate would naturally also require additional teaching resources, which would lead to an increase in costs. Hence, the numbers given above should be perceived as upper bounds on the potential gains. Nevertheless, the total potential gain in subsidies to the department from such an intervention would be considerable. Even with an average effect size of 0.08, the dropout rate will be reduced by around 3.5-4 percentage points for the targeted group, which - depending on the size of the student cohort - is likely to justify a considerable investment in dropout reducing interventions.

When it comes to society's perspective, a number of additional factors come into play; first of all, some dropouts are productive in the sense that students go on to complete alternative educations yielding possibly higher productivity, while others complete lower-level education, and yet others leave the educational system entirely. Second, there is the reduced time spent in the educational system for those switching to alternative educations - this time saving may lead to earlier labour market entry for those students. Third, the returns to education is typically estimated at 6-11% per year, which amounts to 30-55% additional earnings per year for the rest of one's working life, if the alternative is to leave the educational system. On the other hand, the student then has five additional years in the labour market, where he or she adds to GDP. Hence, the calculation from society's perspective is more complex, but we would argue that reductions in dropout rates of the magnitudes discussed above are likely to yield large overall benefits to society as well.

[20]Since we found effect sizes of slightly larger magnitude on six-year graduation rates, albeit on a subset of studies and results, this may not be overly optimistic after all.

We find that the existing literature on interventions aimed at reducing dropout and/or increasing completion rates could be improved and extended, given the perceived benefits from such interventions. First of all, it is quite surprising to us, given that universities consist of researchers, that this is such an under-researched area with so relatively few randomized trials. Second, there is a general lack of studies in the literature scoring high on the study assessment checklist. This is particularly due to lack of pre-registration, lack of power calculations, and lack of information on intervention compliance and fidelity. We would urge future studies to pre-register protocols, conduct ex-ante power calculations, and closely monitor implementation fidelity, as it was done in the study by e.g. Dobronyi, Oreopoulos and Petronijevic (2019). We would also note that we excluded many studies based on matching methods due to inadequate controls (in particular, high school GPA or similar controls), which further adds to the impression that quite many studies in the area has surprisingly low quality.

Third, there is a dramatic shortage of research that distinguishes between study, institution, and system dropout. However, it is important to distinguish between study changes within and between universities as well as switches to lower level education and system dropout, since these all carry different costs. Yet, to our knowledge, no studies have pursued this issue, likely due to data availability.

The present study certainly has some limitations. First, any meta-analysis suffers from sensitivity to the chosen search-string as well as the specified inclusion/exclusion criteria (year of publication, applied methodology, country of study origin etc.). However, both were laid out in advance in the pre-registered protocol, so anyone should be able to reproduce and amend the analysis. Second, we were not able to calculate effect sizes using Glass' Δ in all cases, so in some cases we used Cohen's d instead, which may lead to a (presumably minor) source of uncertainty in the meta-analysis.

This paper aimed to systematically analyze the impact of different interventions on student dropout and graduation rates in higher education. Through a meticulous screening process of approximately 7250 abstracts, we identified 38 (quasi-)experimental studies that met our inclusion criteria. By conducting a narrative systematic review and meta-analysis of these studies, we discovered a positive overall trend, with significant positive effects observed in both first-year persistence and six-year graduation rates.

REFERENCES

Aina, Carmen, Eliana Baici, Giorgia Casalone, and Francesco Pastore. 2018. "The Economics of University Dropouts and Delayed Graduation: A Survey." *IZA Discussion Paper 11421.*

Andrade, Maureen S. 2008. "Learning Communities: Examining Positive Outcomes." *Journal of College Student Retention*, 9(1): 1–20.

Autor, David H. 2014. "Skills, education, and the rise of earnings inequality among the "other 99 percent"." *Science*, 344(6186): 843–851.

Barro, Robert J. 2013. "Education and Economic Growth." *Annals of Economics and Finance*, 14(2): 301–328.

Becker, Gary. 1962. "Investment in Human Capital. A Theoretical Analysis." *Journal of Political Economy*, 70(5): 9–49.

Behr, A., M. Giese, H.D.T. Kamdjou, and K. Theune. 2020. "Dropping out of university: a literature review." *Review of Education*, 8(2): 614–652.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2021. *Introduction to Meta-Analysis, 2nd Ed.* Wiley.

Bradford, Brittany C., Margaret E. Beier, and Frederick L. Oswald. 2021. "A Meta-analysis of University STEM Summer Bridge Program Effectiveness." *CBE-Life Sciences Education*, 20:ar21: 1–14.

Card, David. 1999. "The causal effect of education on earnings." *Handbook of Labor Economics*, 3: 1801–1863.

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

Dawson, R.F., M.S. Kearney, and J.X. Sullivan. 2020. "Comprehensive approaches to increasing student completion in higher education: A survey of the landscape." *NBER Working paper 28046.*

Delnoij, Laurie E. C., Kim J. H. Dirkx, José P.W. Janssen, and Rob L. Martens. 2020. "Predicting and resolving non-completion in higher (online) education – A literature review." *Educational Research Review*, 29: 1–17.

Denning, Jeffrey T., Eric R. Eide, Kevin J. Mumford, Richard W. Patterson, and Merrill Warnick. 2022. "Why Have College Completion Rates Increased?" *American Economic Journal: Applied Economics*, 14(3): 1–29.

Dynarski, Susan, Aizat Nurshatayeva, Lindsay C. Page, and Judith Scott-Clayton. 2023. "Chapter 5 - Addressing nonfinancial barriers to college access and success: Evidence and policy implications." In . Vol. 6 of *Handbook of the Economics of Education*, , ed. Eric A. Hanushek, Stephen Machin and Ludger Woessmann, 319–403. Elsevier.

Dynarski, Susan, Lindsay C Page, and Judith Scott-Clayton. 2022. "College Costs, Financial Aid, and Student Decisions." National Bureau of Economic Research Working Paper 30275.

Egger, Matthias, George Davey Smith, Martin Schneider, and Christoph Minder. 1997. "Bias in Meta-Analysis Detected by a Simple, Graphical Test." *British Medical Journal*, 315(7109): 629–34.

Feygin, Amy, Trey Miller, Eric Bettinger, and Madison Dell. 2022. "Advising for College Success: A Systematic Review of the Evidence." *Institute of Education Sciences*.

Foster, Carly, and Peter Francis. 2020. "A systematic review on the deployment and effectiveness of data analytics in higher education to improve student outcomes." *ASSESSMENT & EVALUATION IN HIGHER EDUCATION*, 45: 822–841.

French, Robert, and Philip Oreopoulos. 2017. "Behavioral barriers transitioning to college." *Labour Economics*, 47: 48–63.

Gurevitch, Jessica, Julia Koricheva, Shinichi Nakagawa, and Gavin Stewart. 2018. "Early exposure to service-learning and college success beyond the freshman year." *Nature*, 555: 175–182.

Havránek, Tomáš, T. D. Stanley, Hristos Doucouliagos, Pedro Bom, Jerome Geyer-Klingeberg, Ichiro Iwasaki, W. Robert Reed, Katja Rost, and R. C. M. van Aert. 2020. "REPORTING GUIDELINES FOR META-ANALYSIS IN ECONOMICS." *Journal of Economic Surveys*, 34(3): 469–475.

Hedges, L.V., E. Tipton, and M.C. Johnson. 2010. "Robust variance estimation in meta-regression with dependent effect size estimates." *Research Synthesis Methods*, 1(1): 39—-65.

Hershbein, Brad, Melissa S. Kearney, and Luke W. Pardue. 2020. "College Attainment, Income Inequality, and Economic Security: A Simulation Exercise." *AEA Papers and Proceedings*, 110: 352–355.

Higgins, Julian PT, and Simon G Thompson. 2002. "Quantifying Heterogeneity in a Meta-Analysis." *Statistics in Medicine*, 21(11): 1539–58.

IntHout, Joanna, John PA Ioannidis, and George F Borm. 2014. "The Hartung-Knapp-Sidik-Jonkman Method for Random Effects Meta-Analysis Is Straightforward and Considerably Outperforms the Standard DerSimonian-Laird Method." *BMC Medical Research Methodology*, 14(1): 25.

Kirp, David. 2019. *The College Dropout Scandal.* New York: Oxford University Press.

Knapp, Guido, and Joachim Hartung. 2003. "Improved Tests for a Random Effects Meta-Regression with a Single Covariate." *Statistics in Medicine*, 22(17): 2693–2710.

Langan, Dean, Julian P.T. Higgins, Dan Jackson, Jack Bowden, Areti Angeliki Veroniki, Evangelos Kontopantelis, Wolfgang Viechtbauer, and Mark Simmonds. 2019. "A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses." *Research Synthesis Methods*, 10(1): 83–98.

Light, Richard J, J Richard, Richard Light, David B Pillemer, et al. 1984. *Summing up: The science of reviewing research.* Harvard University Press.

Lovenheim, Michael, and Jonathan Smith. 2023. "Chapter 4 - Returns to different postsecondary investments: Institution type, academic programs, and credentials." In . Vol. 6 of *Handbook of the Economics of Education*, , ed. Eric A. Hanushek, Stephen Machin and Ludger Woessmann, 187–318. Elsevier.

Mikkelsen, Mai Bjørnskov, and Michael Rosholm. 2018. "Systematic review and meta-analysis of interventions aimed at enhancing return to work for sick-listed workers with common mental disorders, stress-related disorders, somatoform disorders and personality disorders." *Occupational and Environmental Medicine*, 75(9): 675–686.

Mountjoy, Jack. 2022. "Community Colleges and Upward Mobility." *American Economic Review*, 112(8): 2580–2630.

Nguyen, Tuan D., Jenna W. Kramer, and Brent J. Evans. 2018. "The Effects of Grant Aid on Student Persistence and Degree Attainment: A Systematic Review and Meta-Analysis of the Causal Evidence." *Stanford CEPA Working Paper*, 18-04.

OECD. 2019. "Education at a Glance 2019." Policy Paper.

Oreopoulos, Philip. 2021. "What Limits College Success? A Review and Further Analysis of Holzer and Baum's Making College Work." *Journal of Economic Literature*, 59(2): 546–73.

Ost, B., W. Pan, and D. Webber. 2018. "The Returns to College Persistence for Marginal Students: Regression Discontinuity Evidence from University Dismissal Policies." *Journal of Labor Economics*, 36(3): 779–805.

Page, Matthew J, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. 2021. "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews." *BMJ*, 372.

Patton, Lori D., Carla Morelon, Dawn Michele Whitehead, and Don Hossler. 2006. "Campus-Based Retention Initiatives: Does the Emperor Have Clothes." *New Directions for Institutional Research*, 130: 9–24.

Permzadian, Vahe, and Marcus Credé. 2016. "Do First-Year Seminars Improve College Grades and Retention? A Quantitative Review of Their Overall Effectiveness and an Examination of Moderators of Effectiveness." *Review of Educational Research*, 86(1): 277–316.

Robbins, Steven B., Kristy Lauver, Huy Le, Daniel Davis, and Aaron Carlstrom. 2004. "Do Psychosocial and Study Skill Factors Predict College Outcomes? A Meta-Analysis." *Psychological Bulletin*, 120(2): 261–288.

Rosenbaum, James E., and Janet Rosenbaum. 2013. "Beyond BA Blinders: Lessons from Occupational Colleges and Certifi cate Programs for Nontraditional Students." *Journal of Economic Perspectives*, 27(2): 153–172.

Sianesi, B., and J. van Reenen. 2003. "The returns to education: Macroeconomics." *Journal of Economic Surveys*, 17: 157–200.

Sneyers, Eline, and Kristof De Witte. 2018. "Interventions in higher education and their effect on student success: a meta-analysis." *Educational Review*, 70(2): 208–228.

Spence, Michael. 1973. "Job Market Signalling." *Quarterly Journal of Economics*, 87(3): 355–374.

Stanley, Tom D., and C. Doucouliagos. 2012. *Meta-regression analysis in economics and business.* Oxford, England: Routledge.

Stanley, Tom D., and Hristos Doucouliagos. 1997. "Meta-Regression Approximations to Reduce Publication Selection Bias." *Research Synthesis Method*, 5(1): 60—-78.

Stinebrickner, Ralph, and Todd Stinebrickner. 2014. *Journal of Labor Economics*, 32(3): 601–644.

Stinebrickner, Todd, and Ralph Stinebrickner. 2012. "Learning about Academic Ability and the College Dropout Decision." *Journal of Labor Economics*, 30(4): 707–748.

Tinto, Vincent. 1975. "Dropout from Higher Education: A Theoretical Synthesis of Recent Research." *Review of Educational Research*, 45(1).

Tinto, Vincent. 1993. *Leaving College: Rethinking the causes and cures of student attrition.* . 2. ed., Chicago University Press.

Valentine, Jeffrey C., Amy S. Hirschy, Christine D. Bremer, Walter Novillo, Marisa Castellano, and Aaron Banister. 2011. "Keeping At-Risk Students in School: A Systematic Review of College Retention Programs." *Educational Evaluation and Policy Analysis*, 33(2): 214–234.

Vevea, Jack L., Kathleen Coburn, and Alexander Sutton. 2019. "PUBLICATION BIAS." *The Handbook of Research Synthesis and Meta-Analysis*, 383–430. Russell Sage Foundation.

Viechtbauer, Wolfgang. 2005. "Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model." *Journal of Educational and Behavioral Statistics*, 30(3): 261–293.

Zimmerman, S. 2014. "The Returns to College Admission for Academically Marginal Students." *Journal of Labor Economics*, 32(4): 711–754.

References Included in the Systematic Review

[rev1] **Attewell, Paul, and David Monaghan.** 2016. "How Many Credits Should an Undergraduate Take?" *Research in Higher Education*, 57(6): 682–713.

[rev2] **Aulck, Lovenoor, Nishant Velagapudi, Joshua Blumenstock, and Jevin West.** 2016. "Predicting Student Dropout in Higher Education." *Proceedings of the 33rd International Conference on Machine Learning (ICML '16), Workshop on Machine Learning in Social Good Applications.*

[rev3] **Bergey, Bradley W., Rauno K. Parrila, Annie Laroche, and S. Hélène Deacon.** 2019. "Effects of peer-led training on academic self-efficacy, study strategies, and academic performance for first-year university students with and without reading difficulties." *Contemporary Educational Psychology*, 56: 25–39.

[rev4] **Bettinger, Eric P., and Bridget Terry Long.** 2018. "Mass Instruction or Higher Learning? The Impact of College Class Size on Student Retention and Graduation." *Education Finance and Policy*, 13(1): 97–118.

[rev5] **Boatman, Angela, and Bridget Terry Long.** 2018. "Does Remediation Work for All Students? How the Effects of Postsecondary Remedial and Developmental Courses Vary by Level of Academic Preparation." *Educational Evaluation and Policy Analysis*, 40(1): 29–58.

[rev6] **Booij, Adam S., Edwin Leuven, and Hessel Oosterbeek.** 2017. "Ability Peer Effects in University: Evidence from a Randomized Experiment." *Review of Economic Studies*, 84(2): 547–578.

[rev7] **Booth, Alison L., Lina Cardona-Sosa, and Patrick Nolen.** 2018. "Do single-sex classes affect academic achievement? An experiment in a coeducational university." *Journal of Public Economics*, 168: 109–126.

[rev8] **Bowman, Nicholas A., and Joshua M. Holmes.** 2018. "Getting off to a good start? First-year undergraduate research experiences and student outcomes." *Higher Education*, 76(1): 17–33.

[rev9] **Bowman, Nicholas A., and K. C. Culver.** 2018. "When Do Honors Programs Make the Grade? Conditional Effects on College Satisfaction, Achievement, Retention, and Graduation." *Research in Higher Education*, 59(3): 249–272.

[rev10] **Campbell, Toni A., and David E. Campbell.** 2007. "Outcomes of Mentoring At-Risk College Students: Gender and Ethnic Matching Effects." *Mentoring & Tutoring: Partnership in Learning*, 15(2): 135–148.

[rev11] **Carrieri, Vincenzo, Marcello D'Amato, and Roberto Zotti.** 2015. "On the Causal Effects of Selective Admission Policies on Students' Performances: Evidence from a Quasi-experiment in a Large Italian University." *Oxford Economic Papers*, 67(4): 1034–1056.

[rev12] **Casey, Marcus D., Jeffrey Cline, Ben Ost, and Javaeria A. Qureshi.** 2018. "ACADEMIC PROBATION, STUDENT PERFORMANCE, AND STRATEGIC COURSE-TAKING." *Economic Inquiry*, 56(3): 1646–1677.

[rev13] **Chiteng Kot, Felly.** 2014. "The Impact of Centralized Advising on First-Year Academic Performance and Second-Year Enrollment Behavior." *Research in Higher Education*, 55(6): 527–563.

[rev14] **Clark, M. H., and Nicole L. Cundiff.** 2011. "Assessing the Effectiveness of a College Freshman Seminar Using Propensity Score Adjustments." *Research in Higher Education*, 52(6): 616–639.

[rev15] **Culver, K. C., and Nicholas A. Bowman.** 2020. "Is What Glitters Really Gold? A Quasi-Experimental Study of First-Year Seminars and College Student Success." *Research in Higher Education*, 61(2): 167–196.

[rev16] **Deighton, Kevin, Joanne Hudson, Andrew J. Manley, Mariana Kaiseler, Laurie B. Patterson, Zoe H. Rutherford, and Michelle Swainson.** 2019. "Effects of emotional intelligence and supportive text messages on academic outcomes in first-year undergraduates." *Journal of Further and Higher Education*, 43(4): 494–507.

[rev17] **Dennehy, Tara C., and Nilanjana Dasgupta.** 2017. "Female peer mentors early in college increase women's positive academic experiences and retention in engineering." *PNAS Proceedings of the National Academy of Sciences of the United States of America*, 114(23): 5964–5969.

[rev18] **De Paola, Maria, and Vincenzo Scoppa.** 2014. "The Effectiveness of Remedial Courses in Italy: A Fuzzy Regression Discontinuity Design." *Journal of Population Economics*, 27(2): 365–386.

[rev19] **Dobronyi, Christopher R., Philip Oreopoulos, and Uros Petronijevic.** 2019. "Goal Setting, Academic Reminders, and College Success: A Large-Scale Field Experiment." *Journal of Research on Educational Effectiveness*, 12(1): 38–66.

[rev20] **Duchini, Emma.** 2017. "Is College Remedial Education a Worthy Investment? New Evidence from a Sharp Regression Discontinuity Design." *Economics of Education Review*, 60: 36–53.

[rev21] **Ellis, Jimmy R., and Seth Gershenson.** 2020. "Gender, peer advising, and college success." *Labour Economics*, 62: N.PAG–N.PAG.

[rev22] **Findley-Van Nostrand, D., and R. S. Pollenz.** 2017. "Evaluating Psychosocial Mechanisms Underlying STEM Persistence in Undergraduates: Evidence of Impact from a Six-Day Pre-College Engagement STEM Academy Program." *Cbe-Life Sciences Education*, 16(2).

[rev23] **Johnson, Matthew D., Amy E. Sprowles, Katlin R. Goldenberg, Steven T. Margell, and Lisa Castellino.** 2020. "Effect of a Place-Based Learning Community on Belonging, Persistence, and Equity Gaps for First-Year STEM Students." *Innovative Higher Education*, 45(6): 509–531.

[rev24] **Keller, Robert R, and Michael G Lacy.** 2013. "Propensity score analysis of an honors program's contribution to students' retention and graduation outcomes." *Journal of the National Collegiate Honors Council*, 14(2): 73–84.

[rev25] **Lindo, Jason M., Nicholas J. Sanders, and Philip Oreopoulos.** 2010. "Ability, Gender, and Performance Standards: Evidence from Academic Probation." *American Economic Journal: Applied Economics*, 2(2): 95–117.

[rev26] **Masserini, Lucio, and Matilde Bini.** 2021. "Does Joining Social Media Groups Help to Reduce Students' Dropout within the First University Year?" *Socio-Economic Planning Sciences*, 73.

[rev27] **Nix, S., K. L. Roberts, and R. M. Hughes.** 2016. "Female undergraduate stem persistence: A focus on the role of living and learning communities." *Journal of Women and Minorities in Science and Engineering*, 22(4): 349–374.

[rev28] **Nosaka, Taé, and Heather Novak.** 2014. "Against the Odds: The Impact of the Key Communities at Colorado State University on Retention and Graduation for Historically Underrepresented Students." *Learning Communities: Research & Practice*, 2(2): 1–22.

[rev29] **Novak, Heather, Christina Paguyo, and Thomas Siller.** 2016. "Examining the Impact of the Engineering Successful/Unsuccessful Grading (SUG) Program on Student Retention." *Journal of College Student Retention: Research, Theory & Practice*, 18(1): 83–108.

[rev30] **Oosterbeek, Hessel, and Reyn van Ewijk.** 2014. "Gender Peer Effects in University: Evidence from a Randomized Experiment." *Economics of Education Review*, 38: 51–63.

[rev31] **Plak, S., I. Cornelisz, M. Meeter, and C. van Klaveren.** 2020. "Early warning systems for more effective student counselling in higher education: Evidence from a Dutch field experiment." *Higher Education Quarterly*.

[rev32] **Reynolds, C. Lockwood.** 2020. "The Effect of Dormitory Residence during College on Student Outcomes." *Journal of Human Capital*, 14(2): 249–289.

[rev33] **Song, Wei, Andrew Furco, Geoffrey Maruyama, and Isabel Lopez.** 2018. "Early exposure to service-learning and college success beyond the freshman year." *International Journal of Research on Service-Learning & Community Engagement*, 6(1).

[rev34] **Soria, Krista M., and Linnette Werner.** 2018. "Academic Leadership Courses: Catalysts for Students' Retention and Graduation." *Journal of Leadership Education*, 17(3): 26–41.

[rev35] **Stegers-Jager, Karen M., Janke Cohen-Schotanus, and Axel P. N. Themmen.** 2013. "The effect of a short integrated study skills programme for first-year medical students at risk of failure: A randomised controlled trial." *Medical Teacher*, 35(2): 120.

[rev36] **Tafreschi, Darjusch, and Petra Thiemann.** 2016. "Doing It Twice, Getting It Right? The Effects of Grade Retention and Course Repetition in Higher Education." *Economics of Education Review*, 55: 198–219.

[rev37] **Wachen, John, Joshua Pretlow, and Karrie G. Dixon.** 2018. "Building College Readiness: Exploring the Effectiveness of the UNC Academic Summer Bridge Program." *Journal of College Student Retention: Research, Theory & Practice*, 20(1): 116–138.

[rev38] **Zilvinskis, John, Victor Borden, and Jr. Severtis, Ronald E.** 2020. "Employing a Quasi-Experimental Technique to Measure the Effects of Conditional Admissions." *Journal of College Student Retention: Research, Theory & Practice*, 22(1): 141–154.

Appendix

A1. The PICO Model and Search String

The PICO model provides a recognized framework for structuring search strings. The acronym PICO comprises the components: Population, Intervention, Comparison and Outcome measures. Defining and including these variables in the search string secures effective searches.

Table A1—: PICO model

Category	**Search words**
Population	Students Freshmen *Context:* Higher education University College Post-secondary education Further education Tertiary education
Intervention	*We will not restrict ourselves to specific interventions. This will intentionally be left blank.*
Comparators	Meta-analysis Randomized controlled trials Controlled trials Regression discontinuity design Difference-in-differences Propensity score matching Instrumental variables Quasi-experiment Natural experiment Field experiment
Outcome	Dropout Attrition Departure Withdrawal Academic failure Non-completion Graduation Retention Persistence Completion Student outcome Academic outcome Transferring Change of study

Table A2—: Search string

Student* OR freshm*
AND
Universit* OR "higher education" OR HE OR colleg* OR "post-secondary education" OR "further education" OR "tertiary education"
AND
"Meta analys*" OR "meta-analys*" OR "random* control* trial*" OR RCT* OR "control* trial*" OR "random* trial*" OR "regression discontinuity" OR "instrumental variable*" OR "diff* in* diff*" OR "random* experiment*" OR "propensity score*" OR matching OR randomized OR randomised OR "quasi* experiment*" OR "natural experiment*" or "field experiment*"
AND
Dropout* OR "drop* out*" OR attrition* OR depart* OR graduat* OR retention* OR withdraw* OR "academic failur*" OR persist* OR complet* OR "non-completion" OR "student outcome*" OR "academic outcome*" OR transfer* OR "change of stud*"

Table A3—: Study Assessment Checklist

Criteria	**National Institutes of Health Quality Assessment Tool for Controlled Intervention Studies**
(a) Randomised trial	Was the study described as randomized, a randomized trial, a randomized clinical trial, or an RCT?
(b) Controlled trial	Was the study described as a controlled trial?
(c) Acceptable randomisation or control procedure	Adequate randomization: Randomization is adequate if it occurred according to the play of chance (e.g., computer generated sequence in more recent studies, or random number table in older studies). Was the control group matched on relevant variables (age, gender, education, disorder)?
(d) <20% dropout	Was the overall drop-out rate from the study at endpoint 20% or lower of the number allocated to treatment?
(e) Insignificant difference in dropout between groups	Was the differential dropout rate (between groups) at endpoint 15 percentage points or lower? (calculated by subtracting the dropout rate of one group minus the dropout rate of the other group).
(f) Good fidelity	Was there high adherence to the intervention protocols for each treatment group? (defined as 75% attendance or more)
(g) Definition of counterfactual treatment is clear	Were other interventions avoided or similar in the groups (e.g., similar background treatments)?
(h) Valid, realiable, and consistent measurement of outcomes	Were outcomes assessed using valid and reliable measures, implemented consistently across all study participants?
(i) Power calculations	Did the authors report that the sample size was sufficiently large to be able to detect a difference in the main outcome between groups with at least 80% power?
(j) Protocol?	Were outcomes reported or subgroups analyzed prespecified (i.e., identified before analyses were conducted)?
(k) Recognised statistical analyses	For RCTs: Were all randomized participants analyzed in the group to which they were originally assigned, i.e., did they use an intention-to-treat analysis? For controlled studies: Was a recognized statistical method employed? (Recognized methods defined as dif-in-dif, regression discontinuity, propensity score matching, instrumental variables).

Table A4—: Study Assessment Checklist - Included Studies

Author	**Year**	**(a)**	**(b)**	**(c)**	**(d)**	**(e)**	**(f)**	**(g)**	**(h)**	**(i)**	**(j)**	**(k)**	**Score**
Attewell & Monaghan	2016	No	Yes	No	No	No	No	Yes	Yes	No	No	Yes	4
Aulck et al.	2021	No	Yes	No	Yes	No	No	Yes	Yes	No	No	Yes	5
Bergey et al.	2019	Yes	No	Yes	No	No	No	Yes	No	No	No	No	3
Bettinger & Long	2018	No	Yes	Yes	No	No	No	Yes	Yes	No	No	Yes	5
Boatman & Long	2018	No	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No	Yes	7
Booij et al.	2017	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No	No	Yes	7
Booth et al.	2018	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	8
Bowman & Culver	2018	No	Yes	No	No	No	No	Yes	No	No	No	Yes	3
Bowman & Holmes	2018	No	Yes	No	No	No	No	Yes	No	Yes	No	Yes	4
Campbell & Campbell	2007	No	Yes	No	No	No	No	Yes	No	No	No	No	2
Carrieri et al.	2015	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	8
Casey et al.	2018	No	Yes	No	Yes	Yes	Yes	Yes	Yes	No	No	Yes	7
Chiteng Kot	2014	No	Yes	No	No	No	No	No	Yes	No	No	Yes	3
Clark & Cundiff	2011	No	Yes	No	No	No	No	Yes	Yes	No	No	Yes	4
Culver & Bowman	2020	No	Yes	No	No	No	No	No	Yes	No	No	Yes	3
De Paola & Scoppa	2014	No	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	7
Deighton et al.	2019	Yes	No	No	Yes	Yes	Yes	Yes	No	No	No	Yes	6
Dennehy & Dasgupta	2017	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	8
Dobronyi et al.	2019	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	10
Duchini	2017	No	Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes	6
Ellis & Gershenson	2020	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	8
Finley van Nostrand & Pollenz	2017	No	Yes	No	Yes	Yes	No	Yes	Yes	No	No	Yes	6
Johnson et al.	2020	No	Yes	No	Yes	No	No	No	No	No	No	No	2
Keller & Lacy	2013	No	Yes	No	No	No	No	Yes	Yes	No	No	Yes	4
Lindo et al.	2010	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	Yes	7
Masserini & Bini	2021	No	Yes	No	Yes	Yes	No	No	No	No	No	Yes	4
Nix et al.	2016	No	Yes	No	Yes	Yes	No	No	No	No	No	No	3
Nosaka & Novak	2014	No	Yes	No	Yes	Yes	No	No	Yes	No	No	No	4
Novak et al.	2016	No	Yes	No	Yes	Yes	No	No	Yes	No	No	No	4
Oosterbeek & van Ewijk	2014	Yes	No	Yes	No	Yes	Yes	Yes	Yes	No	No	Yes	7
Plak et al.	2020	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	8
Reynolds	2020	No	Yes	No	No	No	No	Yes	Yes	No	No	Yes	4
Song et al.	2018	No	Yes	Yes	Yes	No	No	Yes	Yes	No	No	No	5
Soria & Werner	2018	No	Yes	No	No	No	No	No	Yes	No	No	No	2
Stegers-Jager et al.	2013	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	8
Tafreschi & Thiemann	2016	No	Yes	Yes	Yes	No	Yes	No	No	No	No	Yes	5
Wachen et al.	2018	No	Yes	No	Yes	Yes	No	Yes	No	No	No	Yes	5
Zilvinskis et al.	2020	No	Yes	No	No	No	No	Yes	Yes	No	No	No	3

A2. Robustness Checks

Table A5—: Effect sizes: different τ^2 estimators

	REML		**DerSimonian-Laird**		**Paule-Mandel**	
	(1)	(2)	(3)	(4)	(5)	(6)
Estimate	0.075	0.124	0.075	0.127	0.074	0.122
	(0.028)	(0.044)	(0.050)	(0.060)	(0.027)	(0.040)
p-value	0.012	0.022	0.144	0.068	0.009	0.015
τ^2	0.018	0.009	0.006	0.004	0.008	0.011
I^2	94.43%	87.37%	84.19%	73.43%	95.05%	89.47%
PI	-0.20;0.35	-0.12;0.36	-0.11;0.26	-0.07;0.32	-0.22;0.37	-0.10;0.34
N_m	37	12	37	12	37	12
N_s	31	9	31	9	31	9
	ML		**Empirical Bayes**		**Sidik-Jonkman**	
	(7)	(8)	(9)	(10)	(11)	(12)
Estimate	0.075	0.124	0.074	0.122	0.074	0.121
	(0.029)	(0.045)	(0.026)	(0.040)	(0.024)	(0.036)
p-value	0.015	0.029	0.008	0.015	0.004	0.009
τ^2	0.017	0.008	0.020	0.011	0.023	0.013
I^2	94.17%	85.33%	95.04%	89.47%	95.60%	91.22%
PI	-0.20;0.35	-0.10;0.35	-0.22;0.37	-0.13;0.38	-0.24;0.39	-0.16;0.40
N_m	37	12	37	12	37	12
N_s	31	9	31	9	31	9

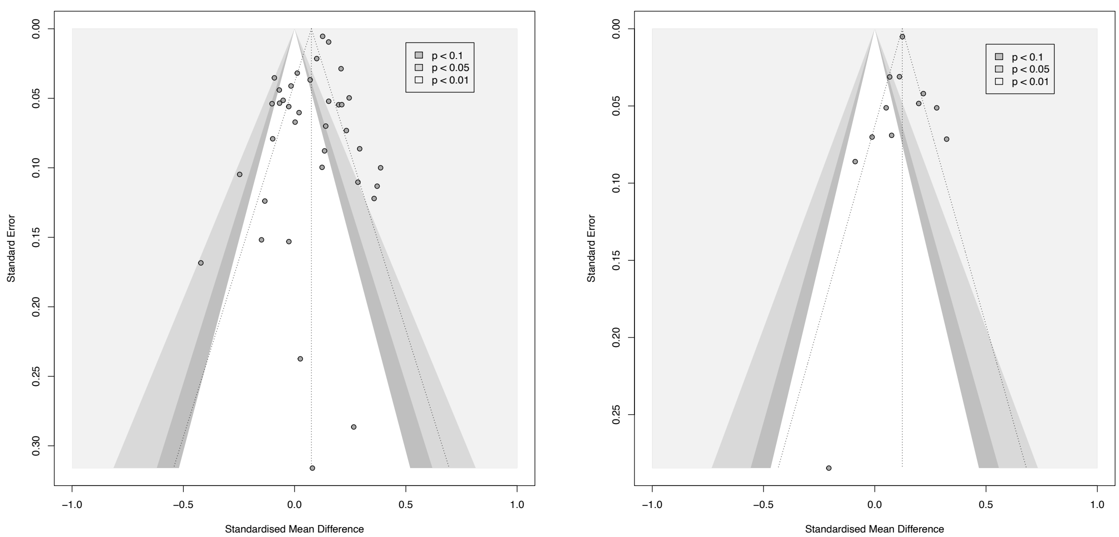
Note: PI refers to the prediction interval, N_m refers to the number of effects sizes, and N_s refers to the number of studies. Odd-numbered columns are for first-year graduation, and even-numbered columns are for six-year graduation.

Table A6—: Effect sizes: specification checks

	Robust & KH		Only robust		Only KH		None	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Estimate	0.075	0.124	0.075	0.124	0.075	0.124	0.075	0.124
	(0.028)	(0.044)	(0.026)	(0.038)	(0.027)	(0.034)	(0.026)	(0.032)
p-value	0.012	0.022	0.007	0.012	0.009	0.004	0.004	0.001
PI	-0.20;0.35	-0.12;0.36	-0.20;0.35	-0.11;0.36	-0.20;0.35	-0.10;0.34	-0.19;0.34	-0.07;0.32
N_m	37	12	37	12	37	12	37	12
N_s	31	9	31	9	31	9	31	9

Note: PI refers to the prediction interval, N_m refers to the number of effects sizes, and N_s refers to the number of studies. Odd-numbered columns are for first-year graduation, and even-numbered columns are for six-year graduation.

Figure A1. : Contour plots



(a) First-year persistence (b) Six-year graduation

Figure A2. : Detailed description of included studies

Author	Year	Country	Background	Intervention	Content	Method	Outcome(s)	Effect	S.A.C.
Attewell & Monaghan	2016		Representative sample of U.S. students who enrolled in 2003 (public or non-profit colleges): BPS (04/09)	Course credits	Taking 15 course credits (corresponding to 30 ECTS) instead of 12 in the first semester. 15 course credits correspond to full time enrollment.	PSM	1: 1st-y per. 2: G within 6 y	1: Null 2: Pos	Medium
Aulck et al.	2021		Large, publicly funded university: University of Washington. 17 cohorts of 1st-time, 1st-year students enrolling 1998-2014.	1st-year interest groups ("FIGs"): two-credit freshman orientation seminars.	All students within the same FIG cluster of classes take a seminar class together (led by an upperclassman) during their 1st academic term. The class focuses on a discussion of students' personal experiences rather than academics to develop a sense of involvement, participation, and community.	PSM	1: 1st-y per. 2: G within 6 y	1: Pos 2: Pos	Medium
Bergey et al.	2019		Students at a large, research-intensive Canadian university.	Peer coaching	Peer-led study strategies training for 1st-year students with the purpose of increasing academic self-efficacy. Consisted of 5 one-hour workshops with activities during the first two months during the 1st year.	RCT	1st-y per.	Null	Low
Bettinger & Long	2018		Students who enrolled in one of Ohio's 11 public, 4-y colleges during Fall 98, Fall 99, or Fall 01 and are < 21 y with ACT records	Class size	Consider size of "lecture" sections in the first semester. This varied between 18 and 760 students. The average class size was 94 students.	IV	1: 1st-y per. 2: G within 6 y in Ohio	1: Neg 2: Null	Medium
Boatman & Long	2018		Students at 6 public universities in TN who enrolled in 2000, are < 21 y, and took the Computer Adaptive Placement Assessment and Support System exam (based on low ACT scores)	Developmental / remedial courses	They differ, in contrast to other studies, between the two. Developmental courses are just below college-level, while remedial courses are a step further down. We report outcomes only for developmental courses: 1: Developmental math, 2: Developmental reading, 3: Developmental Writing	RDD	1st-y per.	1: Neg 2: Null 3: Null	High
Booij et al.	2017		Economics students at University of Amsterdam from 09-11.	Tutorial groups (tracking)	Manipulated the ability composition of tutorial groups for 1st-year students based on GPA from final exams of secondary education (peer effects). 1: Low group, 2: Middle group, 3: High group	Cond. RCT	1st-y per.	1: Pos 2: Pos 3: Null	High
Booth et al.	2018		Economics or Business students at the University of Essex, arriving Fall 10.	Tutorial groups	1 hour a week of single-sex classes (<30 students) in an introductory economics course (compared to classes with on average 30% females and 70% males) within a coeducational environment. Amounts to 8.3 % of total instruction. 1: Females, 2: Males	RCT	G from university	1: Pos 2: Null	High
Bowman & Culver	2018		41 colleges committed to liberal arts education including religiously affiliated, single-sex and minority serving schools	Honors programs	The student has been part of an honors program within the 1st year. Honors programs often provide more interaction with faculty and extra assignments, giving a more thorough understanding of the material.	PSM	1st-y per.	Null	Low
Bowman & Holmes	2018		46 colleges committed to liberal arts education including religiously affiliated, single-sex and minority serving schools	Undergraduate research program	The student has participated in undergraduate research within the 1st year. The authors do not elaborate on what the content specifically consists of.	PSM	1st-y per.	Null	Medium
Campbell & Campbell	2007		Students at a large metropolitan university in CA who entered in 1992-1994. The target group was minority students.	Student-faculty mentoring	Faculty mentors were asked to meet with their mentees at least 3 times during the semester. The average was 7 times with a mean total duration of 120 minutes.	Matching	1st-y per.	Pos	Low

Carrieri et al.	2015		The Economics faculty at University of Salerno (large, public university in the South) from academic year 2005/2006	Decentralized admission policy (peer effects)	The university introduced a test-based selective admission policy where only a certain number of students were admitted based on rank (in contrast to a compulsory non-selective test and free admission).	DID	1st-y per.	Pos	High
Casey et al.	2018		9 cohorts at a large, non-elite, urban U.S. public university from 2004 to 2013	Academic probation (AP)	On AP, students are required to attain at least a GPA of 2.0 in the next semester, have extra meetings with the academic advisors, and raise their cumulative GPA to above 2.0 to be removed from AP. They consider the impact of being placed on AP at the end of the 1st semester.	RDD	1: 1st-y per. 2: G within 6 y	1: Null 2: Null	High
Chiteng Kot	2011		Students who entered a large, metropolitan, public, research university in fall 2010	Academic advising	At least one face-to-face meeting at a centralized academic advising center, which provide different forms of assistance such as academic counselling and discussing goals.	PSM	1st-y per.	Pos	Low
Clark & Cundiff	2014		1st-year students at Southern Illinois University Carbodale (moderate-sized, rural, midwestern university), who was enrolled in an introductory psychology course in 2004-2007	Freshman seminar	The University 101 course (worth 3 credits, corresponding to 6 ECTS, and runs throughout the semester) introduces the knowledge and skills necessary for a positive academic and personal experience. The students e.g. interview faculty members and attend campus events.	PSM	1st-y per.	Pos	Medium
Culver & Bowman	2020		43 colleges committed to liberal arts education including religiously affiliated, single-sex and minority serving schools	First-year/freshman seminars	The student has participated in a seminar designed specifically for first-year students (e.g., First-Year, Freshman Seminar. The authors do not elaborate on what the content specifically consists of.	PSM	1st-y per.	Null	Low
De Paola & Scoppa	2014		Students enrolled at the University of Calabria (a medium-sized public university in the South) in fall 2009	Remedial courses / preparation program	Consisted of 160 h of lectures carried out at the beginning of the academic year (lasting 2 months and covered both mathematics and language skills). These remedial courses do not confer to academic credits. Regular college courses for all students only started after the remedial courses had ended.	RDD	1st-y per.	Pos	High
Deighton et al.	2019		Sport and Exercise Science students at a university in UK in the Fall 2015	Text messages	Three text messages from the faculty student liaison officer (SLO) in potentially stressful periods for the students, e.g., referring to preparing for examinations, considering assessment results, and preparing for the upcoming semester. The text messages also reminded students of the available pastoral, academic and personal support within the faculty and the university and invited students to see an SLO.	RCT(ish)	1st-y per.	Pos	Medium
Dennehy & Dasgupta	2017		Female students intending to major in engineering at University of Massachusetts Amherst (public) in 2011-2014	Peer mentoring	Female students were randomly assigned to male or female peer mentors or no mentors for 1 year (control). They met roughly once a month. 1: Female mentor, 2: Male mentor	RCT	1st-y per.	1: Pos 2: Null	High
Dobronyi et al.	2019		Students at the University of Toronto's satellite campus in Mississauga (commuter campus) who enrolled in introductory economics in 14/15	Goal-setting	Two interventions: 1) Doing a goal-setting exercise (G); 2) Doing a condensed version of the goal-setting exercise and a short mindset exercise (GM). Then, half of each of the two interventions regularly receive personalized goal-oriented reminders through e-mail or text messages (R).	RCT	1st-y per.	Null	High

Duchini	2017		Economics students at a university located in northern Italy who enrolled in 2009-2013	Remedial courses – the authors state that it is a 'nudge-type policy'	This remedial policy combines a short remedial course in math (21 h - 1/3 of the length of a standard course in Italy) with the threat of re-enrolling them in the 1st year in case of failures in the 5 retake opportunities for the remedial exam over the course of the 1st year.	RDD	1st-y per.	1: Null	Medium
Ellis & Gershenson	2020		Students at American University (selective, midsized, private, non-for-profit) at the academic unit of College of Arts and Sciences in 2013 and 2014 cohorts	Same-sex peer advisor	The program provides free one-on-one support services during the first semester. Peer advisors complete a competitive application process, receive extensive training on working with students and handling sensitive issues, must maintain a GPA of 3.0 or higher, and commit to working the entire academic year. Most students only meet peer advisor once, early in the semester. Peer advisor is either male or female. 1: Male advisor, 2: Female advisor	RCT(ish)	1st-y per.	1: Null 2: Null	High
Finley van Nostrand & Pollenz	2017		Science students at the University of South Florida (metropolitan and high research setting) in 2015 who have pre-calculus as the 1st math course and math SAT < 650 or ACT < 29	Preparation program	The STEM Academy (SA) is a weeklong program for incoming freshman students in a small-group setting. The SA program consists of 19 different modules including but not limited to mentorship, career planning, undergraduate research, and math exercises.	PSM	1st-y per.	Null	Medium
Johnson et al.	2020		STEM students at a residential campus at a mid-sized, rural Master's-granting state university in northwestern California for 2015-2017 cohorts	Place-based learning community (LC)	The LC embeds an interdisciplinary focus on the landscape, people, and cultures of the University's location. Supports first-year students by: a summer immersion, peer mentoring, a first-year seminar, and a cohort blocked scheduling of lower division courses required for their chosen STEM major.	PSM	1st-y per.	Pos	Low
Keller & Lacy	2013		Students who initially entered Colorado State University from 2005-2008	Honors programs	The honors program provides access to innovative and small classes, more attention and advising from faculty and staff, residential learning communities, thesis experiences, extra-curricular opportunities, and additional scholarships.	PSM	1: 1st-y per. 2: G within 6 y	1: Pos. 2: Pos.	Medium
Lindo et al.	2010		3 campuses at a large university for 1996-2004 cohorts between the ages of 17 and 21	Academic probation (AP)	Students with a GPA below a certain threshold is placed on AP. These face the threat of suspension for 1 year if their grades do not improve sufficiently. Students on AP can return to good academic standing by bringing their cumulative GPA up to the cutoff. Students receive the notice at the end of the 1st year.	RDD	1: 1st-y per. 2: G within 6 y	1: Neg. 2: Null	High
Masserini & Bini	2021		Freshmen at a major university in central Italy in 2016	(Online) social network	Participation in Facebook groups or pages run by other students, representing a way to promote and strengthen social relationships amongst students but also to help share information on courses / material useful for studying and exams.	PSM	1st-y per.	Null	Medium
Nix et al.	2016		Female, STEM 1st-year students at a very large, public, research-intensive university in the Southeast, who entered in 2001-2006	Living-Learning community (LLC)	Women-only STEM LLC. During the 1st year, all participants live in the same residence hall and attend a weekly colloquium together. The participants have access to academic tutoring and social events. After the 1st year, they move out of the LLC, but can still attend activities.	PSM	Graduation from the institution before 2012	Pos	Low

Nosaka & Novak	2014		Students at Colorado State University who enter in 2005-2011 and typically have lower graduation rates	Living-Learning community (LLC)	In the so-called Key communities program, all 1st-year students live together in a popular and centrally located hall. They participate in one community: Academic; Culture, Communication, and Sport; Explore; Health Professions; or Service. The program employs special Key Seminar courses and mentorship.	PSM	1: 1st-y per. 2: G within 6 y	1: Pos. 2: Pos.	Medium
Novak et al.	2016		1st-time, 1st-year students in the College of Engineering at Colorado State University (large, public university) in 2010-2012	Grading: Successful / Unsuccessful (S/U)	Students had the option to take all or none of their courses using the S/U system. This choice had to be made within the first two weeks of the semester. Students can check their actual grades after the semester.	PSM	1st-y per.	Pos.	Medium
Oosterbeek & van Ewijk	2014		Economics and Business students at the University of Amsterdam who entered in 2007-2008	Tutorial groups	1st-year students were randomly placed in tutorial groups (on average 39 students) where the share of females vary between 0.14 and 0.51. This composition is fixed for the entire 1st year, where more than 60% of all teaching hours take place in these groups.	RCT	1st-y per.	Null	High
Plak et al.	2020		Vrije Universiteit (VU) Amsterdam in AY 16/17	Early Warning Systems (EWS) in student counselling	The developed dashboard shared student-specific risk (based on a machine learning model) and background information with student counsellors in their regular counselling activities. The system allowed for (1) proactive and timely invitations by student counsellors of at-risk students for an individual appointment to perform coaching interventions, and (2) feedback to students to make them better aware of their performance and their risk of dropping out.	RCT	1st-y per.	Null	High
Reynolds	2020		Large, public, research-oriented university in the Midwest. Students are in-state, < 20 years old, and live in a zip code < 80 miles from campus, 99-07	Dormitory residence	Dormitory residence during the first year of college compared to living at home.	IV	1st-y per.	Null	Medium
Song et al.	2018		Students at a large, public, research university, entering in Fall 2011 or 2012	Taking a course with service-learning components	For these courses, partnerships were maintained with more than 300 nonprofit organizations to provide sites for students' service experiences. Minimum requirements: (1) students performed a minimum of 3 hours per week of service-learning related activity in a semester; (2) students were provided opportunities for critical reflection on their service-learning experience. Most courses also incorporated other best-practice principles from the service-learning literature.	PSM	1: 1st-y per. 2: G within 6 y	1: Pos 2: Pos	Medium
Soria & Werner	2018		1st-year students who enrolled at a large, public, research-intensive university in the Midwest in 2012	Academic leadership course	Teaches leadership skills based on 5 themes: Self-awareness of leadership; identities and leading across differences; personal definition of leadership; critical thinking skills; and leadership journey.	PSM	1st-y per.	Pos.	Low
Stegers-Jager et al.	2013		1st-year medical students at risk of failure at Erasmus MC Medical School Rotterdam in 2008 or 2009	Study skills program (SSP)	The SSP students met 5 times for 1.5 h. The focus was on awareness and diagnosis of the individual causes for academic failure and on practicing various study skills, including time management, previewing, creating study guides and test taking.	RCT	1st-y per.	Null	High

Tafreschi & Thiemann	2016		Economics or Business students at the University of St. Gallen who started in 2001-2008 and completed all 1st-year courses	Grade repetition	All 1st -year students who fall short of a pre-defined performance requirement must repeat all 1st-year courses before they can proceed to the second year.	RDD	1st-y per.	Neg	Medium
Wachen et al.	2018		1st-time students who are barely admitted to 5 public universities in North Carolina (UNC)	Summer bridge program	This 5-6 week summer residential program provides support services (e.g. tutoring, mentoring, counselling). Students had to complete college-level math and English with minimum a "C" in both courses.	PSM	1st-y per.	Pos	Medium
Zilvinskis et al.	2020		Marginal students at Indiana University Southeast (public) who entered in 2010 or 2011	Conditional admission	Conditional admission was offered to students who were ineligible for full admission but required applicants to have completed a college preparatory curriculum (labeled the Core 40 diploma in Indiana) and either a 2.3 GPA or a 900 SAT score (19 ACT score). Conditional admits were discouraged, but not prohibited, to register for > 12 credit hours in each of the semesters of their first year.	DID	1st-y per.	Null	Low