Review

Effectiveness of grade retention: A systematic review and meta-analysis

Mieke Goos\textsuperscript{a,1,2,3,*}, Joana Pipa\textsuperscript{b,3,**}, Francisco Peixoto\textsuperscript{b}

\textsuperscript{a} Department of Teacher Training, University College Leuven – Limburg, Hertogstraat 178, 3001, Heverlee, Belgium
\textsuperscript{b} CIE-ISPA (Centre for Educational Research-ISPA), ISPA- Instituto Universitário, Rua Jardim do Tabaco 34, 1149-041, Lisbon, Portugal

\textbf{ARTICLE INFO}

Keywords:
Effectiveness
Meta-analysis
Learning outcomes
Grade retention
K-12 education

\textbf{ABSTRACT}

Research on the effectiveness of grade retention has a long history, yet, has seen an upsurge during the last decade. In this study, we review 84 recent, methodologically sound studies estimating effects of retention in grades K-12 on repeaters’ and non-repeaters’ development, in a variety of countries across the world, disentangling grade and age comparison results. Based on vote counting analysis and three-level metaregression analysis we find grade retention to have an average zero effect, indicating that repeaters and non-repeaters seem to show a similar development, on average. At the same time, we find grade retention effects to differ according to some specific effect and study characteristics. More specifically, grade retention seems less effective in countries applying a mixture of grade retention and tracking to tackle student heterogeneity, and when repeaters are compared with non-repeaters of the same age. Conversely, grade retention seems more effective in countries using strategies such as ability grouping, setting, and streaming to deal with student heterogeneity. Positive effects also seem to arise when studying students’ psychosocial functioning, when investigating short-run effects, when comparing repeaters with their younger non-retained grade-mates, and when evaluating effects via a regression discontinuity method.

1. Introduction

Within any given class, there is typically considerable variation in students’ academic and psychosocial skills. This heterogeneity challenges educational practitioners and policy makers in their attempts to guarantee high educational standards and raises the question how to deal with it. One can use ability grouping, streaming, tracking, individualized instruction, grade retention, peer tutoring, collaborative learning, or a mixture of practices. The practice we focus on in this study is grade retention. Research on the effectiveness of grade retention has a long history, going back to as early as 1908 (Jackson, 1975), yet, has seen an upsurge since 2000, and especially since 2010. The aim of our study is to review recent, methodologically sound studies estimating effects of retention in grades K-12 on repeaters’ and non-repeaters’ development (in terms of academic achievement, psychosocial functioning, school
career, and job career), in a variety of countries across the world, disentangling grade and age comparison results.

### 1.1. Grade retention in practice

Grade retention means that one holds back students who are not keeping up with peers or do not meet a certain predefined level, in the same grade for an extra school year, instead of promoting them to the next grade level (Driessen, 2020). This is one practice that educational practitioners and policy makers can apply to tackle student heterogeneity in grades K-12.

In some countries, grade retention is the most commonly applied practice to tackle student heterogeneity (Dupriez, Dumay, & Vause, 2008). Mons (2007) refers to this type of countries as ‘uniform integration countries’. Examples are France, Spain, and Portugal (see high retention rates in Fig. 1). In another set of countries, grade retention is jointly used with ability grouping and tracking to deal with student heterogeneity (Dupriez et al., 2008). Mons (2007) calls this type of countries, among which Germany, Austria, Hungary, Switzerland, Luxembourg, Belgium, and the Netherlands (see high retention rates in Fig. 1), ‘separation countries’. In some other countries, grade retention is seldom used, and is only used as a last remedy when ability grouping, setting, and streaming appear to insufficiently tackle student heterogeneity (Dupriez et al., 2008). Mons (2007) refers to this type of countries as ‘a la carte integration countries’. Examples are the United States, the United Kingdom, Canada, Australia, and New Zealand (see low to moderate retention rates in Fig. 1). Finally, in a last set of countries, grade retention is not allowed or rarely applied, because student heterogeneity is mainly dealt with using various forms of differentiation and individualized teaching (Dupriez et al., 2008). Mons (2007) calls this type of countries, among which Denmark, Iceland, Finland, Norway, and Sweden (see low retention rates in Fig. 1), ‘individualized integration countries’.

In addition to these broad application differences across country typologies, the practice of grade retention also differs across countries within a given typology, as well as across states and regions within a given country. Each educational context is unique. Each country, and to some extent also each state and region, has its own educational policy (e.g., regulations related to school entry, presence or absence of high-stakes testing and related regulations concerning grade retention decisions, regulations regarding the application of the retention year as ‘mere rehearsal’ versus so-called ‘retention +’ coupling grade retention with many other remedial strategies such as extra tutoring, summer school, etc.), school population, school organization, curriculum, and beliefs about grade retention effectiveness, which altogether influence grade retention application in practice (Agasisti & Cordero, 2017; Demanet & Van Houtte, 2013; Driessen, 2020; Eurydice, 2011; Goos et al., 2013).

![Fig. 1. Percentage of students reporting having repeated at least once by age 15, per OECD member country, as derived from the PISA 2018 student questionnaire.](image-url)
1.2. Theoretical debate

As for the beliefs about grade retention effectiveness, educators from ‘uniform integration countries’, ‘separation countries’, and ‘à la carte integration countries’ retain struggling students with the best intentions; intentions that can be linked to diverse theories in education, psychology, and sociology. First of all, many educators are convinced, inspired by maturationist developmental theory (Gesell, 1940; Piaget & Inhelder, 1962), that mastering the learning content related to a specific grade is necessary to succeed in later grades, from a basic ‘building blocks’ idea, with each grade building upon the previous one. Thus, students who do not yet master specific grade-related knowledge or skills, are in their view best off repeating. The next grade’s learning content would be too difficult for them, setting them up for failure and possibly accumulating in a negative school and job career. Second, many educators believe that classes become more homogeneous when the weakest performing students repeat their grade. As a result, they believe that they can instruct their students more efficiently, cover more advanced learning content, and better satisfy students’ individual needs, boosting the achievement of all students, both repeaters and non-repeaters (cf. sociocultural developmental theory, Vygotsky, 1978). Third, many educators are convinced that repeaters, during their retention year, start with an advantage in academic knowledge and skills over their new, younger classmates, get the opportunity to refresh and experience new successes, and have accumulated more experience in socializing with peers and coping with conflicts at school in comparison to their new, younger classmates. Therefore, many educators expect grade retention to stimulate repeaters’ academic achievement, academic self-concept, self-confidence, pleasure in learning, academic motivation, school well-being, and interpersonal relationships, especially during the retention year (cf. self-efficacy theory, Bandura, 1977; self-determination theory, Deci & Ryan, 2000). Fourth and finally, many educators believe that the threat of grade retention might act like a ‘stick’. By setting standards, being clear that achievement matters, and enforcing negative consequences if goals are not reached, many educators are convinced that all students (both potential repeaters and non-repeaters) will be motivated to work harder and thus achieve better (cf. self-determination theory, Deci & Ryan, 2000).

Holding back struggling students, however, also has potential disadvantages. First, repeating a grade for a big part is simply sheer rehearsal of subjects, while repeaters might already be partly or sufficiently skillful at some subject(s). As a consequence, repeaters – unintentionedly – might not receive the continuous intellectual challenges they need, and thus their achievement progress might in fact be impeded instead of stimulated (cf. contextual developmental theory, Bronfenbrenner, 1980; Vygotsky, 1978). Second, repeaters are left behind by their friends, have to get along with new classmates who might bully or reject them, and might feel like ‘having failed’, seeing all their former classmates advance successfully to the next grade while they have to repeat the grade and its entire subject material. As such, repeating a grade might have a negative impact on repeaters’ academic self-concept, self-confidence, pleasure in learning, academic motivation, school well-being, behavior, and interpersonal relationships, especially during the retention year (cf. self-determination theory, Deci & Ryan, 2000; social comparison theory, Festinger, 1954; social control theory, Hirschi, 2002). Third, the threat of grade retention might cause a lot of stress, among both potential repeaters and non-repeaters, possibly impeding students’ academic progress (self-determination theory, Deci & Ryan, 2000). Fourth and finally, the practice of grade retention incurs extra costs for the government and the (family of the) repeaters, including the expense of providing all repeaters with an additional year of education and the cost in delaying those students’ entry into the labor market by one year. The OECD estimated such costs in Belgium, Spain, the Netherlands, and Portugal to be as high as USD 11,000 or more per repeater (Ikeda & García, 2014).

1.3. Previous research on grade retention effectiveness

Given these disadvantages, do repeaters at least benefit from this practice? This question has been investigated in loads of studies over the last decades, and the results of these studies have in turn been summarized in several meta-analyses (Allen, Chen, Willson, & Hughes, 2009; Bright, 2011; Holmes, 1989; Holmes & Matthews, 1984; Jimerson, 2001) and systematic reviews (Jackson, 1975; Xia & Kirby, 2009). In sum, these meta-analyses and systematic reviews suggest that grade retention is less effective than generally believed by educators, especially in the long run (Driessen, 2020).

These meta-analyses and systematic reviews, however, have five major drawbacks. First, they have almost exclusively focused on grade retention in primary education. This is important, keeping the percentage of primary school repeaters in some countries in mind. Yet, many students also repeat kindergarten or a grade in secondary education. Effects of grade retention in secondary school might be more negative, for example, because the negative experience of broken peer relationships (tied to grade retention) might be more profound at this age (Ikeda & García, 2014), whereas the opposite may hold for kindergarten retention (Xia & Kirby, 2009, p. 30). Somewhat surprisingly, in the meta-analyses and systematic reviews mentioned above (see Allen et al., 2009; Holmes, 1989; Jackson, 1975; Jimerson, 2001; Xia & Kirby, 2009), out of all hundreds of studies summarized, only a small number of studies have assessed effects of repeating kindergarten (28 studies) or a secondary school grade (seven studies).

Second, previous meta-analyses and systematic reviews have predominantly focused on studies regarding grade retention effects in the United States. In the United States, like already pointed out before, student heterogeneity is commonly dealt with using ability grouping, setting, and streaming (Mons, 2007). Having to repeat a grade happens less often (see also Fig. 1) and, if applied, is frequently coupled to additional remedial measures during the retention year (‘retention +’; Schwerdt, West, & Winters, 2017). In such case, grade retention might have positive effects on repeaters’ academic achievement (because of intensive remediation) and negative effects on repeaters’ psychosocial functioning (because of being the sole exception to repeat). As a consequence, results from previous meta-analyses and systematic reviews are not informative for educational practitioners and policy makers from other countries. Out of all hundreds of studies on grade retention effectiveness reviewed so far, only four studies were based on data from students in Canada and Switzerland (see Bright, 2011; Xia & Kirby, 2009), and, surprisingly, none on data from students in any of the countries with the...
highest number of repeaters.

Third, previous meta-analyses and systematic reviews have mainly summarized findings with regard to repeaters’ academic achievement and (to a lesser extent) psychosocial functioning. Insight into these effects is relevant to correctly inform educators and parents who are confronted with a grade retention decision in daily practice. Yet, equally important is knowing how such effects accumulate over time, and influence repeaters’ later school and job career (Reschly & Christenson, 2013). Allowing them extra time, do repeaters have a higher chance of obtaining a high school diploma, graduating from tertiary education, having a paid job etc.? These longer-run effects might be completely different from the shorter-run academic and psychosocial effects, yet, have up till now been summarized in one systematic review (Xia & Kirby, 2009) only, based on 18 studies. Similarly, evenly informative is insight into effects of grade retention on repeaters’ peers (Reschly & Christenson, 2013). Contrary to repeaters, non-repeaters might experience positive effects of grade retention as classrooms might become more homogeneous and thus easier to instruct, to due to retention. Such effects have so far been summarized in no meta-analysis or systematic review at all.

Fourth, previous meta-analyses and systematic reviews have mingled results from grade retention effectiveness studies using a grade approach versus age approach. These two comparison approaches, however, answer very different questions. Grade comparisons compare retainees with their grademates (either their new, younger classmates or their former classmates a school year earlier), and answer the questions how, at the cost of one extra year of education, repeaters develop in comparison to younger students with whom they will eventually graduate and how repeaters affect these younger peers. Age comparisons, by contrast, compare repeaters with their agemates who got promoted to a higher grade, and give an answer to the questions how retainees would have developed and would have affected their peers, had they been promoted to the next grade instead of been held back. Both comparisons and questions are equally relevant for educational practitioners and policy makers, and should never be mingled.

Fifth and finally, previous meta-analyses and systematic reviews have mainly summarized results of methodologically flawed studies. In particular, so far, only three studies reviewed in one meta-analysis (Jackson, 1975) were based on a randomized controlled trial. From an ethical point of view, this is perfectly understandable. It is ethically impossible for a researcher to randomly assign students to an experimental group of grade-retained students versus a control group of grade-promoted students. Therefore, most researchers until now were restricted to following a large cohort of students and registering whom of them were assigned to the experimental group by their teacher(s) and parents, be it based on an a priori defined criterion or not. An important disadvantage of such approach is that under this scenario, retained and promoted students differ on many more characteristics than their retention status alone (e.g., socio-economic status, academic achievement (progress), for an overview, see for instance Xia & Kirby, 2009). Thus, one needs to control for these differences adequately, to not leave open the possibility that these differences rather than grade retention itself cause differences in outcomes to appear. Quasi-experimental methods such as regression discontinuity methods, propensity score methods, instrumental variable methods, difference-in-differences methods, and factor analytic methods are most appropriate for this purpose, yet, have only been used in 19 out of all hundreds of studies reviewed so far (Allen et al., 2009; Bright, 2011; Xia & Kirby, 2009). This has important consequences for the interpretation of findings of previous meta-analyses and systematic reviews, as Allen et al. (2009, p.492) showed that studies using such approaches generally reveal less unfavorable effects on repeaters’ academic achievement than do studies tackling selection bias less effectively.

1.4. The present study

This study aims to summarize the research findings of recent, methodologically sound studies on the effects of retention in grades K-12 on several outcomes, in a variety of countries across the world, disentangling grade and age comparison results. Our research questions are:

(1) RQ1: What is the average effect of grade retention?
(2) RQ2: Does the effect of grade retention differ by
- retention timing (i.e., kindergarten, primary education, versus secondary education);
- retention country heterogeneity management model (i.e., ‘separation’, ‘à la carte integration’, versus ‘uniform integration’);
- retention application (i.e., mere rehearsal versus ‘retention +’);
- outcome domain (i.e., academic achievement, psychosocial functioning, school career, versus later job career of repeaters and non-repeaters);
- outcome timing (i.e., the number of years since the start of the retention year);
- comparison approach (i.e., grade comparison versus age comparison); and
- method (i.e., randomized controlled trial, regression discontinuity method, propensity score method, instrumental variable method, difference-in-differences method, versus factor analytic method)?

2. Method

2.1. Search of studies

Studies on grade retention effects were systematically searched through electronic databases (i.e., Web of Science, PsycINFO, and ERIC) and Google Scholar search engine, using the combination of grade retention, grade repetition, grade failure, repeating a grade, flunking, being held back, repeaters, retainees, or nonpromotion on the one hand, and effect* on the other hand (including corresponding translations in French, German, Dutch, Spanish, and Portuguese). In each database, the selected terms were searched either in the title,
abstract, or keywords. Additionally, hand searches were conducted based on the reference lists of the retrieved studies and the publication lists of known authors working in the field of grade retention effectiveness. This strategy allowed us to find unpublished studies and grey literature, attenuating publication bias.

In total, 628 studies were identified through electronic databases searches and hand searches (see Fig. 2). The studies were exported to EndNote Web software for duplicate removal, and then to Rayan QCRI online tool (Ouzzani, Hammady, Fedorowicz, & Elmagarmid, 2016) for screening.

2.2. Selection of studies

For inclusion in this review, studies had to meet the following five criteria:

(1) Studies needed to consider grade retention, which was defined as repeating a grade after having spent a full year in that grade. Studies examining retention in grades K-12 were included. Studies on retention in tertiary education were excluded given that repeating a grade in tertiary education has distinctive features and procedures as compared to repeating grades K-12. Studies on school delay (in the broad sense) were also excluded as school delay can have multiple causes in addition to grade retention (e.g., delayed school entrance before kindergarten, placement in developmental kindergarten or a transition room, and enrollment in a grade below age expectations because of foreign newcomer status).

(2) Studies needed to evaluate all students, not a specific subcategory of students (e.g., second language learners).

![Flow diagram of study search and selection process based on the PRISMA Statement (Moher et al., 2009).](image-url)
(3) Studies needed to investigate effects of grade retention, not predictors of grade retention. More specifically, studies needed to evaluate effects, among repeaters or non-repeaters, on at least one of the following four outcome domains: (a) academic achievement, (b) psychosocial functioning, (c) school career, and (d) later job career. Academic achievement was operation- alized as scores on (standardized) tests in math, languages, or any other domain, grade point averages, and performance as rated by the teacher. Psychosocial functioning was operationalized very broadly, covering academic self-concept, classroom behavior, school well-being, social skills, and relationships with peers, as judged by students themselves, their teachers, their peers, or external observers. School career was operationalized as decisions regarding students’ school career such as placement in special education, school mobility, high-school completion, etc., as derived from official student records, class and school lists, or teacher questionnaires. Finally, job career was operationalized as income, unemployment status, social security status, etc., as derived from questionnaires or official governmental records.

(4) Studies needed to clearly distinguish an experimental group (grade-retained students) and a control group (grade-promoted students), needed to use an experimental design or (by lack of randomization options) any of the following five types of quasi-experimental methods: (a) regression discontinuity methods, (b) propensity score methods, (c) instrumental variable methods, (d) difference-in-differences methods, or (e) factor analytic methods, and needed to provide a reasonable rationale for the assumptions underlying these methods to be credibly met (e.g., robustness checks, proofs of equivalence of the experimental group and the control group on a considerable amount of observed background characteristics). Studies without a credible control group were excluded. This decision was based on the seminal work of scholars working in the field of causal inference, such as Angrist, Bound, andוד (2010), Imbens and Rubin (2015), and Murnane and Willett (2011, p. 310), as well as on recommendations by the Cochrane Collaboration (Higgins et al., 2019, pp. 321-324), the BEE The Best Evidence Encyclopedia group (n.d.), and the What Works Clearinghouse (2020).

(5) Studies needed to be published between January 2000 and October 2019.

The titles and abstracts of all 496 retrieved unique studies were screened based on these five criteria by the first two authors independently, using the blind review feature in Rayann QCRI tool (with 88% agreement). This resulted in a selection of 132 studies for full text reading (see Fig. 2), which were again screened by the first two authors independently. Finally, 84 studies (Afsa, 2011; Alet, Bonnal, & Favard, 2013; Allensworth, 2005; André, 2009; Andrews, 2014; Appelmans, 2016; Battistin & Schizzerotto, 2012, 2019; Belot & Vandenbergh, 2014; Bhattacharya, 2008; Bietenbeck, 2014; Cham, Hughes, West, & Im, M. H., 2015; Chen, Liu, Zhang, Shi, & Rozelle, 2010; Cockx, Picchio, & Baert, 2019; Cooley- Fruehwirth, Navarro, & Takahashi, 2016; Detienne, 2017; d’Haultfoeuille, 2010; Diris, 2012, 2017; Dong, 2010; Ehmke, Drechsel, & Carstensen, 2010, 2013; Ehmke, Salzer, Pietsch, Drechsel, & Muller, 2017; Eide & Showalter, 2001; Eren, Depew, & Barnes, 2017; Ferreira Sequeda, Golsteyn, & Parra-Cely, 2018; Garcia-Perez, Hidalgo-Hidalgo, & Robles-Zurita, 2014; Gary-Bobo, Gousse, & Robin, 2016; Geng & Rockoff, 2017; Glick & Sahn, 2010; Goos, Van Damme, Onghena, & Petry, 2011; Goos, Van Damme, et al., 2013; Greene & Winters, 2007, 2009; Hamlin, 2013; Hill, Weiss, & Zhai, 2011; Hofer, Farran, Lipsey, Aydogan, & Bilbrey, 2010; Hong & Raudenbush, 2005, 2006; Hong & Yu, 2007, 2008; Hughes, Cao, West, Smith, & Cerda, 2017; Hughes, Chen, Thoemmes, & Kwok, 2015; Hughes, West, Kim, & Bauer, 2016; Huang & Cappella, 2018; Im, Hughes, Kwok, Puckett, & Cerda, 2013; Jacob & Lefgren, 2004, 2009; Keslair, 2007; Klapproth et al., 2016; Koppensteiner, 2014; Kretschmann, Vock, Lüdtke, Jansen, & Gronostaj, 2019; Lamote, Pinxten, Van den Noortgate, & Van Damme, 2014; Leighton, Souza, & Straub, 2019; Lorence, 2014; Lorence & Dworkin, 2006; Mahjoub, 2017; Manacorda, 2012; Mariano & Martorell, 2013; Martorell & Mariano, 2017; Mathys, Véronneau, & Leccocq, 2017; McCombs, Kirby, & Mariano, 2009; Mendez, Kim, Ferron, & Woods, 2015; Moser, West, & Hughes, 2012; Nunes, Balção Reis, & Seabra, 2018; Okurut, 2018; Ou & Reynolds, 2010; Ozek, 2015; Park, Steiner, & Kaplan, 2018; Pereira & Reis, 2014; Peterson & Hughes, 2013; Roderick & Nagaoka, 2005; Schumann, 2012; Schwerdt et al., 2017; Troncos, 2005; Uysal, 2010; Vandecandelaere, Schmitt, Vanlaar, De Fraine, & Van Damme, 2016c; Vandecandelaere, Vanlaar, Goos, De Fraine, & Van Damme, 2011, 2013; Vandecandelaere, Vansteelandt, De Fraine, & Van Damme, 2016a, 2016b; Winters & Greene, 2012; Wu, West, & Hughes, 2008, 2010) covering 1328 single effects were included in our review (see Fig. 2).

2.3. Coding of studies

Following the guidelines from Higgins et al. (2019, p.130), all studies were double-coded. Our coding schema covered the following coding categories: (a) publication information (author, year, and type of publication); (b) general study information (sample size, country, data); (c) grade level in which retention was applied; (d) method; (e) data analysis procedure; (f) outcome examined; (g) grade or age when the outcome was measured; (h) sign of effect; and (i) size of effect.

As for the sign of effect, each single effect was coded taking into consideration its significance and direction (Higgins et al., 2019, p. 328), as significantly negative (detrimental), nonsignificant (null), versus significantly positive (beneficial).

As for the size of effect, each single effect was coded taking into consideration its magnitude, expressed as Hedges’ g. First, Cohen’s $d$ was retrieved or calculated for each single effect, from data directly available in the study (either Cohen’s $d$ or $M$, $SD$, and $N$ for repeaters and non-repeaters; or $t$ and $N$ for repeaters and non-repeaters; or OR; or $\eta^2$) or data available from contacting the study’s authors in case data to calculate Cohen’s $d$ were lacking. Calculations of Cohen’s $d$ and conversions to Cohens’ $d$ from other metrics were done based on formulae provided by Borenstein, Hedges, Higgins, and Rothstein (2009, pp. 37-49). In case no data in addition to $p$-values above .05 were available from the study or its authors, we assigned each single effect a conservative Cohen’s $d$ estimate of 0. Second, as recommended by Lakens (2013) and Higgins et al. (2019, p. 158), we transformed Cohen’s $d$ to Hedges’ g based on a formula provided by Borenstein et al. (2009, p. 39), adjusting our effect size estimates to small sample bias.

As for the outcome examined, it should be noted that, in case of multiple publications based on the same dataset, duplicate
outcomes were excluded. For example, if a first publication evaluated students’ grade point average (GPA) in grade 3 and a second publication evaluated students’ GPA in grades 3–6, grade 3 GPA data were selected from the first publication, and grade 4–6 GPA data from the second publication.

To ensure coding reliability, double-coding was first applied to a subsample of 18 randomly selected studies. Coding results of the two independent coders were confronted to solve discrepancies, until 100% agreement was reached. Second, interrater agreement was calculated for the coded total number of significantly negative, nonsignificant, and significantly positive effects, considering the total number of effects extracted (r = 0.95) and the coded direction of these effects (Cohen’s k = 0.96). Online appendices A and B present the coding results summarized per study. Online appendix C presents the detailed coding results per single effect.

2.4. Synthesis of studies

In order to estimate the average and differential effects of grade retention based on the 84 studies and 1328 single effects, we performed a vote counting analysis, a statistical meta-analysis, and a publication bias analysis. All analyses were performed with Excel software and R 4.0.2. software (package metafor, version 2.4-0, Viechtbauer, 2010). The R script can be found in Online appendix F.

2.4.1. Vote counting analysis

We transformed the numbers of significantly negative, nonsignificant, and significantly positive effects per study (from Online appendix A) to proportions per study (Online appendix B), and averaged these proportions across studies, to estimate the average effect of grade retention (RQ1). We also performed a sign test on the number of significantly positive effects versus significantly negative effects, to test for significance (Higgins et al., 2019, p. 328).

2.4.2. Statistical meta-analysis

We analyzed all 1328 single effect sizes by performing a series of multilevel meta-regressions. A multilevel approach was chosen as most studies (94%) yielded multiple effect sizes, violating the assumption of independence of effect sizes underlying a traditional fixed-effects approach. More specifically, we fitted a series of three-level linear regression models. A three-step sequential modeling approach was used, in which each successive model reflected an increased complexity. In a first step (Model 1), a regression model of grade retention (RQ1). We also performed a sign test on the number of significantly positive effects versus significantly negative effects, to test for significance (Higgins et al., 2019, p. 328).

2.4.3. Publication bias analysis

We checked for publication bias in our sample via two different approaches. First, we used the small sample bias method (Borenstein et al., 2009, pp.277-292; Fernandez-Castilla et al., 2018; Higgins et al., 2019, pp. 362-369). We plotted a funnel plot, with all 1328 effect sizes plotted on the horizontal axis and their standard errors on the vertical axis (see Online Appendix E), and we applied Egger’s weighted regression test to evaluate funnel plot asymmetry. The results indicated an absence of publication bias (z = 0.18, p = .680). Second, we applied an additional moderator regression analysis (Model 2h), similar in approach as those discussed in the previous paragraph (Models 2a to 2g), but now with publication status as study-level dummy predictor added to the fixed part of the baseline model. Again, the results suggested an absence of publication bias, by revealing no differences in grade retention effect sizes between published and unpublished studies ($R^2 (1, 1326) = 0.10, p = .680$).

3. Results

As can be seen in Online appendix A, most studies focused on grade retention in primary education ($k_{studies} = 40$), were conducted in ‘à la carte integration countries’ ($k_{studies} = 42$, which in our case are studies from the United States only), evaluated effects on repeaters’ academic achievement ($k_{studies} = 34$) or a combination of outcomes ($k_{studies} = 29$), used either a grade comparison approach ($k_{studies} = 32$) or an age comparison approach ($k_{studies} = 43$), not both ($k_{studies} = 9$), used a propensity score method ($k_{studies} = 43$), followed students for at least three school years ($k_{studies} = 67$; $M_{duration} = 6.93$ years), examined at least 1000 students ($k_{studies} = 61$; range from 52 students to more than 1,000,000 students), were published in a journal ($k_{studies} = 64$), and were so far not examined in the
Note. Results of three-level moderator analyses of grade retention effects.

3.1. Average effect of grade retention

On average, 35% of effects were significantly negative, 41% non-significant, and 24% significantly positive. Significantly negative effects (favoring non-repeaters) were as present as significantly positive effects (favoring repeaters) according to the vote count sign test, \( z = -0.60, p = .550 \).

The mean effect size estimated in the three-level meta-regression baseline model (Model 1) was \(-0.04 (SE = 0.04, 95\% CI = [-0.11; 0.03], r(t(1327)) = -1.11, p = .266)\), indicating that repeaters and non-repeaters, on average, show a similar development, pooling all outcomes we investigated. The forest plot, showing the effect sizes per study, is presented in Online appendix D.

3.2. Differential effects of grade retention

The differential effect sizes estimated in the three-level meta-regression models with predictors (Models 2a to 2g) are summarized in Table 1. Effect sizes were estimated to differ, on average, by retention country heterogeneity management model \( (F(2, 1294) = 3.37, p = .035) \), retention application \( (F(1, 1326) = 11.57, p < .001) \), outcome domain \( (F(4, 1323) = 5.99, p < .001) \), outcome timing \( (F(1, 1326) = 9.43, p = .002) \), comparison approach \( (F(1, 1326) = 23.60, p < .001) \), and method \( (F(5, 1322) = 3.06, p = .009) \), but not retention timing \( (F(2, 1325) = 1.46, p = .232) \). Pairwise \( t \)-tests (with dummy variables instead of categorical variables) showed that

<table>
<thead>
<tr>
<th>Moderator</th>
<th>( k_{\text{studies}} )</th>
<th>( k_{ES} )</th>
<th>ES (g)</th>
<th>95% CI</th>
<th>Test statistic</th>
<th>( p )</th>
<th>Test statistic</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retention timing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F(2, 1294) = 3.37</td>
<td>.035</td>
<td>t(1290) = 2.14</td>
<td>.033</td>
</tr>
<tr>
<td>- Kindergarten</td>
<td>14</td>
<td>177</td>
<td>-0.13</td>
<td>-0.27</td>
<td>1.46</td>
<td>.232</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>- Primary education</td>
<td>51</td>
<td>794</td>
<td>-0.02</td>
<td>-0.09</td>
<td>0.60</td>
<td>.548</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>- Secondary education</td>
<td>28</td>
<td>357</td>
<td>-0.04</td>
<td>-0.13</td>
<td>0.05</td>
<td>.959</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Retention country heterogeneity management model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F(2, 1294) = 3.37</td>
<td>.035</td>
<td>t(1290) = 2.14</td>
<td>.033</td>
</tr>
<tr>
<td>- Separation</td>
<td>20</td>
<td>322</td>
<td>-0.18</td>
<td>-0.32</td>
<td>-0.03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Uniform integration</td>
<td>20</td>
<td>163</td>
<td>-0.08</td>
<td>-0.23</td>
<td>0.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- A la carte integration</td>
<td>42</td>
<td>809</td>
<td>0.05</td>
<td>-0.05</td>
<td>0.14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retention application</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F(1, 1326) = 11.57</td>
<td>&lt;.001</td>
<td>t(1290) = 1.68</td>
<td>.094</td>
</tr>
<tr>
<td>- Mere rehearsal</td>
<td>69</td>
<td>789</td>
<td>-0.10</td>
<td>-0.17</td>
<td>-0.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Retention +</td>
<td>15</td>
<td>539</td>
<td>0.20</td>
<td>0.04</td>
<td>0.35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome domain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F(4, 1323) = 5.99</td>
<td>&lt;.001</td>
<td>t(1290) = 1.26</td>
<td>.208</td>
</tr>
<tr>
<td>- Repeaters’ academic achievement</td>
<td>58</td>
<td>524</td>
<td>-0.07</td>
<td>-0.14</td>
<td>-0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Repeaters’ psychosocial functioning</td>
<td>22</td>
<td>326</td>
<td>0.08</td>
<td>-0.01</td>
<td>0.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Repeaters’ school career</td>
<td>29</td>
<td>446</td>
<td>-0.10</td>
<td>-0.19</td>
<td>-0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Repeaters’ job career</td>
<td>2</td>
<td>10</td>
<td>-0.07</td>
<td>-0.34</td>
<td>0.19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Non-repeaters’ development(^a)</td>
<td>4</td>
<td>22</td>
<td>0.02</td>
<td>-0.26</td>
<td>0.30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome timing</td>
<td>84</td>
<td>1328</td>
<td></td>
<td></td>
<td>F(1, 1326) = 9.43</td>
<td>.002</td>
<td>t(1290) = -3.30</td>
<td>.001</td>
</tr>
<tr>
<td>- Retention year</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F(1, 1326) = 23.60</td>
<td>&lt;.001</td>
<td>t(1290) = -4.99</td>
<td>.001</td>
</tr>
<tr>
<td>Comparison approach</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F(5, 1322) = 3.06</td>
<td>.009</td>
<td>t(1290) = 1.17</td>
<td>.242</td>
</tr>
<tr>
<td>- Grade</td>
<td>41</td>
<td>559</td>
<td>0.05</td>
<td>-0.03</td>
<td>0.12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Age</td>
<td>52</td>
<td>769</td>
<td>-0.11</td>
<td>-0.18</td>
<td>-0.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method(^b)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F(1, 1326) = 0.10</td>
<td>.753</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>- DID</td>
<td>9</td>
<td>76</td>
<td>-0.05</td>
<td>-0.20</td>
<td>0.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- FAD</td>
<td>3</td>
<td>29</td>
<td>-0.06</td>
<td>-0.43</td>
<td>0.32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- IV</td>
<td>15</td>
<td>74</td>
<td>-0.15</td>
<td>-0.31</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- PS</td>
<td>46</td>
<td>591</td>
<td>-0.09</td>
<td>-0.18</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- RCT</td>
<td>1</td>
<td>17</td>
<td>0.06</td>
<td>-0.56</td>
<td>0.69</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- RD</td>
<td>17</td>
<td>541</td>
<td>0.17</td>
<td>0.03</td>
<td>0.31</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Results of three-level moderator analyses of grade retention effects.

\(^a\) A sensitivity analysis was carried out to check the effect of leaving out subgroups with less than 10 studies. Results (available upon request from the authors) equal those presented here.

\(^b\) Effects regarding non-repeaters’ academic achievement, psychosocial functioning, school career, and job career (retrieved from 4 studies) were collapsed into one category due to small sample size.

\(^c\) DID = difference-in-differences method; FAD = factor analytic method; IV = instrumental variable method; PSM = propensity score method; RCT = randomized controlled trial; RD = regression discontinuity method.
effect sizes and thus effects of grade retention were significantly more negative (a) in ‘separation countries’ (Hedges’ $g = -0.18$), (b) when retention was applied as ‘mere rehearsal’ (Hedges’ $g = -0.10$), and (c) when applying an age comparison approach (Hedges’ $g = -0.11$). By contrast, pairwise t-tests also showed that effect sizes and thus effects of grade retention were significantly more positive (a) in ‘à la carte integration countries’ (Hedges’ $g = 0.05$), (b) when retention was applied as ‘retention +’ (Hedges’ $g = 0.20$), (c) when looking at effects on repeaters’ psychosocial functioning (Hedges’ $g = 0.08$), (d) when applying a grade comparison approach (Hedges’ $g = 0.05$), and (e) when testing effects with a regression discontinuity method (Hedges’ $g = 0.17$), and an omnibus F-test showed that effect sizes and thus effects of grade retention were significantly more positive when looking at effects in the shorter run (with Hedges’ $g = 0.01$ in the retention year, diminishing with 0.01 each year afterwards). Differential effect sizes by retention country heterogeneity management model ($t(1290) = 2.14, p = .033$), outcome timing ($t(1290) = -3.30, p = .001$), and comparison approach ($t(1290) = -4.99, p < .001$) were estimated to keep their significance in a three-level meta-regression model with joint predictors (Model 3, see Table 1), proving their robustness.

4. Discussion

4.1. Summary of the findings

A large body of research has examined the effectiveness of holding students back in grade. We reviewed 84 recent studies on this topic. Three important findings emerge from our study.

First, our study seems to indicate that the field of grade retention effectiveness research has considerably improved during the last two decades. More attention has been paid to examining effects of grade retention in a variety of countries across the world, and especially in countries where grade retention is most frequently applied (39 of the 84 reviewed studies were conducted in ‘separation countries’ and ‘uniform integration countries’, whereas almost none in the past). In addition, more attention has been paid to evaluating effects of grade retention on a broad spectrum of outcomes, instead of academic achievement alone (29 of the 84 reviewed studies looked at a combination of outcomes). And finally, more attention has been paid to evaluating grade retention effects using a range of sound methods, among which a propensity score method currently seems to be used most often, especially among scholars in education and psychology.

Second, our study seemingly shows that, on average, grade retention has no (negative nor positive) effect on repeaters’ and non-repeaters’ development. We find as many significantly negative effects as significantly positive effects, and an average effect size of $-0.04$. This seems contrary to all previously published meta-analyses and systematic reviews that suggested grade retention to be less effective than generally believed by educators, especially in the long run (Allen et al., 2009; Bright, 2011; Holmes, 1989; Holmes & Matthews, 1984; Jackson, 1975; Jimerson, 2001; Xia & Kirby, 2009). It should be kept in mind, however, that we selected studies using more rigorous methods. Allen et al. (2009) obtained a similar zero average effect size when restricting the study sample to high-quality studies.

Third, regardless of an average zero effect, our study at the same time seems to indicate differential effects to be present, under specific circumstances. More specifically, we find our effect sizes to range from $-0.18$ to 0.20. Grade retention seems least effective in Belgium, Germany, Luxembourg, and Uganda (‘separation countries’) and most effective in the United States, especially under the ‘retention +’ regime, nowadays in effect in many US regions. This finding is in line with contextual developmental theory (Bronfenbrenner, 1980; Vygotsky, 1978) stating that students need continuous intellectual challenges, and not sheer rehearsal of subjects, while already being partly or sufficiently skillful at some subject(s). Furthermore, of all outcome domains investigated, repeaters’ psychosocial functioning seems most positively affected by grade retention. This finding is congruent with self-efficacy theory (Bandura, 1977), self-determination theory (Deci & Ryan, 2000), and social comparison theory (Festinger, 1954), altogether predicting that repeaters (as compared to younger non-retained grademates) will show a more positive academic self-concept, more self-confidence, more pleasure in learning, a better academic motivation, a higher school well-being, and more positive interpersonal relationships, especially during the retention year, due to having advantages in academic knowledge and skills at the start of the retention year, having more experience in socializing with peers and coping with conflicts, and getting chances to ‘refresh’ and experience new successes throughout the retention year. Repeaters’ school career and later job career (both long-run effects) seem most negatively affected by grade retention. Compared with similar non-repeaters, repeaters have a higher chance of placement in special education, track downgrading, school absences, school suspensions, school drop-out, non-entry into tertiary education, low income, public aid receipt, etc. Apparently, very small negative effects on repeaters’ academic achievement seem to accumulate with time, creating a negative ‘snowball effect’ on repeaters’ school and job career, a finding in line with that found by Xia and Kirby (2009).

Next, grade retention seems most effective under a grade comparison approach and least effective under an age comparison approach. This finding is congruent with findings of Allen et al. (2009) and Xia and Kirby (2009), and emphasizes, once more, the importance of disentangling grade and age comparison results. Finally, grade retention effects seem most positive when evaluated via a regression discontinuity method. This finding is not surprising as this method selects ‘marginal’ students (i.e., students scoring just below or above a certain predefined achievement level), and thus disregards the lowest achievers who need a wider range of remediation measures. This finding stresses the need for educational practitioners and policy makers to interpret results of grade retention effectiveness studies with care, always keeping in mind the method used.

4.2. Study limitations and suggestions for future research

Although the results of this study provide insights into the effects of grade retention, several limitations should be acknowledged.
4.3. Implications for educational policy and practice

First, looking at our selection procedure, we excluded a large share of studies because they did not clearly distinguish an experimental versus control group, nor used an experimental design or (by lack of randomization options) quasi‐experimental method, nor provided a reasonable rationale for the assumptions underlying these methods to be credibly met (i.e., inclusion criterion 4; cf. recommendations by Antonakis et al., 2016; Higgins et al., 2019; Imbens & Rubin, 2015; Murnane & Willett, 2011, the BEE group, and the What Works Clearinghouse). Including such studies (e.g., observational studies with regression‐based analyses) would have allowed us to test the differential effect of grade retention in yet another method subcategory (in addition to a randomized controlled trial, regression discontinuity method, etc.). Based on findings of Allen et al. (2009, p. 492) focusing on grade retention effects on repeaters’ academic achievement, we could hypothesize that the grade retention effect size among such studies would be 0.22 lower (and thus negative), and that the average grade retention effect size would drop, as a result of including such studies in our meta‐analysis sample. Future research could confirm this hypothesis. We decided not to include such studies, for reasons of methodological rigor. In addition, after applying our five inclusion criteria altogether, 84 studies remained. It appeared that these 84 studies investigated grade retention effects in a still rather fragmentary way: (like older studies reviewed in previous meta‐analyses and systematic reviews) they still less frequently focus on grade retention in kindergarten and secondary education, still seldom evaluate effects on repeaters’ job career and non‐repeaters’ development, and still are not or rarely conducted in countries such as Colombia, Luxembourg, Spain, or Portugal, four countries in the OECD top 5 regarding grade retention rates, nor countries such as Finland or Sweden, two countries in the OECD bottom 10 regarding grade retention rates. To better generalize and contextualize our findings, we believe more research is needed considering a variety of retention timings, outcomes, and contexts.

Second, looking at our coding procedure, we coded our retention application moderator as ‘mere rehearsal’ versus ‘retention +’, which is a very raw categorization, based on general educational policy regulations. In practice, repeaters might get additional support, by different people, and probably for different reasons. Research on this matter is almost absent (Reschly & Christenson, 2013), but highly needed to better understand what happens during the retention year and how this shapes retention effects.

Third, looking at our meta‐analysis procedure, we were able to explore seven potentially important moderators of grade retention effects: (a) retention timing, (b) retention country heterogeneity management model, (c) retention application, (d) outcome domain, (e) outcome timing, (f) comparison approach, and (g) method. Unfortunately, we could not add student characteristics such as ability or second language learning status as an extra set of moderators, because fewer studies so far examined such differential effects and our meta‐analysis sample as a result would be too small to test for significance. With more research forthcoming in the next decade, however, our exploration could be extended in such way, resulting in a better understanding of who does (or does not) benefit from grade retention, highly relevant questions raised by educational policy makers and practitioners.

Fourth, looking at our publication bias procedure, we were constrained in detecting bias via different methods, simply because such methods so far have been hardly developed nor tested in the context of multilevel meta‐analysis (Fernandez‐Castilla et al., 2018). This seems an important venue for future research.

4.3. Implications for educational policy and practice

Our study revealed that repeaters seem worst off in ‘separation countries’, applying grade retention and tracking to tackle student heterogeneity, and best off in the United States, applying ability grouping, setting, and streaming (and to a much lesser extent grade retention) to deal with student heterogeneity. Keeping in mind the costs grade retention brings about to the economy and the (family of the) individual student, we advocate countries, and particularly ‘separation countries’ (e.g., ability grouping, ability tracking), to deal with student heterogeneity. Considering a variety of retention timings, outcomes, and contexts.

First, looking at our selection procedure, we excluded a large share of studies because they did not clearly distinguish an experimental versus control group, nor used an experimental design or (by lack of randomization options) quasi‐experimental method, nor provided a reasonable rationale for the assumptions underlying these methods to be credibly met (i.e., inclusion criterion 4; cf. recommendations by Antonakis et al., 2016; Higgins et al., 2019; Imbens & Rubin, 2015; Murnane & Willett, 2011, the BEE group, and the What Works Clearinghouse). Including such studies (e.g., observational studies with regression‐based analyses) would have allowed us to test the differential effect of grade retention in yet another method subcategory (in addition to a randomized controlled trial, regression discontinuity method, etc.). Based on findings of Allen et al. (2009, p. 492) focusing on grade retention effects on repeaters’ academic achievement, we could hypothesize that the grade retention effect size among such studies would be 0.22 lower (and thus negative), and that the average grade retention effect size would drop, as a result of including such studies in our meta‐analysis sample. Future research could confirm this hypothesis. We decided not to include such studies, for reasons of methodological rigor. In addition, after applying our five inclusion criteria altogether, 84 studies remained. It appeared that these 84 studies investigated grade retention effects in a still rather fragmentary way: (like older studies reviewed in previous meta‐analyses and systematic reviews) they still less frequently focus on grade retention in kindergarten and secondary education, still seldom evaluate effects on repeaters’ job career and non‐repeaters’ development, and still are not or rarely conducted in countries such as Colombia, Luxembourg, Spain, or Portugal, four countries in the OECD top 5 regarding grade retention rates, nor countries such as Finland or Sweden, two countries in the OECD bottom 10 regarding grade retention rates. To better generalize and contextualize our findings, we believe more research is needed considering a variety of retention timings, outcomes, and contexts.

Second, looking at our coding procedure, we coded our retention application moderator as ‘mere rehearsal’ versus ‘retention +’, which is a very raw categorization, based on general educational policy regulations. In practice, repeaters might get additional support, by different people, and probably for different reasons. Research on this matter is almost absent (Reschly & Christenson, 2013), but highly needed to better understand what happens during the retention year and how this shapes retention effects.

Third, looking at our meta‐analysis procedure, we were able to explore seven potentially important moderators of grade retention effects: (a) retention timing, (b) retention country heterogeneity management model, (c) retention application, (d) outcome domain, (e) outcome timing, (f) comparison approach, and (g) method. Unfortunately, we could not add student characteristics such as ability or second language learning status as an extra set of moderators, because fewer studies so far examined such differential effects and our meta‐analysis sample as a result would be too small to test for significance. With more research forthcoming in the next decade, however, our exploration could be extended in such way, resulting in a better understanding of who does (or does not) benefit from grade retention, highly relevant questions raised by educational policy makers and practitioners.

Fourth, looking at our publication bias procedure, we were constrained in detecting bias via different methods, simply because such methods so far have been hardly developed nor tested in the context of multilevel meta‐analysis (Fernandez‐Castilla et al., 2018). This seems an important venue for future research.

4.3. Implications for educational policy and practice

Our study revealed that repeaters seem worst off in ‘separation countries’, applying grade retention and tracking to tackle student heterogeneity, and best off in the United States, applying ability grouping, setting, and streaming (and to a much lesser extent grade retention) to deal with student heterogeneity. Keeping in mind the costs grade retention brings about to the economy and the (family of the) individual student, we advocate countries, and particularly ‘separation countries’, efforts to make available more resources for schools (a) to identify at an earlier stage those students at risk for developing school difficulties, by means of a more regular, careful monitoring of their students’ progress (especially those countries not yet doing so), (b) to provide all at‐risk students with targeted remedial interventions that have been proven to be cost‐effective (e.g., self‐regulation interventions, reading interventions, 1‐to‐1 or small‐group tuition, summer schools; Baye, Inns, Lake, & Slavin, 2019; Higgins et al., 2016; Slavin, Lake, Davis, & Madden, 2011), be it inside or outside the classroom, from the exact first moment difficulties arise, and (c) to pay considerable attention to repeaters and to give them all services needed, if it is decided to hold particular students back in grade, in their best interest, whatever the reason may be (e.g., outlining a specific plan highlighting how the retention year will be different from the previous year, for example, with regard to level of intensity and duration of support).

CRediT authorship contribution statement

Mieke Goos: Conceptualization, Methodology, Formal analysis, Writing – original draft, Visualization, Funding acquisition. Joana Pipa: Conceptualization, Methodology, Formal analysis, Writing – original draft, Visualization. Francisco Peixoto: Supervision, Funding acquisition, Mieke Goos and Joana Pipa contributed equally to this study.

Declaration of competing interest

None.

Acknowledgement

This work was supported by the Ministry of the Flemish Government [Grant OBPWO 10.02 awarded to the first author] and by Portuguese national funds via Fundação para a Ciência e Tecnologia (FCT) [Grant FHO/BD/132195/2017 awarded to the second...
Supplementary data to this article can be found online at https://doi.org/10.1016/j.edurev.2021.100401.
* Park, S., Steiner, P. M., & Kaplan, D. (2018). Identification and sensitivity analysis for average causal mediation effects with time-varying treatments and mediators:
* Pereira, M. C., & Reis, H. J. (2014). Grade retention during basic education in Portugal: Determinants and impact on student achievement.
* Nunes, L. C., Balc


