DOI: 10.1002/cl2.1159

PROTOCOL

Campbell Collaboration WILEY

PROTOCOL: The effects of small class sizes on students' academic achievement, socioemotional development, and well-being in special education

Anja Bondebjerg | Nina T. Dalgaard | Trine Filges | Morten K. Thomsen | Bjørn C. A. Viinholt

VIVE—The Danish Center for Social Science Research, Copenhagen, Denmark

Correspondence
Anja Bondebjerg, VIVE—The Danish Center for Social Science Research, Herluf Trolles Gade 11, Copenhagen 1052, Denmark.
Email: anbo@vive.dk

Abstract

This is the protocol for a Campbell review. The objective of this systematic review is to uncover and synthesise data from studies to assess the impact of small class sizes on the academic achievement, socioemotional development, and well-being of students with special educational needs. Where possible, we will also investigate the extent to which the effects differ among subgroups of students. Furthermore, we will perform a qualitative exploration of the experiences of children, teachers, and parents with special education class sizes.

1 | Background

1.1 | Description of the condition

Class size decreases in general education are some of the most researched educational interventions in social science, yet researchers have not reached any final conclusions regarding their effects. While some researchers point to small and insignificant differences between varying class sizes, others find positive and significant effects of small class sizes on, for example, children's academic outcomes. In a previous Campbell Systematic Review on small class sizes in general education, Filges et al. (2018) found evidence suggesting, at best, a small effect on reading achievement, whereas there was a negative, but statistically insignificant, effect on mathematics.

While research on the relationship between general education class size and student achievement is plentiful, research on class size in special education is scarce (see e.g., McCrea, 1996; Russ et al., 2001; Zarghami & Schnellert, 2004), even though class size issues must be considered particularly important to students with special educational needs. These students compose a highly diverse group in terms of diagnoses, functional levels, and support needs, but the share a common need for special educational accommodations, which often entails additional instructional support in smaller units than what is normally provided in general education. Special education class sizes may vary greatly, both across countries and regions, as well as across different student groups, but will usually be small relative to general education classrooms. In most cases, placement in special education, as opposed to, for example, inclusion in general education, is based exactly in the child's need for close adult support in a smaller unit, where instruction can be tailored to the needs of each child and a calmer, more structured environment can be created. Following this, one may assume that there are advantages to small class sizes in special education, in that children are placed in a suitable environment with the support they need to thrive and learn (for a discussion of perceptions on the benefits of special education, see e.g., Kavale & Forness, 2000). However, there may also be challenges to small class sizes, for example, in terms of the opportunities available for building friendships.

Finally, class size issues, both in general and in special education, are associated with ongoing discussions on educational spending and budgetary constraints. Hence, in school systems imposed with financial constraints, small class sizes in special education settings may be deemed too expensive. As a result, children with special educational needs may be placed in larger units with potential adverse effects on their learning and well-being. At this point, there is however a lack of clarity as to the effects of special education class sizes on student academic achievement and socioemotional development. Inevitably, such lack of clarity is an obstacle for special educators and

policymakers trying to make informed decisions. This highlights the policy relevance of the current systematic review, in which we will examine the effects of small special education class sizes on the academic achievement, socioemotional development, and well-being of children with special educational needs. In working towards this aim, we will apply a mixed methods approach, consisting of (1) a statistical meta-analysis of class size effects in special education on students' academic achievement, socioemotional development, and well-being; and (2) a qualitative exploration of the experiences of children, teachers, and parents with class sizes in special education. We choose to include studies applying a qualitative methodology because the combination of quantitative and qualitative methods allows us to draw on the strengths of both approaches, providing a deeper insight into the complexity of class size questions in special education, including the voices of children and teachers who spend their everyday lives in special education contexts.

1.2 | Description of the intervention

Special education refers to educational settings designed to provide instruction for children with special educational needs. In such settings, both the instructional and physical classroom environment may be adjusted to accommodate the specific needs of the student group, as in the use of individual work tables and visual aids (pictograms) for children on the autism spectrum. Special education may be full time or part time (e.g., in the form of resource rooms attended by students with special educational needs). We will include studies of all kinds of special education.

In this review, it is important to distinguish between the following terms: *class size*, *student–teacher ratio*, and *caseload*. *Class size* refers to the number of students present in a classroom at a given point in time. *Student–teacher ratio* refers to the number of students per teacher within a classroom or an educational setting. Furthermore, some studies may apply the term *caseload* which is typically defined as the number of students with individual education plans (IEPs) for whom a teacher serves as "case manager" (Minnesota Department Children Families & Learning, 2000). In this review, the intervention is *a reduction in class size*. Thus, studies only considering student–teacher ratios or caseloads are not eligible. However, some studies investigating student–teacher ratios may qualify as class size studies, for example, when the number of teachers is held stable, while alternating between different numbers of students (e.g., comparing ratios of 1:3, 1:6, and 1:9). In such cases, it is in fact the number of students, that is, the class size, which is manipulated, hence these studies will be eligible for inclusion.

Our rationale for focusing on class size is based in the belief that although class size and student–teacher ratios or caseloads in special education are related, they involve somewhat different assumptions about how a reduction might change the opportunities for students and teachers. With class size, the mechanism in play is based on assumptions about the dynamics of a smaller group and the belief that with smaller groups, teachers are better able to develop an in-depth understanding of student needs through more focused interactions, better assessment, and fewer disciplinary problems (Ehrenberg et al., 2001; Filges et al., 2018). The size of the group in itself will often be of specific importance to students with special educational needs, for example, students diagnosed with sensory processing disorders, making them sensitive to noise and movement, or students with ASD who struggle with reading social cues in larger groups. For such students, being in a larger class would likely feel overwhelming and stressful, no matter the student/teacher ratio.

Student/teacher ratio and caseload are also of great importance, but do not take in the specific mechanisms of being in a smaller group which we find to be central in special education. We acknowledge the relatedness of these concepts to class size and are as noted aware that terms may in some cases overlap. We will pay attention to this when searching for studies by adding a search term for student/teacher ratio and when screening the studies.

It is possible that the intensity of the intervention, that is, the size of the reduction and the initial class size from which the reduction is made, can play a role in determining the intervention effects. For intensity, the question is: how small does a class have to be in order to optimise the advantage? In general education for example, large gains are attainable when class size is below 20 students (Biddle & Berliner, 2002; Finn, 2002), but gains are also attainable if class size is not below 20 students (Angrist & Lavy, 1999; Borland et al., 2005; Fredriksson et al., 2013; Schanzenbach, 2007). It has been argued that the impact of class size reduction of different sizes and from different baseline class sizes is reasonably stable and more or less linear when measured per student (Angrist & Pischke, 2009; Schanzenbach, 2007). Other researchers argue that the effect of class size is not only non-linear but also non-monotonic, implying that an optimal class size exists (Borland et al., 2005). Thus, the question of whether the size of reduction and initial class size matters for the magnitude of gain from small classes is still an open question. For this reason, we will include intensity (size of reduction and initial class size) as a moderator.

1.3 | How the intervention might work

Due to the specialised and varied nature of special needs provision, issues of class size in this area are likely to be complex (Ahearn, 1995). However, small class sizes may promote student engagement and instructional individualisation, which is of particular importance to students with special educational needs. A research report from 1997 evaluating increases in resource room instructional group size in New York City public schools may serve to illustrate the importance of individualisation in special education (Gottlieb & Alter, 1997). The report indicated that increases in instructional group sizes from five to at most eight students per teacher led to decreases in the reading achievement scores of resource room students. Resource room teachers reported diminished opportunities for sufficiently helping students. Furthermore, observations revealed little time spent on individual instruction.

Small class sizes may be better suited to address the potential physical and psychological challenges of students with special educational needs, for example, by providing closer adult/child interaction, better accommodation of individual needs, and a more focused social interaction with fewer peers. Thus, smaller class sizes in special education may have a positive impact on both academic achievement and socioemotional development as well as on student well-being in school.

On the other hand, small class sizes may limit the possibilities for finding compatible peers with whom to build friendships, hence leading to adverse effects on students' social and personal well-being in school. This may also impact on the options available for building and training social skills, which are vital to, for example, students with autism spectrum disorders. Furthermore, small class sizes may lead to decreased variation in academic and social skills within the class, limiting the potential for positive peer effects on student academic learning and socioemotional development (e.g., learning from peers with more advanced academic skills).

1.4 | Why it is important to do this review

As previously noted, there is a lack of clarity as to the impact of special education class sizes on student academic achievement, socioemotional development and well-being, making it difficult for special educators and policymakers to make informed decisions. Furthermore, class size alterations are associated with ongoing discussions on educational spending and budgetary constraints, highlighting the policy relevance of strengthening the knowledge base through a systematic review of the available literature.

Few authors have tried to review the available literature on special education class sizes, and these reviews have not followed rigorous, systematic frameworks, such as that applied in a Campbell systematic review. In 1996, *Linda D. McCrea* conducted a review on special education and class size including a sample of American studies. These studies pointed to some effects of class size on the learning environment in class as well as on student achievement and behaviour, especially at the elementary level. Furthermore, in an article exploring the class size literature, Zarghami and Schnellert (2004) examined the effects of appropriate class size and caseload on special education student academic achievement. The authors were not able to identify a single best way to determine appropriate class and group sizes for special education instruction. However, they pointed to the existence of well-qualified teachers as an important factor in increasing student achievement. Finally, in a 1995 report, *Eileen M. Ahearn* analysed state special education regulations on class size/caseload in the United States and reviewed research on class size in general education and special education. The report showed that state requirements for class size/caseload in special education programmes were much more specific and complicated than those for general education, and that the specialised nature and variety of the services delivered to students with special educational needs, combined with the restrictions attributable to specific student disabilities, contributed to those complications. In line with the article by Zarghami & Schnellert, Ahearn (1995) concluded that there was no single best way to determine class sizes for special education programmes, adding that the information available was inadequate.

The above-mentioned reviews did not apply the extensive, systematic literature searches and critical appraisals that are performed in a Campbell systematic review. Furthermore, they date back 15 years or more, which means that they do not include newer developments in special education research. Therefore, we find that there is a need for an up-to-date, rigorously performed systematic review of the available literature on the effects of small class sizes in special education. We believe that the present systematic mixed methods review will serve this need by providing a comprehensive overview of the field and a robust synthesis founded in both a statistical meta-analysis and a thematic qualitative synthesis.

2 | OBJECTIVES

The objective of this systematic review is to uncover and synthesise data from studies to assess the impact of small class sizes on the academic achievement, socioemotional development, and well-being of students with special educational needs. Where possible, we will also investigate the extent to which the effects differ among subgroups of students. Furthermore, we will perform a qualitative exploration of the experiences of children, teachers, and parents with special education class sizes.

3 | METHODS

3.1 | Criteria for considering studies for this review

3.1.1 | Types of studies

The proposed project will follow standard procedures for conducting systematic reviews using meta-analysis techniques.

In order to summarise what is known about the possible causal effects of small class sizes in special education, we will include all study designs that use a control group, that is, a group of students in larger special education classes. This is further outlined in the section *Assessment of risk of bias in included studies*, and the methodological appropriateness of the included studies will be assessed according to the risk of bias.

The study designs we will include in the review are:

1. Randomised and quasi-randomised controlled trials (allocated at either the individual level or cluster level, e.g., class/school/geographical area, etc.).
2. Nonrandomised studies (allocation has occurred in the course of usual decisions, not controlled by the researcher, and there is a comparison of two or more groups of participants, that is, at least a treated group and a control group).

Studies using single group pre-post comparisons will not be included. Nonrandomised studies using an instrumental variable approach will not be included—see Appendix E for our rationale for excluding studies of these designs. A further requirement to all types of studies (randomised as well as nonrandomised) is that they are able to identify an intervention effect. Studies where, for example, small classes are present in one school only and the comparison group is larger classes at another school (or more schools for that matter) cannot separate the treatment effect from the school effect. The main control or comparison condition is students in special education classes with more students than in the treatment classes. We will only include studies that use measures of class size and measures of outcome data at the individual or class level. We will exclude studies that rely on measures of class size and measures of outcomes aggregated to a level higher than the class (e.g., school or school district).

In addition to exploring the causal effects of small class sizes in special education, we wish to gain insight into the experiences of children, teachers and parents with class size issues in special education contexts. To this end, we will include all types of empirical qualitative studies that collect primary data and provide descriptions of main methodological issues such as sampling, data collection procedures, and type of data analysis. Eligible qualitative studies may apply a wealth of data collection methods including but not limited to participant observations, in-depth interviews, or focus groups.

3.1.2 | Types of participants

The review will include children with special educational needs in grades K to 12 (or the equivalent in European countries) in special education. Studies that meet inclusion criteria will be accepted from all countries. In this review, we exclude children in home- or preschool as well as children placed in treatment facilities.

Some controversy exists regarding the definition of what constitutes a special educational need (Vehmas, 2010, Wilson, 2002). However, in this review we apply the widely used definition from the US Individuals with Disabilities Education Act (IDEA), in which special needs are divided into 13 different disability categories under which children are eligible for services[1]. These categories are:

- specific learning disability (covers challenges related to a child's ability to read, write, listen, speak or do math, e.g., dyslexia or dyscalculia),
- other health impairment (covers conditions limiting a child's strength, energy, or alertness, e.g., ADHD),
- autism spectrum disorder (ASD),
- emotional disturbance (may include, e.g., anxiety, obsessive-compulsive disorder and depression),
- speech or language impairment (covers difficulties with speech or language, e.g., language problems affecting a child's ability to understand words or express herself),
- visual impairment (covers eyesight problems, including partial sight and blindness),
- deafness (covers instances where a child cannot hear most or all sounds, even with a hearing aid),
- hearing impairment (refers to a hearing loss not covered by the definition of deafness),
- deaf-blindness (covers children suffering from both severe hearing and vision loss),
- orthopaedic impairment (covers instances when a child has problems with bodily function or ability, as in the case of cerebral palsy),
- intellectual disability (covers below-average intellectual ability),
- traumatic brain injury (covers brain injuries caused by accidents or other kinds of physical force),
- multiple disabilities (children with more than one condition covered by the IDEA criteria).

The above-listed criteria are not to be conceived as exhaustive or as clear-cut definitions of what constitutes special educational needs, but should rather be seen as guidance tools in the search for and screening of relevant studies. We acknowledge that existing attempts to define special educational needs, as discussed in Vehmas (2010) and Wilson (2002), are characterised by a lack of clarity, which requires us to be transparent as to our own use of the term throughout the review process.

- For more information on the IDEA Act disability categories, go to: https://sites.ed.gov/idea/regs/b/a/300.8 (the U.S. Department of Education's Individuals with Disabilities Education Act (IDEA) website).

3.1.3 | Types of interventions

The intervention in this review is a reduction in special education class size. The more precise a class size is measured, the more reliable the findings of a study will be. Studies only considering the average class size measured as student–teacher ratio within a school (or at higher levels) will not be eligible. Neither will studies where the intervention is the assignment of an extra teacher (or teaching assistants or other adults) to a class be eligible. The assignment of additional teachers (or teaching assistants or other adults) to a classroom is not the same as reducing the size of the class, and this review focuses exclusively on the effects of reducing class size. We acknowledge that class size can change per subject or eventually vary during the day. The precision of the class size measure will be recorded.

Special education refers to settings where children with special educational needs are taught in classes segregated from general education students. These classes may be composed of children with similar special educational needs (such as classes specifically for children with ASD) or they may consist of mixed groups of children with diverse special educational needs. In such settings, the

instructional environment is adjusted to accommodate the specific needs of the student group. In the present review, *special education* may thus be defined as any given group composition consisting of only children with special educational needs. In some studies, *special education* may also be referred to as, for example, *segregated placement* or *resource room*. Special education may be full time or part time (e.g., in the form of resource rooms attended by students with special educational needs for parts of the day). We will include studies of all kinds of special education.

3.1.4 | Types of outcome measures

For quantitative studies, only valid and reliable outcomes that have been standardised on a different population (and are "objective", that is, not "experimenter-designed") will be included. If it is not clear from the description of outcome measures in the studies whether they are standardised, we will use electronic sources to determine whether a measure is standardised or not. We will not consider measures where researchers have picked a subset of questions from a standardised measure. We will extract the following outcomes:

Academic achievement

Academic achievement outcomes include reading and mathematics as well as measures of other academic subjects and global academic performance. Outcome measures must be standardised measures as in the case of standardised literacy and numeracy tests (testing areas such as reading, writing, mathematical problem-solving, and numeracy). Standardised tests in other academic subjects (e.g., science or second language) will also be included.

The following measures are examples of academic performance tests which may be included in the review:

- Woodcock-Johnson III Tests of Achievement (Mather et al., 2001)
- Stanford Achievement Test (SAT) (The Psychological Corporation, 1990)
- Grade Point Average.

Socioemotional development and adjustment

Socioemotional development and adjustment outcomes refer to validated measures of children's psychological, emotional, and social adjustment, as well as mental health. Examples of relevant measures which may be included are:

- The Strengths and Difficulties Questionnaire (SDQ) (Goodman, 2001)
- The Child Behaviour Checklist (CBCL) (Achenbach & Ruffle, 2000)
- The development and well-being assessment (DAWBA) (Goodman et al., 2000).

Well-being

Well-being refers to measures of children's subjective quality of life, pleasant emotions, happiness, and low levels of stress and negative moods. Examples of relevant measures which may be included in the review are:

- The Perceived Competence Scale for Children (Harter, 1982)
- The Loneliness Scale (Asher et al., 1984)
- The Kidscreen questionnaires (Europe, 2006)
- The Self-Esteem Index (Brown & Alexander, 1991).

3.1.5 | Primary outcomes

Academic achievement, socioemotional development and adjustment, and well-being are primary outcomes.

3.1.6 | Secondary outcomes

In addition to the primary outcomes, we will consider school completion rates as a secondary outcome. Furthermore, we will include validated measures of student classroom behaviour, such as structured observations of student engagement, on-task behaviour, and disruptive behaviour. Examples of relevant measures which may be included in the review are:

- The Code for Instructional Structure and Student Academic Response (CISSAR) (Greenwood et al., 1978)
- The Classroom Environment Scale (CES) (Fisher & Fraser, 1983; Moos & Trickett, 1987; Moos, 1979).

Studies will only be included if they consider at least one of the primary or secondary outcomes.

Duration of follow-up

Follow-up at any given point in time will be included if meaningful based on the objectives for the review. This means that if possible, we will include post-intervention outcomes measured during and after placement in a small class and follow-up data regarding both our primary and secondary outcomes throughout the children's life course. Separate meta-analyses will be carried out by grouping included time points in meaningful intervals such as follow-up less than a year, 1–2-year follow-up, and more than 2-year follow-up.

Qualitative outcomes

For the qualitative analysis, we are interested in exploring the experiences of children, teachers, and parents with special education class sizes, as they present themselves in, for example, in-depth qualitative interviews or participant observations. Relevant data may,

for example, stem from interviews with teachers on their perceptions of children's academic achievement and well-being in small versus large special education classes, or their experiences with ensuring student engagement and attention under different class sizes. We will not define a list of outcomes in advance, but remain open to what presents itself as important to children, teachers, and parents concerning special education class sizes.

Types of settings

In this review, we will include studies of children with special educational needs placed in any special education setting. We will exclude children in home- or preschool as well as children placed in treatment facilities.

3.2 | Search methods for identification of studies

Relevant quantitative and qualitative studies will be identified through searches in electronic databases, governmental and grey literature repositories, hand searches in specific targeted journals, citation tracking, contact to international experts, and Internet search engines. Furthermore, we will search for published and unpublished literature in Danish, Swedish and Norwegian. The search strategy is characterised by an overall focus on sensitivity.

Locating qualitative research may present the reviewer with particular challenges since existing search strategies have largely been developed for and applied to the quantitative literature (Frandsen et al., 2016). As of yet, not all databases have implemented rich qualitative vocabularies or specific structures tailored to accommodate qualitative literature searches. Furthermore, screening on title and abstract may prove challenging since titles and abstracts in qualitative studies are sometimes more focused on content than issues of methodology (Frandsen et al., 2016).

Attempts have been made to develop tools specifically designed for qualitative literature searches as an answer to the perceived difficulties in using such existing tools as the PICO(s) framework (**P**opulation, **I**ntervention, **C**omparison (or control), **O**utcome, and **S**tudy design and type). Cooke et al. (2012), for example, present the SPIDER search strategy which attempts to adapt the PICO components to make them more suitable for qualitative research. The SPIDER strategy contains the following components: **S**ample, **P**henomenon of **I**nterest, **D**esign, **E**valuation, and **R**esearch type. In the study by Cooke, Smith, and Booth, two systematic searches are performed, using first the PICO framework and then the SPIDER tool. The results show that the PICO search strategy generates a large number of hits, while the SPIDER tool leads to fewer hits, with the potential advantage of greater specificity. This means that the SPIDER tool may be more precise and easier to manage in terms of the amount of references for screening, however carrying the risk of missing studies.

In this review, we will apply elements of the PICO(s) framework to search for both quantitative and qualitative studies. This will be done by adding quantitative and qualitative methodological terms in the search string, as well as by carefully looking for both types of studies in our grey literature and hand searches, and so forth. By choosing this strategy, we prioritise the breadth and comprehensiveness of our search (sensitivity) which seems the most appropriate choice given the anticipated low number of studies exploring class size effects particular to special education. Should we manage to locate a large amount of studies for screening, we have the necessary resources to perform this task within the review team.

3.2.1 | Electronic searches

The following bibliographical databases will be searched:

- ERIC (EBSCO)
- Academic Search Premier (EBSCO)
- EconLit (EBSCO)
- PsycINFO (EBSCO)
- SocIndex (EBSCO)
- International Bibliography of the Social Sciences (ProQuest)
- Sociological Abstracts (ProQuest)
- Science Citation Index Expanded (Web Of Science)
- Social Sciences Citation Index (Web Of Science)

Description of search string

The search string is based on the PICO(s)-model, and contains three concepts, of which we have developed three corresponding search facets: population, intervention, and study type/methodology. The search string includes searches in title, abstract and subject terms for each facet. To increase sensitivity of the search, we also searched in full text for the intervention terms. The subject terms in the facets will be selected according to the thesaurus on each database.

Example of a search string

The search string below from the ERIC database exemplifies the search as it will be performed. The search string follows this structure:

- Search 1–4 covers the population
- Search 5–9 covers the intervention
- Search 10–16 covers the study type/methodology terms.
- Search 17 combines the three aspects

Search	Terms
S17	S4 AND S9 AND S16
S16	S10 OR S11 OR S12 OR S13 OR S14 OR S15
S15	DE ("Qualitative Research" OR "Ethnography" OR "Case Studies" OR "Evaluation Methods" OR "Field Studies" OR "Focus Groups" OR "Interviews" OR "Mixed Methods Research" OR "Naturalistic Observation" OR "Participant Observation" OR "Classroom Observation Techniques" OR "Observation" OR "Action Research")
S14	AB (qualitative* OR ethnograp* OR "case stud*" OR evaluation* OR "focus group*" OR interview* OR "mixed method*" OR observation*)
S13	TI (qualitative* OR ethnograp* OR "case stud*" OR evaluation* OR "focus group*" OR interview* OR "mixed method*" OR observation*)
S12	DE ("Effect Size" OR "Control Groups" OR "Experimental Groups" OR "Experiments" OR "Matched Groups" OR "Quasiexperimental Design" OR "Randomized Controlled Trials" OR "Comparative Testing" OR "Intervention")
S11	AB (effect* OR trial* OR experiment* OR "control group*" OR random* OR impact* OR compar* OR difference*)
S10	TI (effect* OR trial* OR experiment* OR "control group*" OR random* OR impact* OR compar* OR difference*)
S9	S5 OR S6 OR S7 OR S8
S8	DE ("Class Size" OR "Small Classes" OR "Teacher Student Ratio")
S7	TX (group* OR class*) N5 (size*)
S6	AB (group* OR class*) AND AB (size* OR ratio*)
S5	TI (group* OR class*) AND TI (size* OR ratio*)
S4	S1 OR S2 OR S3
S3	DE ("Special Needs Students" OR "Special Schools" OR "Residential Schools" OR "Educationally Disadvantaged" OR "Developmental Delays" OR "Students with Disabilities" OR "Special Classes" OR "Special Education" OR "Self Contained Classrooms" OR "Resource Room")
S2	AB (special*) AND AB (need* OR education OR child* OR student* OR pupil*)
S1	TI (special*) AND TI (need* OR education OR child* OR student* OR pupil*)

Limitations of the search string

We will not implement any restrictions to our searches based on, for example, publication date or language.

3.2.2 | Searching other resources

Hand search

We will implement hand searches in key journals in order to identify references that were poorly indexed in the bibliographical databases, as well as covering references that was published, but not yet indexed in the bibliographical databases during the search process. We will hand search the individual table of contents of the respective issues of the journals going back to 01/01/2015.

Our selection of journals to hand search is based on the frequency of the journals identified in our pilot searches during the design phase of the search string. Journals with the highest frequency in the pilot searches were selected for hand search:

- *Behavioral Disorders*
- *Journal of Autism & Developmental Disorders*
- *Exceptional Children*
- *Learning Disability Quarterly*
- *International Journal of Disability, Development & Education*
- *Remedial and Special Education*
- *Journal of Speech, Language, and Hearing Research*
- *British Journal of Special Education*
- *Learning Disabilities Research & Practice*
- *Journal of Intellectual Disability Research*

Searches for unpublished literature

Most of the resources searched for unpublished literature contains multiple types of unpublished literature. For the sake of transparency, we have divided the resources into categories based on the type of literature we expect to be most prevalent in the resource. Further resources might be added during the search process. A final list of resources will be included in the appendix of the review.

Searches for dissertations and theses in English

- ProQuest Dissertations & Theses Global (ProQuest)
- EBSCO Open Dissertations (EBSCO-host)

Searches for working papers and conference proceedings in English

- Open Grey—http://www.opengrey.eu/
- Google Scholar—https://scholar.google.com/
- Social Science Research Network—https://www.ssrn.com/index.cfm/en/
- OECD iLibrary—https://www.oecd-ilibrary.org/
- NBER working paper series—http://www.nber.org
- European Educational Research Association (EERA)—https://eera-ecer.de/
- American Educational Research Association (AERA)—https://www.aera.net/

Search for Reports and ongoing studies in English.

- Google searches—https://www.google.com/
- Best Evidence Encyclopaedia—http://www.bestevidence.org/
- Social Care Online—https://www.scie-socialcareonline.org.uk/

Searches for dissertations, theses, working papers and conference proceedings in other languages.

- Forskning.ku—Academic publications from the university of Copenhagen—https://forskning.ku.dk/soeg/
- AAU Publications—Academic publications from the University of Aarhus—https://pure.au.dk/portal/da/organisations/8000/publications.html
- SwePub—Academic publications at Swedish universities—http://swepub.kb.se/
- NORA—Norwegian Open Research Archives—http://nora.openaccess.no/
- DIVA—Swedish Digital Scientific Archives—http://www.diva-portal.org/smash/
- Skolporten—Swedish Dissertations—https://www.skolporten.se/forskning/

Searches for reports and ongoing studies in other languages.

- CORE—research outputs from international repositories—https://core.ac.uk/
- Google searches—https://www.google.com/

Search for systematic reviews. Prior to this protocol, we developed a specific search string to identify other systematic reviews in the databases listed above. This was done simultaneously with the development of the search string described above, and the identified relevant reviews are considered in this protocol.

We will also search for further systematic reviews on the following resources:

- Campbell Journal of Systematic Reviews—https://campbellcollaboration.org/
- Cochrane Library—https://www.cochranelibrary.com/
- Centre for Reviews and Dissemination Databases—https://www.crd.york.ac.uk/CRDWeb/
- EPPI-Centre database of education research—https://eppi.ioe.ac.uk/webdatabases/Intro.aspx?ID=6

Citation-tracking and snowballing methods of systematic reviews. Systematic reviews identified during the search process will be citation tracked in order to identify additional relevant references. Furthermore, we will utilise forwards citation-tracking methods on key systematic reviews. The systematic reviews selected for citation tracking will be listed in the search reporting section of the systematic review.

Citation-tracking and snowballing methods of individual references. We will select the most recently published, and the most cited key references for citation tracking. We will select studies from the pool of included references after the title/abstract screening is finished. The number of key references we will select is subject to change, but we expect to select approximately 20 (10 recent, 10 most cited). The studies selected for citation tracking/snowballing will be listed in the search reporting section of the systematic review.

Contact to experts. If we find references to or mentions of ongoing studies in screened publications, we will contact the study authors for more information. Furthermore, we will extend our contact to other researchers if the search process points to individual experts or particular institutions with relevant content expertise.

3.3 | Data collection and analysis

The following sections are focused on the quantitative part of the review. The procedures employed for qualitative studies will be presented under the heading: *Treatment of qualitative research*.

3.3.1 | Description of methods used in primary research

An example of a study which may be included in the review is that of MAGI Educational Services (1995), who conducted an evaluation study of a change in New York City special education class regulations allowing increases in the size of self-contained special education classes from 12 to 15 students. The effects of this change were analysed through (1) a descriptive study based on information gathered from special education staff members, administrators, students, and parents, and (2) an experimental study involving direct observation of 753 elementary and secondary students and 203 teachers randomly selected from classes containing either 12 students or 15 students. The experimental study was conducted in New York City Modified Instructional Services (MIS) I classes, which were classes designed for students who required instruction in a special class, with opportunities for mainstreaming. The large majority of these students were classified as learning disabled. MIS I elementary class sizes were designated as 15 students, but space constraints and other factors often limited the class size to 12 students, allowing for the experimental design, in which similar students could be compared under the 12 and 15 student class size options. Observations were performed using two standardised instruments: The Code for Instructional Structure and Student Academic Response (CISSAR) and The Instructional Environment System (TIES II). In the experimental study, one key finding was evident across all analyses: at the elementary level, a larger class size was associated with less time spent on student academic behaviours and more time spent on acting out behaviours such as disruption and inappropriate talking.

Selection of studies

Under the supervision of review authors, two review team assistants will first independently screen titles and abstracts to exclude studies that are clearly irrelevant. Studies considered eligible by at least one assistant or studies where there is insufficient information in the title and abstract to judge eligibility, will be retrieved in full text. The full

texts will then be screened independently by two review team assistants under the supervision of the review authors. Any disagreement of eligibility will be resolved by the review authors. Exclusion of studies that otherwise might be expected to be eligible will be documented and presented in an appendix.

The study inclusion criteria will be piloted by the review authors (see Appendix A). The overall search and screening process will be illustrated in a flow diagram. None of the review authors will be blind to the authors, institutions, or the journals responsible for the publication of the articles.

Data extraction and management

Two review authors will independently code and extract data from included studies. A coding sheet will be piloted on several studies and revised as necessary (see Appendix B). Disagreements will be resolved by consulting a third review author with extensive content and methods expertise. Disagreements resolved by a third reviewer will be reported. Data and information will be extracted on: available characteristics of participants, intervention characteristics and control conditions, research design, sample size, risk of bias and potential confounding factors, outcomes, and results. Extracted data will be stored electronically.

Assessment of risk of bias in included studies

We will assess the risk of bias in randomised studies using Cochrane's revised risk of bias tool, ROB 2 (Higgins et al., 2019).

The tool is structured into five domains, each with a set of signalling questions to be answered for a specific outcome. The five domains cover all types of bias that can affect the results of randomised trials.

The five domains for individually randomised trials are:

(1) bias arising from the randomisation process;
(2) bias due to deviations from intended interventions (separate signalling questions for the effect of assignment and adherence to intervention);
(3) bias due to missing outcome data;
(4) bias in measurement of the outcome;
(5) bias in selection of the reported result.

If we include cluster-randomised trials, an additional domain is included: (1b) Bias arising from identification or recruitment of individual participants within clusters. We will use the latest template for completion (currently it is the version of 15 March 2019 for individually randomised parallel-group trials and 20 October 2016 for cluster randomised parallel-group trials). In the cluster randomised template, however, only the risk of bias due to deviation from the intended intervention (effect of assignment to intervention; intention to treat ITT) is present and the signalling question concerning the appropriateness of the analysis used to estimate the effect is missing. Therefore, for cluster randomised trials, we will only use the signalling questions concerning the bias arising from identification or recruitment of individual participants within clusters from the template for cluster randomised parallel-group trials; otherwise we will use the template and signalling questions for individually randomised parallel-group trials.

We will assess the risk of bias in nonrandomised studies using the model ROBINS-I, developed by members of the Cochrane Bias Methods Group and the Cochrane Non-Randomised Studies Methods Group (Sterne, Hernán, et al., 2016). We will use the latest template for completion (currently it is the version of 19 September 2016).The ROBINS-I tool is based on the Cochrane RoB tool for randomised trials, which was launched in 2008 and modified in 2011 (Higgins et al., 2011).

The ROBINS-I tool covers seven domains (each with a set of signalling questions to be answered for a specific outcome) through which bias might be introduced into nonrandomised studies:

(1) bias due to confounding;
(2) bias in selection of participants;
(3) bias in classification of interventions;
(4) bias due to deviations from intended interventions;
(5) bias due to missing outcome data;
(6) bias in measurement of the outcome;
(7) bias in selection of the reported result.

The first two domains address issues before the start of the interventions and the third domain addresses classification of the interventions themselves. The last four domains address issues after the start of interventions and there is substantial overlap for these four domains between bias in randomised studies and bias in nonrandomised studies (although signalling questions are somewhat different in several places, see Sterne, Higgins, et al., 2016 and Higgins et al., 2019).

Randomised study outcomes are rated on a "Low/Some concerns/High" scale on each domain, whereas nonrandomised study outcomes are rated on a "Low/Moderate/Serious/Critical/No Information" scale on each domain. The level "Critical" means that the study (outcome) is too problematic in this domain to provide any useful evidence on the effects of the intervention and it is excluded from the data synthesis. The same critical level of risk of bias (excluding the result from the data synthesis) is not directly present in the RoB 2 tool, according to the guidance to the tool (Higgins et al., 2019).

In the case of an RCT where there is evidence that the randomisation has gone wrong or is no longer valid, we will assess the risk of bias of the outcome measures using ROBINS-I instead of ROB 2. Examples of reasons for assessing RCTs using the ROBINS-I tool may include studies showing large and systematic differences between treatment conditions while not explaining the randomisation procedure adequately, suggesting that there was a problem with the randomisation process; studies with large-scale differential attrition between conditions in the sample used to estimate the effects; or studies selectively reporting results for some part of the sample or for only some of the measured outcomes. In such cases, differences between the treatment and

control conditions are likely systematically related to other factors than the intervention, and the random assignment is, on its own, unlikely to produce unbiased estimates of the intervention effects. Therefore, as ROBINS-I allows for an assessment of for example confounding, we believe it is more appropriate to assess effect sizes from studies with a compromised randomisation using ROBINS-I rather than ROB 2. We will report this decision as part of the risk of bias assessment of the outcome measure in question. As other effect sizes assessed with ROBINS-I, these effect sizes may receive a "Critical" rating, leading them to be excluded from the data synthesis.

We will stop the assessment of a nonrandomised study outcome as soon as one domain in the ROBINS-I is judged as "Critical". "Serious" risk of bias in multiple domains in the ROBINS-I assessment tool may lead to a decision of an overall judgement of "Critical" risk of bias for that outcome and it will be excluded from the data synthesis.

3.3.2 | Confounding

An important part of the risk of bias assessment of nonrandomised studies is consideration of how the studies deal with confounding factors. Systematic baseline differences between groups can compromise comparability between groups. Baseline differences can be observable (e.g., age and gender) and unobservable (to the researcher; e.g., children's motivation and "ability"). There is no single nonrandomised study design that always solves the selection problem. Different designs represent different approaches to dealing with selection problems under different assumptions, and consequently require different types of data. There can be particularly great variations in how different designs deal with selection on unobservables. The "adequate" method depends on the model generating participation, that is, assumptions about the nature of the process by which participants are selected into a programme.

As there is no universally correct way to construct counterfactuals for nonrandomised designs, we will look for evidence that identification is achieved, and that the authors of the primary studies justify their choice of method in a convincing manner by discussing the assumptions leading to identification (the assumptions that make it possible to identify the counterfactual). Preferably, the authors should make an effort to justify their choice of method and convince the reader that the special needs students exposed to different class sizes are comparable.

In addition to unobservables, we have identified the following observable confounding factors to be most relevant: performance at baseline, age of the child (chronological age and/or developmental age, if reported), category of special educational need and impairment level, and socioeconomic background. In each study, we will assess whether these factors have been considered, and in addition we will assess other factors likely to be a source of confounding within the individual included studies.

3.3.3 | Importance of prespecified confounding factors

The motivation for focusing on performance at baseline, age of the child, category of special educational need and impairment level, and socioeconomic background is outlined below.

Performance at baseline is a highly relevant confounding factor to consider, since students with special educational needs constitute a highly diverse population. There may be large achievement differences between children in special education classes, even when the children are of equal age and enroled in similar special education classes at the same grade level. This is true both when comparing children with different special educational needs profiles and children diagnosed with similar impairments. This highlights the need for researchers to pay close attention to the risk of confounding due to achievement differences present at baseline.

The reason for including age as a prespecified confounder is that the needs of children change as they grow older. Young children are often more dependent on stimulating adult/child interactions and have higher support needs, both academically and in terms of behavioural/emotional support. Therefore, to be sure that an effect estimate is a result from a comparison of groups with no systematic baseline differences it is important to control for the students' age. It will be important here to both consider chronological age and developmental age, if this is reported.

As can be seen in the definition of special educational needs, the categories cover a very broad range of disabilities and impairment levels. It is possible that special education students with some diagnoses or degrees of impairment require, for example, an increased need for individual support and close adult-child interaction, or they may have an inability to cope in larger groups of children due to difficulties in sensory processing. Therefore, the special needs category and impairment level are important confounding variables.

Finally, a large body of research documents the impact of parental socioeconomic background on almost all aspects of children's development (e.g., Renninger et al., 2006), which is why we find it to be common place to include this as a potential confounding factor.

3.3.4 | Effect of primary interest and important co-interventions

We are mainly interested in the effect of actually participating in the intervention (in this case, receiving instruction in a smaller as opposed to a larger special education class), that is, the treatment on the treated effect (TOT). The risk of bias assessments will therefore be in relation to this specific effect. The risk of bias assessments of both randomised trials and nonrandomised studies will consider adherence and differences in additional interventions ("co-interventions") between intervention groups. Important cointerventions we will consider are other types of classroom support available to children with special educational needs, for example, software packages for children suffering from dyslexia. Furthermore,

additional teachers or teacher aides in a classroom will be considered an important co-intervention.

3.3.5 | Assessment

At least two review authors will independently assess the risk of bias for each relevant outcome from the included studies. Any disagreements will be resolved by a third reviewer with content and statistical expertise and will be reported. We will report the risk of bias assessment in risk of bias tables for each included study outcome in the completed review.

3.3.6 | Measures of treatment effect

Continuous outcomes

For continuous outcomes, effects sizes with 95% confidence intervals will be calculated, where means and standard deviations are available. If means and standard deviations are not available, we will calculate standardised mean differences (SMDs) from *F*-ratios, *t*-values, χ^2 values, and correlation coefficients, where available, using the methods suggested by Lipsey and Wilson (2001). If not enough information is yielded, the review authors will request this information from the principal investigators. Hedges' *g* will be used for estimating SMD. Standardised measures of student academic achievement (e.g., reading and math), are examples of relevant continuous outcomes in this review.

Dichotomous outcomes

For dichotomous outcomes, we will calculate odds ratios with 95% confidence intervals. Children who pass or fail an exam is an example of a relevant dichotomous outcome in this review. There are statistical approaches available to re-express dichotomous and continuous data to be pooled together (Sanchez-Meca et al., 2003). In order to calculate common metric odds ratios will be converted to SMD effect sizes using the Cox transformation. We will only transform dichotomous effect sizes to SMD's if appropriate, as may be the case with, for example, outcomes for behaviour problems or psychosocial adjustment, which can be measured with binary data based on clinical cut-offs or with continuous data.

When effect sizes cannot be pooled, study-level effects will be reported in as much detail as possible. Software for storing data and statistical analyses will be RevMan 5.0, Excel, R, and STATA Version 16.

Unit of analysis issues

We will take into account the unit of analysis of the studies to determine whether individuals were randomised in groups (i.e., cluster-randomised trials), whether individuals may have undergone multiple interventions, whether there were multiple treatment groups, and whether several studies are based on the same data source.

Cluster-randomised trials. Errors in statistical analysis can occur when the unit of allocation differs from the unit of analysis. In cluster randomised trials, participants are randomised to treatment and control groups in clusters, either when data from multiple participants in a setting are included (creating a cluster within the school or community setting), or when participants are randomised by treatment locality or school. Nonrandomised studies may also include clustered assignment of treatment. Effect sizes and standard errors from such studies may be biased if the unit-of-analysis is the individual and an appropriate cluster adjustment is not used (Higgins & Green, 2011).

If possible, we will adjust effect sizes individually using the methods suggested by L. V. Hedges (2007a) and information about the intra-cluster correlation coefficient (ICC), realised cluster sizes, and/or estimates of the within and between variances of clusters. If it is not possible to obtain this information, we will adjust effect sizes using estimates from the literature of the ICC (e.g., L. V. Hedges & Hedberg, 2007), and assume equal cluster sizes. To calculate an average cluster size, we will divide the total sample size in a study by the number of clusters (typically the number of classrooms or schools).

Multiple intervention groups and multiple interventions per individual. Studies with multiple intervention groups with different individuals will be included in this review, although only intervention and control groups that meet the eligibility criteria will be used in the data synthesis. To avoid problems with dependence between effect sizes, we will apply robust standard errors (L. V. Hedges et al., 2010) and use the small sample adjustment to the estimator itself (Tipton, 2015). We will use the results in Tanner-Smith and Tipton (2014) and Tipton (2015) to evaluate if there are enough studies for this method to consistently estimate the standard errors. See "Data Synthesis" section for more details about the data synthesis.

If there is an insufficient number of studies, we will use a synthetic effect size (the average) in order to avoid dependence between effect sizes. This method provides an unbiased estimate of the mean effect size parameter but overestimates the standard error. Random-effects models applied when synthetic effect sizes are involved actually perform better in terms of standard errors than do fixed effects models (L. V. Hedges, 2007b). However, tests of heterogeneity when synthetic effect sizes are included are rejected less often than nominal.

If pooling is not appropriate (e.g., if multiple interventions and/or control groups include the same individuals), only one intervention group will be coded and compared to the control group to avoid overlapping samples. The choice of which estimate to include will be based on our risk of bias assessment. We will choose the estimate that we judge to have the least risk of bias (primarily, confounding bias and in case of equal scoring, the missing outcome data domain will be used).

Multiple studies using the same sample of data. In some cases, several studies may have used the same sample of data or some studies may

have used only a subset of a sample used in another study. We will review all such studies, but in the meta-analysis, we will only include one estimate of the effect from each sample of data. This will be done to avoid dependencies between the "observations" (i.e., the estimates of the effect) in the meta-analysis. The choice of which estimate to include will be based on our risk of bias assessment of the studies. We will choose the estimate from the study that we judge to have the least risk of bias (primarily confounding bias). If two (or more) studies are judged to have the same risk of bias and one of the studies (or more) uses a subset of a sample used in another study (or studies), we will include the study using the full set of participants.

Multiple time points. When the results are measured at multiple time points, each outcome at each time point will be analysed in a separate meta-analysis with other comparable studies taking measurements at a similar time point. As a general guideline, these will be grouped together as follows: follow-up less than a year, 1–2-year follow-up, and more than 2-year follow-up. However, should the studies provide viable reasons for an adjusted choice of relevant and meaningful duration intervals for the analysis of outcomes, we will adjust the grouping.

Dealing with missing data

Missing data in the individual studies will be assessed using the risk of bias tool. Studies must permit calculation of a numeric effect size for the outcomes to be eligible for inclusion in the meta-analysis. Where studies have missing summary data, such as missing standard deviations, we will derive these where possible from, for example, *F*-ratios, *t*-values, χ^2 values, and correlation coefficients using the methods suggested by Lipsey and Wilson (2001). If these statistics are also missing, the review authors will request information from the study investigators.

If missing summary data necessary for the calculation of effect sizes cannot be derived or retrieved, the study results will be reported in as much detail as possible, that is, the study will be included in the review but excluded from the meta-analysis.

Assessment of heterogeneity

Heterogeneity among primary outcome studies will be assessed with χ^2 (Q) test, and the I^2, and τ^2 statistics (Higgins et al., 2003). Any interpretation of the χ^2 test will be made cautiously on account of its low statistical power.

Assessment of reporting biases

Reporting bias refers to both publication bias and selective reporting of outcome data and results. Here, we state how we will assess publication bias. We will use funnel plots for information about possible publication bias if we find a sufficient number of studies (Higgins & Green, 2011). However, asymmetric funnel plots are not necessarily caused by publication bias (and publication bias does not necessarily cause asymmetry in a funnel plot). In general, asymmetry is a sign of small-study effects, of which there can be many causes beside publication bias (Sterne et al., 2005).

Instead of trying to interpret the funnel plots as direct evidence of publication bias, or the lack thereof, we will perform sensitivity analyses for publication bias in meta-analyses as suggested by Mathur and VanderWeele (2020). This method gives a value of how large ratios of publication probabilities (that is the likelihood of affirmative results to be published relative to nonaffirmative results) would have to be to alter the results and therefore indicate how robust the meta-analysis is to publication bias.

Data synthesis

The proposed quantitative data synthesis will follow standard procedures for conducting systematic reviews using meta-analytic techniques.

The overall data synthesis will be conducted where effect sizes are available or can be calculated, and where studies are similar in terms of the outcome measured. Meta-analysis of outcomes will be conducted on each metric separately (as outlined in Section 3.1.4).

As different computational methods may produce effect sizes that are not comparable, we will be transparent about all methods used in the primary studies (research design and analytical strategies) and use caution when synthesising effect sizes. Special caution will be taken concerning studies using regression discontinuity (RD) to estimate a local average treatment effect (LATE) (Angrist & Pischke, 2009). Such studies will be included, but may be subject to a separate analysis depending on the comparability between the LATE's and the effects from other studies. We will check the sensitivity of our results to the inclusion of RD studies. In addition, we will discuss the limitation in generalisation of the results obtained from these types of studies.

When the effect sizes used in the data synthesis are odds ratios, they will be log transformed before being analysed. The reason for this is that ratio summary statistics all have the common feature that the lowest value they can take is 0, that the value 1 corresponds with no intervention effect, and the highest value an odds ratio can ever take is infinity. This scale is not symmetric. The log transformation makes the scale symmetric: the log of 0 is minus infinity, the log of 1 is zero, and the log of infinity is infinity.

Studies that have been coded with a Critical risk of bias will not be included in the data synthesis.

As the intervention deals with diverse populations of participants, and we expect heterogeneity among primary study outcomes, all analyses of the overall effect will be inverse-variance weighted using random-effects statistical models that incorporate both the sampling variance and between-study variance components into the study-level weights. Random-effects weighted mean effect sizes will be calculated using 95% confidence intervals, and we will provide a graphical display (forest plot) of effect sizes. Graphical displays for meta-analysis performed on ratio scales sometimes use a log scale, as the confidence intervals then appear symmetric. This is however not the case for the software Revman 5 which we plan to use in this review. The graphical displays using odds ratios and the mean effect size will be reported as an odds ratio. Heterogeneity among primary outcome studies will be assessed with χ^2 (Q) test, and the I^2, and τ^2

statistics (Higgins et al., 2003). Any interpretation of the χ^2 test will be made cautiously on account of its low statistical power.

For subsequent analyses of moderator variables that may contribute to systematic variations, we will use the mixed-effects regression model. This model is appropriate if a predictor explaining some between-studies variation is available but there is a need to account for the remaining uncertainty (L. W. Hedges & Pigott, 2004; Konstantopoulos, 2006).

Several studies may have used the same sample of data. We will review all such studies, but in the meta-analysis, we will only include one estimate of the effect from each sample of data. This will be done to avoid dependencies between the "observations" (i.e., the estimates of the effect) in the meta-analysis. The choice of which estimate to include will be based on our quality assessment of the studies. We will choose the estimate from the study that we judge to have the least risk of bias, with particular attention paid to confounding bias.

Studies may provide results separated by for example age and/or gender. We will include results for all age and gender groups. To take into account the dependence between such multiple effect sizes from the same study, we will apply the robust variance estimation (RVE) approach (L. V. Hedges et al., 2010). An important feature of this analysis is that the results are valid regardless of the weights used. For efficiency purposes, we will calculate the weights using a method proposed by L. V. Hedges et al. (2010). This method assumes a simple random-effects model in which study average effect sizes vary across studies (τ^2), and the effect sizes within each study are equicorrelated (ρ). The method is approximately efficient, since it uses approximate inverse-variance weights: they are approximate given that ρ is, in fact, unknown, and the correlation structure may be more complex. We will calculate weights using estimates of τ^2, setting $\rho = 0.80$ and conduct sensitivity tests using a variety of ρ values to assess if the general results and estimates of the heterogeneity are robust to the choice of ρ. We will use the small sample adjustment to the residuals used in RVE as proposed by Bell and McCaffrey (2002) and extended by McCaffrey et al. (2001) and by Tipton (2015). We will use the Satterthwaite degrees of freedom (Satterthwaite, 1946) for tests as proposed by Bell and McCaffrey (2002) and extended by Tipton (2015). We will use the guidelines provided in Tanner-Smith and Tipton (2014) to evaluate if there are enough studies for this method to consistently estimate the standard errors.

If there is an insufficient number of studies to use RVE, we will conduct a data synthesis using a synthetic effect size (the average) in order to avoid dependence between effect sizes.

If we apply robust variance estimation, the analysis will be conducted in STATA or R, as robust variance estimation is not implemented in Revman 5.

Subgroup analysis and investigation of heterogeneity

We will investigate the following factors with the aim of explaining potential observed heterogeneity: study-level summaries of participant characteristics (e.g., age group or studies where separate effects for low/high socioeconomic status are available) and different types of special education as well as intervention intensity (size of class size reduction and initial class size).

If the number of included studies is sufficient and given that there is variation in the covariates, we will perform moderator analyses (multiple meta-regression using the mixed model) to explore how observed variables are related to heterogeneity.

If there is a sufficient number of studies, we will apply the RVE approach and use approximately inverse-variance weights calculated using a method proposed by L. V. Hedges et al. (2010). This technique calculates standard errors using an empirical estimate of the variance: it does not require any assumptions regarding the distribution of the effect size estimates. The assumptions that are required to meet the regularity conditions are minimal and generally met in practice. This more robust technique is beneficial because it takes into account the possible correlation between effect sizes separated by the covariates within the same study and allows all of the effect size estimates to be included in meta-regression. We will calculate weights using estimates of τ^2, setting $\rho = 0.80$ and conduct sensitivity tests using a variety of ρ values; to assess if the general results are robust to the choice of ρ. We will use the small sample adjustment to the residuals used in RVE and the Satterthwaite degrees of freedom (Satterthwaite, 1946) for tests (Tipton, 2015). The results in Tipton (2015) suggest that the degrees of freedom depend not only on the number of studies but also on the type of covariates included in the meta-regression. The degrees of freedom can be small, even when the number of studies is large, if a covariate is highly unbalanced or a covariate with very high leverage is included. The degrees of freedom will vary from coefficient to coefficient. The corrections to the degrees of freedom enable us to assess when the RVE method performs well. As suggested by Tanner-Smith and Tipton (2014) and Tipton (2015), if the degrees of freedom are smaller than four, the RVE results should not be trusted.

We will report 95% confidence intervals for regression parameters. We will estimate the correlations between the covariates and consider the possibility of confounding. Conclusions from meta-regression analysis will be cautiously drawn and will not solely be based on significance tests. The magnitude of the coefficients and width of the confidence intervals will be taken into account as well. Otherwise, single factor subgroup analysis will be performed. The assessment of any difference between subgroups will be based on 95% confidence intervals. Interpretation of relationships will be cautious, as they are based on subdivision of studies and indirect comparisons.

In general, the strength of inference regarding differences in treatment effects among subgroups is controversial. However, making inferences about different effect sizes among subgroups on the basis of between-study differences entails a higher risk compared to inferences made on the basis of within-study differences (see Oxman & Guyatt, 1992). We will therefore use within-study differences where possible.

We will also consider the degree of consistency of differences, as making inferences about different effect sizes among subgroups

entails a higher risk when the differences are not consistent within the studies (Oxman & Guyatt, 1992).

Sensitivity analysis

Sensitivity analysis will be carried out by restricting the meta-analysis to a subset of all studies included in the original meta-analysis and will be used to evaluate whether the pooled effect sizes are robust across components of risk of bias. We will consider sensitivity analysis for each domain of the risk of bias checklists and restrict the analysis to studies with a low risk of bias.

Sensitivity analyses with regard to research design and analytical strategies in the primary studies are important elements of the analysis to ensure that different methods produce consistent results.

Treatment of qualitative research

As mentioned previously, we will include all types of empirical qualitative studies that collect primary data and provide descriptions of main methodological issues such as sampling, data collection procedures, and type of data analysis. If an included quantitative study contains relevant qualitative data, these will be treated in the same way as other qualitative studies and will be considered for inclusion in the qualitative synthesis.

We will use findings from qualitative studies to address and extend questions related to our effectiveness review, broadening the scope of the review to also include the lived experiences of children, teachers, and parents who spend their everyday lives in special education settings under different class size arrangements. The qualitative analysis in this review will be performed as a thematic synthesis.

Critical appraisal of qualitative studies. All qualitative studies will be independently appraised by two reviewers in order to assess whether or not they should be included in the thematic synthesis. This means that studies will be double coded, after which the two reviewers will discuss their assessments and reach a final conclusion on whether to include a given study in the synthesis. In case of disagreements that cannot be reconciled between the two reviewers, a third reviewer will assess the study and make the final decision on inclusion. We will only include studies for synthesis that pay sufficient attention to qualitative research standards for credibility, transferability, dependability, and confirmability (Hannes, 2011). We will critically appraise qualitative studies using an adapted version of the JBI Critical Appraisal Checklist for Qualitative Research, developed by the Joanna Briggs Institute (Joanna Briggs Institute, 2017; Lockwood et al., 2015). This checklist includes 10 questions that lead to an overall appraisal of "include", "exclude", or "seek further info". The 10 questions take integral parts of the qualitative methodological process into consideration, such as the congruity between the choice of research methodology and the research objectives, the influence of the researcher on the research, and the flow of conclusions from the analysis or interpretation of data. In the original checklist, the questions are checked in boxes indicating "yes", "no", "unclear" or "not applicable". In this review, reviewers will further be required to justify their choice of "yes", "no", "unclear" or "not applicable" in a comment box. This is done by importing the checklist into EPPI-Reviewer 4 and adding comment boxes. Reviewers will also be required to justify their overall appraisal assessment. The reason for demanding justifications in addition to ticking the boxes is founded in a wish to both ensure high methodological rigour and detail in the assessment and facilitate discussion between reviewers on whether to include or exclude studies for synthesis. All critical appraisals of qualitative studies will be performed in EPPI-Reviewer, where comparisons can also be made between reviewer assessments.

Data extraction. We will extract data for the thematic synthesis using a data extraction form that will be developed by the research team and imported into EPPI-Reviewer 4. The information extracted will concern the study context and participants, the design and methods used, as well as the research findings. As discussed by Thomas and Harden (2008) and Noyes & Lewin, (2011a, 2011b), determining what constitutes "findings" in qualitative studies is not always straight forward. It is for the researcher to decide on clear criteria for what is to be considered "a finding". In Thomas and Harden (2008), a choice is made to take findings to be all text labelled as *"Results"* or *"Findings"* and to import all such text into a qualitative analysis software package. In the current review, we will define "findings" in a similar way to Thomas and Harden (2008) and draw on the functionalities developed in EPPI-Reviewer 4 for inductive coding of textual data.

Thematic synthesis. Given a sufficient amount of included qualitative studies, we will conduct a thematic synthesis following the procedures presented in Thomas and Harden (2008) and applied in other Campbell systematic reviews such as that of Snilstveit et al., 2019). A thematic synthesis has three stages, which are interwoven and to an extent overlapping. In the first stage, research findings are subjected to free inductive line-by-line coding, informed by usual guidelines for thematic analysis in primary qualitative research. In this process, every sentence is applied with one or more codes, and with each new study, reviewers can draw on already existing codes, or add new ones, leading to the production of a "code bank" and the beginning of a translation of concepts between studies (Thomas & Harden, 2008). The inductive coding will be performed by two review authors in the following manner: Both authors individually code all eligible studies using the line-by-line coding functionality in EPPI-Reviewer 4. In cases where authors have used similar codes, a common wording may be chosen, and a united code bank is produced which includes the total amount of inductive codes generated by the review authors. Before completing this stage of the synthesis, the authors will examine all text supplied with a given code in order to check for coding consistency and to add additional codes if needed (Thomas & Harden, 2008).

Examples of the inductive coding of studies will be provided to enhance analytical transparency.

In stage two of the thematic synthesis, the review authors will group the inductive codes into related areas in order to construct

descriptive themes, staying close to the primary data. This will be done first individually by each of the two review authors, and then in unison. This same procedure will be used in the final stage, where the descriptive themes are translated into higher-order analytical themes that go beyond the primary data, allowing for the generation of new understandings and hypotheses (Thomas & Harden, 2008). The goal here will be to generate analytical themes that will help us gain insight into children, teacher, and parent perspectives on class size in special education. How do children, for example, describe their experiences with being placed in small class units? How do teachers assess their opportunities for helping children thrive and learn under different class sizes? What are the perspectives of parents as to what constitutes optimal class sizes considering the particular difficulties of different groups of children with special educational needs? We will present and discuss the generated analytical themes, drawing on examples from the included studies, in order to present answers and new perspectives on the review questions. Tables will be provided that exemplify the flow from descriptive codes to analytical themes, such that our analytical work remains as transparent as possible.

Integrated discussion of findings from quantitative and qualitative studies. After separately completing (1) the statistical meta-analysis, and (2) the qualitative thematic synthesis, it is our aim to integrate the findings of each analysis narratively in a final discussion. With a focus on discussing and drawing up different perspectives on class size issues in special education, we will use the results of each synthesis to extend one another, adding greater depth and complexity to the final review conclusions.

Summary of findings and assessment of the certainty of the evidence
Findings of the review will be summarised and the certainty of the evidence will be assessed as outlined in the above sections.

AUTHOR CONTRIBUTIONS

Anja Bondebjerg holds a Master's degree in Sociology and has worked extensively with systematic reviews and research mappings in the fields of education and early childhood education and care. She is knowledgeable regarding school provision for children with special educational needs. In addition to this review, Anja is currently involved as a lead or coauthor on a number of other Campbell Systematic Reviews.

Nina T. Dalgaard is a psychologist, Ph.D. Nina has previously worked as both an educational psychologist within a primary school setting and as a clinical child psychologist and thus has knowledge about the socioemotional and cognitive development of children. In addition to this review, Nina is currently involved as a lead or coauthor on a number of other Campbell Systematic Reviews.

Trine Filges holds a Ph.D. in Economics and has extensive experience as a systematic reviewer and methodologist, having completed a number of systematic reviews in social welfare topic areas as well as in the field of education. Trine has published thirteen Campbell Systematic reviews, is currently the lead reviewer on three Campbell Systematic Reviews, is further involved as a reviewer in two Campbell Systematic Reviews and has published systematic and meta-analytic reviews in high-impact journals. Trine's fields of expertise are systematic review methods and statistical analysis.

Morten K. Thomsen holds an M.Sc. in Psychology and an MPhil in Social and Developmental Psychology (Cantab). Morten has previously worked as a clinical psychologist in the Danish health care sector, doing psychological assessments and attachment-focused therapy. Furthermore, Morten has experience from research positions in large-scale Danish and UK cohort studies. Morten is trained in the methodology of meta-analytic reviews, including risk of bias assessment and advanced statistical analyses. In addition to this review, Morten is currently lead-authoring two other Campbell Systematic Reviews.

Bjørn C. A. Viinholt (information specialist) has four years of experience in developing and writing systematic reviews. As a part of undertaking systematic reviews, Bjørn has experience in developing systematic search strategies and processes of reference management. Bjørn will contribute to the development of the systematic search strategy and the execution of searches as well as providing assistance with reference management and grey literature searches. Bjørn will also assist with aspects relating to systematic literature searches in Campbell review methodology.

DECLARATIONS OF INTEREST

None of the review authors have conflicts of interest related to this review.

SOURCES OF SUPPORT

Internal sources

- VIVE, The Danish Center for Social Science Research, Denmark.

REFERENCES

Other references

Additional references

Achenbach, T. M., & Ruffle, T. M. (2000). The Child Behavior Checklist and related forms for assessing behavioral/emotional problems and competencies. *Pediatrics in Review*, *21*(8), 265–271.

Ahearn, E. M. (1995). *Caseload/class size in special education: A brief analysis of state regulations*. Final Report. Alexandria, VA: National Association of State Directors of Special Education.

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

Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Asher, S. R., Hymel, S., & Renshaw, P. D. (1984). Loneliness in children. *Child Development*, *55*, 1456–1464.

Bell, R. M., & McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, *28*(2), 169–181.

Biddle, B. J., & Berliner, D. C. (2002). Small class size and its effects. *Educational Leadership*, *59*(5), 12–23.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons, Ltd.

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

Brown, L., & Alexander, J. (1991). *Self-esteem index examiner's manual*. PRO-ED.

Cooke, A., Smith, D., & Booth, A. (2012). Beyond PICO: The SPIDER tool for qualitative evidence synthesis. *Qualitative Health Research*, *22*(10), 1435–1443.

Donner, A., Piaggio, G., & Villar, J. (2001). Statistical methods for the meta-analysis of cluster randomized trials. *Statistical Methods in Medical Research*, *10*(5), 325–338.

Ehrenberg, R. G., Brewer, D. J., Gamoran, A., & Willms, J. D. (2001). Class size and student achievement. *Psychological Science and the Public Interest*, *2*(1), 1–30.

Europe, T. K. G. (2006). *The Kidscreen questionnaires: Quality of life questionnaires for children and adolescents*. Pabst Science Publishers.

Filges, T., Sonne-Schmidt, C. S., & Nielsen, B. C. V. (2018). Small class sizes for improving student achievement in primary and secondary schools. *Campbell Systematic Reviews*, *2018*(10), 1–107.

Finn, J. D. (2002). Small classes in American schools: Research, practice and politics. *Phi Delta Kappan*, *83*(7), 551–560.

Fisher, D. L., & Fraser, B. J. (1983). Validity and use of Classroom Environment Scale. *Educational Evaluation and Policy Analysis*, *5*, 261–271.

Forness, S. R., & Kavale, K. A. (1985). Effects of class size on attention, communication, and disruption of mildly mentally retarded children. *American Educational Research Journal*, *22*(3), 403–412.

Frandsen, T. F., Christensen, J. B., & Ølholm, A. M. (2016). Systematisk søgning efter kvalitativ litteratur kan styrkes. *Ugeskriftet Læger*, *178*, V06160384.

Fraser, B. J. (1998). Classroom environment instruments: development, validity and applications. *Learning Environments Research*, *1*, 7–34.

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

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*(11), 1337–1345.

Goodman, R., Ford, T., Richards, H., Gatward, R., & Meltzer, H. (2000). The development and well-being assessment: Description and initial validation of an integrated assessment of child and adolescent psychopathology. *Journal of Child Psychology and Psychiatry*, *41*(5), 645–665.

Gottlieb, J., & Alter, M. (1997). *An evaluation study of the impact of modifying instructional group sizes in resource rooms and related service groups in New York City. Final report. Revised*. Department of Teaching and Learning, School of Education, New York University.

Gough, D., Thomas, J., & Oliver, S. (2012). Clarifying differences between review designs and methods. *Systematic Reviews*, *1*, 28.

Greenwood, C. R., Delquadri, J., & Hall, R. V. (1978). *Code for instructional structure and student academic response: CISSAR*. University of Kansas, Bureau of Child Research, Juniper Gardens Children's Project.

Hannes, K. (2011). Chapter 4: Critical appraisal of qualitative research. In J. Noyes, A. Booth, K. Hannes, A. Harden, J. Harris, S. Lewin & C. Lockwood (Eds.), *Supplementary guidance for inclusion of qualitative research in Cochrane systematic reviews of interventions. Version 1 (updated August 2011)* (p. 2011). Cochrane Collaboration Qualitative Methods Group.

Harden, A., & Thomas, J. (2005). Methodological issues in combining diverse study types in systematic reviews. *International Journal of Social Research Methodology*, *8*(3), 257–271.

Harter, S. (1982). The Perceived Competence Scale for Children. *Child Development*, *53*(1), 87–97.

Heckman, J. J., & Urzua, S. (2010). Comparing IV with structural models: What simple IV can and cannot identify. *Journal of Econometrics*, *156*, 27–37.

Heckman, J. J., Urzua, S., & Vytlacil, E. (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, *88*(3), 389–432.

Hedges, L. V. (2007a). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370.

Hedges, L. V. (2007b). Meta-analysis. In C. R. Rao (Ed.), *The handbook of statistics* (pp. 919–953). Elsevier.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group randomized trials in education. *Education Evaluation and Policy Analysis*, *29*(1), 60–87.

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

Hedges, L. W., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, *9*(4), 426–445.

Higgins, J. P. T., Altman, D. G., Gotzsche, P. C., Juni, P., Moher, D., Oxman, A. D., Savovic, J., Schulz, K. F., Weeks, L., & Sterne, J. A. C. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, *343*(d5928), d5928.

Higgins, J. P. T., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions. Version 5.1.0 [updated March 2011]*. Wiley-Blackwell, The Cochrane Collaboration.

Higgins, J. P. T., Savovic, J., Page, M. J., & Sterne, J. A. C. (2019). *Revised Cochrane risk-of-bias tool for randomized trials (RoB 2): Detailed guidance*. Updated 15 March. http://www.riskofbias.info

Higgins, J. P. T., Sterne, J. A. C., Savovic, J., Page, M. J., Hrobjartsson, A., Boutron, I., Reeves, B., & Eldridge, S. (2016). A revised tool for assessing risk of bias in randomized trials. In J. Chandler, J. McKenzie, I. Boutron, V. Welch (Eds.). *Cochrane Methods. Cochrane Database of Systematic Reviews*, Issue 10 (Suppl. 1). https://doi.org/10.1002/14651858.CD201601

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*(7414), 557–560.

Joanna Briggs Institute. (2017). *JBI critical appraisal checklist for qualitative research*. https://joannabriggs.org/sites/default/files/2019-05/JBI_Critical_Appraisal_Checklist_for_Qualitative_Research2017_0.pdf

Kavale, K. A., & Forness, S. R. (2000). History, rhetoric, and reality: Analysis of the inclusion debate. *Remedial and Special Education*, *21*(5), 279–296.

Keith, P., Keith, T., Young, D., & Fortune, J. (1993). *Investigating the influences of class size and class mix on special education student outcomes: Phase One Results*. Paper presented at the Annual Meeting of the Eastern Educational Research Association. Clearwater, FL.

Konstantopoulos, S. (2006). *Fixed and mixed effects models in meta-analysis* (IZA DP no. 2198). Northwestern University & IZA Bonn.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis. Applied social research methods series* (Vol. 49).SAGE Publications.

Lockwood, C., Munn, Z., & Porritt, K. (2015). Qualitative research synthesis: Methodological guidance for systematic reviewers utilizing meta-aggregation. *International Journal of Evidence-Based Healthcare*, *13*(3), 179–187.

MAGI Educational Services Inc. (1995). Class size research bulletin. Results of a statewide research study on the effects of class size in special education. New York State Education Department.

Mather, N., Wendling, B. J., & Woodcock, R. W. (2001). *Essentials of WJ III [TM] tests of achievement assessment. Essentials of psychological assessment series*. John Wiley & Sons, Inc.

Mathur, M. B., & VanderWeele, T. J. (2020). Sensitivity analysis for publication bias in meta-analyses. *Applied Statistics*, *69*(5), 1091–1119.

McCaffrey, D. F., Bell, R. M., & Botts, C. H. (2001). *Generalizations of biased reduced linearization*. In: Proceedings of the Annual Meeting of the American Statistical Association, August 5–9.

McCrea, L. (1996). *A review of literature: Special education and class size* (ERIC Document Reproduction Services No. ED 407 387). Lansing: Michigan State Board of Education.

Minnesota Department, Children, Families & Learning. (2000). *Issues in special education caseload/class size policy. Report summary*. https://mn.gov/mnddc/past/pdf/00s/00/00-ISE-MDE.pdf

Moos, R. H., & Trickett, E. J. (1987). *Classroom Environment Scale Manual*. Consulting Psychologists Press.

Moos, R. H. (1979). *Evaluating educational environments: Procedures, measures, findings and policy implications*. Jossey-Bass.

Noyes, J., Booth, A., Cargo, M., Flemming, K., Harden, A., Harris, J., Garside, R., Hannes, K., Pantoja, T., & Thomas, J. (2019). Chapter 21: Qualitative evidence. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page & V. A. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions version 6.0 (updated July 2019)*. Cochrane.

Noyes, J., & Lewin, S. (2011a). Chapter 6: Supplemental guidance on selecting a method of qualitative evidence synthesis, and integrating qualitative evidence with Cochrane intervention reviews. In J. Noyes, A. Booth, K. Hannes, A. Harden, J. Harris, S. Lewin & C. Lockwood (Eds.), *Supplementary guidance for inclusion of qualitative research in Cochrane systematic reviews of interventions. Version 1 (updated August 2011)*. Cochrane Collaboration Qualitative Methods Group.

Noyes, J., & Lewin, S. (2011b). Chapter 5: Extracting qualitative evidence. In J. Noyes, A. Booth, K. Hannes, A. Harden, J. Harris, S. Lewin & C. Lockwood (Eds.), *Supplementary guidance for inclusion of qualitative research in Cochrane systematic reviews of interventions. Version 1 (updated August 2011)*. Cochrane Collaboration Qualitative Methods Group.

Oxman, A. D., & Guyatt, G. H. (1992). A consumers guide to subgroup analyses. *Annals of Internal Medicine*, *116*(1), 78–84.

Project FORUM. (2000). *Special education issues in caseload/class size*. National Association of State Directors of Special Education (NASDSE).

Project FORUM. (2003). *Caseload/class size in special education*. National Association of State Directors of Special Education. (NASDSE).

Renninger, A., Sigel, I. E., Damon, W., & Lerner, R. M. (2006). *Handbook of child psychology, Child psychology in practice*. John Wiley & Sons Inc.

Russ, S., Chiang, B., Rylance, B. J., & Bongers, J. (2001). Caseload in special education: An integration of research findings. *Exceptional Children*, *67*(2), 161–172.

Sanchez-Meca, J., Marin-Martinez, F., & Chacon-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, *8*(4), 448–467.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics*, *2*, 110–114.

Schanzenbach, D. W. (2007). *What have researchers learned from Project STAR? Brookings Papers on Education Policy*. Brookings Institution.

Snilstveit, B., Stevenson, J., Langer, L., Polanin, J., Shemilt, I., Eyers, J., & Ferraro, P. J. (2018). Protocol: Incentives for climate mitigation in the land use sector: A mixed-methods systematic review of the effectiveness of payment for environment services (PES) on environmental and socio-economic outcomes in low- and middle-income countries. *The Campbell Collaboration*, *14*, 1–77.

Snilstveit, B., Stevenson, J., Langer, L., Tannous, N., Ravat, Z., Nduku, P., Polanin, J., Shemilt, I., Eyers, J., & Ferraro, P. J. (2019). Incentives for climate mitigation in the land use sector—The effects of payment for environmental services on environmental and socioeconomic outcomes in low- and middle-income countries: A mixed-methods systematic review. *Campbell Systematic Reviews*, *15*, e1045.

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 75–98). John Wiley & Sons, Ltd.

Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A. W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., … Higgins, J. P. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomized studies of interventions. *BMJ*, *355*(i4919), i4919.

Sterne, J. A. C., Higgins, J. P. T., Elbers, R. G., Reeves, B. C., & The Development Group for ROBINS-I. (2016). *Risk of bias in non-randomized studies of interventions (ROBINS-I): Detailed guidance*. Updated 12 October. http://www.riskofbias.info

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods*, *5*(1), 13–30.

The Psychological Corporation. (1990). *Stanford achievement test series: Technical data report*. Harcourt Brace Jovanovich.

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, *8*, 45.

Thurlow, M. L., Ysseldyke, J. E., Wotruba, J. W., & Algozzine, B. (1993). Instruction in special education classrooms under varying student-teacher ratios. *The Elementary School Journal*, *93*(3), 305–320.

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*(3), 375–393.

Vehmas, S. (2010). Special needs: A philosophical analysis. *International Journal of Inclusive Education*, *14*(1), 87–96.

Whittemore, R., Chao, A., Jang, M., Minges, K. E., & Park, C. (2014). Methods for knowledge synthesis: An overview. *Heart & Lung: The Journal of Acute and Critical Care*, *43*(5), 453–461.

Wilson, J. (2002). Defining "special needs". *European Journal of Special Needs Education*, *17*(1), 61–66.

Ysseldyke, J. E. (1988). *Student-teacher ratios and their relationship to instruction and achievement for mildly handicapped students. Final project report*. University of Minnesota.

Ysseldyke, J. E., Christenson, S. L., McVicar, R., Bakewell, D., & Thurlow, M. L. (1986). *Instructional environment scale: Scale development and training procedures*. University of Minnesota, Instructional Alternatives Project.

Zarghami, F., & Schnellert, G. (2004). Class size reduction: No silver bullet for special education students' achievement. *International Journal of Special Education*, *19*(1), 89–96.

How to cite this article: Bondebjerg, A., Dalgaard, N. T., Filges, T., Thomsen, M. K., & Viinholt, B. C. A. PROTOCOL: The effects of small class sizes on students' academic achievement, socioemotional development, and well-being in special education. *Campbell Systematic Reviews*. 2021, e1159. https://doi.org/10.1002/cl2.1159

APPENDIX A: FIRST- AND SECOND-LEVEL SCREENING

First-level screening is on the basis of titles and abstracts. Second level is on the basis of full texts.

Reference id. no.:

Reviewers initials:

Source:

Year of publication:

Country/countries of origin:

Author(s):

The study will be excluded if one or more of the answers to Questions 1–4 are "No". If the answers to Questions 1– 4 are "Yes" or "Uncertain", then the full text of the study will be retrieved to assess second-level eligibility. All unanswered questions need to be posed again on the basis of the full text. If insufficient information is available, or if the study details are unclear, the authors of the study will be contacted if possible.

Screening questions:

1. *Does the study measure the effects of special education (may be referred to as e.g. segregated placement, special class, self-contained special education classes, or resource rooms)?*
   Yes—include
   No—stop here and exclude
   Uncertain—include
   Question 1 guidance:
   Special education refers to educational settings catering exclusively to children with special educational needs, i.e., groups or classes contain only special education students. Placement in a special education setting may be full time or part time.
2. *Does the study measure effects for students with special needs?*
   Yes—include
   No—stop here and exclude
   Uncertain—include
   Question 2 guidance:

The population of this review are children with special educational needs in grades K to 12 (or the equivalent in European countries) in special education. Studies that meet inclusion criteria will be accepted from all countries. In this review we apply the widely used definition from the US Individuals with Disabilities Education Act (IDEA), in which special needs are divided into 13 different disability categories under which children are eligible for services. These categories are:

- specific learning disability (covers challenges related to a child's ability to read, write, listen, speak or do math, e.g., dyslexia or dyscalculia),
- other health impairment (covers conditions limiting a child's strength, energy, or alertness, e.g., ADHD),
- autism spectrum disorder (ASD),
- emotional disturbance (may include, e.g., anxiety, obsessive-compulsive disorder, and depression),
- speech or language impairment (covers difficulties with speech or language, e.g., language problems affecting a child's ability to understand words or express herself),
- visual impairment (covers eyesight problems, including partial sight and blindness),
- deafness (covers instances where a child cannot hear most or all sounds, even with a hearing aid),
- hearing impairment (refers to a hearing loss not covered by the definition of deafness),
- deaf-blindness (covers children suffering from both severe hearing and vision loss),
- orthopaedic impairment (covers instances when a child has problems with bodily function or ability, as in the case of cerebral palsy),
- intellectual disability (covers below-average intellectual ability),
- traumatic brain injury (covers brain injuries caused by accidents or other kinds of physical force),
- multiple disabilities (children with more than one condition covered by the IDEA criteria).

Note that the above categories should not be seen as exhaustive, but as guiding tools. Other definitions of special needs than the above mentioned will also be eligible. If in doubt, include study for full-text screening.

3. *Is the report/article a quantitative study with a comparison condition, or a qualitative study collecting empirical data?*
   Yes—include
   No—stop here and exclude
   Uncertain—include
   Question 3 guidance:

Quantitative studies: We are only interested in primary quantitative studies with a control or comparison group. Eligible study designs are randomised controlled trials (RCTs), quasi-randomised controlled trial designs (QRCTs), quasi-experimental studies (QES), and repeated-measures experimental designs in which the same caregiver and/or children are observed under different conditions within a short time span. Studies reporting associations in cohort, cross-sectional and longitudinal study designs without a comparison group are not eligible.

Qualitative studies: We will include all types of empirical qualitative studies that collect primary data and provide descriptions of main methodological issues such as sampling, data collection procedures, and type of data analysis. A qualitative study may apply a wealth of data collection methods, with participant observation, in-depth interviews, or focus groups being examples of possible methods we may encounter in the included studies.

Note: We are not interested in theoretical papers on the topic, or surveys/reviews of studies of the topic. (This question may be difficult to answer on the base of titles and abstracts alone). If in doubt, include for second-level screening on full text.

APPENDIX B: DATA EXTRACTION (QUANTITATIVE STUDIES)

Names of author(s)	
Title	
Language	
Journal	
Year	
Country	
Type of school setting (including grade level)	
Programme feature: *Study design* (brief description)	
Programme feature: *Intervention* (type of special education setting, such as full or part time)	
Programme feature *Outcomes* (academic achievement, social emotional learning, or wellbeing)	
Programme feature *Participants* (type of special educational need/ disability category, age, gender, SES)	
Programme feature *Teacher characteristics*, (number of teachers, educational background, years of experience, continuous professional development)	
Type of data used in study (independent observation, questionnaire, other (specify))	
Level of aggregation (individual and/or setting)	
Time period covered by analysis (divide into intervention and follow up)	
Sample size (divide into treated/comparison)	

APPENDIX C: OUTCOME MEASURES

Dichotomous outcome data

OUTCOME	TIME POINT (s) (record exact time from participation, there may be more than one, record them all)	SOURCE	VALID Ns	CASES	NON-CASES	STATISTICS	Pg. # & NOTES
Comparison	Comparison	Questionnaire Admin data Other (specify) Unclear	Participation Comparison	Participation	Participation	RR (risk ratio) OR (odds ratio) SE (standard error) 95% CI DF P- value (enter exact p value if available) Chi2 Other	

*Repeat as needed.

Continuous outcome data

OUTCOME	TIME POINT (s) (record exact time from participation, there may be more than one, record them all)	SOURCE (specify)	VALID Ns	Means	SDs	STATISTICS	Pg. # & NOTES
Comparison	Comparison	Questionnaire Admin data Other (specify) Unclear	Participation Comparison	Participation	Participation	P t F Df ES Other	

*Repeat as needed.

APPENDIX D: ASSESSMENT OF RISK BIAS IN INCLUDED STUDIES

User guide for unobservables

Systematic baseline differences between groups can compromise comparability between groups. Baseline differences can be observable (e.g., age and gender) and unobservable (to the researcher; e.g., motivation and "ability"). There is no single nonrandomised study design that always solves the selection problem. Different designs solve the selection problem under different assumptions and require different types of data. Especially how different designs deal with selection on unobservables varies. The "right" method depends on the model generating participation, that is, assumptions about the nature of the process by which participants are selected into a programme.

As there is no universally correct way to construct counterfactuals, we will assess the extent to which the identifying assumptions (the assumption that makes it possible to identify the counterfactual) are explained and discussed (preferably the authors should make an effort to justify their choice of method). We will look for evidence that authors are using, e.g. (this is NOT an exhaustive list):

Natural experiments

Discuss whether they face a truly random allocation of participants and that there is no change of behaviour in anticipation of, e.g., policy rules.

Matching (including propensity scores)

Explain and discuss the assumption that there is no selection on unobservables, only selection on observables.

(Multivariate, multiple) Regression

Explain and discuss the assumption that there is no selection on unobservables, only selection on observables. Further discuss the extent to which they compare comparable people.

Regression discontinuity (RD)

Explain and discuss the assumption that there is a (strict!) RD treatment rule. It must not be changeable by the agent in an effort to obtain or avoid treatment. Continuity in the expected impact at the discontinuity is required.

Difference-in-difference (treatment-control-before-after)

Explain and discuss the assumption that the trends in treatment and control groups would have been parallel, had the treatment not occurred.

APPENDIX E: JUSTIFICATION OF EXCLUSION OF STUDIES USING AND INSTRUMENTAL VARIABLE (IV) APPROACH

Studies using IV for causal inference in nonrandomised studies will not be included as the interpretation of IV estimates is challenging. IV only provides an estimate for a specific group, namely people whose behaviour changes due to changes in the particular instrument used. It is not informative about effects on never-takers and always-takers because the instrument does not affect their treatment status. The estimated effect is thus applicable only to the subpopulation whose treatment status is affected by the instrument. As a consequence, the effects differ for different IVs and care has to be taken as to whether they provide useful information. The effect is

interesting when the instrument it is based on is interesting in the sense that it corresponds to a policy instrument of interest. Further, if those that are affected by the instrument are not affected in the same way, the IV estimate is an average of the impacts of changing treatment status in both directions, and cannot be interpreted as a treatment effect. To turn the IV estimate into a LATE requires a monotonicity assumption. The movements induced by the instrument go in one direction only, from no treatment to treatment. The IV estimate, interpreted as a LATE, is only applicable to the complier population, those that are affected by the instrument in the "right way". It is not possible to characterise the complier population as an observation's subpopulation cannot be determined and defiers do not exist by assumption.

In the binary-treatment–binary-instrument context, the IV estimate can, given monotonicity, be interpreted as a LATE; that is, the average treatment effect for the subpopulation of compliers. If treatment or instruments are not binary, interpretation becomes more complicated. In the binary-treatment–multivalued-instrument (ordered to take values from 0 to J) context, the IV estimate, given monotonicity, is a weighted average of pairwise LATE parameters (comparing subgroup j with subgroup $j-1$). The IV estimate can thus be interpreted as the weighted average of average treatment effects in each of the J subgroups of compliers. In the multivalued-treatment (ordered to take values from 0 to T)—multivalued-instrument (ordered to take values from 0 to J) context, the IV estimate for *each pair of instrument values*, given monotonicity, is a weighted average of the effects from going from $t-1$ to t for persons induced by the change in the value of the instrument to move from any level below t to the level t or any level above. Persons can be counted multiple times in forming the weights.