What do we know about the effectiveness of group work in Japan's secondary English education? A systematic review of an 'active learning' technique

Lee Harvey Alexander
St Anne's College

Dissertation submitted in part-fulfilment of the requirements for the degree of Master of Science in Applied Linguistics and Second Language Acquisition

University of Oxford
Trinity Term, 2019

日本の中等英語教育におけるグループワークの有効性に関して今まで何が分かったか？
「アクティブラーニング」技法の一つをシステマティックレビュー

Abstract

With its 2017/18 revised curriculum guidelines for junior high school and high school (*Courses of Study*; to be implemented in 2020), Japan's education ministry (MEXT) requires that teachers adopt 'active learning' techniques; such as 'discovery learning' and group work. MEXT (2012, p. 37) promotes group work as an "effective" method for teaching English to Japan's junior high school and high school students; in preference to what the ministry terms "unidirectional lecture-style" teaching (i.e. Japan's version of grammar-translation and other teacher-centred approaches).

This systematic review is an assessment of the empirical basis for using group work over alternatives in Japan's secondary English education. It reviews experimental studies of group work in junior high school and high school English lessons in Japan, that measured linguistic proficiency or academic attainment between 2009 and 2019. A systematic search of 8 bibliographic databases plus web searches, conducted in May 2019, revealed only 8 studies that met inclusion criteria. Of these, 4 favoured group work, 3 found no difference between group work and alternatives such as individual work, and 1 was not interpretable. Findings of a narrative, best-evidence synthesis concerning the effects of group work in this context were equivocal. The evidence base was small and lacked robust designs. On measures of linguistic proficiency and academic attainment, it was not clear if group work differs meaningfully from alternatives such as individual work. For English teachers at junior high schools and high schools in Japan to have an evidence-based understanding of the effects of group work in their context, more intervention research is needed.

【要約】 平成 29･30 年改訂（令和 2 年実施予定）中･高等学校学習指導要領により、文部科学省は教師が「発見学習」やグループワークなどの「アクティブラーニング」（主体的･対話的で深い学び）技法の採用を義務付けています。文部科学省は、「一方向的な講義形式」の指導（つまり、文法訳読法などの教師中心のアプローチ）よりも、日本の中学生や高校生に英語を教えるのに「有効な」方法としてグループワークを推進しています。このシステマティックレビューは、日本の中等英語教育における他の方法よりもグループワークを使用するための根拠を検討します。

2009 年から 2019 年の間の、言語能力や学力を測定した日本の中学校および高等学校の英語授業でのグループワークの実験的研究をレビューしました。 2019 年 5 月に行われた 8 つの書誌データベースとウェブ検索の系統的な検索では、基準を満たした研究は 8 つだけであった。これらのうち、4 つの研究はグループワークの有効性を支持し、3 つはグループワークと個人ワークのような他の方法との間に違いが見つからず、1 つは解釈不能でした。定性的ベスト･エビデンス統合（best-evidence synthesis）における日本の中等英語教育でのグループワークの効果に関して、知見はあいまいであった。証拠基盤が小さく、堅牢なリサーチ･デザインが欠けていました。言語能力と学力の尺度に関しては、グループワークが個人ワークのような他の方法と有意に異なるかどうかは明らかではなかった。日本の中学校および高等学校の英語教師がグループワークの効果についてのエビデンスに基づいた理解を持つためには、さらなる介入研究が必要です。

Structured summary

Background

In 2020, English teachers at Japan's junior high schools and high schools will be required to adopt group work, and other 'active learning' techniques. Japan's education ministry (MEXT, 2012, p. 37) promotes group work as "effective." Although group work is thought to be beneficial, effects in Japan's junior high school and high school English lessons have yet to be verified.

Objectives

To assess the empirical basis for using group work over alternatives in Japan's junior high school and high school English lessons, the aim of this systematic review is to critically evaluate the nature and extent of empirical research on effects of group work in secondary English education in Japan.

Methods

This review includes experimental studies of group work in junior high school and high school English lessons in Japan, that measured linguistic proficiency or academic attainment. Risk of bias is assessed using the *Effective Public Health Practice Project Quality Assessment Tool for Quantitative Studies* (Thomas et al., 2004), a validated tool adapted to education by Chalmers (2019).

Results

A systematic search of 8 bibliographic databases plus web searches, conducted in May 2019, revealed only 8 studies; including 1 randomized trial, 5 non-randomized comparisons, and 2 studies of single groups. Participants ranged from 25 to 162. Of these 8 studies, 4 favoured group work, 3 found no difference between group work and alternatives such as individual work, and 1 was not interpretable. Findings of a narrative, best-evidence synthesis concerning the effects of group work in this context were equivocal. The evidence base was small and lacked robust designs.

Interpretation

On measures of linguistic proficiency and academic attainment, it is not clear if group work differs meaningfully from alternatives such as individual work in this context. For English teachers at junior high schools and high schools in Japan to have an evidence-based understanding of the effects of group work in their context, more intervention research is needed.

Acknowledgements

Thanks to Ryoko, who supported me in coming to Oxford, as you support me in so much. Thank you to Lena and Ray for putting up with my 'essays' and always putting a smile on my face. Thanks to Mum, Dad, Glenn, Shaun, Craig, and Bea. I love you all.

Thank you to Hamish Chalmers. If this dissertation has any good parts, you deserve the credit. Thank you for inspiring me to do this project and being a great supervisor and friend.

Thanks to Abbey Palmer for moral support with this project, both as a fellow systematic reviewer and a fellow parent. We did it!

Thank you to everyone at the Oxford University Department of Education. I leave with many fond memories. Keep on making the world a better place.

And thank you to all my teachers over the years. Your hard work sent a precariat boy to the University of Oxford.

Contents

Figures

Tables

Chapter 1 Introduction

Approximately every ten years, Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT)[1] issues revisions to its national curriculum guidelines (Tahira, 2012); known as the *Courses of Study*.[2] The latest guidelines for junior high schools and high schools were published in July 2017 and July 2018, respectively. Full implementation is set for 2020. These 2017/18 guidelines take 'active learning'[3] as their basis, with English teachers expected to adopt techniques such as 'discovery learning' and group work (McMurray, 2018). This systematic review focuses on one element of 'active learning': group work. This chapter introduces the rationale (1.1) and objectives (1.2) of this review.

1.1 Rationale

The ministry has stated that group work is "effective" (MEXT, 2012, p. 37). Group work is thought to benefit students. Beneficial effects of cooperative techniques on students' academic attainment have been reported since at least the 1970s (Johnson, Johnson, Johnson, & Anderson, 1976). Both systematic reviews and non-systematic literature reviews of group work have contributed to a picture of improved academic attainment, interpersonal skills, attitudes to learning, and self-esteem (Cohen, 1994; Johnson & Johnson, 1989; Slavin, 1980, 1990). Yet, studies from the USA disproportionately feature in the evidence base, with studies from non-western contexts almost entirely absent from the literature. This is important because differing cultural, ethnolinguistic and institutional factors are thought to affect outcomes in second language acquisition (SLA; Dörnyei & Csizér, 1998).

Collaborative techniques are not widely used in Japan's junior high school and high school English lessons; where teacher-centred techniques such as grammar-

[1] In this work I use 'Japan's education ministry', 'the ministry', and 'MEXT' interchangeably. The ministry's name in Japanese (in footnotes hereafter: JP) is 文部科学省 *Monbukagaku-shō*. Monbusho has also been used in English (in footnotes hereafter: EN).

[2] JP: 学習指導要領 *gakushū shidō yōryō*.

[3] MEXT has used 'voluntary, interactive, and in-depth learning' (MEXT, 2015; the OECD used 'proactive, interactive and authentic learning;' OECD, 2018) as an English-language translation of JP: 主体的・対話的で深い学び *shutai-teki taiwa-teki de fukai manabi*, a term the ministry uses in place of active learning in its outreach material. I have collapsed all of these into the term 'active learning'.

translation dominate practice (Gorsuch, 1998; Hino, 1988; Wada, 2013; Hayashi, 2017). Teachers and parents in Japan have tended to resist collaborative techniques (Sugie, 1995), perhaps because allowing students to construct their own knowledge conflicts with a traditional conception of teachers as knowledge-givers (Hofstede, Hofstede, & Minkov, 2010; Nguyễn, 2008). Or because emphasizing process over content, or meaning over form, does not fit Japan's culture of learning (Samimy & Kobayashi, 2004). On a more practical note, a near absence of formal pre- and in-service teacher training may prevent Japanese teachers from effectively implementing group work (Kizuka, 2006), and pressure to teach to the tests in limited lesson time might cause teachers to prefer direct, teacher-centred instruction (Littlewood, 2007; Hayashi, 2017). Teachers not having the confidence in their own English to facilitate group work and handle students' unforeseen needs is also a concern (Samimy & Kobayashi, 2004).

It is worth considering whether the benefits attributed to group work based on research in the USA translate into Japan's English as a foreign language context. Since the effectiveness of collaborative techniques in Japan's secondary English education has yet to be verified (Wada, 2013), a systematic assessment of the empirical case for effects of group work in Japan's junior high school and high school English lessons is warranted.

1.2 Objectives

The aim of this systematic review is to critically assess the nature and extent of empirical research on effects of group work in secondary English education in Japan. To evaluate the Japanese education ministry's claim that group work is effective (MEXT, 2012), I focus on studies that report on the effects of group work on linguistic proficiency and academic attainment. The parameters of this systematic review were informed by the PICO framework (Richardson, Wilson, Nishikawa, & Hayward, 1995). The specifications for the review are shown in Table 1.1.

Table 1.1 PICO

Population	Students of English as a foreign language who attend a junior high school, high school, or other secondary school in Japan
Intervention	Group work; including pair work, collaborative learning, and cooperative learning techniques
Comparison	Other classroom interventions, or no stated intervention
Outcome	Measures of linguistic proficiency or academic attainment

This systematic review is guided by the *Preferred reporting items for systematic reviews and meta-analyses* (PRISMA; Moher, Liberati, Tetzlaff, & Altman, 2009).

Chapter 2 Literature review

This chapter attempts to show that an assessment of the empirical basis for using group work over alternatives in Japan's secondary English education is called for. To put Japan's 2017/18 revised curriculum guidelines into context, I review Japan's post-war curriculum reforms to-date (2.1). I also discuss some reasons why group work is thought to be helpful (2.2). I finish (2.3) by discussing why conclusions about effects of group work in the Japanese context should not be drawn from the body of empirical evidence for group work outside Japan (Cohen, 1994; Johnson & Johnson, 1989; Slavin, 1980, 1990).

2.1 An overview of Japan's post-war curriculum reforms

Every decade since the end of the second world war, Japan's education ministry has issued revisions to its national curriculum guidelines (Tahira, 2012). The latest guidelines for junior high schools and high schools were made public in July 2017 and July 2018, respectively. Full implementation is set for 2020. These 2017/18 guidelines take 'active learning' as their basis, with English teachers expected to adopt techniques such as 'discovery learning' and group work (McMurray, 2018). To give a sense of how Japan got to its new active learning curriculum, I review Japan's post-war curriculum reforms below.

2.1.1 Experiential education

The curriculum guidelines first published in 1947, during the American occupation of Japan, were based on the ideas of American philosopher and educator John Dewey (Mizuhara, 2010; cited in Koreto, 2017). Dewey's experiential education promoted 'learning by doing' over 'traditional' methods (Dewey, 1997). Japan's junior high school and high school teachers were expected to abandon techniques such as rote learning in favour of more expressive activities. In English teaching, the guidelines emphasized listening and speaking skills, and were interpreted as favouring the audiolingual method (Tahira, 2012). These early guidelines can be characterized as an attempt to westernize Japan's education policy.

2.1.2 Grammar-translation

The American-influenced curriculum of the occupation era came to be regarded as a failure (Beauchamp, 1987). It was blamed for declining test performance and was dropped by the ministry in the first curriculum revisions after Japan regained its sovereignty in 1952 (Mizuhara, 2010). The 1958/60 guidelines were an attempt to make the curriculum more consistent with Japan's educational traditions (Beauchamp, 1987). Ensuring uniformity across Japan, the guidelines for English stipulated a compulsory word list of 5,700[4] words to be covered from the start of junior high school (the start of English education for most students; English is not compulsory at the primary level in Japan) to the end of high school. Grammar was also re-emphasized. The guidelines were interpreted as favouring grammar-translation (Tahira, 2012), and a localized version of this method[5] came to dominate Japan's secondary English education (Hino, 1988). Another revision (the 1969/70 guidelines) was issued, but the curriculum remained grammar-driven into the 1970s (Tahira, 2012).

2.1.3 Yutori education

Faced with mounting domestic criticism over youth violence and classroom discipline, the ministry changed direction (Koreto, 2017). Under the theme of 'pressure-free and fulfilling'[6] education, also known as yutori (meaning pressure-free) education, the 1977/78 guidelines sought to make Japanese education less rigid and more individualist (Iwaki, 2004). To promote student engagement, more time was made for elective subjects at the secondary level. This was achieved by cutting lesson hours across core subjects, with English lessons reduced from five to four hours a week as standard. The guidelines also reduced prescribed content. The ministry met opposition from teachers concerned about reduced lesson hours but pressed ahead; a ministry official asserting that even three hours a week would be enough to teach students English (Koreto, 2017). Linguist and educator Yasuteru Otani (1981) came out in opposition to the reduced hours for English, predicting that results would decline. Events were to prove him right.

---

[4] Later reduced to 4,900 words.

[5] In the grammar-translation method as applied in Japan, translating into Japanese has tended to make up most translation work; with minimal translation into the target language (Hino, 1988).

[6] JP: ゆとりと充実 *yutori to jūjitsu*

2.1.4 Communicative language teaching

As Otani had predicted, test performance among Japanese students of English declined rapidly and the ministry came under pressure to abandon yutori education. In its 1989 guidelines, MEXT jettisoned many of the previous reforms. However, instead of simply returning to earlier principles, the 1989 guidelines emphasized communication. These guidelines have been characterized as a pivot toward communicative language teaching (CLT). Drawing on American linguist and anthropologist Dell Hymes' concept of communicative competence, CLT stresses communication as both means and goal of language learning. The 1989 guidelines declared that developing students' communicative competence was the main purpose of secondary English education in Japan (Yoshida, 2003). This curriculum reform started a sustained period in which CLT dominated Japan's secondary English curriculum. Both the 1998 guidelines and the 2008/09 guidelines maintained and widened the ministry's support for CLT (Tahira, 2012).

Implementation of the ministry's guidelines remained half-hearted in the 1990s and 2000s (Kikuchi & Browne, 2009). Despite three decades of policy favouring CLT, as it stands, communicative activities tend to be side-lined in Japan's junior high school and high school English classes (Tahira, 2012); with teachers preferring teacher-centred techniques such as grammar-translation (Gorsuch, 1998; Hino, 1988; Wada, 2013). An entrenched gap between policy and practice in Japan's secondary English education has been observed (Kikuchi & Browne, 2009). Three major obstacles are argued to impede curriculum reform in Japan (Gorsuch, 2001); these being 1) a near absence of formal pre- and in-service teacher training (Kizuka, 2006), 2) washback from university entrance exams that focus on translation and receptive skills (Kikuchi, 2006), and 3) junior high school and high school English teachers preference for, and reliance on, Japan's version of grammar-translation (Gorsuch, 1998; Hino, 1988; Wada, 2013). It is into this environment that the latest curriculum reforms come.

2.1.5 Active learning

The latest curriculum guidelines for upper- and lower secondary schools were made public in the period July 2017 to July 2018. Full implementation of these 2018/19 guidelines is set for 2020. After three decades of CLT-based reforms, the new guidelines take 'active learning' as their basis (McMurray, 2018). Indeed, the OECD (2018) has said Japan needs to implement active learning to achieve its curriculum goals. The ministry's promotion of active learning in Japan can be

traced back to 2012 (Mio, 2017), when it published a call for reform. That report contained the following definition of active learning (MEXT, 2012, p. 37; author's translation; emphasis added):

> A general term for teaching and learning methods that include students actively participating in their learning, unlike unidirectional lecture-style education by teachers ... Discovery learning ... etc. are included, but *group discussion, debate, group work* and so on in the classroom are also *effective* active learning methods.

First, this section of the report clarifies that various forms of group work[7] are important elements of the ministry's pivot toward active learning. Since assessing the evidence base of all the 'active learning' techniques that the ministry recommends was beyond the scope of this study, this review focuses on group work alone; as one of the components of the suite of techniques promoted by MEXT. Second, it is apparent that the ministry promotes group work, among other techniques, in the belief that these techniques are more effective than existing practice. This systematic review seeks to understand the nature and extent of empirical research that has informed this position.

2.2 A theoretical framework for group work

Group work is thought to benefit students in multiple ways. Beneficial effects of cooperative techniques on students' academic attainment have been reported since at least the 1970s (Johnson et al., 1976). Both systematic reviews and non-systematic literature reviews of group work have contributed to a picture of improved academic attainment, interpersonal skills, attitudes to learning, and self-esteem (Cohen, 1994; Johnson & Johnson, 1989; Slavin, 1980, 1990). This section discusses social constructivism; theory that might explain beneficial effects of group work.

Social constructivism, influenced by the ideas of Soviet psychologist Lev Vygotsky, gives weight to the learner's active involvement in the learning process. Learners are thought to construct knowledge, rather than simply mirroring what they see, hear or read (Von Glasersfeld, 1989). This process of knowledge-making is held to be social; learners co-construct knowledge through interaction and, as

[7] For the purposes of this systematic review, group work is two or more students working together. It is recognized that the terms pair work, group work, collaborative learning, and cooperative learning can be distinguished.

individuals, appropriate this socially constructed knowledge (Bruning, Schraw, & Ronning, 1999). By sharing individual perspectives, members of a group are thought to be able to come to an understanding that would be impossible alone (Van Meter & Stevens, 2000). A shared, deeper understanding of truth in a field is held to be possible when learners collaborate, bringing their differing skills and insights from different backgrounds to the discussion (Duffy & Jonassen, 1992).

Yet, in practice, group dynamics are crucial. It can be difficult to facilitate group work in which all members cooperate (Maiden & Perry, 2011). Working in groups can be an unpleasant experience for some students (Hall & Buzwell, 2012), and assessment of group work is thought to be a cause of anxiety (Maiden & Perry, 2011). In group work, a reduced sense of accountability may lead to social loafing, where students put less effort into a task when they are a member of a group than when they work alone (Hall & Buzwell, 2012). Free-riding, where some members leave most of the work to more diligent members, is a risk (Levin, 2003). Free-riding is thought to lower group morale and overall productivity (McArdle, Clements, & Hutchinson-Lendi, 2005). So, despite theoretical and empirical support for group work, it may not be beneficial in all contexts.

2.3 The problem of context

In contrast to many English as a foreign language settings, collaborative techniques are not widely used in Japan's junior high school and high school English lessons; where teacher-centred techniques, most prominently grammar-translation, have long dominated practice (Gorsuch, 1998; Hino, 1988; Wada, 2013; Hayashi, 2017). It is also notable that the ministry's promotion of group work comes despite teachers and parents in Japan tending to resist collaborative techniques (Sugie, 1995). There are several possible reasons for opposition to group work in this context. Allowing students to construct their own knowledge may conflict with a traditional conception of teachers as knowledge-givers (Hofstede et al., 2010; Nguyễn, 2008). Emphasizing process over content, or meaning over form, may not fit Japan's culture of learning (Samimy & Kobayashi, 2004). On a more practical note, a near absence of formal pre- and in-service teacher training may prevent Japanese teachers from effectively implementing group work (Kizuka, 2006). Pressure to teach to the tests in limited lesson time might cause teachers to prefer direct, teacher-centred instruction (Littlewood, 2007; Hayashi, 2017). Teachers' self-esteem is also relevant. Teachers may not have the

confidence in their own English to facilitate group work and handle students' unforeseen needs (Samimy & Kobayashi, 2004).

Despite support for group work outside Japan (Cohen, 1994; Johnson & Johnson, 1989; Slavin, 1980, 1990), it is worth considering whether group work is helpful in Japan's English as a foreign language context. Swan (2018, p. 250) puts it well:

> It is not enough for a teaching approach to conform to current SLA thinking, however persuasive the supporting arguments. It must also fit the relevant context, offering solutions that match the problems inherent in the situation.

What works in one place under one set of conditions does not necessarily work in another place or under a different set of conditions. It is recognized that the evidence synthesized by Slavin (1980, 1990) makes a persuasive case for group work in relation to academic attainment. Yet Slavin concluded that "achievement results, though usually positive, seem to depend on the particular techniques, settings ... or other characteristics" (1980, p. 333). So, we should not assume that group work is effective under all conditions. Slavin's systematic reviews also demonstrated that studies from the USA disproportionately feature in the evidence base. This is important because cultural, ethnolinguistic and institutional differences are thought to affect outcomes in second language acquisition (SLA; Dörnyei & Csizér, 1998). "No cause has its effect apart from some larger context involving other variables" (Chambliss & Schutt, 2018, p. 110). Evidence from countries that differ markedly from Japan in terms of cultural, ethnolinguistic and institutional factors does not establish effects of pedagogical techniques in the Japanese context. Since the effectiveness of collaborative techniques in Japan's secondary English education has yet to be verified (Wada, 2013), a systematic assessment of the empirical case for effects of group work in Japan's junior high school and high school English lessons is justified.

Chapter 3 Methods

This chapter describes the methods used to gather, appraise, and synthesize research. I present the research question (3.1) and describe the design of the research (3.2). I also discuss the protocol (3.3), a document which acted as a 'blueprint' for my research. I recount where I chose to search (3.4), the eligibility criteria that were set before I started searching (3.5), and the search itself (3.6). The study selection process is also addressed (3.7). I describe how I extracted data from the included studies and what data I recorded (3.8). I explain how risk of bias for the included study was assessed individually (3.9). How I synthesized the results is also set out (3.10). The chapter concludes with a description of risk of bias assessment for the body of included studies (3.11).

3.1 Research question

The aim of this systematic review is to critically assess the nature and extent of empirical research on effects of group work in secondary English education in Japan. To evaluate the Japanese education ministry's claim that group work is effective (MEXT, 2012), I researched effects on linguistic proficiency and academic attainment. To that end, my research question was:

What is the nature and extent of research investigating effects of group work on linguistic proficiency and academic attainment in English as a foreign language lessons in Japan's secondary schools?

3.2 Research design

To capture the nature and extent of empirical research on effects of group work in Japan's junior high school and high school English lessons, I chose to conduct a systematic review. In contrast to traditional literature reviews (of which Chapter 2 is an example), systematic reviews seek to bring the totality of empirical evidence to bear on a defined research question. With the aim of minimizing bias, search strategies are explicit, searches are conducted systematically, and prespecified eligibility criteria are applied in the screening process (Boland, Cherry, & Dickson, 2017). Systematic reviews are held to provide "more reliable findings

from which conclusions can be drawn and decisions made" (Cochrane Collaboration, 2017).

The following considerations informed my decision to select a systematic review as the most appropriate research design for my question:

- Systematic reviews are seen as the 'gold standard' in locating, evaluating, and synthesizing research on a specific research question (Boland et al., 2017).
- It is important to fully evaluate interventions, to ensure they provide optimal outcomes (Boland et al., 2017). Recommendations resulting from a systematic review are predicated on the totality of best available evidence.
- Compared to other forms of review, the explicit and transparent methods of systematic reviews reduce risk of bias, adding weight to any findings (Cooper, 2010).
- By providing an overview of what is known about a specific question, systematic reviews can not only provide answers but also throw light on gaps in research (Boland et al., 2017). They can therefore inform future research; potentially opening new threads of enquiry.
- As a teacher, I am interested in research informing policy and practice. The generalizability and replicability of a systematic review offer a strong basis for applying any findings to policy and practice in my context (Gough, Oliver, Thomas, & Hobbs, 2013).

3.3 Protocol

After considering the research design, I composed a protocol which acted as a 'blueprint' for this systematic review. In the protocol I addressed the relevance, rationale, justification, and specification of my research; specifying the population, intervention, comparisons, and outcome measures that I sought (i.e. the PICO framework; Richardson et al., 1995). I set out my search strategy, specified the study selection process and eligibility criteria, and considered quality assessment, data extraction, and data synthesis. I also considered the resources that I would need to conduct this systematic review, its timeframe, and potential avenues for disseminating the findings. I did not publish the protocol, and I did not register the title of my systematic review at any time.

It is considered best practice to register prospective systematic reviews, both to provide public notice and to prevent duplication of effort (PLoS Medicine Editors, 2011). Additionally, publishing the protocol for a systematic review provides a measure of transparency. The Campbell Collaboration (n.d.) offer a means to register titles and publish protocols. However, the peer review process and the international focus of Campbell systematic reviews make them ill-suited to a systematic review conducted as part of a master's degree that is also focused on one specific context. Where systematic reviews have a health-related outcome, the *International Prospective Register of Systematic Reviews* (PROSPERO; University of York, n.d.) allows researchers to check for duplicates and register new systematic reviews quickly. I talked with my supervisor and other academics, and searched the internet, but I did not find a well-established resource equivalent to PROSPERO for registering systematic reviews that evaluate the efficacy of interventions with non-health outcomes. Since this systematic review was conducted without registering the title, the possibility that it might duplicate the work of others cannot be ruled out.

3.4 Information sources

Eight electronic bibliographic databases were searched. These are shown in Table 3.1.

Table 3.1 Electronic databases

	Name	Discipline	Platform/Interface
1	Applied Social Sciences Index and Abstracts (ASSIA)	Social sciences	ProQuest
2	British Education Index	Education	EBSCO
3	CiNii (covers Japanese academic libraries)	Multidisciplinary	https://ci.nii.ac.jp/
4	Education Collection (includes ERIC)	Education	ProQuest
5	Linguistics and Language Behavior Abstracts (LLBA)	Linguistics	ProQuest
6	ProQuest Dissertations & Theses Global	Multidisciplinary	ProQuest
7	Scopus	Multidisciplinary	SciVerse
8	Web of Science (formally Web of Knowledge)	Multidisciplinary	Thomson Reuters

Except for CiNii, these databases were accessed through the web portal of the Bodleian Libraries. CiNii is open access.

In addition, I conducted backward citation searches. Where studies met all inclusion criteria, I consulted the reference sections in search of other potentially eligible studies. I also sought suggestions from two Japan-based researchers whose work is related to components of active learning in English education, including group work. Understanding that selective publication of research can influence the results of research synthesis (Song et al., 2010), I searched the internet using Google's search engine for eligible grey literature.[8] Using the English search terms in Table 3.3 and the Japanese search terms in Table 3.4, I searched Google using the verbatim search tool. I also ran those same searches limiting by file type to 'pdf' only.

[8] The Twelfth International Conference on Grey Literature defined this term as the "manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights, of sufficient quality to be collected and preserved by libraries and institutional repositories, but not controlled by commercial publishers; i.e. where publishing is not the primary activity of the producing body." (Schöpfel, 2010; cited in Chalmers, 2019).

3.5 Eligibility

The aim of my research was to evaluate the empirical basis for guidelines requiring use of group work, given by MEXT to teachers of English at Japan's junior high schools and high schools. To that end, I compiled a list of eight eligibility criteria for use in the selection process of my systematic review. The criteria *excluded* studies that:

1) Had incomplete bibliographic records
2) Were not reported in English or Japanese
3) Were irrelevant to assessing group work in teaching English as a foreign language
4) Had an ineligible study design (for explanation see Table 3.1, Include 4)
5) Did not include students of English as a foreign language attending a secondary school
6) Did not include typical school populations
7) Were conducted entirely outside Japan
8) Contained no measures of linguistic proficiency or academic attainment

These criteria, and their rationale, are shown in Table 3.2.

Table 3.2 Eligibility criteria and rationale

	Eligibility criteria	**Rationale**
Bibliographic records	*Include 1:* Bibliographic record is complete	I expected that locating full reports from incomplete bibliographic records would not be practical.
	Exclude 1: Bibliographic record is incomplete	
Language	*Include 2:* Available in English or Japanese	The potential for bias in limiting studies to only those published in Japanese and English is recognized. In the context of this systematic review, I did not have the resources to translate articles that were not in English or Japanese.
	Exclude 2: Not available in English or Japanese	
Relevance	*Include 3:* Relevant to assessing group work in teaching English as a foreign language	It is recognized that collaborative learning and cooperative learning can be distinguished. For the purposes of this systematic review, any group intervention (one with two or more students working together) met this criterion.
	Exclude 3: Not relevant to assessing group work in teaching English as a foreign language	

Design	*Include 4:* [9] If formal evaluations are found, include only these. If no formal evaluations are found, include other designs with preference for designs higher on the 'hierarchy of evidence'.[10]	I wanted to give preference to study designs capable of addressing effects of group work, such as randomized controlled trials. I also recognized that few if any formal evaluations might exist. I therefore framed this criterion to capture the best evidence available, in line with Slavin's (1986) best evidence approach.
	Exclude 4: If formal evaluations are found, exclude other designs. If no formal evaluations are found, only exclude lower designs where there are at least 3 rigorous designs higher on the 'hierarchy of evidence'.	
Population and context	*Include 5:* Students of English as a foreign language attending secondary schools	Age effects have been found in SLA research (DeKeyser, Alfi-Shabtay, & Ravid, 2010). Since this review aimed to assess guidance on how to teach English to junior high school and high school students, including other ages might have confounded findings.
	Exclude 5: Not students of English as a foreign language attending secondary schools	
	Include 6: Includes typical school populations	The MEXT guidelines for junior high schools and high schools is aimed at mainstream teachers. Evidence must reflect that general population.
	Exclude 6: Solely includes students with special educational needs, learning difficulties or language disorders.	
	Include 7: Conducted in Japan in whole or in part	This review aimed to assess guidelines given by MEXT to teachers in Japan. Studies conducted exclusively outside Japan were not considered relevant to the question posed.
	Exclude 7: Conducted exclusively outside Japan	
Outcomes	*Include 8:* Reports measures of linguistic proficiency or academic attainment	Only these measures were considered relevant to an assessment of effectiveness. Measures of *perceived* proficiency or attainment were excluded.
	Exclude 8: Does not report measures of linguistic proficiency or academic attainment	

[9] Where the following literature did not contain empirical examples, it was excluded: Methodological studies, studies of statistical associations or relationships between variables, white papers, working papers, planning documents, article reviews, book reviews, theory or guidance pieces, editorials, letters, commentaries, opinion or exhortation pieces, news reports, textbooks, bibliographies, indexes or content pages.

[10] The hierarchy of evidence, also called the study design pyramid, ranks study designs based on the extent to which they control for possible biases. It is considered "the gold standard for evaluating effectiveness" (Petticrew & Roberts, 2003, p. 528).

The criterion for study design (Table 3.2, Include 4) deserves justification. This systematic review sought to assess the empirical basis for MEXT's promotion of group work as *effective*. The nature of research in education is such that among the spectrum of study designs, bias-reducing designs such as randomized trials are used proportionally less (Paul Connolly, Keenan, & Urbanska, 2018). This low proportion presents a problem if such designs are considered the 'gold standard' for assessing the effectiveness of interventions (Meldrum, 2000). I therefore followed Slavin's (1986) best evidence approach. This approach gives due prominence (other things being equal) to robust, bias-reducing designs while recognizing the merit of a cautious synthesis of studies with designs that are more susceptible to bias. The value of this approach is that cautious conclusions can potentially be drawn from research synthesis even where studies with more trustworthy designs cannot be found. Such synthesis may also provide an overview of the state-of-the-art, which could highlight new directions for research.

3.6 Search

As a first step in developing my search strategy, I compiled a list of potential English search terms from terminology that I was already familiar with from the literature on second language acquisition. Since the aim of my research was to evaluate the empirical basis for MEXT's guidelines recommending that *group work* be used in *secondary* schools to teach *English* in *Japan*, my list of search terms naturally divided into four sets of related words. These were 1) terms related to group work, 2) terms related to secondary education, 3) terms related to English, and 4) terms related to Japan. After soliciting suggested additions to the list from colleagues, I ran scoping searches of the databases listed in Table 3.1. The list was augmented with additional synonyms that emerged in the process of these scoping searches. The scoping searches, and my supervisor, also brought my attention to instances of potential redundancy in the preliminary list, such as "junior high school" (redundant because "high school" was also included). I ran scoping searches with and without the potentially redundant terms and eliminated them where I found that the number of returns remained substantially the same. As a final step, I sought the help of a departmental librarian with experience refining searches for systematic reviews. She suggested additions that I had overlooked, and these were included in the final list of English-language search terms.

That list formed the basis for the search strings used to search all databases except CiNii. These search terms are listed in Table 3.3.

Table 3.3 English-language search terms

Field 1[11]	**AND**[12]	**Field 2**	**AND**	**Field 3**	**AND**	**Field 4**
"group work*"	"	secondary	"	English	"	Japan*
"pair work*"	"	"high school*"	"	EFL	"	
"collaborative learning"	"	"middle school*"	"	TEFL	"	
"cooperative learning"	"		"	ELL	"	
	"		"	EAL	"	

Based on the results of my scoping searches, I opted to treat CiNii differently from the other databases. CiNii is a database of material from Japanese academic libraries, the bulk of which is in Japanese. Since CiNii is searchable in both English and Japanese, I did exploratory searches using equivalent terms in both languages. These exploratory searches showed a difference in the number of returned records. Searching in Japanese returned more results. A look at the results showed that, of the equivalent searches in the two languages, the Japanese-language search returned more Japanese records without missing out English-language records. In other words, searching in Japanese appeared to capture everything that an equivalent English-language search captured, as well as capturing more Japanese-language material. Therefore, I decided to search CiNii in Japanese only. Since CiNii exclusively covers material written in Japanese or material related to Japan, including search terms limiting the search to this context was redundant. That left three clusters of search terms used to search CiNii: 1)

[11] Search terms in the same field were linked using the Boolean operator OR.

[12] Groups of search terms in fields 1-4 were linked by the Boolean operator AND.

terms related to group work, 2) terms related to secondary education, and 3) terms related to English.

To compile the list of Japanese-language search terms, I first translated terms from the English-language list. After adding known synonyms, I solicited suggestions from Japanese colleagues. I finalized the list by including overlooked terms that emerged during scoping searches. That list was the basis for the search string used to search CiNii. These Japanese-language search terms are listed below in Table 3.4.

Table 3.4 Japanese-language search terms

Field 1	**AND**	**Field 2**	**AND**	**Field 3**
グループワーク [group work][13]	"	中等教育 [secondary education]	"	英語 [English language]
ペアワーク [pair work]	"	中学校 [junior high school]	"	EFL
協働的学習 [collaborative learning]	"	高等学校 [high school]	"	TEFL
協調学習 [collaborative learning, synonym]	"	高校 [high school, synonym]	"	ELL
協同学習 [cooperative learning]	"		"	EAL

In search strategies, there is a trade-off between sensitivity, also known as recall, and precision, also known as specificity (Brunton, Stansfield, & Thomas, 2012).

[13] In this table the contents of square brackets [] are EN translations provided to assist the reader. Only the JP terms (outside the square brackets) were used to search CiNii.

A sensitive search maximizes capture of relevant records at the expense of including a high proportion of irrelevant records. A high proportion of irrelevant records increases the time and effort required for screening, potentially by orders of magnitude. Precise searches avoid this problem by maximizing the proportion of relevant returned records. Of course, a precise search risks failing to capture the breadth of relevant research, potentially undermining findings. In the case of this systematic review, the question was in any case framed tightly. I chose to look at only one aspect of MEXT's guidelines, and as set out in 2.3, reviewing research from specifically the Japanese context is crucial to evaluating the empirical basis for the ministry's advice. As such, the search terms that logically stemmed from my research question resulted in a manageable number of returned records.

Where the database's interface allowed searches confined to abstracts, I limited each group of terms (Fields 1-4 in Table 3.3, and Fields 1-3 in Table 3.4) to abstracts only. I did this for all databases except Web of Science. Web of Science does not have a function for limiting search terms to abstracts only. In the case of Web of Science, I limited by topic. The topic limiter on Web of Science searches the title, abstract, and keywords. These limiters were applied to increase the precision of my search. Scoping searches without these limiters returned records in the tens of thousands, most of which were records of irrelevant studies that only mentioned the search terms in passing.

The search was augmented by looking at citations from included studies and searching the internet using Google's search engine. The Google searches returned many results. The English-language search string limited to pdfs returned 275,000 results, for example. Therefore, unlike the database search, my screening of the Google search results was not exhaustive. The Google search engine ranks returns by relevance. Starting at the top of the first page,[14] I screened each listing in turn, clicking on the listings and reviewing the full contents unless they could be excluded based solely on the snippet (the content summary that comes with each listing). I continued screening all ten listings on each page until I found that I could exclude all listings based solely on the snippet for 30 listings in a row (i.e. three continuous pages). I stopped once I reached this point.

[14] Here, page refers to a search engine results page (SERP).

3.7 Study selection

The study selection process was 1) removing duplicates from the records that were returned by the database searches described in 3.6, 2) screening the titles and abstracts of the records to exclude ineligible studies, 3) locating the full reports where studies had not been excluded based on the titles and abstracts, and 4) screening the full reports. I describe each part of the study selection process below, as well as discussing the reliability of the process.

3.7.1 Removing duplicates

After conducting the electronic database search described in 3.6, I uploaded the returned records to *Rayyan* (Ouzzani, Hammady, Fedorowicz, & Elmagarmid, 2016), a web-based application for deduplicating and screening studies for systematic reviews. The application highlighted five sets of potential duplicates, of which only two sets proved to be duplicates. I deleted one copy of each duplicated study and moved to title and abstract screening. During screening, I identified one further set of duplicates that the application had not detected. I labelled one copy as a duplicate, marked it for exclusion, and eliminated it from my final count of excluded studies.

3.7.2 Screening titles and abstracts

After resolving duplicates, I screened each title and abstract against the eligibility criteria presented in 3.5. Where it was clear from the contents of the title and abstract that a study met one or more of the exclusion criteria (see Table 3.2), I marked it for exclusion and recorded a reason.[15] Where it was not clear from the contents of the title and abstract that a study met one or more of the exclusion criteria, I provisionally marked it for inclusion. Once I had screened all records, I created a separate list of provisional includes, for which I located the full reports.

3.7.3 Locating full reports

The bibliographic records from the list of provisional includes were used to locate full reports. I retrieved most of the English-language reports from journal repositories accessed through the web portal of the Bodleian Libraries. Others were

[15] I only recorded reasons for the first exclusion criterion that each study met. Some studies met multiple exclusion criteria, but I did not record any further criteria.

retrieved from academic websites (such as ResearchGate) or personal websites maintained by the authors. I retrieved some of the Japanese-language reports from institutional journal repositories accessed through the CiNii website. In many cases, CiNii did not contain a link to a full report. In these cases, I took the following steps: 1) I searched academic or personal websites, and where that method failed 2) I ordered photocopies of the articles from Japan's national library.[16] Government and news reports were obtained from their associated websites.

3.7.4 Screening full reports

After locating full reports for the studies that I marked for inclusion in screening of titles and abstracts, I reviewed each report in turn. I started by reading the methods chapter, if there was one. If I could not exclude a study based on the methods chapter alone, I read the full report. Where the contents indicated that the study met one or more of the exclusion criteria I marked it for exclusion, recording a reason. Where the studies met all the inclusion criteria, they were marked for inclusion.

3.7.5 Reliability

It was important to minimize any effects of personal bias or the selected screening method on the screening process (Boland et al., 2017). In systematic reviews, it is common for members of a review team to carry out simultaneous screening, followed by comparison of the members' decisions (Gough et al., 2013). Having two or more raters agree on the eligibility of studies helps demonstrate the validity of a review's study selection process. As a lone researcher, I opted to have a portion of the returned reports independently screened. A fellow master's student with experience working on systematic reviews screened 10% of the records at the titles and abstracts screening phase (the phase described in 3.7.2). Using the eligibility criteria discussed in 3.5, she independently assessed the eligibility of this sample blind to my inclusion/exclusion decisions. After she had finished, I compared her inclusion/exclusion decisions to my own decisions. There was 100% agreement in our decisions.

[16] EN: National Diet Library, JP: 国立国会図書館 *Kokuritsu kokkai toshokan*.

3.8 Data extraction

I extracted data from the included studies using a coding sheet designed for the purpose. The coding sheet recorded the following:

1) How the report was identified
2) The language the report was published in
3) Where the study took place
4) The type of school in which the study was conducted
5) What the language teaching provision was
6) Who the participants were
7) What the research question was
8) What the group work intervention was
9) What the intervention was compared to
10) Who delivered the intervention and to how many participants at a time
11) What the independent and dependent variables were
12) What study design was used
13) Whether the outcome measures reflected linguistic proficiency or academic attainment
14) How many participants there were
15) The length of the study
16) The study's risk of bias rating

The coding sheet is presented in Appendix A.

3.9 Risk of bias within studies

Before forming conclusions based on a study included in a systematic review, it is first important to assess both the study's rigour and relevance (Greenhalgh & Brown, 2017): Are the studies reliable, and do they answer the review's research question? Failing to address these questions risks combining 'strong' evidence (which has a higher chance of reflecting the real-world outcomes of interventions) with 'weak' evidence (that is less likely to reflect real-world outcomes; for example, evidence that may be susceptible to bias or may fail to account for an important confounding variable). Conclusions drawn from poor quality studies risk being invalid and may mislead policy-makers and practitioners (Gorard, See, & Siddiqui, 2017).

In line with a best evidence approach (Slavin, 1986), I chose to focus my quality assessment on risk of bias. I sought an appropriate validated risk of bias assessment tool to draw out evidence for each study's trustworthiness, as advised by Drucker, Fleming, & Chan (2016). I did not find a validated tool specifically designed for risk of bias assessment of applied research in second language acquisition. I used Chalmers' (2019) adaptation of the *Effective Public Health Practice Project Quality Assessment Tool for Quantitative Studies* (hereafter: EPHPP tool; Thomas et al., 2004). The EPHPP tool is a validated risk of bias assessment tool, applicable to a variety of study designs, that is used in the field of public health. Chalmers adapted the EPHPP tool to language teaching on the premise that public health, more than other fields in healthcare, shares many commonalities with education; language teaching included. Use of risk of bias tools from healthcare is common in research synthesis in the social sciences, and the EPHPP tool is recommended as such by Petticrew and Roberts (2006).

Chalmers' adapted EPHPP tool adheres closely to the original. Substantive changes were confined to one item (c), reflecting the problem of "unrecorded, unknown or unknowable" (2019, p. 89) confounders. He also changed terminology to reflect educational rather than clinical settings. For more detail on the changes, and their rationale, see Chalmers (2019). The adapted EPHPP tool covered the following:

a) How representative the participants were of the target population, and how many of the selected participants agreed to take part
b) The extent to which the study design controlled for possible biases
c) What means were used to control for potential confounders
d) Whether participants knew the research question or whether the researchers knew the participants' group allocations when outcomes were assessed
e) What kinds of tools were used to collect the data
f) The proportion of participants who withdrew from the study
g) Any indication of contamination from other interventions or how consistently the intervention was applied
h) Whether the analysis suited the question posed

The EPHPP tool uses three ratings: strong, moderate, and weak. Hereafter, I will call strong 'high' and weak 'low'. Where no information relevant to an item was reported, I rated the risk of bias for that item as 'moderate'. Where no control groups were reported, and where the report contained no discussion of this

limitation, I rated the risk of bias for study design (c) as 'high'. Where the target population was not reported, I rated risk of selection bias as 'moderate'. Where the EPHPP tool lacked 'not applicable' as a response, instances where the question did not apply were recorded as 'can't tell'. The EPHPP tool is presented in Appendix B.

3.10 Synthesis of results

One method for synthesizing quantitative data is meta-analysis. The method typically employs a variation "on a weighted average of the effect estimates from ... different studies" (Higgins & Green, 2011, Chapter 9, para. 4). By statistically pooling data in this way, a meta-analysis increases sample size, which offers the potential advantage of decreased sampling error. By essentially creating a larger study, a meta-analysis should give more precise estimates for effect size (Russo, 2007). The method has its limitations. Bakker, van Dijk, & Wicherts (2012) have drawn attention to the problem of publication bias leading to overestimated effect sizes, for example. Nevertheless, meta-analysis is fundamental to evidence-based practice (Cochrane, 2019).

In some of the studies included in this systematic review, an intervention was not compared to any alternative. Since meta-analysis focuses "on pair-wise comparisons of interventions" (Higgins & Green, 2011, 9.1.2, para. 2), it would not have been appropriate to synthesize results using this method. I therefore opted for a narrative synthesis. Narrative synthesis is recommended by the *NIHR Complex Reviews Support Unit* (NIHR CRSU, n.d.) for synthesis of quantitative data where statistical synthesis is inappropriate. Drawing on the data extracted from included studies and the results of risk of bias assessment, I summarized the scope, themes, and quality of the studies included in this systematic review.

3.11 Risk of bias across studies

Risk of bias for the included studies was considered collectively. From the EPHPP tool results, I assessed the proportion of studies that had a 'high' overall risk of bias rating. I also considered which risk of bias rating on the EPHPP tool was most common among the included studies, and what that proportion was, for the following points: 1) study design, 2) data collection, and 3) withdrawals and dropouts. In addition to the EPHPP tool ratings, risk of bias across studies was assessed in the form of: 1) the proportion of studies that used control groups, 2)

the proportion of studies in which participants' group allocations were concealed at outcome assessment, 3) the proportion of studies that prospectively generated groups, and 4) the proportion of studies that randomly assigned interventions to groups.

Chapter 4 Results

This chapter reports the results of study selection (4.1) and lists the studies that were included in this systematic review (4.2). Characteristics of the included studies are summarized, and their findings are presented (4.3). I report results of risk of bias assessment within studies (4.4) and across studies (4.5). The included studies are also summarized individually (4.6). The chapter concludes with a synthesis of results (4.7).

4.1 Study selection

The database searches described in 3.6 returned 173 records. I identified 13 additional records through Google searches and backward citation searches. After removing all duplicates, there remained 183 distinct records. After screening titles and abstracts, 100 records were excluded based on the information contained therein. The reasons for exclusion were:[17]

- Exclude 7: conducted outside Japan, n = 46
- Exclude 3: not relevant to assessing group work, n = 44
- Exclude 5: only investigated ineligible learners, n = 10

The remaining 83 records could not be excluded based on information contained in the titles and abstracts, so full reports of these studies were sought. Of these, I was able to access 82 full-text articles. I assessed these against the exclusion criteria, and was able to exclude 74 on the following bases:

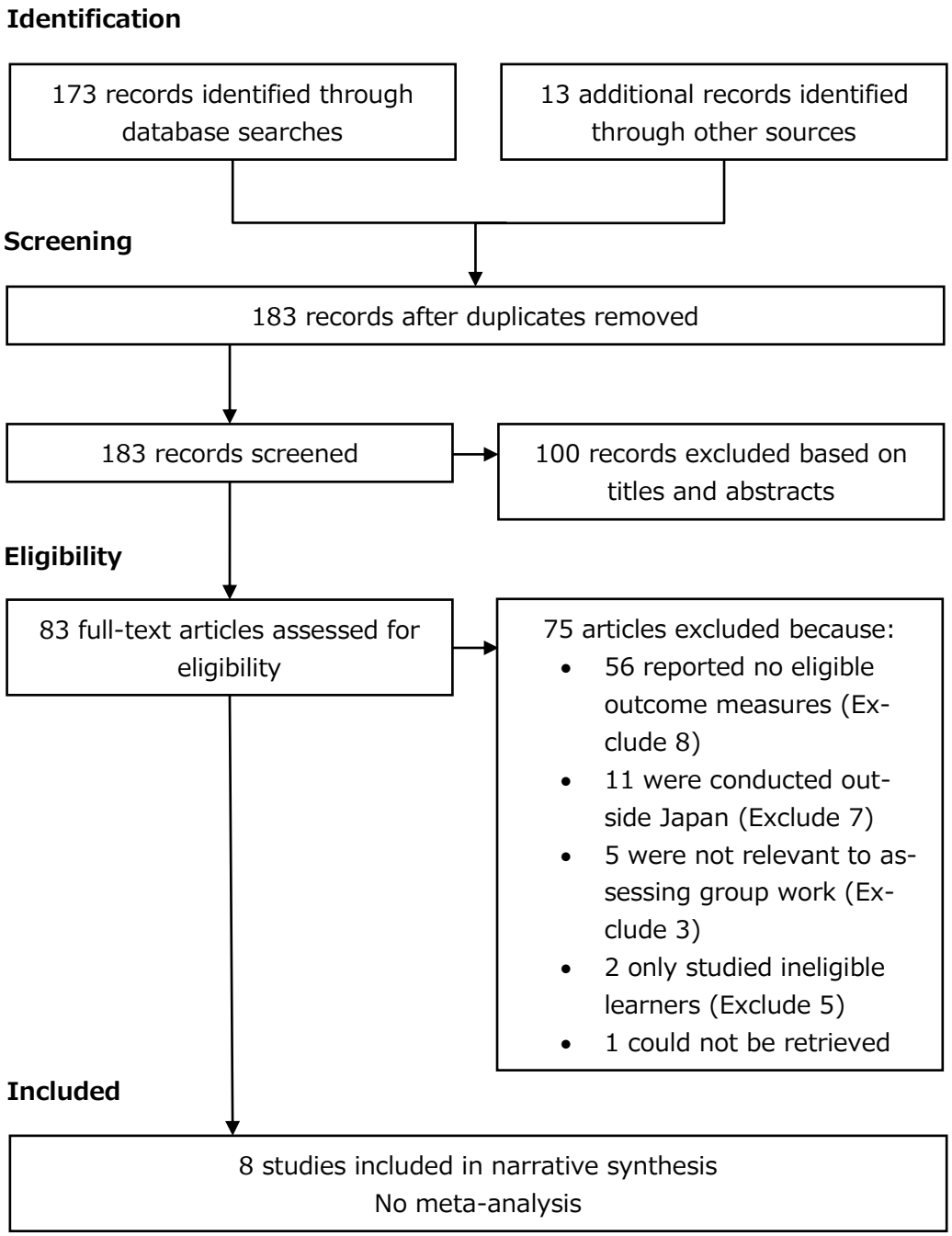
- Exclude 8: no eligible outcome measures reported, n = 56
- Exclude 7: conducted outside Japan, n = 11
- Exclude 3: not relevant to assessing group work, n = 5
- Exclude 5: only investigated ineligible learners, n = 2

Additionally, I excluded one master's dissertation that I was not able to obtain (Onishi, 2012). No electronic copy could be accessed straightforwardly, and the

[17] Some studies were excludable on multiple criteria. I only report one exclusion criterion for each study (only one reason was recorded).

author's university library did not have a hard copy. Thus, a total of 75 studies were excluded at this stage, leaving eight studies that met all inclusion criteria. These studies are listed in Section 0. An overview of the study selection process is shown in the PRISMA flow diagram (Moher et al., 2009), in Figure 4.1.

Figure 4.1 Study selection process



4.2 Included studies

References for the included studies:

1) Kurihara, N. (2017). Peer review in an EFL classroom: Impact on the improvement of student writing abilities. *Asian Journal of Applied Linguistics*, *4*(1), 58–72.
2) Wada, N. (2013). Chūgakkō eigo rīdingu jugyō ni okeru kyōdō gakushū no kōka ni kansuru kenkyū. *Kindaihimejidaigaku Kyōiku Gakubu Kiyō*, (6), 103–112.
3) Nemoto, A. (2012). Cooperative writing in junior high school English class: A comparison between cooperative learning and individual learning. *Tsukuba Eigo Kyōiku*, (33), 117–128.
4) Ishihara, Y. (2016). Adopting 'active learning' (A practical report) in English classes - Peer-tutoring for 11th graders in 'English Communication Ⅱ'. *Hiroshimadaigaku Fuzoku-Chū Kōtō Gakkō Chūtō Kyōiku Kenkyū Kiyō*, *63*, 67–74.
5) Fujishiro, N., & Miyaji, I. (2009). The effectiveness of blended instruction in class on the skills of oral reading and speaking in English. *Nihon Kyōiku Kōgakukai Ronbunshi*, *32*(4), 395–404.
6) Fujishiro, N., & Miyaji, I. (2010). A study on the assessment criteria for the development of speaking skill through utilization of e-learning in English class. *Denshi Jōhō Tsūshin Gakkai Gijutsu Kenkyū Hōkoku. ET, Kyōiku Kōgaku*, *110*(263), 15–18.
7) Takagi, H. (2009). Kyōdō gakushū ga kaeru manabi no katachi: Kōbedaigaku fuzoku Sumiyoshi chūgakkō no jissen o tōshite. *Proceedings of the Annual Meeting of the Japan Society for Educational Information*, *25*, 132–135.
8) Takazawa, I. (2011). Eigo gakuryoku kōjō no tame no ondoku shidō no kufū: Saitotoransurēshon to kyōdō gakushū no shuhō o mochiita ondoku o tōshite. *Kyōiku Jissen Kenkyū*, *21*, 203–208.

4.3 Study characteristics

Characteristics of the included studies are summarized below. Numbers in parentheses identify the included studies by the number assigned in 0. I cover risk of bias assessment results in the following sections. Individual study summaries are presented in 4.6.

4.3.1 Location

Four studies were conducted in urban parts of Japan, including two (2, 7) in Kobe, one (4) in Hiroshima, and one (3) in Ibaraki Prefecture. Three more studies were done in rural parts of Japan, including two (5, 6) in Okayama Prefecture and one (8) in Niigata Prefecture. The location of one study (1) is simply reported to be Japan.

4.3.2 Type of school

Four studies (2, 3, 6, 8) were conducted at public[18] junior high schools, the category of schools attended by a majority of lower secondary[19] students in Japan. Three studies (1, 4, 5) were done at high schools; one public (5) and two private (1, 4). Both public and private schools are common at the upper secondary level[20] in Japan. One study (7) was conducted at a university-affiliated secondary school, catering to both lower- and upper secondary students; an atypical structure in Japan.[21]

4.3.3 Type of publication

Five (3, 5, 6, 7, 8) are research reports published by scholarly societies. One of these studies was prepared for publication from a master's dissertation (3). Two reports (2, 4) came from in-house journals of Japanese academic institutions,[22] journals that have not traditionally circulated outside the publishing institution

[18] Managed by local public bodies. Contrasted with private schools.

[19] International Standard Classification of Education (ISCED) Level 2; Grade 7 (ages 12 and 13) to Grade 9 (ages 14 to 15).

[20] ISCED Level 3; Grade 10 (ages 15 to 16) to Grade 12 (ages 18 to 19).

[21] Education in Japan is based on a 6-3-3 system: six years of compulsory primary education, three years of compulsory, lower secondary education (which I refer to as junior high school), and three years of voluntary, upper secondary education (which I call high school).

[22] CiNii uses the term 'departmental bulletins.' Kamada (2007) called these 'kiyo journals.' JP: 紀要 kiyō

and its affiliates (Kamada, 2007). One study (1) was published in a conventionally-circulated peer-reviewed journal.

4.3.4 Intervention

Two studies (5, 6) delivered a blend of web-based training, pair work, and teacher-centred instruction. Two studies implemented peer review or peer tutoring (1, 4). One study (2) delivered a fixed sequence of pair and group work characterized as collaborative learning. One study (3) delivered a fixed sequence of group work characterized as cooperative learning. One study (7) implemented group presentations. Another study (8) focused on sight translation in groups.

4.3.5 Duration

The interventions ranged from four days (3) to three months (8). All outcomes were measured immediately following the intervention, with no delayed post-tests reported.

4.3.6 Design

Six studies (2, 3, 4, 5, 6, 8) compared the effects of two alternative conditions on intact classes. No prospective generation of groups, random or otherwise, was reported for these studies. In one of these studies (3), intact classes were assigned to conditions by means of random allocation. The method of random allocation is not reported. Two studies (1, 7) investigated one intervention across a single group of participants and compared the results of pre- and post-tests. In six studies (1, 2, 3, 4, 5, 6) interventions were delivered to whole classes. In two studies (7, 8) interventions were delivered to groups made up of a subset of students from a class.

4.3.7 Outcome

Two studies (4, 8) measured English language proficiency. Two (5, 6) assessed unstructured English speech with an interview test; of which one (5) also assessed structured speech with a recitation test. One study (1) assessed essay-writing performance. One (2) measured comprehension. Another study (3) counted correct sentences or clauses (error-free T-units) in English writing. One study (7) assessed performance on a test of English third person singular verb forms in the present tense. Outcome measures are itemized in Table 4.2.

4.3.8 Size

The study with the least participants (1) involved 25 students, whereas the study with the most reported participants (6) involved 162 students.

4.3.9 Findings

Four studies (1, 5, 6, 7) found in favour of the group work intervention. Three studies (3, 4, 8) did not detect a difference between a group work intervention and an alternative. The findings of one study (2) were not clear. In this last study, the author analysed the 20-hour intervention as 2 sets of 10 hours. Analysis of the first 10 hours revealed a statistically significant difference between groups in favour of group work. No statistically significant difference was detected in the second 10 hours. No overall analysis was reported, and clear conclusions were not drawn. A summary of descriptive data for the studies, including findings, is presented in Table 4.1. A summary of statistical data is presented in Table 4.2, including effect sizes where reported.

Table 4.1 Summary of descriptive data from the included studies

Reference	Design	Intervention	School	Participants	Grade	Duration	Results
Kurihara (2017)	Single group, pre/post	Peer review	High school	25	Grade 12	5 periods (1 period = 50 minutes)	Favoured group work
Wada (2013)	Non-randomized comparison	Collaborative learning	Public junior high school	117	Grade 9	20 hours	Not clear[23]
Nemoto (2012)	Randomized trial	Cooperative learning	Public junior high school	116	Grade 9	4 days	No difference detected
Ishihara (2016)	Non-randomized comparison	Peer tutoring	University-affiliated high school	Not reported (5 classes)	Grade 11	4 periods	No difference detected
Fujishiro and Miyaji (2009)	Non-randomized comparison	Web-based training with pair work	Public (prefectural) high school	39	Grade 11	6 periods	Favoured group work
Fujishiro and Miyaji (2010)	Non-randomized comparison	Web-based training with pair work	Public (municipal) junior high school	162	Grade 9	6 periods	Favoured group work
Takagi (2009)	Single group, pre/post	Group presentations	University-affiliated secondary school	29	Grade 7	6 periods	Favoured group work
Takazawa (2011)	Non-randomized comparison	Sight translation in groups	Public junior high school	26	Grade 9	3 months	No difference detected

[23] Wada analysed the 20-hour intervention as 2 sets of 10 hours. Analysis of the first 10 hours revealed a statistically significant difference between groups in favour of group work. No statistically significant difference was detected in the second 10 hours. No overall analysis was reported.

Table 4.2 Summary of statistical data from the included studies

Reference	**Comparator 1: Mean (SD)***	**Comparator 2: Mean (SD)***	**Effect size***
Kurihara (2017)	Essay-writing with peer feedback	n/a	
	Essays		
	Overall scores		Δ = 0.57 (Glass's delta)
	Pre-test: 3.32 (1.11)		
	Post-test: 3.96 (1.07)		
	Relevance		Δ = 1.02
	Pre-test: 1.56 (0.63)		
	Post-test: 2.2 (0.57)		
	Organization		Δ = 0.9
	Pre-test: 1.76 (0.49)		
	Post-test: 2.2 (0.58)		
	Cohesion		Δ = 0.64
	Pre-test: 1.72 (0.45)		
	Post-test: 2.06 (0.45)		
	Vocabulary		Δ = 0.92
	Pre-test: 1.36 (0.66)		
	Post-test: 1.96 (0.58)		
	Grammar		Δ = 0.71
	Pre-test: 1.2 (0.49)		
	Post-test: 1.54 (0.49)		
	Punctuation		Δ = 0.41
	Pre-test: 2.46 (0.58)		
	Post-test: 2.7 (0.45)		
	Spelling		Δ = 0.27
	Pre-test: 2.76 (0.45)		
	Post-test: 2.88 (0.33)		
Wada (2013)	Fixed sequence of pair and group work	Teacher-centred grammar translation	
	Comprehension	*Comprehension*	
	Vocabulary	*Vocabulary*	
	Mid-test:** 12.46 (4.9)	Mid-test: 9.08 (4.96)	
	Post-test: 10.57 (2.91)	Post-test: 9.7 (4.06)	
	Sentence	*Sentence*	
	Mid-test: 13.05 (5.85)	Mid-test: 9.76 (5.37)	
	Post-test: 12.23 (5.14)	Post-test: 13.2 (5.93)	

	Content Mid-test: 11.14 (5.52) Post-test: 22.39 (8.59) *Cohesion* Mid-test: 13.55 (6.3) Post-test: 9.18 (5.51)	*Content* Mid-test: 8.84 (4.39) Post-test: 19.23 (10.03) *Cohesion* Mid-test: 11.84 (6.15) Post-test: 10.6 (6)	
Nemoto (2012)	Cooperative learning *Number of error-free T-units* *Class 1* Pre-test: ≈ 11.5*** Post-test: ≈ 15 *Class 2* Pre-test: ≈ 11.75 Post-test: ≈ 11.25	Individual learning *Number of error-free T-units* *Class (control)* Pre-test: ≈ 11.5 Post-test: ≈ 13.75	η_p^2 = .042
Ishihara (2016)	Peer tutoring *English language proficiency* *Class (experimental)* First pre-test: 65.3 Second pre-test: 56.8 Post-test: 51	Business as usual *English language proficiency* *Class A* First pre-test: 72 Second pre-test: 62 Post-test: 45 *Class B* First pre-test: 69.3 Second pre-test: 64.5 Post-test: 56.3 *Class C* First pre-test: 65.5 Second pre-test: 57.6 Post-test: 45.8 *Class D* First pre-test: 60 Second pre-test: 50.9 Post-test: 40.6	
Fujishiro and Miyaji (2009)	Web-based training, pair work, and teacher-centred instruction *Recitation (post-test results)* *Pronunciation* Word: 3.6 (0.6) Sentence: 3.5 (0.6) Clause: 3.6 (0.6)	Business as usual *Recitation (post-test results)* *Pronunciation* Word: 3.1 (0.6) Sentence: 3.1 (0.6) Clause: 3.1 (0.6)	

	Intonation Word: 3.1 (0.6) Sentence: 3 (0.6) Clause: 3 (0.6) *Pauses* Sentence: 3.1 (0.6) Clause: 3.1 (0.7) *Liaison* Sentence: 3 (0.5) Clause: 3 (0.5) *Stress* Word: 3.3 (0.6) Sentence: 3.2 (0.6) Clause: 3.3 (0.6) *Sentence Stress* Sentence: 3.1 (0.5) Clause: 3.1 (0.6) *Overall* Rater's impression: 3.4 (0.5) *Interview test* *Informativeness* Pre-test: 3.5 (0.9) Post-test: 3.9 (0.4) *Ideas* Pre-test: 3.6 (0.9) Post-test: 3.9 (0.4) *Thematic conformity* Pre-test: 3.8 (0.8) Post-test: 4 (0.4) *Pronunciation* Pre-test: 3.0 (0.9) Post-test: 3.4 (0.5) *Volume* Pre-test: 3.3 (0.7) Post-test: 3.6 (0.7)	*Intonation* Word: 2.7 (0.5) Sentence: 2.6 (0.5) Clause: 2.5 (0.6) *Pauses* Sentence: 2.8 (0.5) Clause: 2.6 (0.7) *Liaison* Sentence: 2.5 (0.5) Clause: 2.6 (0.4) *Stress* Word: 2.9 (0.6) Sentence: 2.9 (0.6) Clause: 2.8 (0.6) *Sentence Stress* Sentence: 2.8 (0.6) Clause: 2.7 (0.5) *Overall* Rater's impression: 2.9 (0.5) *Interview test* *Informativeness* Pre-test: 3.3 (0.5) Post-test: 3.6 (0.4) *Ideas* Pre-test: 3.3 (0.4) Post-test: 3.6 (0.4) *Thematic conformity* Pre-test: 3.4 (0.5) Post-test: 3.7 (0.4) *Pronunciation* Pre-test: 2.7 (0.4) Post-test: 3 (0.3) *Volume* Pre-test: 3 (0.5) Post-test: 3.3 (0.4)
Fujishiro and Miyaji (2010)	Web-based training, pair work, and teacher-centred instruction *Interview test* *Vocabulary* Pre-test: 2.4 (1.2) Post-test: 2.6 (1.2)	Teacher-centred listening instruction

	Pronunciation Pre-test: 2.3 (1.1) Post-test: 2.6 (1.1) *Stress* Pre-test: 2.1 (0.9) Post-test: 2.4 (1) *Sentence stress* Pre-test: 2.1 (0.9) Post-test: 2.3 (0.9) *Intonation* Pre-test: 2.1 (0.9) Post-test: 2.3 (0.9) *Informativeness* Pre-test: 2.5 (1.3) Post-test: 2.8 (1.2) *Ideas* Pre-test: 2.5 (1.2) Post-test: 2.8 (1.2) *Thematic conformity* Pre-test: 2.4 (1.2) Post-test: 2.7 (1.2) *Overall impression* Pre-test: 2.3 (1.1) Post-test: 2.6 (1.2)	
Takagi (2009)	Group work *Tests of English third person singular verb forms in the present tense* *Group 1* Pre-test: 53.6 Post-test: 92.9 *Group 2* Pre-test: 57.1 Post-test: 89.3 *Group 3* Pre-test: 67.9 Post-test: 96.4	n/a
Taka-zawa (2011)	Sight translation in groups *English language proficiency* Pre-test: 76.5 (7.6) Post-test: 85.1 (6.7)	Grammar-translation *English language proficiency* Pre-test: 76.8 (7.7) Post-test: 81.1 (8.3)

Component test	*Component test*
Vocabulary	*Vocabulary*
Pre-test (Lesson 1): 8.7 (2.2)	Pre-test (Lesson 1): 9 (2.5)
Post-test (Lesson 2): 12.2 (2.3)	Post-test (Lesson 2): 10.4 (2.5)
Post-test (Lesson 3): 9.4 (4.8)	Post-test (Lesson 3): 7.7 (2.3)
Grammar	*Grammar*
Pre-test (Lesson 1): 9.1 (2.5)	Pre-test (Lesson 1): 8.3 (2.4)
Post-test (Lesson 2): 9.5 (1.3)	Post-test (Lesson 2): 4.7 (4.3)
Post-test (Lesson 3): 8.8 (1.3)	Post-test (Lesson 3): 5.4 (3.8)
Scrambled sentences	*Scrambled sentences*
Pre-test (Lesson 1): 10.4 (4.1)	Pre-test (Lesson 1): 8.9 (3.8)
Post-test (Lesson 2): 11.7 (0.5)	Post-test (Lesson 2): 10 (3.1)
Post-test (Lesson 3): 9.2 (0.8)	Post-test (Lesson 3): 6.8 (3.5)
Pronunciation	*Pronunciation*
Pre-test (Lesson 1): 12.3 (2.5)	Pre-test (Lesson 1): 12.5 (1.8)
Post-test (Lesson 2): 13.3 (1.4)	Post-test (Lesson 2): 13.1 (2)
Post-test (Lesson 3): 14.3 (1)	Post-test (Lesson 3): 13.6 (1)
Accent	*Accent*
Pre-test (Lesson 1): 10.2 (2.2)	Pre-test (Lesson 1): 10.5 (2.6)
Post-test (Lesson 2): 11.5 (1.2)	Post-test (Lesson 2): 10.9 (2.4)
Post-test (Lesson 3): 11.7 (1.3)	Post-test (Lesson 3): 10.5 (1.7)

	Comprehension Pre-test (Lesson 1): 8.5 (5.3) Post-test (Lesson 2): 13.3 (0.4) Post-test (Lesson 3): 10.6 (1.5)	*Comprehension* Pre-test (Lesson 1): 8.6 (4.6) Post-test (Lesson 2): 13.1 (2.3) Post-test (Lesson 3): 9.8 (2)

**Where applicable and reported*

***Wada reported a test conducted after 10 hours of a 20-hour intervention. I have called that test 'mid-test'.*

****The data were not reported as text. These results were interpreted from a bar chart and should be treated with caution.*

4.4 Risk of bias within studies

Results of within-studies risk of bias assessment are presented in this section. Risk of bias assessment was conducted using the EPHPP tool. I report results organized in the format of the EPHPP tool. Numbers in parentheses identify the included studies by the number assigned in 0.

4.4.1 Selection bias

Assessing the extent to which researchers had tried to avoid selection bias was difficult. No reports stated the target population. It is not possible to objectively assess the representativeness of participants relative to an unknown target population (Chalmers, 2019). Since no reports stated the target population, I rated risk of selection bias as 'moderate' (as reported in 3.9) for all included studies. Thus, a meaningful assessment of risk of selection bias was not possible in this review, due to incomplete reporting. This serves to stress the benefits of using reporting guidelines such as STROBE (for observational studies; von Elm et al., 2008), CONSORT (for randomized trials; Moher et al., 2012), and PRISMA (for systematic reviews; Moher et al., 2009).

4.4.2 Study design

Six studies (2, 3, 4, 5, 6, 8) were comparisons of pre-existing groups. Of these, one study (3) randomly assigned alternative interventions to pre-existing groups. This randomized design had a 'low' risk of bias rating. The remaining five comparisons (2, 4, 5, 6, 8) were rated 'moderate' for risk of bias in study design. Two studies (1, 7) investigated one intervention across a single group of participants and compared the results of pre- and post-tests. Lacking comparisons, the risk of bias in the design of these studies was rated 'high'.

4.4.3 Confounders

The problem of "unrecorded, unknown or unknowable" (Chalmers, 2019, p. 89) confounders means that it is impossible to directly assess the extent to which researchers have controlled for them. Evaluating study design provides an approximation of risk of bias from confounders. Relative to non-randomized designs, other things being equal randomized designs reduce the risk of bias from confounders (Miettinen, 1983). For this reason, I rated the risk of bias from

confounders of one randomized trial (3) as ‘low’ and of the remaining unrandomized studies (1, 2, 4, 5, 6, 7, 8) as ‘high’.

4.4.4 Blinding and allocation concealment

No reports discussed blinding, whether blinding of assessors at outcome assessment or blinding of participants to research questions. Allocation concealment was also not discussed in any of the reports. Guidance for the EPHPP tool states that where blinding and allocation concealment are not reported, this item should be rated ‘moderate’. In line with the guidance (B. H. Thomas, Ciliska, Dobbins, & Micucci, 2010), I rated risk of bias for blinding and allocation concealment as ‘moderate’ for all included studies.

4.4.5 Data collection method

One study reported test–retest reliability. The risk of bias in data collection for this study (5) was rated ‘moderate’. One study used a count of error-free T-units, a measure that has been demonstrated to be unreliable because of the difficulty of reliably defining an error (Hirano, 1988). The risk of bias in data collection for this study (3) was rated ‘high’. The remaining studies used internal examinations (designed by schools) or researcher-designed tools. The risk of bias in data collection for these studies (1, 2, 4, 6, 7, 8) was rated ‘high’.

4.4.6 Withdrawals and dropouts

Over four-fifths (81%) of participants completed one study. The risk of bias from withdrawals and dropouts for this study (3) was rated ‘low’. For the remaining studies, a retention rate could not be calculated because the necessary figures were not reported. Guidance for the EPHPP tool states that where withdrawals and dropouts are not reported, this item should be rated ‘high’. In line with the guidance, I rated risk of bias for withdrawals and dropouts as ‘high’ for the remaining studies (1, 2, 4, 5, 6, 7, 8).

4.4.7 Overall

The overall risk of bias for one study (3) was rated ‘moderate’ on the EPHPP tool. All remaining studies (1, 2, 4, 5, 6, 7, 8) were rated ‘high’ for overall risk of bias. In line with the guidance, studies were rated ‘high’ overall where the ratings for

other items contained no 'low' ratings. A summary of risk of bias within studies is presented in Table 4.3.

Table 4.3 Risk of bias within studies

Reference	**Selection bias**	**Study design**	**Confounders**	**Blinding**	**Data collection**	**Withdrawals and dropouts**	**Overall**
Kurihara (2017)	Moderate	High	High	Moderate	High	High	*High*
Wada (2013)	Moderate	Moderate	High	Moderate	High	High	*High*
Nemoto (2012)	Moderate	Low	Low	Moderate	High	Low	*Moderate*
Ishihara (2016)	Moderate	Moderate	High	Moderate	High	High	*High*
Fujishiro and Miyaji (2009)	Moderate	Moderate	High	Moderate	Moderate	High	*High*
Fujishiro and Miyaji (2010)	Moderate	Moderate	High	Moderate	High	High	*High*
Takagi (2009)	Moderate	High	High	Moderate	High	High	*High*
Takazawa (2011)	Moderate	Moderate	High	Moderate	High	High	*High*

4.5 Risk of bias across studies

The overall ratings on the EPHPP tool indicate a strong risk of bias for the included studies taken together. Seven out of eight studies were rated 'high' for overall risk of bias. Yet, the overall ratings on the EPHPP tool should not stand alone. Overall ratings are simply derived by counting the number of 'low' ratings on the other items. The items are not weighted in any way, despite arguably not being equal. Unbiased selection of participants and blinding of assessors at outcome assessment add little to the trustworthiness of a study if a study's design is not fit for purpose or data collection methods are invalid or unreliable (Gorard, 2014). Gorard places the most weight on study design, data collection methods, and withdrawals and dropouts. These three are considered in turn. On risk of bias in study design, the greatest portion of the included studies, five out of eight, were rated 'moderate'. On risk of bias in data collection, seven out of eight studies were rated 'high'. On risk of bias for withdrawals and dropouts, seven out of eight studies were rated 'high'.

In addition to the EPHPP tool ratings, I present risk of bias across studies in the form of the following proportions:

1) Six out of eight studies used control groups.
2) The proportion of studies in which participants' group allocations were concealed at outcome assessment cannot be estimated, due to incomplete reporting.
3) No studies used prospectively generated groups.
4) One out of eight studies randomly assigned pre-existing groups to conditions.

4.6 Individual study summaries

This section gives a summary of each included study. It outlines each study according to the PICO format (participants, interventions, comparisons and outcomes), summarizes the main results, and gives the conclusions of the study's authors. Tables Table 4.4 to Table 4.11 show the risk of bias assessment results for each study, with reasons for each rating.

4.6.1 Kurihara: Students review each other's writing

Reference

Kurihara, N. (2017). Peer review in an EFL classroom: Impact on the improvement of student writing abilities. *Asian Journal of Applied Linguistics*, *4*(1), 58–72.

Research Question

Kurihara (2017, p. 60):

1. Does peer review affect student writing abilities?
2. If so, what aspects of writing does peer review affect?
3. Is there a relationship between student attitude toward peer review and improvement in student writing performance?

Participants

Kurihara selected one class of 25 students (8 boys and 17 girls) aged 17-18 in Grade 12 at an academic high school in Japan.

Interventions and comparisons

No comparison was reported. All participants received peer feedback training for five periods of 50 minutes each. They did two rounds of essay-writing, each followed by written and oral feedback in pairs. The feedback sessions lasted 40 minutes each. The students also received teacher feedback.

Outcomes

Kurihara administered pre- and post-essay tests selected from past Japanese university entrance examinations.

Results

Paired *t*-tests comparing the pre- and post-test scores found statistically significant differences. Average essay-writing performance improved from pre- to post-test.

Author's conclusions

Kurihara (2017, p. 60):

> The study ... found that incorporating peer review in addition to teacher feedback contributes to the improvement of students' ability to write a new text ... It was also found that the degree of student improvement in writing performance seems to be related to how much they trust peer feedback.

Table 4.4 Risk of bias assessment for Study 1

Selection bias	**Study design**	**Confounders**	**Blinding**	**Data collection**	**Withdrawals and dropouts**
Moderate	High	High	Moderate	High	High
Target population not reported	No comparison	No randomization	Blinding and allocation concealment not reported	Validity and reliability not reported	Numbers not reported in text or figures

4.6.2 Wada: Read together or listen to the teacher?

Reference

Wada, N. (2013). Chūgakkō eigo rīdingu jugyō ni okeru kyōdō gakushū no kōka ni kansuru kenkyū. *Kindaihimejidaigaku Kyōiku Gakubu Kiyō*, (6), 103–112.

Research Question

Wada (2013, p. 104; author's translation):

1) In junior high school English reading lessons, which is more effective, collaborative learning or teacher-centred grammar-translation?
2) How is the influence of collaborative learning and teacher-centred grammar-translation exerted on high-, medium-, and low-performing learners?

Participants

Wada selected three classes totalling 117 students (54 boys and 63 girls) in Grade 9 at a public junior high school in Kobe.

Interventions and comparisons

Two conditions were compared. The experimental group was two classes totalling 77 students who received a fixed sequence of pair and group work in two interventions of 10 hours each. The control group was one class of 38 students who for the same time received instruction characterized as teacher-centred grammar-translation.

Outcomes

Wada administered four internal English comprehension tests: a vocabulary test, a sentence test, a cohesion test, and a content test.

Results

In the first 10-hour intervention, paired *t*-tests found statistically significant differences between the two groups on the vocabulary, sentence, and content tests, but not on the cohesion test. No statistically significant differences were found in the second 10-hour intervention.

Author's conclusions

Based on the comprehension test results for the first intervention, Wada linked collaborative techniques with better text comprehension compared to grammar-translation. He suggested that students being absent because of sickness and external exams might have affected the results of the second 10-hour intervention.

Table 4.5 Risk of bias assessment for Study 2

Selection bias	**Study design**	**Confounders**	**Blinding**	**Data collection**	**Withdrawals and dropouts**
Moderate	Moderate	High	Moderate	High	High
Target population not reported	Groups not prospectively generated, nor interventions randomly assigned	No randomization	Blinding and allocation concealment not reported	Validity and reliability not reported	Numbers not reported in text or figures

4.6.3 Nemoto: Write together or write alone?

Reference

Nemoto, A. (2012). Cooperative writing in junior high school English class: A comparison between cooperative learning and individual learning. *Tsukuba Eigo Kyōiku*, (33), 117–128.

Research Question

Nemoto (2012, p. 119; author's translation):

1) Compared to individual learning, is cooperative learning an effective learning method for writing English?
2) Does cooperative learning affect the junior high school students' willingness to learn in writing activities?

Participants

Nemoto selected three classes totalling 116 students in Grade 9 at a public junior high school in Ibaraki.

Interventions and comparisons

Two conditions were compared. The experimental group was two classes who received a fixed sequence of group work over four days. The control group was one class who received the same fixed sequence except that group work was replaced by individual work, over the same four days. Nemoto randomly assigned pre-existing classes to the group work condition and the individual work condition. The random assignment method is not reported.

Outcomes

Nemoto administered descriptive writing tasks scored for number of correct sentences or clauses (error-free T-units).

Results

On a two-way analysis of variance examining the influence of group work and individual work on the pre- and post-test results (number of error-free T-units), main effects for both technique (group or individual work) and test results were not statistically significant.

Author's conclusions

Nemoto deemed the results inconclusive; recommending a longer study.

Table 4.6 Risk of bias assessment for Study 3

Selection bias	**Study design**	**Confounders**	**Blinding**	**Data collection**	**Withdrawals and dropouts**
Moderate	Low	Low	Moderate	High	Low
Target population not reported	Interventions randomly assigned to pre-existing groups	Randomization	Blinding and allocation concealment not reported	Problematic tool (count of error-free T-units)	81% of participants completed study

4.6.4 Ishihara: Students tutor each other

Reference

Ishihara, Y. (2016). Adopting 'active learning' (A practical report) in English classes - Peer-tutoring for 11th graders in 'English Communication Ⅱ'. *Hiroshimadaigaku Fuzoku-Chū Kōtō Gakkō Chūtō Kyōiku Kenkyū Kiyō*, *63*, 67–74.

Research Question

Not stated. I infer that Ishihara sought to test if peer tutoring activities would be more effective than business as usual in terms of high school students' performance on English proficiency tests.

Participants

Ishihara selected five classes of students in Grade 11 (a complete year group) at Hiroshima University Senior High School, a public university-affiliated high school in Hiroshima. Participant numbers are not reported. I have estimated between 105 and 200[24] students.

Interventions and comparisons

Two conditions were compared. The experimental group was one class of 43 participants who did peer tutoring activities in groups of four or five for six periods.[25] The control group was the other four classes in the year group. I infer that during the same period the teachers of these classes used their usual techniques. There is not enough information to infer what those techniques were.

Outcomes

Ishihara reported averages for the five classes on three consecutive internal tests of English language proficiency.

[24] The legal maximum class size in Japan's secondary schools is 40 students. Since two classes of 20 students or less can be merged into a single class, I assumed that each class would have between 21 and 40 students. I multiplied these numbers by the number of classes (5) to derive upper and lower estimates.

[25] Length not reported. Lessons at the secondary level in Japan typically last 50 minutes.

Results

A *t*-test which I interpret as having compared the pre- and post-test scores for the experimental group found no statistically significant differences. On the two internal tests of English language proficiency prior to the intervention, the class that formed the experimental group had ranked fourth out of the five classes in the year group. On the internal test after the intervention, the class improved relative to the other classes; moving up to second place.

Author's conclusions

Ishihara did not come to a firm conclusion but noted the relative improvement of the class that did peer tutoring activities.

Table 4.7 Risk of bias assessment for Study 4

Selection bias	**Study design**	**Confounders**	**Blinding**	**Data collection**	**Withdrawals and dropouts**
Moderate	Moderate	High	Moderate	High	High
Target population not reported	Groups not prospectively generated, nor interventions randomly assigned	No randomization	Blinding and allocation concealment not reported	Validity and reliability not reported	Numbers not reported in text or figures

Reviewer's note

Ishihara's report failed to set out a research question, report participant numbers, describe the comparator condition, or say what the *t*-test measured. Consequently, a large proportion of the above summary relies on inference. As such, it should be treated with caution. Ishihara's report serves to highlight the importance of reporting guidelines. One guideline that is relevant to intervention studies such as this is the *Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide* (Hoffmann et al., 2014).

4.6.5 Fujishiro and Miyaji: Web-based training and pair work or business as usual?

Reference

Fujishiro, N., & Miyaji, I. (2009). The effectiveness of blended instruction in class on the skills of oral reading and speaking in English. *Nihon Kyōiku Kōgakukai Ronbunshi*, *32*(4), 395–404.

Research Question

Fujishiro and Miyaji sought to test if a blend of web-based training (WBT), pair work, and teacher-centred instruction would affect oral ability in a high school EFL setting.

Participants

Fujishiro and Miyaji selected four classes totalling 39 students who had elected to take 'English Expressions' in Grade 11 at a prefectural high school in Okayama.

Interventions and comparisons

Two conditions were compared. The experimental group was two classes totalling 18 students who received a fixed sequence of WBT, pair work, and teacher-centred instruction for six periods over one month. The control group was two classes totalling 21 students who were taught the same content using the teacher's usual techniques. The report does not say what those techniques were.

Outcomes

Fujishiro and Miyaji administered an interview test, a recitation test, and an English language proficiency test.

Results

Paired *t*-tests comparing the pre- and post-test recitation scores found statistically significant results on all items for the experimental group but only found statistically significant results on rater's general impression for the control group. Paired *t*-tests found statistically significant differences between the two groups on 11 of 16 items on the interview test results. The English language proficiency test results were not reported.

Author's conclusions

Fujishiro and Miyaji concluded that a blend of WBT, pair work, and teacher-centred instruction 1) improved recitation, 2) deepened the content of, and improved accuracy in, unstructured speech, 3) improved listening skills, and 4) made classes easier to understand.

Table 4.8 Risk of bias assessment for Study 5

Selection bias	**Study design**	**Confounders**	**Blinding**	**Data collection**	**Withdrawals and dropouts**
Moderate	Moderate	High	Moderate	Moderate	High
Target population not reported	Groups not prospectively generated, nor interventions randomly assigned	No randomization	Blinding and allocation concealment not reported	Test–retest reliability reported. Recitation ICC > .91, Interview ICC > .98	Numbers not reported in text or figures

4.6.6 Fujishiro and Miyaji: WBT and pair work or teacher-centred listening?

Reference

Fujishiro, N., & Miyaji, I. (2010). A study on the assessment criteria for the development of speaking skill through utilization of e-learning in English class. *Denshi Jōhō Tsūshin Gakkai Gijutsu Kenkyū Hōkoku. ET, Kyōiku Kōgaku*, *110*(263), 15–18.

Research Question

Fujishiro and Miyaji sought to test if a blend of web-based training (WBT), pair work, and teacher-centred instruction would affect speaking ability in a junior high school EFL setting.

Participants

Fujishiro and Miyaji selected five classes totalling 162 students in Grade 9 at a municipal junior high school in Okayama.

Interventions and comparisons

Two conditions were compared. The experimental group was three classes totalling 94 students who received a fixed sequence of WBT, pair work, and teacher-centred instruction for six periods over three weeks. The control group was two classes totalling 68 students who received teacher-centred listening instruction using a CD.

Outcomes

Fujishiro and Miyaji administered an interview test.

Results

Paired *t*-tests comparing the pre- and post-test interview scores found statistically significant results on 10 out of 16 items for the experimental group but only found statistically significant results on 1 item for the control group.

Author's conclusions

Fujishiro and Miyaji (2010, p. 18; author's translation):

> Among the 10 items that met statistical significance, the five items of vocabulary, pronunciation, stress, sentence stress, and intonation can be inferred to be the result of individual learning using e-learning materials. In addition, the three items of informativeness, opinion, and theme suitability are considered to be the result of collaborative learning.

Table 4.9 Risk of bias assessment for Study 6

Selection bias	**Study design**	**Confounders**	**Blinding**	**Data collection**	**Withdrawals and dropouts**
Moderate	Moderate	High	Moderate	High	High
Target population not reported	Groups not prospectively generated, nor interventions randomly assigned	No randomization	Blinding and allocation concealment not reported	Validity and reliability not reported	Numbers not reported in text or figures

4.6.7 Takagi: Learning vocabulary together

Reference

Takagi, H. (2009). Kyōdō gakushū ga kaeru manabi no katachi: Kōbedaigaku fuzoku Sumiyoshi chūgakkō no jissen o tōshite. *Proceedings of the Annual Meeting of the Japan Society for Educational Information*, *25*, 132–135.

Research Question

Not stated. It is inferred that Takagi aimed to test if group work changed test performance and attitudes.

Participants

Takagi selected 29 students from one class of 37 in Grade 7 at Kobe University Secondary School, a public university-affiliated, combined junior- and senior high school in Kobe.

Interventions and comparisons

No comparison was reported. All participants did group work, including group presentations using computers and projectors, for six periods focusing on English third person singular verb forms in the present tense.

Outcomes

Before and after the intervention, Takagi administered internal tests of English third person singular verb forms in the present tense.

Results

Performance on the tests of English third person singular verb forms in the present tense improved from pre- to post-test. No statistical analysis was reported.

Author's conclusions

Takagi concluded that group work, including group presentations using computers and projectors, leads to deeper understanding and improvements in attitude to group work but causes lower performing students to rely more on higher performing classmates.

Table 4.10 Risk of bias assessment for Study 7

Selection bias	**Study design**	**Confounders**	**Blinding**	**Data collection**	**Withdrawals and dropouts**
Moderate	High	High	Moderate	High	High
Target population not reported	No comparison	No randomization	Blinding and allocation concealment not reported	Validity and reliability not reported	Numbers not reported in text or figures

Reviewer's note

Takagi's report did not discuss a research question. It is important to report information as basic as the question the researcher intends to address. A reporting guideline that might be relevant is the *Guidelines for conducting and reporting mixed research in the field of counseling and beyond* (Leech & Onwuegbuzie, 2010).

4.6.8 Takazawa: Translate together or listen to the teacher?

Reference

Takazawa, I. (2011). Eigo gakuryoku kōjō no tame no ondoku shidō no kufū: Saitotoransurēshon to kyōdō gakushū no shuhō o mochiita ondoku o tōshite. *Kyōiku Jissen Kenkyū*, *21*, 203–208.

Research Question

Takazawa (2011, p. 204; author's translation):

> Hypothesis: If sight translation and cooperative learning methods are used for recitation training with junior high school students, their willingness to learn will increase, their understanding of English meaning will deepen, and English language skills such as vocabulary, grammar and content comprehension will improve.

Participants

Takazawa selected higher performing students from two classes, totalling 26 students, in Grade 9 at a public junior high school in Niigata.

Interventions and comparisons

Two conditions were compared. The experimental group was 12 students from one class who did sight translation in groups as recitation training over three months. The control group was 14 students from another class who for the same time received instruction characterized as grammar-translation.

Outcomes

Takazawa administered internal English language proficiency tests.

Results

A paired *t*-test found no statistically significant difference between the two groups overall on the language proficiency tests. Two out of six test components; vocabulary and grammar, met significance.

Author's conclusions

Noting that the results were not generalizable beyond higher performing students, Takazawa concluded that it is not clear if sight translation in groups for recitation training improves reading comprehension.

Table 4.11 Risk of bias assessment for Study 8

Selection bias	**Study design**	**Confounders**	**Blinding**	**Data collection**	**Withdrawals and dropouts**
Moderate	Moderate	High	Moderate	High	High
Target population not reported	Groups not prospectively generated, nor interventions randomly assigned	No randomization	Blinding and allocation concealment not reported	Validity and reliability not reported	Numbers not reported in text or figures

4.7 Synthesis of results

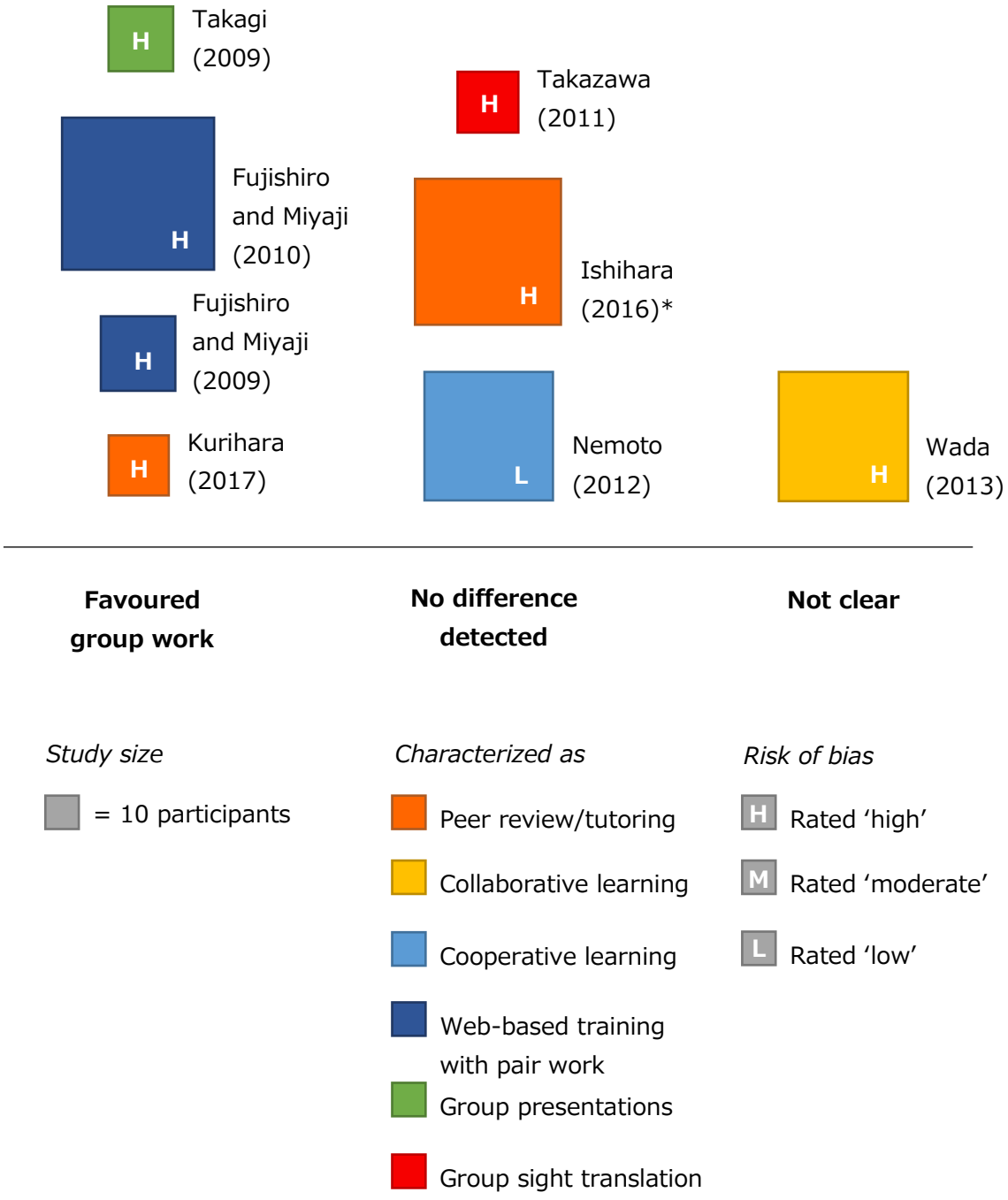
It was impossible to synthesize results statistically because data necessary for meta-analysis was not reported. I have therefore prepared a narrative synthesis. A summary of the findings is presented in Figure 4.2, in the form of data on outcomes, study size, how the author's characterized the intervention of interest, and risk of bias assessment for the included studies. As Figure 4.2 shows, four studies found in favour of group work. One of these studies (Fujishiro & Miyaji, 2010) had 162 participants.[26] The remaining three studies had fewer than 40 participants. All four were rated 'high' for overall risk of bias. Three studies did not detect a difference between the group work intervention and an alternative. Two of these studies are counted as having more than one hundred participants; but since the number of participants for Ishihara's (2016) study has been estimated, that number should be treated with caution. The remaining study that did not detect a difference (Takazawa, 2011) had 26 participants. One of the studies that did not detect a difference (Nemoto, 2012) was rated 'moderate' for overall risk of bias. The remaining two studies had 'high' overall risk of bias ratings. For one study (Wada, 2013), I could not determine whether the findings favoured group work, favoured an alternative, or did not find a difference. The author did not come to a clear conclusion and a finding could not be inferred from the reported data. The overall risk of bias rating was 'high' for that study.

The highest number of included studies (four out of eight) found in favour of group work. Since no studies found in favour of an alternative, there appears cause for optimism regarding the effectiveness of group work in this context. However, Figure 4.2 helps to illustrate that there is only a small collective difference between studies favouring group work and studies detecting no difference. Collectively, they do not differ substantially in terms of the number of participants. All studies that favoured group work and most studies that detected no difference had a 'high' overall risk of bias rating. The lack of a clear difference, both in terms of number of participants and risk of bias, between studies that favoured group work and studies that detected no difference means that these results should be treated as equivocal. This interpretation can only be tentative, however. The body

[26] Sample sizes were not reported in any studies. In place of sample size, I have reported the number of participants that researchers set out to study. These numbers do not account for withdrawals and dropouts. They should not be treated as an accurate reflection of the weight of any findings.

of studies lacks rigorous designs and reports are incomplete; both of which complicate interpretation.

Figure 4.2 Outcome, size, characterization, and risk of bias for the included studies



Number of participants estimated here as 150. See 4.6.4, Participants.

Chapter 5 Discussion

This chapter discusses the evidence synthesized in this systematic review. In Section 5.1, I group studies by design and discuss the findings in those terms. Limitations of this review are considered in Section 5.2. I address how I searched and how I did not search, as well as what steps I took to assure the quality of this review, and what was beyond the scope of this review. I also discuss problems of publication bias, the report that I did not retrieve, and information missing from reports. Limitations of the included studies are explored in Section 5.3. I discuss study design and its relation to selection bias. Problems with outcome measures used in the included studies are addressed in terms of validity and reliability. I also discuss small sample sizes and the problem of fidelity of intervention delivery not being reported. Finally, I explore conclusions in Section 5.4. I discuss the included studies in terms of how the group work technique was characterized and whether they found in favour of group work or detected no difference. I also divide the body of evidence by methodological suitability for making causal inferences and discuss the balance of findings in each group. I consider the extent of the research and close the chapter with implications for practice and for future research.

5.1 Summary of evidence

This systematic review takes a best-evidence approach (Slavin, 1986), attempting to account for the methodological merits of each study. To provide a meaningful overview while giving more weight to studies that are better suited to making causal inferences, I group studies with comparably trustworthy designs and discuss the findings in those terms. Nevertheless, study design is not the only consideration. Systematic reviews "will be invalid if each study is merely given equal weight" (Gorard, 2014, p. 48). Studies with reduced bias deserve more weight than more biased studies, even where these more biased studies use more robust designs (Campbell & Stanley, 1963). A large sample, low attrition and blinding of assessors to group allocation are also important considerations. As such, these are also discussed.

5.1.1 Randomized trials

One study (Nemoto, 2012) used a form of random assignment. Random assignment is considered the most trustworthy technique for generating comparison

groups that are fair approximations of each other. All else being equal, they are therefore considered to be the most trustworthy design for making causal inferences (Campbell & Stanley, 1963; P Connolly, Biggart, Miller, O'Hare, & Thurston, 2017). This study minimized systematic differences between groups by randomly assigning pre-existing classes to conditions. The sample size was large (116 participants) and attrition was low (19%). No blinding of assessors to group allocation was reported. Comparing group work and individual work, this randomized trial did not detect a differential effect of group work on performance in descriptive writing tasks.

A single study rarely provides a definitive answer to a question. There are points of methodology in Nemoto's study that could be criticized; lack of blinding of assessors being one example. Yet, Nemoto's study is to-date the most robust single, published piece of evidence that reflects on the effects of group work in Japan's secondary English education. That a randomized trial with 116 participants and low attrition failed to detect an effect of group work over individual work in Japan's junior high school English lessons suggests that in this context a) perhaps no differential effect exists, or b) effects might be so small as to make switching to use group work not worth teachers' time. In the absence of robust contradictory evidence, it seems too early to advise English teachers at junior high schools and high schools in Japan to prefer group work over individual work. In light of Nemoto's study, it is perhaps surprising that MEXT has given precisely this advice.

5.1.2 Non-randomized comparisons

Taking a best-evidence approach (Slavin, 1986), in the absence of more randomized trials consideration should be given to the next most robust studies. Five studies (Wada, 2013; Ishihara, 2016; Fujishiro & Miyaji, 2009, 2010; Takazawa, 2011) were comparisons of pre-existing groups where the method for assigning participants to conditions was not random. It was not clear how participants were allocated to conditions in these studies. Bias caused by systematic (i.e. non-random) differences between groups cannot be ruled out. For all these studies, attrition and blinding of assessors to group allocation were not reported. The lack of information on the methodological rigour of these studies sounds a note of caution. Unable to confirm that any meaningful proportion of the participants completed the study, or that assessors did not in some way favour one group over another, teachers and researchers should be sceptical about whether the findings of these studies reflect real-world outcomes.

Two non-randomized comparisons found in favour of group work. Fujishiro and Miyaji (2009; 39 participants) found better speaking ability from using a sequence of web-based training, pair work, and teacher-centred instruction than from using the teacher's usual techniques; though it is not clear what those techniques were. In another study, Fujishiro and Miyaji (2010; 162 participants) found better speaking ability from using the above sequence than from teacher-centred listening instruction. Of the six studies that used control groups in this body of evidence (an evidence base of only eight studies), these are the only studies that found in favour of group work. It is notable that such a small proportion of the studies that are suited to making causal inferences favoured group work. The body of evidence is small, and these are in any case studies that have not taken steps to control for bias, so it is impossible to draw firm conclusions. Yet, in itself, the small proportion favouring group work suggests that teachers and researchers should be cautious about purported effects of group work on linguistic proficiency or academic attainment. Also, quite apart from the proportion of the body of evidence that Fujishiro and Miyaji's studies represent, the nature of their intervention of interest also deserves attention. Their favourable findings are for a sequence of web-based training, pair work, and teacher-centred instruction over alternatives. They did not assess group work by itself. While group work may have contributed to the detected effects, it may not. There appears nothing to say that a similar sequence of web-based training and teacher-centred instruction *without* group work would not have produced similar results. Being unable to untangle effects of group work from those of web-based training and teacher-centred instruction, teachers and researchers cannot take anything concrete from Fujishiro and Miyaji's studies about effects of group work in Japan's secondary English education.

Two studies did not detect a difference between group work and an alternative. Compared to business as usual, Ishihara (2016; estimated 105-200 participants) did not detect differential effects of peer tutoring on language proficiency. Compared to instruction characterized as grammar-translation, Takazawa (2011; 26 participants) did not detect a differential effect of group sight translation on language proficiency. These studies have not taken steps to control for bias, so it is impossible to draw firm conclusions. Nevertheless, it can be noted that the intervention of interest in these two studies was group work; they did not combine group work with other non-cooperative techniques. As such, the findings of these studies are more relevant to this systematic review than those of Fujishiro and

Miyaji's studies. These are two non-randomized comparisons, one large and one small, for which attrition and blinding of assessors to group allocation was not reported. Yet, it can be noted that their findings align with those of Nemoto (2012). It could be speculated that together, Ishihara, Takazawa, and Nemoto's studies might represent the beginnings of a case to be made that group work is not any more helpful than some major alternatives in English lessons at the secondary level in Japan. While that case remains to be made, it appears clear that more intervention research, particularly using bias-reducing designs, is needed before teachers and researchers can have an evidence-based understanding of effects of group work in English lessons at Japan's junior high schools and high schools.

Findings of one study were unclear. Wada (2013; 117 participants) analysed the results of a 20-hour intervention as two sets of 10 hours. After the first 10 hours, a statistically significant improvement in comprehension for a fixed sequence of pair and group work relative to instruction characterized as teacher-centred grammar-translation was observed. This improvement was not observed after the second 10-hour period. These results suggest that effects of group work, if they exist, might be short-lived. Perhaps students' comprehension initially improves with group work, but these gains are lost as the novelty of studying cooperatively wears off. Research using a variety of measures sustained over a longer period would help build an understanding of any effects varying over time.

Taken together, these non-randomized comparisons do not give a clear picture of a differential effect of group work on language proficiency. However, the improved speaking ability from web-based training, pair work, and teacher-centred instruction over an unspecified alternative and teacher-centred listening instruction using a CD (Fujishiro & Miyaji, 2009, 2010), suggests a warrant for assessing this combination of techniques using bias-reducing designs. Designs with multiple conditions, including web-based training, pair work, and teacher-centred instruction both together and separately, would help tease apart any differential effects of these techniques.

5.1.3 Studies of single groups

Two studies (Kurihara, 2017; Takagi, 2009) investigated one intervention across a single group of participants and compared the results of pre- and post-tests. Using only one group and comparing data from pre- and post-tests is seen as the second weakest design for making causal inferences (Campbell & Stanley, 1963).

With peer feedback training, Kurihara (2017) found improvements in essay-writing performance over time. With group presentations, Takagi (2009) found improved performance over time on tests of English third person singular verb forms in the present tense. Despite favourable findings, this body of evidence does not give a clear picture of the effects of group work. With this experimental design, the effects of the intervention cannot be disentangled from any number of other changes that could have happened during the study (Campbell & Stanley, 1963).

5.1.4 Overall

When the included studies that have comparably trustworthy designs are grouped, and the findings are considered in light of this trustworthiness, no clear picture of a differential effect of group work on linguistic proficiency or academic attainment emerges. The single randomized trial detected no difference, and findings for the non-randomized comparisons were mixed. Only the studies with the least robust designs (studies of single groups) offer collective support for group work in Japan's secondary English education. There appears to be no clear evidence to say that group work is more effective than major alternatives in Japan's secondary English education. Yet, lacking empirical backing for their claims in this context, Japan's education ministry (MEXT) has explicitly stated that group work is "effective" (MEXT, 2012, p. 37). In the same year that MEXT made this unfounded claim, Nemoto (2012) reported on a randomized trial with 116 participants and low attrition that failed to detect an effect of group work over individual work in Japan's junior high school English lessons. MEXT could have reassessed its position in light of Nemoto's study; a study that is to-date the most robust single, published piece of evidence that reflects on the effects of group work in Japan's secondary English education. Yet, the ministry did not. MEXT went on to make group work a key pillar of its 2017/18 revised curriculum guidelines. In this case, it appears that MEXT has ignored research to promote a technique that evidence suggests might be no better than major alternatives. As a member of the teaching corps in Japan, I would feel more comfortable switching from one mode of delivery to another if there were a robust empirical case for doing so. I would urge MEXT to engage with the research community to build evidence-based curriculum guidelines; guidelines built on demonstrable effects.

5.2 Limitations of this review

Limitations of this systematic review are discussed in this section. See 5.2.5 for limitations of the studies that are included in this review.

5.2.1 Search

With the resources available to me, I made every effort to find relevant research. However, I may have missed some reports. I did not search hard copies of journals by hand. Except for contacting two Japan-based researchers whose work is related to components of active learning in English education, I did not solicit suggestions from academics in the field. The search was limited to electronic databases, Google's search engine, and reference lists of included studies (a backward citation search).

The search was only conducted in English and Japanese. Relevant reports in other languages will have been missed if they have not included English or Japanese titles or abstracts. While among peer reviewed literature, translating titles and abstracts into English appears common, the same may not be true of grey literature. Yet the grey literature searches (Google search) were limited to English and Japanese. The grey literature searches were also not exhaustive. They relied on Google's search engine ranking records by relevance. In those searches (of which the search with the lowest number of hits returned 275,000 results limited to pdfs only), I stopped screening once listings had proved irrelevant for three continuous pages. Relevant research may have been listed on subsequent pages. Since these searches were not exhaustive, relevant reports may have been missed.

Finally, the search for this systematic review was conducted in May 2019. New evidence, that might affect the results of this review, may have been published since then.

5.2.2 Quality assurance

It is considered best practice in reducing bias in systematic reviewing to double screen titles and abstracts, and eligible full papers. In title and abstract screening, I was able to find a helper to independently screen a portion of the records that I screened. However, I was not able to recruit further help. No portion of the full reports underwent double screening. Similarly, no part of data extraction was independently duplicated. Consequently, the possibility that I am a source of bias in reviewing the literature cannot be ruled out. If this study is repeated, a second reviewer should be recruited to help reduce such potential bias.

5.2.3 Publication bias

Publication bias can threaten the validity of systematic reviews (C. Torgerson, 2006). A grey literature search can alleviate positive-results bias, a type of publication bias stemming from the systematic overrepresentation of results that show an effect for an intervention (a so-called positive result), in peer-reviewed literature (McAuley, Pham, Tugwell, & Moher, 2000). While grey literature searches (Google searches) were conducted, they were not exhaustive.

This systematic review may inevitably be affected by the file-drawer problem; the effect of researchers declining to publish results. This review may also be affected by outcome-reporting bias: systematic underreporting of statistically nonsignificant outcome measures. In a comparison of dissertations and published versions of the same research on educational interventions, Pigott, Valentine, Polanin, Williams, and Canada (2013) found that statistically nonsignificant outcome measures were 30% more likely to be excluded from published reports than statistically significant outcomes. These problems may have influenced the results of this systematic review, but it is impossible to be sure.

5.2.4 Incomplete retrieval of identified research

I was unable to locate one study that appeared to meet the inclusion criteria for the review; a master's dissertation (Onishi, 2012). It was excluded without assessing its eligibility for this systematic review. It may have been relevant. I recognize that, had the dissertation been retrieved and found relevant, the results of this systematic review may have been different.

5.2.5 Incomplete reporting

There are several limitations related to incomplete reports of the included studies. A meaningful assessment of risk of selection bias was not possible, because no reports stated the target population. Withdrawals and dropouts were not reported in the text of any reports, and a retention rate could not be calculated for seven out of eight studies because the necessary figures were not reported. Consequently, a meaningful assessment of risk of bias associated with withdrawals and dropouts could not be made. Finally, Ishihara (2016) did not report participant numbers, so these were estimated. As a result, the size of that study shown in Figure 4.2 is likely to be inaccurate.

5.3 Limitations of included studies

Concerning risk of bias, limitations of the included studies were covered in sections 4.4 and 4.5. Study design is also summarized in 5.1. Additional limitations in outcome measures, sample size, and fidelity of delivery are addressed below.

5.3.1 Outcome measures

Validity of outcome measures was not reported for any of the included studies. Reliability was only reported for one study (Fujishiro & Miyaji, 2009; test–retest reliability). It could be inferred that four out of eight studies (Wada, 2013; Ishihara, 2016; Takagi, 2009; Takazawa, 2011) used tests designed by the schools where the studies were conducted. Two studies used researcher-designed tools (Fujishiro & Miyaji, 2009, 2010). Nemoto (2012) used a count of error-free T-units, a measure that has been demonstrated to be unreliable because of the difficulty of reliably defining an error (Hirano, 1988). It is also worth noting that the outcome measure used in one study (Takagi, 2009; a test of English third person singular verb forms in the present tense) was perhaps too narrow to contribute to a meaningful assessment of effects of group work. In this body of evidence, a range of dissimilar outcome measures, and differing or no statistical analysis, complicates comparison. Without the opportunity to do a statistical synthesis, it is difficult to meaningfully compare one intervention with another, and it is impossible to calculate an overall outcome for the effects of group work.

5.3.2 Sample size

Sampling uncertainty is a problem in estimating the effectiveness of interventions. In general, a larger sample size provides greater certainty that findings are representative (C. Torgerson, D. Torgerson, & Styles, 2013). In this body of evidence, the findings of half the studies rest on fewer than 40 participants each. The small size of this portion of the included studies means that any estimates of effects of group work in Japan's secondary English education drawn from this body of evidence might be unreliable.

5.3.3 Fidelity of delivery

The interventions ranged from four days to three months. No included studies reported on the fidelity of intervention delivery. There is no way to confirm from the reports that group work was consistently administered over the course of the interventions. Kurihara (2017), for example, found improved essay-writing

performance with peer review. Even if the report contained a clear account of the form of peer review that Kurihara set out to facilitate (it does not), it would be hard to attribute improvements to the intended form of peer review because the nature of the intervention may have changed over the course of the study.

5.4 Conclusions

Japan's education ministry (MEXT, 2012, p. 37) promotes group work as an "effective" method for teaching English to Japan's junior high school and high school students, mandating this technique in its 2017/18 revised curriculum guidelines. Japan's English teachers are expected to overhaul their teaching to incorporate group work, among other techniques. No specific training scheme has been announced. It can be assumed that teachers will use their usual work schedules, or time outside of working hours, to prepare for and implement a switch in their mode of delivery. As such, Japan's curriculum reform is not without costs. It might be expected that MEXT made its changes with good reason, and that there is strong empirical backing for MEXT requiring teachers to use group work in their English lessons. However, this review has not found such backing for group work; suggesting that MEXT chose group work arbitrarily. The best evidence available (as far as this review has been able to determine) suggests that group work is no more helpful than individual work in junior high school and high school English lessons in Japan (Nemoto, 2012). MEXT is advised to engage with the research community to build evidence-based curriculum guidelines, otherwise it risks wasting teachers' time.

This systematic review has synthesized experimental studies of group work in junior high school and high school English lessons in Japan, that measured linguistic proficiency or academic attainment between 2009 and 2019. Findings concerning the effects of group work in this context were equivocal. The evidence base was small and lacked robust designs. On measures of linguistic proficiency and academic attainment, it was not clear if group work differs meaningfully from alternatives such as individual work. I interpret the results of this assessment and discuss implications for practitioners and researchers below.

5.4.1 Group work techniques

Group work techniques explored in this body of evidence were a blend of web-based training, pair work, and teacher-centred instruction characterized as blended learning, peer review or peer tutoring, fixed sequences of group work

characterized as collaborative learning or cooperative learning, group presentations and sight translation in groups. It is difficult to say anything concrete about these group work techniques in Japan's junior high school and high school English lessons, because of methodological flaws (discussed throughout this chapter) of the studies included in this review.

5.4.2 Extent of the research

The body of empirical research on effects of group work in Japan's secondary English education is small; this systematic review found only eight relevant studies. Of these, more than half (five out of eight) were reported in grey literature; in this case, research reports by scholarly societies. One of these was prepared for publication from a master's dissertation. Another two studies were reported in in-house journals of Japanese academic institutions,[27] journals that have not traditionally circulated outside the publishing institution and its affiliates (Kamada, 2007). Only one study was published in a conventionally-circulated peer-reviewed journal. It is apparent that the Japanese education ministry's advice to implement group work in English lessons at Japan's junior high schools and high schools (MEXT, 2018) is not matched by a solid body of research into effects of group work in this context.

5.4.3 Implications for practice

English teachers at junior high schools and high schools in Japan who wish to improve their students' linguistic proficiency or academic attainment have no firm reason to implement group work for that purpose. The body of empirical research on effects of group work in Japan's secondary English education is small, and as a whole, the results are equivocal. There is simply not enough research (especially not enough rigorous research) to be sure that group work is any better for improving linguistic proficiency and academic attainment than alternatives in English lessons at this level in Japan.

As a member of the teaching corps in Japan, I am disappointed that in this instance MEXT has failed to provide evidence-based guidance. In their efforts to improve students' linguistic proficiency and academic attainment, Japan's English teachers deserve the ministry's support. If MEXT opts to tell teachers to teach in a certain way, I expect them to have made that choice based on evidence. The

[27] Kamada (2007) called these 'kiyo journals.' JP: 紀要 kiyō

findings of this systematic review suggest that, on the contrary, MEXT has imposed group work as policy without a clear understanding of its effects in this context. In my opinion, the MEXT report that stated group work is effective (MEXT, 2012), did not give the question the thorough consideration that would be expected of an important public body. In 2020, English teachers must implement the new national curriculum guidelines, which include group work. Consequently, evaluation of students and teachers alike will hinge on implementing a technique that lacks sound empirical backing. Teachers' time and taxpayers' money will be expended switching to a technique that may be no more helpful than prominent alternatives.

5.4.4 Implications for future research

For English teachers at junior high schools and high schools in Japan to have an evidence-based understanding of the effects of group work in their context, more intervention research is needed. In its 2017/18 curriculum reform, Japan's education ministry backed 'active learning' techniques (MEXT, 2018); explicitly recommending group work in supporting material for its curriculum guidelines. The ministry promotes group work as an effective method for teaching English at junior high schools and high schools in Japan (MEXT, 2012). While other possible effects of group work, such as motivation or perceived ability, were beyond the scope of this systematic review, in terms of linguistic proficiency and academic attainment, the results of this review demonstrate that the case for effects of group work in English lessons at this level in Japan remains to be made.

A combination of web-based training, pair work, and teacher-centred instruction showed promise in non-randomized comparisons assessing structured and unstructured English speech (Fujishiro & Miyaji, 2009, 2010), warranting further evaluation using bias-reducing designs. Designs with multiple conditions, including web-based training, pair work, and teacher-centred instruction both together and separately, would help tease apart any differential effects of these techniques.

Studies that principally evaluate the effects of group work in reading and listening are notably absent in this context. For example, research in this context that assesses effects of having students discuss an audio recording in groups before hearing it a second time might valuably add to the evidence base.

Only two studies of group work techniques that measured English language proficiency were found in this context. While neither detected a difference, one had

only 26 participants, and the intervention in the other was delivered for fewer than six hours. Perhaps an effect might be detectable in an intervention of, for example, three months with 100 participants.

Chapter 6 Conclusion

No clear picture of a differential effect of group work on linguistic proficiency or academic attainment in Japan's junior high school and high school English lessons has emerged from this systematic review. There appears to be no clear evidence to say that group work is more effective than major alternatives in Japan's secondary English education. Moreover, as far as this review has been able to determine, the best evidence available suggests that group work may be no more helpful than individual work in junior high school and high school English lessons in Japan (Nemoto, 2012). Yet, without empirical backing for effects of group work in this context, Japan's education ministry (MEXT) has made group work a key element of its curriculum reform coming into effect in 2020. In this case, it is unclear what has underpinned MEXT's policy decision. As a dedicated and professional member of the teaching corps in Japan, I am committed to doing the best by my students. This means basing my practice on robust empirical evidence where it is available. The cost, in terms of time, energy, effort and resources of switching from one mode of delivery to another must be weighed against the benefits to students. In this case it is not clear that MEXT has done that, and as such may be encouraging wasteful expenditure of resources for no good purpose. I would urge MEXT to engage with the research community to build evidence-based curriculum guidelines; guidelines built on demonstrable effects. The junior high school and high school students of Japan deserve no less.

English teachers at junior high schools and high schools in Japan who wish to improve their students' linguistic proficiency or academic attainment have no firm reason to implement group work for that purpose. The body of empirical research on effects of group work in Japan's secondary English education is small, and as a whole, the results of this systematic review are equivocal. There is simply not enough research (especially not enough rigorous research) to be sure that group work is any better for improving linguistic proficiency and academic attainment than alternatives in English lessons at this level in Japan.

References

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Beauchamp, E. R. (1987). The development of Japanese educational policy, 1945-85. *History of Education Quarterly*, *27*(3), 299–324. https://doi.org/10.2307/368630

Boland, A., Cherry, M. G., & Dickson, R. (Eds.). (2017). *Doing a systematic review: A student's guide* (2nd ed.). London, England: Sage.

Bruning, R. H., Schraw, G. J., & Ronning, R. R. (1999). *Cognitive psychology and instruction* (3rd ed.). Englewood Cliffs, New Jersey: Prentice-Hall.

Brunton, G., Stansfield, C., & Thomas, J. (2012). Finding relevant studies. In *An Introduction to Systematic Reviews* (pp. 107–134). London, England: Sage.

Campbell Collaboration. (n.d.). So you want to write a Campbell systematic review? Retrieved 10 June 2019, from https://campbellcollaboration.org/research-resources/writing-a-campbell-systematic-review.html

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, Illinois: Rand McNally. https://doi.org/10.1037/022808

Chalmers, H. W. (2019). *Leveraging the L1: The role of EAL learners' first language in their acquisition of English vocabulary*. Oxford Brookes University, England.

Chambliss, D. F., & Schutt, R. K. (2018). *Making sense of the social world: Methods of invesitigation* (6th ed.). New York, NY: Sage.

Cochrane. (2017). What is a systematic review? Retrieved 22 May 2019, from https://community.cochrane.org/handbook-sri/chapter-1-introduction/11-cochrane/12-systematic-reviews/122-what-systematic-review

Cochrane. (2019). Meta-analysis: What, why, and how. Retrieved 20 June 2019, from https://uk.cochrane.org/news/meta-analysis-what-why-and-how

Cohen, E. G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research*, *64*(1), 1–35. https://doi.org/10.3102/00346543064001001

Connolly, P, Biggart, A., Miller, S., O'Hare, L., & Thurston, A. (2017). *Using randomised controlled trials in education*. London, England: Sage.

Connolly, Paul, Keenan, C., & Urbanska, K. (2018). The trials of evidence-based practice in education: A systematic review of randomised controlled trials in education research 1980–2016. *Educational Research*, *60*(3), 276–291. https://doi.org/10.1080/00131881.2018.1493353

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach*. London, England: Sage.

DeKeyser, R. M., Alfi-Shabtay, I., & Ravid, D. (2010). Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics*, *31*(3), 413–438. https://doi.org//10.1017/S0142716410000056

Dewey, J. (1997). *Education and experience*. New York, NY: Touchstone.

Dörnyei, Z., & Csizér, K. (1998). Ten commandments for motivating language learners: Results of an empirical study. *Language Teaching Research*, *2*(3), 203–229. https://doi.org/10.1177/136216889800200303

Drucker, A. M., Fleming, P., & Chan, A.-W. (2016). Research techniques made simple: Assessing risk of bias in systematic reviews. *Journal of Investigative Dermatology*, *136*(11), e109–e114. https://doi.org/https://doi.org/10.1016/j.jid.2016.08.021

Duffy, T. M., & Jonassen, D. (Eds.). (1992). *Constructivism and the technology of instruction*. New York, NY: Routledge.

Gorard, S. (2014). A proposal for judging the trustworthiness of research findings. *Radical Statistics*, *110*, 47–59.

Gorard, S., See, B. H., & Siddiqui, N. (2017). *The trials of evidence-based education: The promises, opportunities and problems of trials in education*. London, England: Routledge.

Gorsuch, G. (1998). Yakudoku EFL instruction in two Japanese high school classrooms: An exploratory study. *JALT Journal*, *20*(1), 33–48.

Gorsuch, G. (2001). Japanese EFL teachers' perceptions of communicative, audiolingual and yakudoku activities: The plan versus the reality. *Education Policy Analysis Archives*, *9*(10), 1–27. https://doi.org/10.14507/epaa.v9n10.2001

Gough, D., Oliver, S., Thomas, J., & Hobbs, A. (2013). *Learning from research: Systematic reviews for informing policy decisions: A quick guide*. *Alliance for Useful Evidence*. London, England: Nesta. https://doi.org/10.1080/00288306.1983.10422513

Greenhalgh, J., & Brown, T. (2017). Quality assessment: Where do I begin? In A. Boland, M. G. Cherry, & R. Dickson (Eds.), *Doing a systematic review: A student's guide* (pp. 107–130). London, England: Sage.

Hall, D., & Buzwell, S. (2012). The problem of free-riding in group projects: Looking beyond social loafing as reason for non-contribution. *Active Learning in Higher Education*, *14*(1), 37–49. https://doi.org/10.1177/1469787412467123

Higgins, J. P. T., & Green, S. (Eds.). (2011). Cochrane handbook for systematic reviews of interventions Version 5.1.0 [updated March 2011]. Retrieved 14 June 2019, from www.handbook.cochrane.org

Hino, N. (1988). Yakudoku: Japan's dominant tradition in foreign language learning. *JALT Journal*, *10*(1), 45–55.

Hirano, K. (1988). Research on T-unit measures in ESL. *Journal of Child Language Acquisition and Development*, *8*(2), 14–15.

Hoffmann, T., Glasziou, P., Boutron, I., Milne, R., Perera, R., Moher, D., … Michie, S. (2014). Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. *BMJ*, 348-g1687.

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). *Cultures and organizations: Intercultural cooperation and its importance for survival*. New York, NY: McGraw Hill. https://doi.org/10.1007/s11569-007-0005-8

Johnson, D. W., & Johnson, R. T. (1989). *Cooperation and competition: Theory and research*. Edina, Minnesota: Interaction Book Company.

Johnson, D. W., Johnson, R. T., Johnson, J., & Anderson, D. (1976). Effects of cooperative versus individualized instruction on student prosocial behaviour, attitudes toward learning, and achievement. *Journal of Educational Psychology*, *68*(4), 446–452. https://doi.org/10.1037/0022-0663.68.4.446

Kamada, H. (2007). Kiyo journals and scholarly communication in Japan. *Portal: Libraries and the Academy*, *7*(3), 375–383. https://doi.org/10.1353/pla.2007.0032

Kikuchi, K. (2006). Revisiting English entrance examinations at Japanese universities after a decade. *JALT Journal*, *28*(1), 77–96.

Kikuchi, K., & Browne, C. M. (2009). English educational policy for high schools in Japan: Ideals vs. reality. *RELC Journal*, *40*(2), 172–191. https://doi.org/10.1177/0033688209105865

Kizuka, M. (2006). Professionalism in English language education in Japan. *English Language Teacher Education and Development*, *9*(1), 55–62.

Kurihara, N. (2017). Peer review in an EFL classroom: Impact on the improvement of student writing abilities. *Asian Journal of Applied Linguistics*, *4*(1), 58–72.

Leech, N. L., & Onwuegbuzie, A. J. (2010). Guidelines for conducting and reporting mixed research in the field of counseling and beyond. *Journal of Counseling & Development*, *88*(1), 61–69. https://doi.org/10.1002/j.1556-6678.2010.tb00151.x

Levin, P. (2003). Running group projects: Dealing with the free-rider problem. *Planet*, *9*(1), 7–8. https://doi.org/10.11120/plan.2003.00090007

Littlewood, W. (2007). Communicative and task-based language teaching in East Asian classrooms. *Language Teaching*, *40*(3), 243–249. https://doi.org/10.1017/S0261444807004363

Maiden, B., & Perry, B. (2011). Dealing with free-riders in assessed group work: Results from a study at a UK university. *Assessment and Evaluation in Higher Education*, *36*(4), 451–464. https://doi.org/10.1080/02602930903429302

McArdle, G., Clements, K. D., & Hutchinson-Lendi, K. (2005). The free rider and cooperative learning groups: Perspectives from faculty members. In *Twelfth Academy of Human Resource Development International Conference* (pp. 529–535). Estes Park, Colorado.

McAuley, L., Pham, B., Tugwell, P., & Moher, D. (2000). Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analysis? *Lancet*, *356*, 1228–1231.

McMurray, D. (2018). MEXT's new course of study guidelines to rely on active learning. *The Language Teacher*, *42*(3), 27–29.

Meldrum, M. L. (2000). A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. *Hematology/Oncology Clinics of North America*, *14*(4), 745–760. https://doi.org/https://doi.org/10.1016/S0889-8588(05)70309-9

MEXT. (2012). 新たな未来を築くための大学教育の質的転換に向けて～生涯学び続け、主体的に考える力を育成する大学へ～（答申）*Aratana mirai o kizuku tame no daigaku kyōiku no shitsuteki tenkan ni mukete: Shōgai manabi tsudzuke, shutai-teki ni kangaeru chikara o ikusei-suru daigaku e (Tōshin)*. Tokyo, Japan: MEXT.

MEXT. (2016). Overview of the Ministry of Education, Culture, Sports, Science and Technology. Tokyo, Japan: MEXT.

MEXT. (2018). 中学校学習指導要領「外国語」英訳版（仮訳）*Chūgakkō gakushū shidō yōryō 'gaikoku-go' eiyaku-ban (Kayaku)*. Tokyo, Japan: MEXT.

Miettinen, O. S. (1983). The need for randomization in the study of intended effects. *Statistics in Medicine*, *2*(2), 267–271. https://doi.org/10.1002/sim.4780020222

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux,

P. J., ⋯ Altman, D. G. (2012). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *International Journal of Surgery*, *10*(1), 28–55. https://doi.org/10.1016/j.ijsu.2011.10.001

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ*, *339*, b2535. https://doi.org/10.1371/journal.pmed1000097

Nguyễn, P.-M. (2008). *Culture and cooperation: Cooperative learning in Asian Confucian heritage cultures - The case of Viet Nam*. Utrecht University.

NIHR CRSU. (n.d.). Narrative synthesis. Retrieved from http://www.nihrcrsu.org/guidance/narrative_synthesis/

OECD. (2018). *Education policy in Japan: Building bridges towards 2030*. *Reviews of National Policies for Education*. Paris, France: OECD. https://doi.org/10.1787/9789264302402-en.

Onishi, T. (2012). *Pilot cooperative learning in Japanese secondary school EFL contexts: What are the students' perceptions?* University of Edinburgh.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan — a web and mobile app for systematic reviews. *Systematic Reviews*, *5*(210), 1–10.

Petticrew, M., & Roberts, H. (2003). Evidence, hierarchies, and typologies: Horses for courses. *Journal of Epidemiology and Community Health*, *57*, 527–529.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Oxford, England: Blackwell.

Pigott, T. D., Valentine, J. C., Polanin, J. R., Williams, R. T., & Canada, D. D. (2013). Outcome-reporting bias in education research. *Educational Researcher*, *42*(8), 424–432. https://doi.org/10.3102/0013189x13507104

PLoS Medicine Editors. (2011). Best practice in systematic reviews: The importance of protocols and registration. *PLoS Medicine*, *8*(2), e1001009. https://doi.org/10.1371/journal.pmed.1001009

Richardson, W. S., Wilson, M. C., Nishikawa, J., & Hayward, R. S. (1995). The well-built clinical question: A key to evidence-based decisions. *ACP Journal Club*, *123*(3), A12–A12. https://doi.org/10.7326/ ACPJC-1995-123-3-A12

Russo, M. W. (2007). How to review a meta-analysis. *Journal of Gastroenterology and Hepatology*, *3*, 637–642.

Samimy, K. K., & Kobayashi, C. (2004). Toward the development of intercultural communicative competence: Theoretical and pedagogical implications for Japanese English teachers. *JALT Journal*, *26*(2), 245–261.

Schöpfel, J. (2010). Towards a Prague definition of grey literature. In *Twelfth International Conference on Grey Literature: Transparency in grey literature. Grey tech approaches to high tech issues* (pp. 11–26). Prague, Czech Republic.

Slavin, R. E. (1980). Cooperative learning. *Review of Educational Research*, *50*(2), 315–342.

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher*, *15*(9), 5–11.

Slavin, R. E. (1990). *Cooperative learning: Theory, research, and practice*. Englewood Cliffs, New Jersey: Prentice Hall.

Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., … Harvey, I. (2010). Dissemination and publication of research findings: An updated review of related biases. *Health Technology Assessment*, *14*(8), 1–220. https://doi.org/10.3310/hta14080

Sugie, S. (1995). Cooperative learning in Japan. *International Journal of Educational Research*, *23*(3), 215–225. https://doi.org/10.4324/9781315092171-19

Swan, M. (2018). Applied linguistics: A consumer's view. *Language Teaching*, *51*(2), 246–261. https://doi.org/10.1017/S0261444818000058

Tahira, M. (2012). Behind MEXT's new course of study guidelines. *The Language Teacher*, *36*(3), 3–8.

Thomas, B. H., Ciliska, D., Dobbins, M., & Micucci, S. (2010). Dictionary for quality assessment tool for quantitative studies (EPHPP). Retrieved 12 July 2019, from https://merst.ca/wp-content/uploads/2018/02/qualilty-assessment-dictionary_2017.pdf

Thomas, J., Harden, A., Oakley, A., Oliver, S., Sutcliffe, K., Rees, R., … Kavanagh, J. (2004). Integrating qualitative research with trials in systematic reviews. *The BMJ*, *328*, 1010–1012.

Torgerson, C. (2006). Publication bias: The Achilles' heel of systematic reviews? *British Journal of Educational Studies*, *54*(1), 89–102. https://doi.org/10.1111/j.1467-8527.2006.00332.x

Torgerson, C., Torgerson, D., & Styles, B. (2013). *Randomised trials in education: An introductory handbook*. London, England: Educational Endowment Foundation.

University of York. (n.d.). About PROSPERO. Retrieved 10 June 2019, from https://www.crd.york.ac.uk/prospero/#aboutpage

Van Meter, P., & Stevens, R. J. (2000). The role of theory in the study of peer collaboration. *Journal of Experimental Education*, *69*(1), 113–127.

https://doi.org/10.1080/00220970009600652
Von Elm, E., Altman, D., Egger, M., Pocock, S. J., Gotzsche, P. C., & Vandenbroucke, J. P. (2008). The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *International Journal of Public Health*, *53*(1), 3–4. https://doi.org/10.1007/s00038-007-0239-9
Von Glasersfeld, E. (1989). Cognition, construction of knowledge, and teaching. *Synthese*, *80*(1), 121–140. https://doi.org/10.1007/BF00869951
Yoshida, K. (2003). Language education policy in Japan: The problem of espoused objectives versus practice. *Modern Language Journal*, *87*(2), 291–293.
三尾真琴 Mio, M. (2017). アクティブ・ラーニングを学ぶ授業実践：「主体的」「対話的」「深い学び」の習得に向けた試み Lesson practice for active learning: Practice and method for 'subjective' 'interactive' 'deep learning'. 帝京科学大学教育・教職研究 *Teikyōkagakudaigaku Kyōiku Kyōshoku Kenkyū Kiyō*, *4*(1), 21–28.
和田憲明 Wada, N. (2013). 中学校英語リーディング授業における協同学習の効果に関する研究 Chūgakkō eigo rīdingu jugyō ni okeru kyōdō gakushū no kōka ni kansuru kenkyū. 近大姫路大学教育学部紀要 *Kindaihimejidaigaku Kyōiku Gakubu Kiyō*, (6), 103–112.
大谷泰照 Otani, Y. (1981). この状況をどう考えるべきか Kono jōkyō o dō kangaerubeki ka. 英語教育 *The English Teachers' Magazine*, *13*, 3–6.
岩木秀夫 Iwaki, H. (2004). ゆとり教育から個性浪費社会へ *Yutori kyōiku kara kosei rōhi shakai e*. Tokyo, Japan: ちくま新書 Chikuma Shinsho.
惟任泰裕 Koreto, Y. (2017). 学習指導要領改訂にみる戦後日本の英語教育史 The history of English language education in postwar Japan: Focusing on the revision of the course of study. 教育科学論集 *Kyōiku Kagaku Ronshū*, *20*, 1–12.
林伸昭 Hayashi, N. (2017). 日本人英語学習者に適した英語教授法・指導法：PCPP 、AL 、教材 The methods and teaching technique which are suitable to Japanese EFL learners: PCPP, AL and materials. 宮崎公立大学人文学部紀要 *Miyazakikōritsudaigaku Jinbungakubu Kiyō*, *24*(1), 155–172.
根本章子 Nemoto, A. (2012). 中学校での協同学習による「書く活動」についての考察：協同学習と個別学習との比較から Cooperative writing in junior high school English class: A comparison between cooperative learning and individual learning. 筑波英語教育 *Tsukuba Eigo Kyōiku*, (33), 117–128.
水原克敏 Mizuhara, K. (2010). 学習指導要領は国民形成の設計書：その能力観と人間像の歴史的変遷 *Gakushū shidō yōryō wa kokumin keisei no sekkei-sho: Sono nōryoku-kan to ningenzō no rekishi-teki hensen*. Tokyo, Japan: 東北大学出版

会 Tōhokudaigaku shuppan-kai.
石原義文 Ishihara, Y. (2016). 英語科におけるアクティブ・ラーニング型の授業を目指して（実践報告）Adopting ‘active learning’ (A practical report) in English classes - Peer-tutoring for 11th graders in ‘English Communication Ⅱ’. 広島大学附属中・高等学校中等教育研究紀要 *Hiroshimadaigaku Fuzoku-Chū Kōtō Gakkō Chūtō Kyōiku Kenkyū Kiyō*, *63*, 67–74.
藤代昇丈 Fujishiro, N., & 宮地功 Miyaji, I. (2009). ブレンド型授業による英語の音読力と自由発話力に及ぼす効果 The effectiveness of blended instruction in class on the skills of oral reading and speaking in English. 日本教育工学会論文誌 *Nihon Kyōiku Kōgakukai Ronbunshi*, *32*(4), 395–404.
藤代昇丈 Fujishiro, N., & 宮地功 Miyaji, I. (2010). eラーニングの授業活用における英語発話力の変容測定のための評価項目に関する一検討 A study on the assessment criteria for the development of speaking skill through utilization of e-learning in English class. 電子情報通信学会技術研究報告*. ET,* 教育工学 *Denshi Jōhō Tsūshin Gakkai Gijutsu Kenkyū Hōkoku. ET, Kyōiku Kōgaku*, *110*(263), 15–18.
高木浩志 Takagi, H. (2009).「協同学習が変える学びのかたち」: 神戸大学附属住吉中学校の実践を通して Kyōdō gakushū ga kaeru manabi no katachi: Kōbedaigaku fuzoku Sumiyoshi chūgakkō no jissen o tōshite. 日本教育情報学会年会論文集 *Proceedings of the Annual Meeting of the Japan Society for Educational Information*, *25*, 132–135.
髙澤郁男 Takazawa, I. (2011). 英語学力向上のための音読指導の工夫 : サイトトランスレーションと協同学習の手法を用いた音読を通して Eigo gakuryoku kōjō no tame no ondoku shidō no kufū: Saitotoransurēshon to kyōdō gakushū no shuhō o mochiita ondoku o tōshite. 教育実践研究 *Kyōiku Jissen Kenkyū*, *21*, 203–208.

Appendices

Appendix A: Coding sheet

Coding sheet

	Item	Response	Guidance
Administration	1. How was the report identified?		Was the report identified through an electronic database? Forward/backward citation search? Google search? Introduced by colleague?
	2. What was the language of publication?		In what language is the full text of the report written?
Context	3. Where was the study conducted?		What details have the authors given about the geographical location of the study (quote)? What can be inferred (specify 'reviewer interpretation')? If none, code as 'not stated'.
	4. In what type of school was the study conducted?		What details have the authors given about the type of school (quote)? What can be inferred (specify 'reviewer interpretation')? If none, code as 'not stated'.
	5. What term best characterizes the language teaching provision?		What details have the authors given about the language teaching provision (quote)? What can be inferred (specify 'reviewer interpretation')? If none, code as 'not stated'.
	6. Who were the participants and how are they characterized?		What is the age, sex, socioeconomic status, linguistic background, or other characteristics of the participants?

Description	7. What is the research question?		What details have the authors given about the research question (quote)? What can be inferred (specify 'reviewer interpretation')?
	8. What was the group work intervention?		Describe the group work/pair work/cooperative learning/collaborative learning intervention.
	9. What was the group work intervention compared to?		What details have the authors given about any comparators (quote)? What can be inferred (specify 'reviewer interpretation')? If none, 'no comparison'.
	10. Who delivered the intervention and to how many participants at a time?		Delivered by usual teacher? Other teacher? Researcher? A combination? Delivered to one student at a time? Small groups? Whole classes? (Class size?)
	11. What were the independent and dependent variables?		Specify. List IVs and DVs separately.
Methods	12. What study design was used?		RCT, non-randomized comparison, cohort analytic, case-control, cohort, interrupted time series, case study, action research, other? Specify.
	13. Did the outcome measures reflect linguistic proficiency or academic attainment?		One or both? Record all measures

	14. How many participants were there?		Specify participants at start of study. How many completed the study? If possible, calculate retention rate.
	15. How long was the study?		In reported units.
	16. What is the study's quality assessment rating?		Complete this section after final decision of both reviewers on risk of bias assessment.

Appendix B: Adapted EPPHP risk of bias assessment tool (Chalmers, 2019)

QUALITY ASSESSMENT TOOL FOR QUANTITATIVE STUDIES

Reference:

COMPONENT RATINGS

A) SELECTION BIAS

(Q1) Are the individuals selected to participate in the study likely to be representative of the target population?

1. Very likely
2. Somewhat likely
3. Not likely
4. Can't tell

(Q2) What percentage of selected individuals agreed to participate?

1. 80 - 100% agreement
2. 60 – 79% agreement
3. less than 60% agreement
4. Not applicable
5. Can't tell

RATE THIS SECTION	STRONG	MODERATE	WEAK
See dictionary	**1**	**2**	**3**

B) STUDY DESIGN

Indicate the study design

1. Randomized controlled trial
2. Non-randomized comparison (comparison groups prospectively generated by a means other than random allocation)
3. Cohort analytic (two group pre + post)
4. Case-control

5. Cohort (one group pre + post (before and after))
6. Interrupted time series
7. Other specify _____
8. Can't tell

Was the study described as randomized? If NO, go to Component C.

No Yes

If Yes, was the method of randomization described? (See dictionary)

No Yes

If Yes, was the method appropriate? (See dictionary)

No Yes

RATE THIS SECTION	STRONG	MODERATE	WEAK
See dictionary	**1**	**2**	**3**

C) CONFOUNDERS

(Q1) Were there known important differences between groups prior to the intervention?

1. Yes
2. No
3. Can't tell

The following are examples of confounders:

1. Race
2. Sex
3. Marital status/family
4. Age
5. SES (income or class)
6. Education
7. Health status
8. Pre-intervention score on outcome measure

(Q2) If yes, indicate the percentage of relevant confounders that were controlled (either in the design (e.g. stratification, matching) or analysis)?

1. 80 – 100% (most)
2. 60 – 79% (some)
3. Less than 60% (few or none)
4. Can't Tell

RATE THIS SECTION	STRONG	MODERATE	WEAK
See dictionary	1	2	3

D) BLINDING

(Q1) Was (were) the outcome assessor(s) aware of the intervention or exposure status of participants?

1. Yes
2. No
3. Can't tell

(Q2) Were the study participants aware of the research question?

1. Yes
2. No
3. Can't tell

RATE THIS SECTION	STRONG	MODERATE	WEAK
See dictionary	1	2	3

E) DATA COLLECTION METHODS

(Q1) Were data collection tools shown to be valid?

1. Yes
2. No
3. Can't tell

(Q2) Were data collection tools shown to be reliable?

1. Yes
2. No
3. Can't tell

RATE THIS SECTION	STRONG	MODERATE	WEAK
See dictionary	**1**	**2**	**3**

F) WITHDRAWALS AND DROP-OUTS

(Q1) Were withdrawals and drop-outs reported in terms of numbers and/or reasons per group?

1. Yes
2. No
3. Can't tell
4. Not Applicable (i.e. one time surveys or interviews)

(Q2) Indicate the percentage of participants completing the study. (If the percentage differs by groups, record the lowest).

1. 80% - 100%
2. 60% - 79%
3. Less than 60%
4. Can't tell
5. Not Applicable (i.e. Retrospective case-control)

RATE THIS SECTION	STRONG	MODERATE	WEAK
See dictionary	**1**	**2**	**3**

G) INTERVENTION INTEGRITY

(Q1) What percentage of participants received the allocated intervention or exposure of interest?

1. 80% - 100%
2. 60% - 79%
3. Less than 60%

4. Can’t tell

(Q2) Was the consistency of the intervention measured?

1. Yes
2. No
3. Can’t tell

(Q3) Is it likely that subjects received an unintended intervention (contamination or co-intervention) that may influence the results?

4. Yes
5. No
6. Can’t tell

H) ANALYSES

(Q1) Indicate the unit of allocation (circle one)

Local authority :: school :: class/teaching group :: pupil

(Q2) Indicate the unit of analysis (circle one)

Local authority :: school :: class/teaching group :: pupil

(Q3) Are the statistical methods appropriate for the study design?

1. Yes
2. No
3. Can’t tell

(Q4) Is the analysis performed by intervention allocation status (i.e. intention to treat) rather than the actual intervention received?

1. Yes
2. No
3. Can’t tell

GLOBAL RATING

COMPONENT RATINGS

Please transcribe the information from the gray boxes on pages 1-4 onto this page. See dictionary on how to rate this section.

A	**SELECTION BIAS**	**STRONG**	**MODERATE**	**WEAK**	
		1	**2**	**3**	
B	**STUDY DESIGN**	**STRONG**	**MODERATE**	**WEAK**	
		1	**2**	**3**	
C	**CONFOUNDERS**	**STRONG**	**MODERATE**	**WEAK**	
		1	**2**	**3**	
D	**BLINDING**	**STRONG**	**MODERATE**	**WEAK**	
		1	**2**	**3**	
E	**DATA COLLECTION METHOD**	**STRONG**	**MODERATE**	**WEAK**	
		1	**2**	**3**	
F	**WITHDRAWALS AND DROPOUT**	**STRONG**	**MODERATE**	**WEAK**	
		1	**2**	**3**	**NOT APPLICABLE**

GLOBAL RATING FOR THIS PAPER (circle one):

1	**STRONG**	**(no WEAK ratings)**
2	**MODERATE**	**(one WEAK rating)**
3	**WEAK**	**(two or more WEAK ratings)**

With both reviewers discussing the ratings:

Is there a discrepancy between the two reviewers with respect to the component (A-F) ratings?

No Yes

If yes, indicate the reason for the discrepancy

1. Oversight
2. Differences in interpretation of criteria
3. Differences in interpretation of study

Final decision of both reviewers (circle one):

1. **STRONG**
2. **MODERATE**
3. **WEAK**