Running Head: DO JIGSAW CLASSROOMS IMPROVE LEARNING?

Do Jigsaw classrooms improve learning outcomes? Five experiments and an internal meta-analysis

Arnaud Stanczak[1]

Céline Darnon[1]

Anaïs Robert[1]

Marie Demolliens[1]

Camille Sanrey[2]

Pascal Bressoux[3]

Pascal Huguet[1]

Céline Buchs[4]

Fabrizio Butera[5]

and PROFAN Consortium

[1]Université Clermont Auvergne, LAPSCO, Laboratoire de Psychologie Sociale et Cognitive. France

[2] Université Paris Descartes, LPS Laboratoire de Psychologie Sociale. France

[3] Université de Grenoble, LaRAC, Laboratoire de Recherche sur les Apprentissages en Contexte. France

[4] Université de Genève. Swizerland

[5] Université de Lausanne, UNILAPS Laboratoire de Psychologie Sociale. Swizerland

Authors' note

This research was supported financially by the « Ministère de l'éducation et de la jeunesse / Ministère de l'enseignement supérieur, de la recherche et de l'innovation / Mission Monteil pour le numérique éducatif / Programme d'investissements d'avenir, expérimentation « ProFAN ».

PROFAN Consortium: Anatolia Batruch (Université de Lausanne, UNILAPS Laboratoire de Psychologie Sociale, Swizerland); Marinette Bouet, Carlos Cepeda, Théo Ducros, Ruben Martinez, Vincent Mazenod, Benoit Petitcollot, Farouk Toumani (Université Clermont Auvergne, LIMOS, Laboratoire d'Informatique de Modélisation et d'Optimisation des Systèmes, France); Anne-Laure De Place, Pascal Pansu, Mathilde Riant (Université de Grenoble, LaRAC, Laboratoire de Recherche sur les Apprentissages en Contexte, France); Genavee Brown, Luc Goron, Eric Jamet, Estelle Michinov, Nicolas Michinov, Laurine Peter (Université de Rennes, LP3C, Laboratoire de Psychologie, Cognition, Comportement, Communication, France); Olivier Desrichard, Nathalie Mella (Université de Genève, GREPS, Groupe de recherche en psychologie de la santé, Switzerland); Marco Bressan, Céline Poletti, Isabelle Régner, Eva Vives (Université de Aix-Marseille, LPC, Laboratoire de Psychologie Cognitive, France); Emilio Paolo Visintin (Département de Sciences Humaines, Université de Ferrara, Italy).

We wish to thank Sebastien Baron, Stephanie Delpirou, Elodie Eple, Fabrice Taupin and Marie-Claude Borion for their involvement in data collection. Their commitment was highly appreciated. We also thank Boris Quétard and Medhi Marot for their help with data analysis.

The authors declare that there are no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Correspondence should be addressed to Céline Darnon, Laboratoire de Psychologie Cognitive et Sociale, LAPSCO, UMR6024, CNRS, Université Clermont Auvergne, 34 Avenue Carnot, 63037 Clermont-Ferrand Cedex, France (celine.darnon@uca.fr).

Abstract

"Jigsaw" is a peer learning procedure derived from social interdependence theory, which suggests that individuals positively linked by a common goal can benefit from positive and promotive social interactions (Aronson & Patnoe, 2011). Although jigsaw has often been presented as an efficient way to promote learning, empirical research testing its effect on learning remains relatively scarce. The goal of the present research is to test the hypothesis that a jigsaw intervention would yield a meaningful effect size ($d = 0.40$) on learning outcomes, in five randomized experiments conducted among 6th graders. The jigsaw intervention was compared to an "individualistic" ($N_{\text{Experiment 1}} = 252$; $N_{\text{Exp 2}} = 313$) or a "teaching as usual" ($N_{\text{Exp 3A}} = 110$; $N_{\text{Exp 3B}} = 74$; $N_{\text{Exp 3C}} = 101$) approach on the same pedagogical content. Across the five experiments, we did not find empirical support for this hypothesis. Internal meta-analytic estimates ($ES = 0.00$, 95% CI [-0.10, 0.09]) showed that, overall, the jigsaw intervention did not produce the expected positive effects on learning. The reasons why jigsaw classrooms may not always prove beneficial for learning are discussed.

Keywords: Jigsaw classroom, cooperative learning, internal meta-analysis

Educational Impact And Implications Statement

Although the "jigsaw classroom" is a relatively popular cooperative method (Aronson & Patnoe, 2011) which has often been presented as an efficient way to promote learning, empirical research testing its effect on learning remains relatively scarce and debated (Roseth et al., 2019). Poor evidence can be misleading for teachers. Across five randomized experiments conducted on French 6th graders, the present research showed that the "jigsaw classroom" did not yield any significant gain in learning outcomes as compared to "individualistic" or "teaching as usual" conditions of learning. The reasons why the jigsaw intervention did not produce the expected positive effects on learning are discussed.

Do Jigsaw Classrooms Improve Learning Outcomes? Five Experiments and an Internal Meta-analysis

Many students have experienced cooperative learning during schooling. Slavin (2011) defined cooperative learning as instructional methods in which teachers organize students in small groups, which then work together to help one another learn academic content. Most teachers consider cooperative learning to be a very useful strategy to promote learning and other related positive social skills among students (e.g., interpersonal relations; Saborit et al., 2016). The consensus is so large that cooperative learning is often presented as one of the major topics that should guide education reforms (Baloche & Brody, 2017; Hattie, 2010; Johnson & Johnson, 2009; Slavin, 2008), although it continues to be used in quite a limited manner in practice (Pianta et al., 2007). Still, researchers agree that not all cooperative learning techniques are equivalent and that empirical evidence has to be clearly established before making recommendations to practitioners (Slavin, 2008). In fact, cooperative learning can take several forms which can be articulated around different goals or procedures (see Sharan, 1999; Slavin, 2011 for a review). Even though this literature reveals to the existence of an overall positive effects of cooperative learning on learning outcomes, there are significant variations in the effects of different cooperative learning methods on learning outcomes (Johnson & Johnson, 2005; Slavin, 2011). Debates notably remain about the effects of the jigsaw method, a very popular cooperative learning method (e.g., more than 4000 citations of the first edition of the book presenting the method) that structures positive interdependence by distributing complementary resources between group members in the classroom. Although this method has been the object of a recent surge of interest in the scientific community (Roseth et al., 2019), its structure has often been questioned (e.g., Johnson & Johnson, 2002) and only a relatively small number of studies have tested the general claim that jigsaw promotes learning (Aronson & Patnoe, 2011).

Benefits of Cooperative Learning

Positive interdependence theory (Johnson & Johnson, 2009) suggests that individuals engage in the task and help each other because they identify with the group and want others to succeed. From this perspective, the beneficial effects of cooperation emerge when group members feel responsible for their own learning as well as that of others. According to positive interdependence theory (Johnson & Johnson, 2009), several elements are essential for structuring cooperation in classrooms. First, positive interdependence supposedly emerges when the achievement of a group member is positively linked to others. This positive interdependence generates opportunities for students to engage in promotive interactions (e.g., group facilitation, help giving and seeking, feeling of responsibility toward the group), which are in turn likely to improve learning (Johnson & Johnson, 2005). Positive interdependence may refer to group outcomes and emerge from common goals (Hwong et al., 1993) or collective rewards (Buchs et al., 2011; Slavin et al., 2013). It may also refer to means interdependence and emerge from complementary tasks, resources, and roles (see Butera & Buchs, 2019, and Topping et al., 2017, for reviews). Second, in order to decrease the tendency for many — although not all — individuals to expend less effort when working collectively than when working individually (i.e., social loafing; Freeman & Greenacre, 2011; Huguet et al., 1999; Karau & Williams, 1991; Voyles et al., 2015), in cooperative learning, the individuals' contribution to the group's product must be visible and accountable. Individual accountability can be achieved in various ways, including distributing complementary resources to each group member. When resources are interdependently distributed, each group member is responsible for explaining the part of the resource he or she possesses or masters to the other group members. Thus, resource interdependence allows group members to become both positively interdependent and individually accountable (Buchs & Butera, 2015). In addition to these main components (interdependence and accountability), in order for

cooperation to reach optimal results, teachers are usually advised to provide feedback at both the group and individual levels, giving learners opportunities to reflect on group processing (Johnson & Johnson, 1989). According to authors in the area, when these elements are present, cooperative learning can occur and favor learning-related outcomes.

Extensive literature has reported the beneficial effects of cooperative learning on academic achievement, in addition to benefits for motivation, self-esteem, and other psychological variables, such as feelings of competency and attitudes toward group work (Fernandez-Rio et al., 2017; Johnson & Johnson, 2009; Kyndt et al., 2013; Nokes-Malach et al., 2015; Puzio & Colby, 2013; Roseth et al., 2008; Springer, 1999). For example, in one of their reviews, Johnson and Johnson (1989) reported positive and medium associations between positive interdependence and academic achievement (d = 0.64) and self-esteem (d = 0.44) when comparing cooperative to individualistic learning. In the same vein, in a synthesis of several meta-analyses related to the effect of various interventions on academic achievement, Hattie (2010) estimated that cooperative learning, in general, has positive and "medium" effects on learning when compared to both individualistic (d = 0.59, n = 774 studies) and competitive (d = 0.54, n = 1024 studies) learning. However, as mentioned in the introduction of the present article, cooperative learning can be seen as an umbrella term encompassing several cooperative techniques, with various effects on learning (Kyndt et al., 2013). As noted by some authors, more empirical research and evidence are critically needed to determine both the extent to which each of these cooperative techniques is effective and the conditions under which they are (Nokes-Malach et al., 2015).

In the present research, we test the effect of the jigsaw technique on learning (Aronson et al., 1978; Aronson & Bridgeman, 1979; Aronson & Patnoe, 2011; Blaney et al., 1977). Despite its popularity among teachers (e.g., the official website jigsaw.org reports more than 5 million page views since its creation), the structure of this technique has recently been

questioned, notably, because it contains both cooperative and competitive elements (Roseth et al., 2019). In addition, little research has thus far empirically assessed its effectiveness on learning, especially when compared to other cooperative learning methods (e.g., dyadic peer learning; Ginsburg-Block et al., 2006; Rohrbeck et al., 2003). In their review, Johnson and Johnson (2002) reported that only 14 empirical studies examined the effects of jigsaw on academic achievement, compared to competitive learning (n = 9 studies) or individualistic learning (n = 5 studies). They found that the effect sizes of jigsaw were respectively d = 0.29 when compared to a competitive control group (i.e., the presence of negative goal or reward interdependence), and d = 0.13 when compared to an individualistic control group (i.e., individualistic work, no social interdependence between participants). They note that this effect size is consistently lower than the one obtained with other cooperative methods. Indeed, the median effect size of other cooperative methods was d = 0.43 when compared to a competitive and d = 0.46 when compared to an individualistic control group. As a consequence, and since conclusions made on the basis on few studies could be misleading, these authors urge researchers to conduct more research on methods like jigsaw (Johnson & Johnson, 2002).

Hence, in the context of the replicability crisis, and given the incentives for cumulative evidence research in education (Open Science Collaboration, 2015; Świątkowski & Dompnier, 2017), relying on appropriate experimental designs to address the efficiency of teaching methods is particularly important. As noted by some authors, a part of educational psychology research could be particularly at risk of being biased towards positive results (Gage et al., 2017; Götz et al., 2021). Recently, educational researchers have argued that their discipline must seek more transparent, replicable and open science practices to increase the credibility of their findings (Fleming et al., 2021; Plucker & Makel, 2021). According to Patall (2021), “A system that is biased against null, perplexing, or replicated findings in

primary research and in turn, encourages the use of problematic researcher practices in order to thrive, distorts the scientific literature" (p. 147). Thus, a cumulative approach, providing robust and reliable results is therefore particularly recommended in this context.

Jigsaw Classroom: A Highly Structured Cooperative Learning Technique

The jigsaw classroom is a cooperative learning technique that was developed in the early 1970s by Aronson and colleagues, initially to reduce interethnic tensions between students (Aronson et al., 1978). In the jigsaw classroom, students work in groups on pedagogical content divided into several subtopics. The classical procedure usually follows three major steps. In the first step, jigsaw groups are formed based on the number of the course's subtopics. For example, a sixth-grade course on biodiversity could be divided into four subtopics on specific environments (e.g., plains, mountains, deserts, and forests), which all relate to a common problematic: the characteristics of these environments and the repartition of the species within it. In these jigsaw groups, students work individually (e.g., reading a text, taking notes, and answering questions) on their subtopic. Then, in the second step, students assigned to the same subtopics are grouped together to form expert groups, in which students are expected to discuss the main points of their topic and make sure everyone has understood the content. Finally, in the third step, students go back to their initial jigsaw groups to teach their topic to the other group members. Thus, students in a jigsaw classroom are highly interdependent; they work on complementary resources with the objective of ensuring each other's learning. Positive interdependence is ensured by both the means (i.e., working with complementary resources) and the goal (i.e., ensuring that everyone has understood the content; Johnson et al., 1998).

Since its creation, the positive effects of the jigsaw classroom have been documented on important variables, such as prejudice reduction (Aronson & Patnoe, 2011; Desforges et

al., 1991; Sharan, 1980; Walker & Crogan, 1998), self-esteem (Lazarowitz et al., 1994), and self-efficacy (Crone & Portillo, 2013; Darnon et al., 2012). Some research suggests that jigsaw classrooms might also improve learning (Berger & Hänze, 2009; Doymus, 2008; Ghaith & El-Malak, 2004; Hänze & Berger, 2007; Şahin, 2011). For example, Doymus (2008) tested whether the jigsaw approach could increase undergraduates' learning, compared to individual learning, in a chemistry class. Over a period of five weeks, the participants studied the same pedagogical content, either using the jigsaw procedure or individually. Learning was assessed with a comprehension test before and after the course. The results indicated larger gains for students in the jigsaw condition than in the individual condition. In another study, Şahin (2011) tested the effectiveness of the jigsaw approach, compared to business as usual, in a Turkish course (over six weeks) at the sixth-grade level. First, each child was assigned to a "home group" (i.e., jigsaw group). Then each of them had to work as an expert on a specific part of the course (e.g., use of phrases and sentence units, correct use of voice and gerunds) and prepare the assignments (e.g., portfolios and guide questions). In the final step, they had to present the assignments in their respective jigsaw groups. Compared to the control group, which was taught following the usual methods and without cooperation, children in the jigsaw group outperformed the others on a written expression test. Positive effects of Jigsaw have also been found on learning outcomes of 4th graders, as part as a five-week experiment (5 hours in total) in an English as a foreign language course (Ghaith & El-Malak, 2004). Shaaban (2006) and Gömleksi'z (2007) report similar results on literary subjects, among fifth-graders and university students respectively, on experiment during approximatively four weeks (8 hours in total), although it is worth noting that these studies were conducted on relatively small samples ($N_{\text{Ghaith}} = 48$; $N_{\text{Shaaban}} = 44$; $N_{\text{Gömleksi'z}} = 66$).

In more recent research, Roseth et al. (2019) showed that the jigsaw approach yielded positive outcomes on academic achievement when compared to the teaching-as-usual

condition during an entire semester of an undergraduate psychology course. Although the jigsaw method was associated with gains on learning outcomes, it did not significantly increase cooperative efforts between individuals, nor did it increase interest or perceived competence over time when compared to the business-as-usual condition, which contrasts with previous empirical research. Rather than considering the jigsaw method to be purely cooperative, these authors argue that the three-steps procedure generates different dynamics between members. According to them, the "expert" step would be linked to individualistic dynamics as members of the group would be independent in terms of resources (Slavin, 2011) and the "jigsaw" step would involve both positive and negative dynamics, because of the simultaneous presence of resource interdependence and independence of rewards. In fact, mixed results can be found in other experimental and large-scale studies, even after controlling for teacher training in the jigsaw method and the quality of the implementation (Moskowitz et al., 1983; 1985). For example, Moskowitz et al. (1983, 1985) did not find any support for the claim that the jigsaw method improved learning, even when selecting schools where the technique was supposedly well-implemented. Some research also reports null results on academic achievement (Crone & Portillo, 2013; Law, 2011; Moreno, 2009) across various school levels (e.g., from primary school to tertiary education) and across different disciplines (e.g., language, mathematics, sciences). In some cases, researchers even observed a negative effect of jigsaw on learning when compared to a teacher-centered approach. For example, Souvignier and Kronenberger (2007) showed that third graders who studied in a jigsaw classroom in their geometry and astronomy sequences performed worse than those who were in the teacher-centered condition. Taken together, such results call for further evidence. Indeed, as noted by some authors, empirical research on the jigsaw method is rather scarce (Johnson & Johnson, 2002). Furthermore, the vast majority of previous research testing jigsaw effects on learning has been conducted on undergraduate students, whereas very few

articles have tested jigsaw at the secondary level (for exceptions, see Şahin, 2011; Souvignier & Kronenberger, 2007; Tarhan et al., 2013).This is an important issue because concluding on the effectiveness of an intervention with only a few empirical studies, mostly conducted on the same age groups, could be highly misleading for practice (Cheung & Slavin, 2016).

More recently, the "mega-analysis" synthesis led by John Hattie (2017) has included the jigsaw method in the top 10 most effective academic interventions, with an estimated effect size of $d = 1.20$. However, this estimate comes from one meta-analysis of 11 studies all conducted in Turkey between 2005-2012, with an average sample size of 109 participants. This large effect size (Cohen, 1962; Kraft, 2020) is unusual, considering both previous reviews (Johnson & Johnson, 2002; Newmann & Thompson, 1987) and the mean estimates in educational psychology (i.e., $d = 0.33$, see Gall et al., 1996). More generally speaking, it seems that most of the recent empirical research testing the effects of jigsaw on learning was carried out on relatively small samples. Indeed, analyzing 22 empirical articles published between 2000 and 2010, Robert et al. (2021) observed a median total sample size of 84 participants. In more recent articles (between 2010 and 2019), the median total sample size was estimated at 60, based on 77 articles. According to Slavin and Smith (2009): "Small sample effects have significant potential to undermine the scientific validity and the practical utility of program effectiveness reviews in education." (p. 505). This risk could be particularly important in the jigsaw literature because of the overreliance of small sample studies, with important implications for teachers, policy makers and practitioners (Cheung & Slavin, 2016; Kraft, 2020). To sum up, recent literature suggests a strong main effect on learning whereas pioneering work arguing that Jigsaw had little to no significant effect on student learning (Newmann & Thompson, 1987; Moskowitz et al., 1983; 1985). Thus, providing empirical evidence regarding whether Jigsaw significantly impacts learning still represents a particularly important challenge in this literature.

Thus, the purpose of the present research is to contribute to the empirical evidence testing the effect of the jigsaw approach on learning outcomes amongst sixth graders across five randomized and well-powered experiments.

Overview

The goal of the present research is to test the effectiveness of a jigsaw intervention on learning outcomes compared to individualistic (experiments 1 and 2) or teaching-as-usual (experiments 3A, 3B, and 3C) conditions. Across five experiments, we hypothesized that jigsaw should increase learning as compared to the control condition. Following several authors who consider an effect size of $d = 0.40$ as a "threshold of practical significance" (Gall et al., 1996; Hattie, 2010; Springer et al., 1999), we expected the size of the differences between the two conditions to be close to $d = 0.40$. Thus, we performed an a priori power analysis with G*Power 3.0.10 (Faul et al., 2007), considering an effect size of $d = 0.40$ as both a practical significance threshold (Hattie, 2010) and the smallest effect size of interest (i.e., SESOI, Lakens, 2017). We also used the equivalence test procedure (i.e., "Two One-Sided Test" or "TOST", Lakens et al., 2018) to confirm or reject the presence of such effects in our studies. In the present research, we chose to focus specifically on sixth-graders for two reasons. First, it closely matches the population investigated in the seminal studies on the jigsaw classroom (Aronson & Bridgeman, 1979; Blaney et al., 1977; Lucker et al., 1976). Second, this population has not yet been exposed to the consecutive selection process. Indeed, in France, where the studies were all conducted, neither selection nor differential orientation occur before 8th grade. This is important because in some of the recent research, the effects of Jigsaw were examined at the higher education level, namely, on a sub-population of students who might have specific profiles (e.g., high academic level, self-regulation skills, etc). That is not the case of 6th graders who correspond to the general, unfiltered, population.

The power calculation was performed for a one-tailed independent *t*-test for E1 and E2 (i.e., single session studies, between-participants conditions) and a paired *t*-test for E3A, E3B, and E3C (i.e., semester-long studies, within-participant conditions). The estimated sample sizes were $N_{E1\&E2} = 216$ and $N_{E3A,B,C} = 55$, respectively.[1] Table 1 presents a summary of the characteristics of the five experiments. For each of these experiments, the agreement of the head teacher, parents, and the ethical committee (IRB-UCA Ethics Committee of Research IRB00011540-2019-08) was obtained. All experimental materials and databases are accessible at https://osf.io/4pwzy/?view_only=bfa42ad3c076490eaee0fb7bc1a137d7[2]

Finally, following Goh et al. (2016) recommendations (see also Borenstein et al., 2010; Cumming, 2014), we conducted an internal mini meta-analysis to estimate the overall effect size in the five experiments and its dispersion. Our goal was not to establish generalities about jigsaw teaching, but rather provide a synthesis of the results obtained in the present five experiments. Indeed, as argued by these authors, there are multiple advantages to perform a mini meta-analysis on one own's studies. First, an internal meta-analysis allows the results of several studies to be combined and the estimators obtained via this procedure are theoretically more powerful and accurate than results taken independently. It also switches the focus from the statistical significance of the results and their corresponding *p*-values, to the sizes of the effects, which are more relevant indicators (Funder & Ozer, 2019). This issue is particularly important in the context of the jigsaw studies, since, as developed above, most recent research testing the effect of Jigsaw has been conducted on low powered studies and publication bias may occur. Moreover, an internal meta-analysis can be valuable if the cumulative evidence of one's studies fail to reject the null hypothesis, or if a negligible effect size is found.

[1] As the experiments were conducted in real classroom settings, we decided to keep all pupils for which we obtained permission. This explains why the sample sizes of the experiments are often higher than the sample estimated in the a priori power calculation.

[2] The number of classes examined in each of the five studies was not sufficient to conduct multi-level analyses. However, we calculated the intraclass correlation coefficient (ICC) for each experiment. All ICCs were between .00 and .11 (median = .03), suggesting rather weak variations between classes (Bressoux, 2020).

Combining several studies can also address whether the effects observed are heterogeneous. Finally, internal meta-analyses also increase transparency and reduce the file-drawer effect (Rosenthal, 1979) by encouraging authors to report all the studies on a research question instead of reporting only statistically significant results (Cumming, 2014).

Table 1

Summary of the Key Characteristics of the Five Experiments.

Experiment	N	Topic	Duration	Control group	Learning content	Learning test	Statistical test
1	252	Mathematics	1 session of 2h	Individualistic learning	Created by the researchers for the experiment	1 test (8 problem-solving exercises, 17 multiple choice questions)	Independent *t*-test
2	313	Earth and life sciences				1 test (9 multiple choice questions, 2 open-ended questions)	Independent *t*-test[3]
3A	110	Physics and chemistry	16 sessions of 1h each	Teaching as usual	Created by teachers, adapted by the researchers to follow a jigsaw procedure	4 tests (problem-solving exercises, multiple choice questions, true or false questions)	Paired *t*-test
3B	74	Earth and life sciences	12 sessions of 1h30 each			5 tests (problem-solving exercises, multiple choice questions, true or false questions)	
3C	101						

[3] The data did not follow a normal distribution so we performed a non-parametric one-way ANOVA (Kruskal-Willis).

Experiment 1

Method

Participants

Experiment 1 was conducted on 264 sixth graders from twelve classes that included approximately 22 children each (M = 21.67, SD = 3.33) in two junior high schools. We removed 12 participants from the sample as data were missing (i.e., 10 did not answer the entire questionnaire and 2 did not answer the learning test). With the remaining 252 participants, we performed a median-absolute detection (MAD) analysis to detect outliers (Leys et al., 2013). The test revealed that no participants had a variation larger than 2.5 times the median absolute deviation (i.e., moderately conservative criterion), suggesting that we had no true outliers in the sample. Hence, the final sample consisted of 252 participants (121 were categorized as low socioeconomic status (SES) and 131 as high SES children using their parents' occupation, see Smeding et al., 2013 for a similar categorization), 147 girls and 105 boys (M_{age} = 11.14 years, SD = 0.41, min = 10, max = 13), equally distributed into the two experimental conditions, $\chi^2_{gender} < 1$, $p = .983$ and $\chi^2_{SES} = 2.27$, $p = .132$.

Procedure

The experiment took place during class time and lasted two hours in total. Before the experiment started, the classes were randomly assigned to one of the two conditions: jigsaw (n = 127) or individualistic (n = 125). Three experimenters (graduate students) were in charge of the lesson and were trained to teach the lesson as identically as possible. The whole lesson was handled by the experimenters and teachers were not involved. During the first 10 minutes, the experimenters presented the general procedure of the lesson. They also explained that an individual learning test would occur at the end of the learning session.

Participants then worked either individually (i.e., control group) or in jigsaw groups (i.e., experimental group). In the jigsaw condition, the experimenters arranged the classroom to make group work possible by setting five chairs around the tables. They followed the classical three-step procedure indicated by Aronson and Patnoe (2011). First, children were randomly assigned to one of the jigsaw groups. Each of them then received a different part of the pedagogical content and, during the first 10 minutes (step 1), they read their assigned text and answered the guide questions individually. Next, expert groups (step 2) were formed by bringing together participants who had worked on the same text during step 1. They had 10 minutes to discuss their topic, check understanding, and prepare to explain it to the other members of their jigsaw groups. Finally, participants went back to their initial jigsaw group (step 3) and took turns teaching the part they studied to the other members of their group, for a total duration of 20 minutes. There was no competition between groups. In the control group, participants had a similar amount of time (40 minutes) to read and study the three parts of the lessons (texts and guide questions) individually. In the two conditions, participants then had a five-minute break. After the break, they took the learning test and answered a questionnaire measuring several socio-affective and demographic variables[4]). Finally, the experimenter explained the goal of the study, debriefed participants, and provided the corrections of the test.

Material

Learning material. The learning material dealt with mathematical concepts (i.e., arithmetic, proportionality, diagrams, fractions, and geometry). It consisted of several sheets of approximatively 550 words that included both a lesson (e.g., problem solving and

[4] These variables will not be discussed further in the present article, in which we chose to focus exclusively on the learning outcomes. The complete dataset including these measures is available on OSF.

calculation procedures) and guide questions under the form of exercises. The full learning material is presented on the OSF page of the project.

Learning test. The learning test contained 25 exercises which each tap into different competences addressed during the two-hour lesson. Indeed, the learning material included several different mathematical concepts (i.e., arithmetic, geometry, proportionality, diagrams and fractions). Thus, the learning test included 17 multiple choice and 8 problem-solving questions, with each of these questions referring to one or the other of these concepts (e.g., "*Three-fifths of 270 is written as*: …" referred to fractions). These exercises were graded by the experimenter, who was blind to the experimental condition. Then, we averaged the exercises scores into a mean learning score. This score ranged from 0 (lowest grade) to 20 (highest grade, M = 10.92, SD = 4.20, *min* = 0.80, *max* = 19.60), which corresponds to the traditional grade in French schools[5]. It is worth noting that this learning test was taken individually and the final score did not count for the final mathematical trimester grade.

Prior performance in the subject. Participants' prior performance in the subject was measured by their previous quarterly grade in mathematics. This grade could range from 0 (lowest grade) to 20 (highest grade; M = 13.85, SD = 3.84, *min* = 3.20, *max* = 20) and was obtained before the implementation of the experiment through the school head-teacher.

[5] To ensure the reliability of the learning measures in Experiments 1 and 2, we conducted a hierarchical CFA where subscale factors were indicated by items in that subscale, and then a second-order factor was indicated by the subscale factors. Maximal reliability was satisfactory for each first- and second-order factor in Experiment 2 (0.659 and 0.651 respectively), but could suggest item redundancy for Experiment 1 (0.974 and 1.149 respectively) as indicated by the values being greater than .90 (Aguirre-Urreta et al., 2019; Tavakol & Dennick, 2011). In further analyses, second-order latent factors scores were saved for each participant, and we rerun all analyses using these scores as dependent variables. These analyses led to similar conclusions as no positive effect of jigsaw on learning were found.

Results

Assumption Checks

To check for potential grouping bias, we performed an independent *t*-test on prior performance in the subject and did not find any differences between the experimental groups $t(250) = 0.03$, $p = .988$, $d = 0.00$, 95% CI [-0.25, 0.24]. The normality test showed partial support for the hypothesis of a normal distribution of residuals (Shapiro-Wilk = .99 $p = .009$, Kolmogorov-Smirnov = .05, $p = .600$), and Levene's test of the equality of variances suggests a homogeneous variance between groups, $F(1, 250) = 1.41$ $p = .236$.

Learning Outcomes

We used an independent Student's *t*-test to compare the scores on the learning test between the two groups. Contrary to our expectation, participants in the jigsaw condition ($M = 10.83$, $SD = 4.28$) did not outperform participants in the individualistic condition ($M = 11.01$, $SD = 4.13$), $t(250) < 1$, $p = .628$, $d = -0.04$, 95 % CI [-0.29, 0.20].[6] We checked for statistical equivalence using the two one-sided test procedure (Lakens et al., 2018). The results demonstrate that both the upper and lower bounds are rejected under the assumption that the null hypothesis is true, indicating that the observed effect size can be considered as statistically equivalent to zero ($p_{\text{upper}} = .004$ and $p_{\text{lower}} < .001$, respectively).

Discussion

In this first experiment, contrary to our hypothesis, participants working with the jigsaw technique did not outperform those working individually on the learning test (both performed at the same level). However, several limitations can be noted and should be addressed to further test whether the jigsaw approach improves learning. Notably, the

[6] For all the experiments, covariance analyses including gender and prior performance in the subject in the analyses are reported in the Supplementary Material (see Table S4, S5, S6, S7 and S8 respectively). The main effect of the condition remained non-significant and small in size even when these covariates were entered in the analyses.

randomization took place at the whole class level. Even if we checked for the existence of potential discrepancies in prior performance in mathematics between the two conditions and did not find any significant differences, one cannot exclude that the lack of significant differences between the two groups on learning might be due to the characteristics of the classes involved in each condition.

In particular, informal discussions with teachers indicated that some parts of the learning material used in the experiment may have had been discussed in certain classes but not others. Experiment 2 made two main changes to address this issue. First, the children in the classes were fully scrambled in the two conditions; thus, each child was randomly assigned to one of the conditions. Second, the learning materials were constructed in such a way that they dealt with issues that were not part of the program for the sixth graders, meaning they were new for all participants.

Experiment 2

Method

Participants

The head-masters of two junior high schools (different from those used in Experiment 1) agreed to participate in Experiment 2. Fourteen classes including approximately 22 children each (M = 22.57, SD = 4.18), with a total sample of 319 children, participated in this experiment. We dropped six participants for whom data were missing (i.e., three missing questionnaires and three demographic variables). As in Experiment 1 (E1), we performed a MAD test to detect outliers on the learning score. Two participants had scores that varied from than 2.5 times from the absolute median of the sample but did not substantially influence the data (Cook's distance < 0.03). Consequently, we retained these two participants in the analyses. The final sample consisted of 313 participants (137 low SES and 176 high SES

children; 169 girls, 144 boys, M_{age} = 11.60 years, SD = 0.56), equally distributed between the two conditions, $\chi^2_{gender} = 2.269$, $p = .132$ and $\chi^2_{SES} < 1$, $p = .335$.

Procedure

The procedure was similar to the first experiment (E1), with some minor adjustments. First, only one experimenter (a graduate student) handled the experimentation. Second, instead of having a whole class assigned to the jigsaw or individualistic condition, we asked the head-teacher to scramble the classes together in new groups, which we randomly distributed into the two experimental conditions.

Material

Learning material. The learning material dealt with the sleep cycle, a life sciences-oriented subject. It consisted of a six-page text related to sleep that contained several pictures and guide questions. This text was divided into three sub-topics of approximatively 400 words each and originated from a free and reproducible website: http://lecerveau.mcgill.ca/flash/d/d_11/d_11_p/d_11_p_cyc/d_11_p_cyc.html

Learning test. The learning test contained 11 exercises: nine multiple choice questions with three possible answers and two open-ended questions related to the whole learning material (i.e., the three sub-sections). As in Experiment 1, the test was assessed individually and did not count for the children's final semester grade. It was corrected by the experimenter, who was blind to the experimental condition. The scores obtained to each exercise were averaged into a composite learning score. This score initially ranged from 0 (lowest score) to 8.5 (highest score), but was transposed into a score ranging from 0 to 20 (M = 9.12, SD = 3.34, min = 2.22, max = 20, see Table S2 in Supplementary Materials for details).

Prior performance in the subject. Prior performance in the subject was measured by children' quarterly mean grade in life sciences. This grade could range from 0 (lowest grade) to 20 (highest grade; M = 14.03, SD = 2.80, min = 5.00, max = 20.00).

Results

Assumption Checks

As with Experiment 1, we did not expect to observe differences between the two conditions before the experiment started. To check this assumption, we performed an independent t-test prior subject performance. We observed a small mean difference in favor of the jigsaw group ($M_{\text{difference}}$ = 0.47, SE = 0.30), although it did not reach statistical significance $t(311) = 1.54$, $p = .124$, $d = 0.17$, 95% CI [-0.02, 0.21]. The normality test rejected the hypothesis of the normal distribution of residuals (Shapiro-Wilk = .97, $p < .001$, Kolmogorov-Smirnov = .11, $p < .001$) and Levene's test of equality of variances was non-significant $F(1, 311) = 0.12$, $p = .727$. Consequently, we tested our hypothesis with a non-parametric one-way ANOVA.[7]

Learning Outcomes

Again, the jigsaw technique made no difference: Participants in the jigsaw condition (M = 9.12, SD = 3.43) performed as well as the participants in the individualistic condition (M = 9.11, SD = 3.27), $t(311) = 0.03$, $p = .979$, $d = 0.00$, 95% CI [-0.22, 0.22]. The TOST procedure indicated that the t-test for the observed effect size was not statistically different from zero, ($p_{\text{upper}} < .001$ and $p_{\text{lower}} < .001$, respectively). Thus, as in Experiment 1, we did not find any empirical support for our hypothesis that the jigsaw approach improves learning.

[7] It is worth noting that using parametric tests leads to similar conclusions.

Discussion

In Experiment 2, both the randomization procedure and the novelty of the learning material were improved. However, we still did not find support for a beneficial effect of the jigsaw technique. The observed mean differences were very close to zero and, thus, were far from meeting our expected SESOI of $d = 0.40$. However, important limitations of E1 and E2 may be noted. First, our measure of learning outcomes showed relatively poor psychometric validity. Second, a major limitation of both E1 and E2 is that the learning conditions were quite different from usual ecological (real-class) learning conditions. Indeed, the time devoted to the lesson was relatively short (two hours). Most researchers suggest that cooperative learning, and particularly jigsaw approaches, need several sessions before being successfully implemented because of its challenging structure (Roseth et al., 2019). In addition, the two-hour lessons conducted in Experiments 1 and 2 were decontextualized from the regular class lessons. Such a decontextualized lesson, as compared to real-class lessons, may lack pedagogical and tangible goals — or relevance — for the children (Topping et al., 2017). The relevance of the common goal (e.g., being responsible for each other's learning) is particularly important in such cooperative settings because it reinforces positive interdependence between learners, a primary condition for cooperative learning to be effective (Johnson & Johnson, 2009). To address these important limitations, Experiments 3A, 3B, and 3C were conducted in real classroom contexts and with voluntary real junior high school teachers instead of experimenters. These teachers were recruited and accompanied by the experimenter to shift their pedagogical class scenario into a jigsaw teaching scenario. They agreed to use the jigsaw scenario for half of their groups and to maintain their usual practices for the other half of their groups during the first part of the semester, with the approaches being reversed for the second part of the semester

Experiment 3A

Method

Participants

A physics and chemistry high school teacher volunteered to participate. The sample of the experiment included 122 children in six different classes (N = 20.33 per class, SD = 2.67) and their teacher. The experiment took place during the usual one-hour physics and chemistry lesson that occurred each week for a period of 16 weeks. We removed 11 participants from the analyses because of missing data (i.e., 10 demographic variables and one learning test). To detect outliers, we computed the overall score on the four measures of learning and used the MAD technique. One participant had a median variation greater than 2.50 with a potential influence of data (Cook's distance > 0.05). Consequently, this participant was also removed from the analyses. Hence, the final sample consisted of 110 participants (76 low SES and 34 high SES children), 52 girls and 58 boys, equally distributed in the two conditions, $\chi^2_{gender} < 1$, $p = .687$ and $\chi^2_{SES} < 1$, $p = .325$.

Procedure

Randomization. In the first phase of the experiment (January to March, sequences 1 and 2), half of the classes (k = 6) were assigned to the jigsaw condition (n = 53) whereas the other half was assigned to the teaching-as-usual (TAU) condition (n = 57). In the second phase, after a two-week school break, the groups were reversed (March to May, sequences 3 and 4). The children who initially worked in jigsaw conditions then worked under teaching as usual conditions and vice versa (see Table 2 for a summary).

Table 2

Distribution of the Classes Within the Experimental Design.

		First step: Weeks 1 to 6			Second step: Weeks 7 to 16	
		Test 1	*Test 2*		*Test 3*	*Test 4*
Classes	2, 3, 6	Jigsaw (*treatment*)		School Break	Teaching as usual (*control*)	
	1, 4, 5	Teaching as usual (*control*)			Jigsaw (*treatment*)	

The teacher, who received an individual training in cooperative education, was responsible for teaching the six classes. Before the experiment began, each of the classes was randomly assigned to the jigsaw or TAU condition. As the teacher committed to follow the jigsaw versus TAU procedure, it was not possible to assign participants of a same class group to different experimental conditions. Thus, randomization occurred at the whole class level. The researchers accompanied the teacher in adapting the learning material for the jigsaw method (e.g., dividing each topic into several sub-topics, implementing experimentations and exercises). As in the first two experiments, all the participants studied exactly the same learning material, but in the jigsaw condition, the teacher followed the three-step jigsaw procedure as developed by Aronson et al. (2011), whereas in the TAU condition the teacher committed to not changing his usual practice. This usual practice combined unstructured group work in pairs, some larger group discussions, and individual work. At the end of each of the four sequences, children took the learning test corresponding to the sequence. A summary of the procedure is provided in Table 3. It is worth noting that we pre-registered the hypothesis of a main effect of Jigsaw as well as the general procedure for the experiments 3A, 3B and 3C. The details can be found at: https://osf.io/z7rd3

Material

Learning material. The class dealt with the theme "Matter, Movement, Energy and Information." The theme was divided into sequences covering several lessons that followed a

similar structure. First, the teacher introduced the sequence and children had to make a hypothesis/prediction about a topic (e.g., sequence 3: “How much sugar can be dissolved in water?”). The children then implemented a scientific experiment with manipulations and observations to test the hypothesis (e.g., “Mix 10g of sugar with 20cl of water”). Finally, they were asked to answer a list of questions in order to communicate the results to the teacher, who provided corrections (e.g., “Was the sugar (10g) entirely dissolved in water for group 3?”). In the jigsaw condition, the topic and its corresponding hypothesis were translated into several activities (e.g., mix a certain amount of sugar, sand, or vinegar in water). The first step of the procedure included individual time in small groups created by the teacher (i.e., jigsaw groups) to discover the topic, read the instructions, and identify the group members. In the expert groups, children had to test a specific prediction — a sub-topic of the general theme — and implement an experiment following some guide questions. These activities were related to the main topic and consisted of several practical examples (e.g., for sequence 3 on “Mass and Matter,” the children had to learn about mass conservation). To do so, the teacher implemented four activities in which the children had to mix several contents (e.g., sugar, sand, salt, or vinegar) in water, measure the mass with the corresponding tools, and report their results in the form of a test of hypothesis. In the final step of the jigsaw procedure, the children who studied a particular prediction had to present their observations and results to the others to acknowledge the differences and similarities of their experiments and how they related to the main topic presented at the beginning of the course. The teacher then gave a collective correction to the class.

In the TAU condition, the same pedagogical material was taught, but the teacher did not split the class into jigsaw and expert groups. The activities were implemented with teacher-centered instruction and unstructured groupwork: the children had to manipulate and use the same guidelines as the ones studying in the jigsaw condition. When the children

worked on different contents (e.g., different types of mixtures), they did not communicate their results with the others as in the expert groups. Instead, the teacher gave the correction collectively.

Learning test. At the end of each sequence, children took learning tests. The first two tests corresponded to the first part of the learning material, and the other two corresponded to the second part of the learning material. These tests were designed by the teachers and consisted of asking for definitions, completing problem-solving exercises, and drawing diagrams that covered the different topics addressed during the lessons. Thus, each of the exercises tap into different competences that were all addressed during the class and were graded by the teacher. Then the scores obtained to these exercises were averaged into a final composite score which ranged from 0 (i.e., lowest grade) to 20 (i.e., highest grade). Unlike Experiment 1 and Experiment 2, these grades counted in the children's final trimester grade in the topic and the teacher was not blind to the conditions[8].

In order to control for the heterogeneity of scores, we computed the mean grade obtained in the jigsaw condition (i.e., xJigsaw) and the mean grade obtained in the TAU condition (i.e., xTAU) by averaging the corresponding tests. Thus, each participant had a mean score for the learning tests covered during the jigsaw period and another mean score for the learning tests covered during the TAU period. For example, the mean score of xJigsaw corresponds to the mean of the first two tests for children who began the experiment assigned in this condition. Consequently, for these children, xTAU is computed by aggregating their

[8] The learning tests used in Experiments 3a, 3b and 3c were designed and corrected by the teachers. Consequently, we were only able to collect subtest scores (but not item level scores) and thus, we could not compute the alpha score at the item level. Alphas computed at the subtest level were satisfactory, with the exception of Experiment 3a ($\alpha = .483$ for Exp.3a; $\alpha = .723$ for Exp.3b; $\alpha = .797$ for Exp.3c). We believe the fact that these scores were designed, corrected and used in final grading by the teachers supports their ecological validity, the low alpha value of Experiment 3a notwithstanding.

scores for Tests 3 and 4 (see Table S3 in Supplementary Material for the descriptive statistics of each of the tests separately).

Prior performance in the subject. Children' quarterly mean grade in physics and chemistry was used to measure prior academic level (M = 13.07, SD = 3.53, *min* = 4.38, *max* = 19.25).

Results

Assumption Checks

The independent t-test showed that initial differences between the two order groups (jigsaw first vs. TAU first) were non-significant, $t(108) = .027$, $p = .785$, $d = -0.05$, 95% CI [-0.43, 0.32]. The normality test showed full support for the hypothesis of a normal distribution of residuals (Shapiro-Wilk = .99 p = .907, Kolmogorov-Smirnov = .04, p = .992).

Learning Outcomes

In the present experiment, participants had two learning scores: one computed for the jigsaw lessons and another obtained for the TAU lessons. Consequently, learning scores were used as a within-subject variable, and our hypothesis was tested with a paired *t*-test. The results show that — again, contrary to our expectations — learning outcomes on the learning test were not higher when participants were in the jigsaw condition (M = 10.78, SD = 3.58) compared to the teaching-as-usual condition (M = 11.15, SD = 3.34), $t(110) = 0.76$, $p = .777$, $d = -0.07$, 95% CI [-0.26, 0.11]. The TOST procedure indicated that the *t*-test for the observed effect size was not statistically different from zero, ($p_{upper} < .001$ and $p_{lower} = .001$, respectively).

Experiment 3B

Method

Participants

The sample comprised 85 children who all had the same teacher and were distributed across four classes (N = 21.25 per class, SD = 0.75), who met once in a week for a 1.5-hour lesson in earth and life sciences over 16 weeks. We removed 10 participants from the analyses because of missing data (i.e., nine demographic variables and one learning test) and one participant (outlier) who showed potential influence on the data (MAD = 2.46 and Cook's distance = 0.12). The final sample consisted of 74 participants (16 low SES and 58 high SES children), 35 girls, 36 boys, and 3 unreported, equally distributed in the two conditions, $\chi^2_{gender} < 1$, $p = .718$ and $\chi^2_{SES} < 1$, $p = .658$.

Procedure

We used the same design as in Experiment 3A (see Table 2).

Material

Learning material. The pedagogical content corresponded to the theme of environment and biodiversity, which was divided into four sequences covering several aspects of the theme. Each sequence consisted of two or three activities, and a learning test was given to the children once the sequence was over.

The structure of the course was quite similar to E3A. First, the teacher introduced the sequence and a general subject at the whole-class level. For example, in sequence 2, the problematic was "How to explain the distribution of species in their environment?" The children had to make hypotheses about this question and were then introduced to an activity (e.g., study the behavior of animals during winter) that could be done using either the jigsaw (i.e., jigsaw condition) or usual procedures (i.e., TAU condition). In the jigsaw condition,

different activities were distributed to the children to create jigsaw groups. The first step consisted of working individually on a specific example (e.g., In sequence 2, a certain animal: deer, locust, snake, or bird). Children then met in expert groups to answer questions about their examples (e.g., "What provokes such behavior in the snake during winter?"). In the last step, children in the jigsaw groups had to complete a task where all examples were needed (e.g., filling a table with the corresponding animal and its characteristics). In the TAU condition, there was no cooperation; children had to work individually on the same examples sed in the jigsaw groups. Thus, there was no interdependence between children. In both cases, the teacher corrected the activity with the whole class.

Learning test. The same procedure as in Experiment 3A was used to compute the mean grade obtained in the jigsaw condition (i.e., xJigsaw) and the mean grade obtained in the TAU condition (i.e., xTAU).

Prior performance in the subject. The experiment took place during the first trimester and, consequently, we could not have children's prior performance in earth and life sciences. Instead, we computed a prior academic performance in French and mathematics from their fifth-grade scores (M = 12.49, SD = 3.13, min = 4.37, max = 19.32). This score could range from 0 (lowest grade) to 20 (highest grade).

Results

Assumption Checks

The independent t-test showed that initial differences between the two order groups (jigsaw first vs. TAU first) were non-significant, $t(72) = 0.48$, $p = .634$, $d = 0.11$, 95% CI [-0.35, 0.57]. The normality test showed support for the hypothesis of a normal distribution of residuals (Shapiro-Wilk = .99, $p = .702$, Kolmogorov-Smirnov = .08, $p = .725$).

Learning Outcomes

We did not observe statistically significant differences between participants in the jigsaw (M = 15.75, SD = 2.54) and TAU conditions (M = 15.63, SD = 2.81), $t(73) = 0.39$, p = .349, d = 0.05, 95% CI [-0.18, 0.27]. The TOST procedure rejected both upper and lower bounds under the assumption that the null hypothesis was true ($p_{upper} < .001$ and $p_{lower} = .002$, respectively).

Experiment 3C

Method

Participants

The sample included 108 children distributed across five classes (N = 21.80 per class, SD = 1.52) who met once a week for a 1.5-hour lesson over 16 weeks. Two teachers were in charge of respectively three and two classes. Each of the two teachers taught both jigsaw and TAU conditions. We removed six participants from the analyses because of missing data (i.e., absence for more than half of the evaluations). One participant was also removed from the analyses using the same procedure and threshold as before (MAD = 1.70 and Cook's distance = 0.10). Hence, the final sample consisted of 101 participants (55 females, 44 males, and 2 not reported) equally distributed in the two conditions, $\chi^2_{gender} = 2.43$, p = .119.

Procedure

We followed the same procedure as in Experiment 3A and 3B, with two minor changes. First, we did not obtain the school's agreement to gather data on the initial academic achievement of the children. Consequently, and unfortunately, in this experiment, we could not check for initial differences prior to the experiment. Moreover, unlike in Experiments 3A and 3B, in this experiment, teachers taught five (not four) sequences. Learning scores were

thus based on five (not four) tests. Two teachers took part in the experiment and were in charge of three and two classes each, respectively.

Material

Learning material. The learning material (i.e., pedagogical activities, learning tests) was similar to, although slightly different from the content taught in Experiment 3B. Each sequence corresponded to several activities. In the jigsaw condition, these activities were separated into different parts, each of them exploring a specific part of the topic. For example, the first sequence dealt with the identification of different tree leaves. In the first step, jigsaw groups were formed with children studying different examples of leaves individually (e.g., maple, oak, elm, birch). Then they acted as experts to analyze specific characteristics of these examples (e.g., disposition of the leaves on the branch, the outline and sharpness of the leaves, the composition of the leaves). Once back in the jigsaw groups, children had to describe their examples with respect to the specific characteristics to which they were exposed in their expert groups. In the TAU condition, the children did not work cooperatively but studied similar examples; thus, the same kinds of competencies and knowledge were assessed in both conditions.

Learning test. Three tests evaluated children's learning on environment and biodiversity, and two other tests covered the theme "From seed to plant." The learning tests were quite similar as the ones used in Experiment 3A and 3B (i.e., definitions, problem-solving exercises, and drawing of diagrams) and were graded from 0 (i.e., lowest grade) to 20 (i.e., highest grade) by the teachers. We used the same procedure to compute the mean scores obtained in the jigsaw and teaching-as-usual conditions as in Experiments 3A and 3B.

Results

Assumption Checks

As previously noted, we could not check for initial differences in prior performance in earth and life sciences between classes. Nevertheless, the normality test showed support for the hypothesis of a normal distribution of residuals (Shapiro-Wilk = .98 p = .354, Kolmogorov-Smirnov = .07, p = .738).

Learning Outcomes

Performance on the learning test was not significantly higher when participants were in the jigsaw condition (M = 15.51, SD = 3.13) compared to the TAU condition (M = 15.33, SD = 3.31), $t(100) < 1$, p = .293, d = 0.05, 95% CI [-0.14, 0.25]. Once more, TOST procedure indicated that both upper and lower bounds are rejected under the assumption that the null hypothesis is true, indicating that the observed effects size can be considered as statistically equivalent to zero ($p_{\text{upper}} < .001$ and $p_{\text{lower}} < .001$).

Internal Mini Meta-analysis

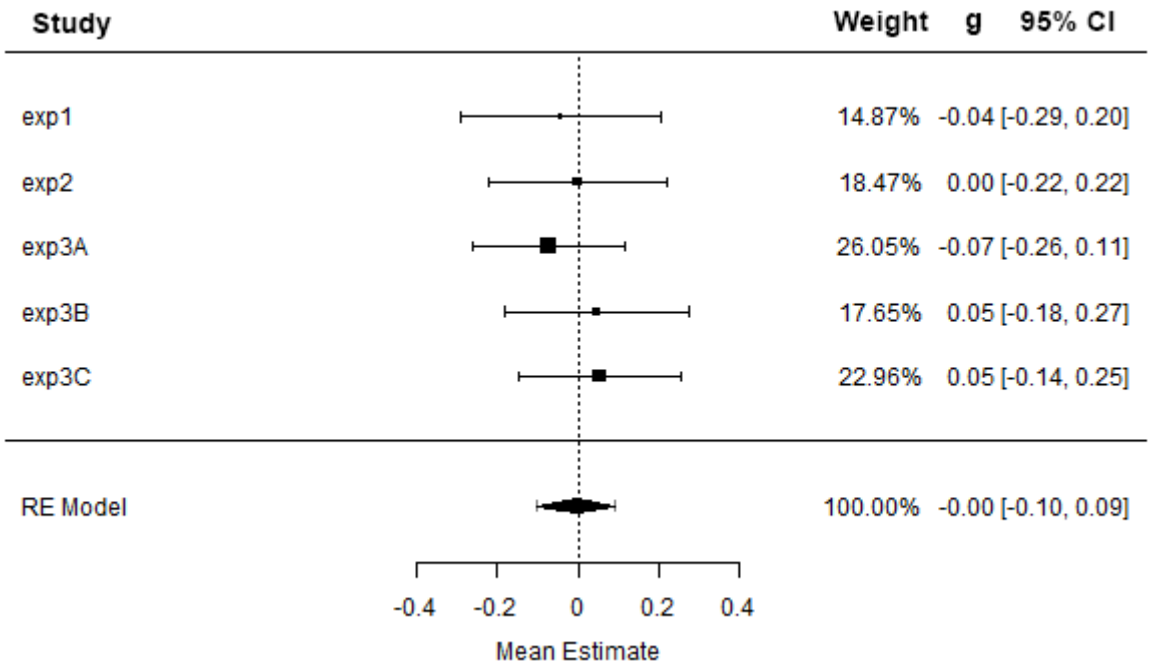
Overall, we did not find support for the hypothesis that jigsaw teaching increased learning when compared to a teaching-as-usual condition. Even in a real classroom environment, for a duration covering the whole trimester and with a real teacher known by children, we did not observe the expected positive effects of jigsaw. To summarize the findings across the five experiments, we decided to perform an internal mini meta-analysis to obtain a summary effect of jigsaw on learning (Borenstein et al., 2010; Cumming, 2014; Goh et al., 2016). We performed our analysis with Jamovi 1.1.8.0 and used a random-effect model with the "MAJOR" extension. As Goh et al. (2016) stated, this approach is usually conservative if a few studies are available, but it affords greater generalizability. We used the

Der-Simonian-Laird method for the estimate's computation, because it is one of the most frequently and simple approach to implement (Veroniki et al., 2016).

The results showed that the summary effect of the jigsaw condition on learning was not statistically significant and not different from zero, $Z = 0.10$, $p = .921$, $ES = 0.00$, 95% CI [-0.10, 0.09]. There was no heterogeneity in the effect sizes across the different experiments, $Q(4) = 1.12$, $p = .890$). At the same time, the proportion of observed variance (I^2) was also zero, as no differences between the experiments were observed. The forest plot of observed effect sizes is presented in Figure 1. Considering this very low heterogeneity between effect sizes, we can assume that the null-hypothesis was not rejected (Borenstein et al., 2010). In other words, such results demonstrate that the five experiments shared a common effect size of zero and the variation between the observed effect sizes was mostly spurious. Although we cannot conclude the absence of an effect, we consistently showed across five experiments that the jigsaw method did not increase learning outcomes to an extent deemed of practical interest (Gall et al., 1996; Hattie, 2010).

Figure 1

Forest Plot of Observed Effect Sizes



General Discussion

The goal of the present research was to test the effectiveness of the jigsaw classroom — a cooperative method that structures positive interdependence by allocating complementary resources to group members — on sixth graders' learning outcome in randomized and well-powered experiments.

To this end, we compared classes in which the jigsaw procedure was implemented to control classes in which participants worked on the exact same pedagogical content either individually (individualistic conditions: Experiments 1 and 2) or under usual conditions of teaching (teaching-as-usual conditions: Experiments 3A, 3B, and 3C). Across the five experiments, we did not find any evidence that the jigsaw condition significantly increased learning: the summary effect size was smaller than our expected SESOI, close to zero, and not statistically significant. Because we used an a priori SESOI of d =0 .40, disposed of enough statistical power to detect an effect of this size and rejected the presence of such effects with the TOST procedure (Lakens et al, 2018), we believe we can conclude on the fact that the

conditions did not affect learning in the present set of experiments. Kraft (2020) recently suggested that a benchmark ranging from $d = 0.05$ (small effects) to $d = 0.20$ (large effects) should be preferred, in experimental designs, over the traditional guidelines by Cohen (1962) when assessing the effects of educational interventions on learning because it reflects more closely research conducted in the field of educational psychology. Nevertheless, in our set of studies, the overall effect of jigsaw on learning outcomes in our five experiments was close to zero. Thus, we think we can be confident in rejecting the presence of a positive effect of jigsaw on learning in the present experiments.

Although it is difficult to conclude why the jigsaw approach did not produce the expected positive effects on learning in the five experiments conducted, we think this lack of significant differences raises important questions at both the theoretical and practical levels. In particular, we next discuss four possible explanations of this lack of significant difference: perhaps the jigsaw approach did produce the expected effects on learning but our dependent variable was not accurate enough to capture this positive effect; perhaps the jigsaw approach is an efficient technique to increase learning but only in some conditions that were not met in the present experiments; perhaps the jigsaw approach produced positive effects on learning in the five experiments but for some reasons, the control groups were also efficient; and previous research might have overestimated the size of the effect and the current, true effect of jigsaw on learning is either nonexistent or at least of a much lower size than initially claimed. We now discuss each of these possibilities and provide some recommendations for future jigsaw research.

Measure of Learning

The first possibility questions the tests used in our research to assess the effects of the jigsaw approach on learning. Although the learning tests used in the five experiments were constructed with teachers who were highly familiar with the sixth-grade curriculum and in all

experiments the test scores positively correlated with children' academic grades (see Supplementary Material, Table S1), it is still possible that these tests were not discriminating enough to capture the expected beneficial effects of the jigsaw approach. In particular, these tests were not validated learning tests, which is a limitation of the present research. The low internal reliability of our learning outcome measures is particularly of concern (e.g, low Cronbach's α in experiment 3A, little item-variance in experiment 2, scale redundancy in Experiment 1, see footnotes 5 and 8), and future research should test the effect of jigsaw on validated learning outcome measures (Robert et al., 2021). However, it is important to note that our results do not show any positive effects of jigsaw on learning even if the analyses were performed at the subscales (E1 and E2) or the subtests level (E3A, E3B, E3C). In addition, although the learning tests used in the present research were quite similar to that used in previous research in the field (Robert et al., 2021), we exclusively measured short-term learning, which leaves open the possibility of the beneficial effects of the jigsaw approach on long-term learning (e.g., at the end of the year or semester).

Of particular relevance here, teachers involved in the conducted experiments often reported that their classes' climate improved after their jigsaw experience. According to them, children seemed more likely to cooperate with each other, offer spontaneous help to those experiencing difficulty, and transfer the jigsaw structure of cooperative learning to informal group work. These unformal reports suggest that the jigsaw technique — although it made no difference in children's level of learning — may have had beneficial effects on complementary dimensions that were not assessed in our research (Desforges et al., 1991; Walker & Crogan, 1998). For example, Walker and Crogan (1998) reported that jigsaw reduced ethnic prejudices and more positive attitudes toward peers (e.g., number of friends). However, such effects of jigsaw on prejudice reduction are not systematically observed (Bratt, 2008). Similarly, Roseth et al. (2019) recently showed that undergraduate students in the

jigsaw condition initially reported higher initial levels of individualistic efforts than students in a teaching-as-usual condition. More precisely, participants in the jigsaw condition initially reported high levels of competition, as well as low levels of cooperation, when compared to control condition. Thus, we believe other variables (e.g., social skills, self-efficacy, intrinsic motivation; Nokes-Malach et al., 2015; Roseth et al., 2019; Voyles et al., 2015) should be systematically investigated in future research to test whether jigsaw affects not only learning outcomes but also other variables related, notably, to the classroom climate or social relationships among peers.

Existence of Hidden Moderators

It is important to note that the overall null effect of jigsaw on learning outcomes observed in the present studies could be due to the existence of hidden and unidentified moderators. Notably, the five experiments were carried out with sixth-graders, on specifics scientifical contents and in a limited period of time, particularly for the first two experiments. As far as age is concerned, according to Aronson and Patnoe (2011), students must have sufficient reading skills to benefit from Jigsaw teaching, and a certain level of cognitive development, which is usually the case after 4th grade. Similarly, according to Blaney and colleagues, fifth-grade students are "mature enough to function without close teacher supervision and yet may not be so conditioned by years of competitive schooling as to preclude learning cooperative classroom behavior" (1977, p. 123). Thus, children of the present experiments, who were all 6th graders should be old enough to benefit from the jigsaw teaching. However, in a recent meta-analysis, Kyndt et al. (2013) suggested that the effects of cooperative learning may depend on the age of the participants. In particular, since the success of the group (i.e., understanding a course) directly depends on the participants' capacity to learn, synthetize, and explain their topic to their classmates, sixth graders — more so than older students — may not be able to automatically benefit from jigsaw (Buchs &

Butera, 2007; Buchs et al., 2015; Topping et al., 2017). This could explain why in the present experiments, children did not benefit from the jigsaw teaching. However, it is important to note that the extent to which age moderates the effect of cooperative learning on learning outcomes is still under debate in the literature. For example, Kyndt et al. (2013) observed more positive effects of cooperative learning in elementary and tertiary education, compared to secondary education (see Lou et al., 1996, for similar results); yet, other meta-analyses do not report a significant moderation of the effect of cooperative learning by grade level (Cheung & Slavin, 2016; Qin et al., 1995), and Johnson et al. suggest that the effect of Jigsaw is robust whatever the participants' age (Johnson & Johnson, 2002). On the whole, because of these inconsistencies, and since so far, most of the research has been conducted at the university level (Robert et al., 2021), more research is needed in order to examine whether jigsaw, or other cooperative learning methods can produce beneficial effects on learning at the secondary level of teaching.

Surprisingly, we did not find any differences between the short-term experiments (E1 and E2) and the longer-term (semester-based) ones (E3A, 3B, 3C) in the internal meta-analysis. Robert et al. (2021) noted that the median time for the jigsaw interventions duration in recent literature was of nine hours, distributed in four to five weeks. To our knowledge, the impact of the duration of jigsaw intervention has not been addressed yet. In a meta-analysis on peer-assisted learning, Rohrbeck et al. (2003) also did not report significant differences between shorter (i.e., < 20 hours) and longer (i.e., > 20 hours) interventions on academic achievement. However, one cannot exclude that a few weeks are not enough to integrate the jigsaw method's structure and benefit from it, especially with sixth graders (Aronson & Patnoe, 2011). In the same vein, Roseth et al. (2019) observed that, over a whole semester, perceived competition decreased and perceived cooperation increased within jigsaw groups

with time, which may result in larger gains in academic achievement when compared to a business-as-usual condition.

Finally, in terms of pedagogical content, or subject domain, Aronson et al. (1978; 2011) suggest that the subjects best suited for the jigsaw method are those emphasizing narrative and writing skill such as history, geography, and humanities. However, it can be noted that a significant part of the jigsaw literature focuses on scientific content, such as mathematics, physical-chemistry, and life and earth sciences (Robert et al., 2021). Beyond the discipline per se, other researchers argue that both task (Cohen, 1994) and content (Deiglmayr & Schalk, 2015) structure may play an important role in explaining discrepancies in the effect of cooperative learning on learning outcomes. Of particular relevance, Deiglmayr and Schalk (2015) develop the notion of "knowledge interdependence" (i.e., the proximity of concepts in the complementary resources distributed among jigsaw groups) and demonstrate that students working cooperatively on closely related examples (i.e., weak knowledge interdependence) performed better than students working on different concepts (i.e., strong knowledge interdependence). This issue should be examined in future research. Ultimately, enough data should be available to conclude on the conditions under which the jigsaw method does, or does not, improve learning.

Beneficial Control Group?

The lack of a significant difference between the two conditions across the five experiments might also be due to the control groups of the present experiments. Indeed, in previous research, control groups mostly consisted of an individualistic learning condition (Lazarowitz et al., 1994; Şahin, 2011; Tarhan et al., 2013). The teaching-as-usual conditions used as control in Experiments 3A, 3B, and 3C also included some sort of unstructured group work and, thus, were more hybrid than purely individualistic. Thus, the differences between cooperative and hybrid learning may be smaller than $d = 0.40$ (Kraft, 2020). This is

particularly true for the present experiments, which were conducted with teachers characterized by a high motivation profile (i.e., they all volunteered to take part in these costly experiments). On the one hand, this could explain why, in Experiments 3A, 3B, and 3C, we did not obtain significant differences between the control and experimental conditions. On the other hand, we believe that, by doing so, we were in good conditions for testing the mere effects of the jigsaw approach, while maintaining other constant elements of the lesson, including the teachers' profile (thereby avoiding confounds). Conducting similar research with larger sample sizes could be a good way to test whether jigsaw produces smaller effects than those initially expected.

Quality of Previous Evidence Regarding the Effects of Jigsaw on Learning

The final possibility may be that, in previous research, the effect size of the jigsaw approach has been overestimated, and its actual effect on learning is either nonexistent or at least of a much lower size than initially claimed. This point is related to the quality of existing evidence regarding the positive effects of the jigsaw approach on learning. In particular, as previously argued, randomized experiments examining the effect of a jigsaw intervention on learning are actually quite scarce and at risk of overestimating the effect size with small samples (Slavin & Smith, 2009). Indeed, most of this research has been conducted on quite small samples with a high degree of freedom in statistical procedures (e.g., MANCOVAs, multiple comparisons without correction) and often report surprisingly large effects considering the rest of the literature.[9] The large and highly heterogeneous observed effect sizes contrast with previous claims that consider the jigsaw technique as "the least effective approach to learning" (Newmann & Thompson, 1987, p. 6). Consequently, an expected size

[9] For example, $d = 1.74$ from Doymus, 2008; $d = 2.33$ from Gömleksiz (2007); $d = 2.52$ from Tarhan et al., 2013.

of interest of $d = 0.40$ might correspond to an overestimation of the true effect of the jigsaw approach.

This could explain why the current five experiments, failed to produce the expected positive effects. As previously argued, powerful and appropriate tests of the effect of an intervention in randomized and controlled experiments are extremely important before any advice is formulated to teachers and practitioners (Hattie, 2010; Slavin, 2008). The present experiments support the idea that, thus far, the empirical evidence is not strong enough to conclude that a beneficial effect of the jigsaw method on learning exists. If this effect exists, it might be smaller than initially thought (Lakens, 2017).

Conclusions

Poor evidence can be highly misleading for teachers and hinder the process of science itself (Kraft, 2020, Plucker & Makel, 2021). As argued herein, the only way to know whether jigsaw is or is not an efficient technique that should be used in class to increase students' learning is to conduct high powered and randomized controlled experiments (Connoly et al., 2018). Of course, more research is needed to draw clear and straightforward conclusions on the potential for the jigsaw technique to improve learning. However, what our data suggest is that these effects, if any, are very unlikely to apply for sixth-grade children. As science should build on cumulative and empirical evidence (Patall, 2021), we urge researchers to implement sufficiently powered experiments to determine the conditions under which the jigsaw approach does or does not produce positive effects on learning and other variables (Aronson & Patnoe, 2011; Nokes-Malach, 2015; Slavin, 2011). Meanwhile, we encourage teachers to focus instead on interventions based on more solid evidence than jigsaw methods. Let us conclude with a final statement from Johnson & Johnson (2002):

> It is somewhat surprising that so few methods have been evaluated. While any teacher may develop a version of cooperative learning that is very effective, without research studies it is unknown whether other teachers can expect reliable results when the method is used. The unevaluated cooperative learning methods, therefore, should be used with some caution. In addition, there is a need for a new generation of researcher-developers who formulate new operationalization of cooperation for classroom and school use and who subject their formulations to rigorous empirical evaluation. (p. 15)

References

Aguirre-Urreta, M. I., Rönkkö, M., & McIntosh, C. N. (2019). A cautionary note on the finite sample behavior of maximal reliability. *Psychological Methods, 24*(2), 236–252. https://doi.org/10.1037/met0000176

Aronson, E., Blaney, N., Stephan, C., Sikes, J., & Snapp, M. (1978). *The jigsaw classroom.* Sage.

Aronson, E., & Bridgeman, D. (1979). Jigsaw groups and the desegregated classroom: In pursuit of common goals. *Personality and Social Psychology Bulletin*, *5*(4), 438–446. https://doi.org/10.1177/014616727900500405

Aronson, E., & Patnoe, S. (2011). *Cooperation in the classroom: The Jigsaw Method*. Pinter & Martin.

Baloche, L., & Brody, C. M. (2017). Cooperative learning: Exploring challenges, crafting innovations. *Journal of Education for Teaching*, *43*(3), 274–283. https://doi.org/10.1080/02607476.2017.1319513

Blaney, N. T., Stephan, C., Rosenfield, D., Aronson, E., & Sikes, J. (1977). Interdependence in the classroom: A field study. *Journal of Educational Psychology*, *69*(2), 121–128. https://doi.org/10.1037/0022-0663.69.2.121

Berger, R., & Hänze, M. (2009). Comparison of two small-group learning methods in 12th-grade physics classes focusing on intrinsic motivation and academic performance. *International Journal of Science Education*, *31*(11), 1511–1527. https://doi.org/10.1080/09500690802116289

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (Eds.). (2010). *Introduction to meta-analysis* (Reprinted). Wiley.

Buchs, C., & Butera, F. (2015). Cooperative learning and social skills development. In R. Gillies (Ed.), *Collaborative Learning: Developments in Research and Practice* (pp. 201-217). Nova Science.

Buchs, C., Gilles, I., Antonietti, J.-P., & Butera, F. (2016). Why students need to be prepared to cooperate: A cooperative nudge in statistics learning at university. *Educational Psychology*, *36*(5), 956–974. https://doi.org/10.1080/01443410.2015.1075963

Buchs, C., Gilles, I., Dutrévis, M., & Butera, F. (2011). Pressure to cooperate: Is positive reward interdependence really needed in cooperative learning? *British Journal of Educational Psychology, 81,* 135–146. https://doi.org/10.1348/000709910X504799

Bressoux, P. (2020). Using multilevel models is not just a matter of statistical adjustment. Illustrations in the educational field. *L'Année Psychologique/ Topic in Cognitive Psychology*, *120*(1), 5. https://doi.org/10.3917/anpsy1.201.0005

Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45*(5), 283–292. https://doi.org/10.3102/0013189X16656615

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

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology, 65*(3), 145-153. https://doi.org/10.1037/h0045186

Connolly, P., 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

Crone, T. S., & Portillo, M. C. (2013). Jigsaw variations and attitudes about learning and the self in cognitive psychology. *Teaching of Psychology*, *40*(3), 246–251. https://doi.org/10.1177/0098628313487451

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Darnon, C., Buchs, C., & Desbar, D. (2012). The jigsaw technique and self-efficacy of vocational training students: A practice report. *European Journal of Psychology of Education*, *27*(3), 439–449. https://doi.org/10.1007/s10212-011-0091-4

Deiglmayr, A., & Schalk, L. (2015). Weak versus strong knowledge interdependence: A comparison of two rationales for distributing information among learners in collaborative learning settings. *Learning and Instruction*, *40*, 69–78. https://doi.org/10.1016/j.learninstruc.2015.08.003

Desforges, D. M., Lord, C. G., Ramsey, S. L., Mason, J. A., Van Leeuwen, M. D., West, S. C., & Lepper, M. R. (1991). Effects of structured cooperative contact on changing negative attitudes toward stigmatized social groups. *Journal of Personality and Social Psychology*, *60*(4), 531–544. https://doi.org/10.1037/0022-3514.60.4.531

Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A.-M. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research*, *87*(2), 243–282. https://doi.org/10.3102/0034654316687036

Doymus, K. (2008). Teaching chemical bonding through jigsaw cooperative learning. *Research in Science & Technological Education*, *26*(1), 47–57. https://doi.org/10.1080/02635140701847470

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fernandez-Rio, J., Cecchini, J. A., Méndez-Gimenez, A., Mendez-Alonso, D., & Prieto, J. A. (2017). Self-regulation, cooperative learning, and academic self-efficacy: Interactions to

prevent school failure. *Frontiers in Psychology*, *8*, Article 22. https://doi.org/10.3389/fpsyg.2017.00022

Fleming, J. I., Wilson, S. E., Hart, S. A., Therrien, W. J., & Cook, B. G. (2021). Open accessibility in education research: Enhancing the credibility, equity, impact, and efficiency of research. *Educational Psychologist, 56*(2), 110–121. https://doi.org/10.1080/00461520.2021.1897593

Freeman, L., & Greenacre, L. (2011). An examination of socially destructive behaviors in group work. *Journal of Marketing Education*, *33*(1), 5–17. https://doi.org/10.1177/0273475310389150

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156–168. https://doi.org/10.1177%2F2515245919847202

Gage, N. A., Cook, B. G., & Reichow, B. (2017). Publication bias in special education meta-analyses. *Exceptional Children, 83*(4), 428–445. https://doi.org/10.1177/0014402917691016

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction* (6th ed.). Longman Publishing.

Ghaith, G., & El-Malak, M. A. (2004). Effect of Jigsaw II on literal and higher order EFL reading comprehension. *Educational Research and Evaluation*, *10*(2), 105–115. https://doi.org/10.1076/edre.10.2.105.27906

Ginsburg-Block, M. D., Rohrbeck, C. A., & Fantuzzo, J. W. (2006). A meta-analytic review of social, self-concept, and behavioral outcomes of peer-assisted learning. *Journal of Educational Psychology*, *98*(4), 732–749. https://doi.org/10.1037/0022-0663.98.4.732

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how: Mini meta-analysis. *Social and Personality Psychology Compass*, *10*(10), 535–549. https://doi.org/10.1111/spc3.12267

Gömleksi˙z, M. N. (2007). Effectiveness of cooperative learning (jigsaw II) method in teaching English as a foreign language to engineering students (Case of Firat University, Turkey). *European Journal of Engineering Education*, *32*(5), 613–625. https://doi.org/10.1080/03043790701433343

Götz, M., O'Boyle, E. H., Gonzalez-Mulé, E., Banks, G. C., & Bollmann, S. S. (2021). The "Goldilocks Zone": (Too) many confidence intervals in tests of mediation just exclude zero. *Psychological Bulletin, 147*(1), 95–114. https://doi.org/10.1037/bul0000315

Hänze, M., & Berger, R. (2007). Cooperative learning, motivational effects, and student characteristics: An experimental study comparing cooperative learning and direct instruction in 12th grade physics classes. *Learning and Instruction*, *17*(1), 29–41. https://doi.org/10.1016/j.learninstruc.2006.11.004

Hattie, J. (2010). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hattie J. (2017). *256 Influences Related To Achievement*. Visible Learning. https://visible-learning.org/hattie-ranking-influences-effect-sizes-learning-achievement/.

Huguet, P., Charbonnier, E., & Monteil, J.M. (1999). Productivity loss in performance groups: People who see themselves as average do not engage in social loafing. *Group Dynamics: Theory, Research, and Practice, 3*(2)*,* 118–131. https://doi.org/10.1037/1089-2699.3.2.118

Hwong, N., Caswell, A., Johnson, D. W., & Johnson, R. (1993). Effects of cooperative and individualistic learning on prospective elementary teachers' music achievement and attitudes. *Journal of Social Psychology, 133*(1), 53–64. https://doi.org/10.1080/00224545.1993.9712118

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

Johnson, D. W., Johnson, R. T., & Holubec, E. (1998). *Cooperation in the classroom* (6th ed.). Interaction Book Company.

Johnson, D. W., & Johnson, R. T. (2002). Cooperative learning methods: A meta-analysis. *Journal of Research in Education, 12*(1), 5-24.

Johnson, D. W., & Johnson, R. T. (2002). Cooperative learning and social interdependence theory. In Tindale, R.S., Heath, L., Edwards, J., Posavac, E.J., Bryant, F.B., Myers, J., Suarez-Balcazar, Y., and Handerson-King, E. (Eds), *Theory and research on small groups* (pp. 9–35). Springer.

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational Researcher*, *38*(5), 365–379. https://doi.org/10.3102/0013189X09339057

Karacop, A., & Doymus, K. (2013). Effects of Jigsaw cooperative learning and animation techniques on students' understanding of chemical bonding and their conceptions of the particulate nature of matter. *Journal of Science Education and Technology*, *22*(2), 186–203. https://doi.org/10.1007/s10956-012-9385-9

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, *65*(4), 681–706. https://doi.org/10.1037/0022-3514.65.4.681

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLOS ONE*, *9*(9), Article e105825. https://doi.org/10.1371/journal.pone.0105825

Kyndt, E., Raes, E., Lismont, B., Timmers, F., Cascallar, E., & Dochy, F. (2013). A meta-analysis of the effects of face-to-face cooperative learning. Do recent studies falsify or verify earlier

findings? *Educational Research Review*, *10*, 133–149. https://doi.org/10.1016/j.edurev.2013.02.002

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Law, Y.-K. (2011). The effects of cooperative learning on enhancing Hong Kong fifth graders' achievement goals, autonomous motivation and reading proficiency. *Journal of Research in Reading*, *34*(4), 402–425. https://doi.org/10.1111/j.1467-9817.2010.01445.x

Lazarowitz, R., Hertz-Lazarowitz, R., & Baird, J. H. (1994). Learning science in a cooperative setting: Academic achievement and affective outcomes. *Journal of Research in Science Teaching*, *31*(10), 1121–1131. https://doi.org/10.1002/tea.3660311006

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

Lucker, G. W., Rosenfield, D., Sikes, J., & Aronson, E. (1976). Performance in the interdependent classroom: A field study. *American Educational Research Journal*, *13*(2), 115–123. https://doi.org/10.3102/00028312013002115

Moreno, R. (2009). Constructing knowledge with an agent-based instructional program: A comparison of cooperative and individual meaning making. *Learning and Instruction*, *19*(5), 433–444. https://doi.org/10.1016/j.learninstruc.2009.02.018

Moskowitz, J. M., Malvin, J. H., Schaeffer, G. A., & Schaps, E. (1983). Evaluation of a cooperative learning strategy. *American Educational Research Journal*, *20*(4), 687–696. https://doi.org/10.3102/00028312020004687

Moskowitz, J. M., Malvin, J. H., Schaeffer, G. A., & Schaps, E. (1985). Evaluation of jigsaw, a cooperative learning technique. *Contemporary Educational Psychology*, *10*(2), 104–112. https://doi.org/10.1016/0361-476X(85)90011-6

Newmann, F. M., & Thompson, J. A. (1987). *Effects of Cooperative Learning on Achievement in Secondary Schools: A Summary of Research*. (ED288853). ERIC. https://eric.ed.gov/?id=ED288853

Nokes-Malach, T. J., Richey, J. E., & Gadgil, S. (2015). When is it better to learn together? Insights from research on collaborative learning. *Educational Psychology Review*, *27*(4), 645–656. https://doi.org/10.1007/s10648-015-9312-8

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Patall, E. A. (2021). Implications of the open science era for educational psychology research syntheses. *Educational Psychologist*, *56*(2), 142–160. https://doi.org/10.1080/00461520.2021.1897009

Plucker, J. A., & Makel, M. C. (2021). Replication is important for educational psychology: Recent developments and key issues. *Educational Psychologist*, *56*(2), 90–100. https://doi.org/10.1080/00461520.2021.1895796

Pianta, R. C., Belsky, J., Houts, R., & Morrison, F. (2007). Opportunities to learn in America's elementary classrooms. *Science*, *315*(5820), Article 1795. https://doi.org/10.1126/science.1139719

Puzio, K., & Colby, G. T. (2013). Cooperative learning and literacy: A meta-analytic review. *Journal of Research on Educational Effectiveness*, *6*(4), 339–360. https://doi.org/10.1080/19345747.2013.775683

Robert, A., Darnon, C., Consortium ProFAN, & Huguet, P. (2021). *La méthode Jigsaw au microscope : une revue critique de la littérature* [The Jigsaw method under scan: a critical review] [Manuscript in preparation].

Rohrbeck, C. A., Ginsburg-Block, M. D., Fantuzzo, J. W., & Miller, T. R. (2003). Peer-assisted learning interventions with elementary school students: A meta-analytic review. *Journal of Educational Psychology*, *95*(2), 240–257. https://doi.org/10.1037/0022-0663.95.2.240

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Roseth, C. J., Johnson, D. W., & Johnson, R. T. (2008). Promoting early adolescents' achievement and peer relationships: The effects of cooperative, competitive, and individualistic goal structures. *Psychological Bulletin*, *134*(2), 223–246. https://doi.org/10.1037/0033-2909.134.2.223

Roseth, C. J., Lee, Y., & Saltarelli, W. A. (2019). Reconsidering jigsaw social psychology: Longitudinal effects on social interdependence, sociocognitive conflict regulation, motivation, and achievement. *Journal of Educational Psychology*, *111*(1), 149–169. https://doi.org/10.1037/edu0000257

Saborit, J. A. P., Fernández-Río, J., Cecchini Estrada, J. A., Méndez-Giménez, A., & Alonso, D. M. (2016). Teachers' attitude and perception towards cooperative learning implementation: Influence of continuing training. *Teaching and Teacher Education*, *59*, 438–445. https://doi.org/10.1016/j.tate.2016.07.020

Şahin, A. (2011). Effects of Jigsaw III technique on achievement in written expression. *Asia Pacific Education Review*, *12*(3), 427–435. https://doi.org/10.1007/s12564-010-9135-8

Shaaban, K. (2006). An initial study of the effects of cooperative learning on reading comprehension, vocabulary acquisition, and motivation to read. *Reading Psychology*, *27*(5), 377–403. https://doi.org/10.1080/02702710600846613

Sharan, S. (1980). Cooperative learning in small groups: Recent methods and effects on achievement, attitudes, and ethnic relations. *Review of Educational Research*, *50*(2), 241–271. https://doi.org/10.3102/00346543050002241

Sharan, S. (1999). *Handbook of cooperative learning methods*. Westport: Greenwood publishing group.

Slavin, R. E. (2008). Cooperative learning, success for all, and evidence-based reform in education. *Éducation et didactique*, *2*(2), 149-157. https://doi.org/10.4000/educationdidactique.334

Slavin, R. E., Hurley, E. A., & Chamberlain, A. (2013). Cooperative learning and achievement: Theory and research. In W. Reynolds, G. Miller, & I. Weiner (Eds.), *Handbook of psychology* (2 ed., Vol. 7, pp. 199–212). Hoboken, NJ: Wiley.

Slavin, R., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis, 31*(4), 500–506. https://doi.org/10.3102/0162373709352369

Slavin, R. E. (2011). Instruction based on cooperative learning. *Handbook of Research on Learning and Instruction*, *4*. https://doi.org/10.4324/9780203839089.ch17

Smeding, A., Darnon, C., Souchal, C., Toczek-Capelle, M.-C., & Butera, F. (2013). Reducing the socio-economic status achievement gap at university by promoting mastery-oriented assessment. *PLoS ONE*, *8*(8), Article e71678. https://doi.org/10.1371/journal.pone.0071678

Souvignier, E., & Kronenberger, J. (2007). Cooperative learning in third graders' jigsaw groups for mathematics and science with and without questioning training. *British Journal of Educational Psychology*, *77*(4), 755–771. https://doi.org/10.1348/000709906X173297

Springer, L., Stanne, M. E., & Donovan, S. S. (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of Educational Research*, *69*(1), 21–51. https://doi.org/10.3102/00346543069001021

Świątkowski, W., & Dompnier, B. (2017). Replicability crisis in social psychology: Looking at the past to find new pathways for the future. *International Review of Social Psychology*, *30*(1), 111–124. https://doi.org/10.5334/irsp.66

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International journal of medical education*, *2*, 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Tarhan, L., Ayyıldız, Y., Ogunc, A., & Sesen, B. A. (2013). A jigsaw cooperative learning application in elementary science and technology lessons: Physical and chemical changes. *Research in Science & Technological Education*, *31*(2), 184–203. https://doi.org/10.1080/02635143.2013.811404

Topping, K., Buchs, C., Duran, D., & Van Keer, H. (2017). *Effective peer learning: From principles to practical implementation*. Taylor & Francis.

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, *7*(1), 55–79. https://doi.org/10.1002/jrsm.1164

Voyles, E. C., Bailey, S. F., & Durik, A. M. (2015). New pieces of the jigsaw classroom: Increasing accountability to reduce social loafing in student group projects. *The New School Psychology Bulletin*, *13*(1), 11–20. http://www.nspb.net/index.php/nspb/article/view/264

Walker, I., & Crogan, M. (1998). Academic performance, prejudice, and the jigsaw classroom: New pieces to the puzzle. *Journal of Community & Applied Social Psychology*, *8*(6), 381–393. https://doi.org/10.1002/(SICI)1099-1298(199811/12)8:6<381::AID-CASP457>3.0.CO;2-