N R
R N

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-123-0

Improving Reading Comprehension and Summarising Skills in Primary School: A Quasi-Experimental Study

Antonio Calvani [1] - **Antonio Marzano** [2]
Lorena Montesano [3] - **Marta Pellegrini** [4]
Amalia Lavinia Rizzo [5] - **Marianna Traversetti** [6]
Giuliano Vivanet [4]

[1] *Società per l'Apprendimento e l'Istruzione Informati da Evidenza – S.Ap.I.E. (Italy)*

[2] *Università degli Studi di Salerno - Department of Human, Philosophical and Educational Sciences (Italy)*

[3] *Università della Calabria - Department of Mathematics and Computer Science (Italy)*

[4] *Università degli Studi di Cagliari - Department of Pedagogy, Psychology, Philosophy (Italy)*

[5] *Università degli Studi Roma Tre - Department of Education (Italy)*

[6] *Sapienza Università di Roma - Department of Developmental and Social Psychology (Italy)*

DOI: https://doi.org/10.7358/ecps-2023-028-calv

antonio@calvani.it
amarzano@unisa.it
lorena.montesano@unical.it
marta.pellegrini@unica.it
amalia.rizzo@uniroma3.it
marianna.traversetti@uniroma1.it
giuliano.vivanet@unica.it

MIGLIORARE LA COMPRENSIONE DEL TESTO E LE CAPACITÀ DI SINTESI NELLA SCUOLA PRIMARIA: UNO STUDIO QUASI-SPERIMENTALE

ECPS Journal – 28/2023 - https://www.ledonline.it/ECPS-Journal/
Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-123-0

Abstract

The paper reports the results of the evaluation of a programme aimed at improving reading comprehension and summarising skills of fourth graders through a quasi-experimental study conducted with 671 students (421 in the experimental group and 250 in the control group) in Italian schools. Students assigned to the experimental group received three months of the intervention while students in the control group continued with regular teacher practice. Results showed a statistically significant difference between the two conditions on reading comprehension and summarising skills. Students included in the experimental group outperformed students in the control group in both the measures used (d_{ppc2} = 0.32 in the Summarising Test and d_{ppc2} = 0.54 in the Summary Qualitative Assessment). No differences were found between students with different proficiency vocabulary levels in the experimental group. The implications for research practice and limitations of the study are discussed.

Keywords: Primary school; Quasi-experimental design; Reading comprehension; Reciprocal teaching; Summarising skills.

1. Introduction

Reading skills are basic elements that every individual needs for an active citizenship in today's society. A low level of these essential skills can severely restrict the possibility of acquiring new competences in the lifelong learning perspective and to keep up with the fast-paced changes of the job market. The OECD PISA 2018 [1] findings showed that a high percentage of students did not reach a minimum level of reading skills in Europe, with a mean of 21.7% of 15-year-olds underachieving in reading.

Underachievement in reading is not just a European issue but a concern shared with several countries outside Europe. In the United States, for instance, 34% of fourth graders score below the expected level of reading for their age at the National Assessment of Educational Progress [2] (NAEP, 2019). These findings are alarming considering that children reading below grade level in the third grade are four times as likely as other students to drop out before high school graduation (National Center for Education Statistics, 2019).

Reading comprehension is a debated topic that has been investigated by many researchers (van Dijk & Kintsch, 1978; Kintsch, 1998). In recent

[1] https://www.oecd.org/pisa/publications/pisa-2018-results.htm
[2] https://www.nationsreportcard.gov/highlights/reading/2019/

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-123-0

years several research reviews were carried out with the purpose of evaluating the effectiveness of different instructional strategies and programmes for the development of reading comprehension skills (NICHD, 2000; de Boer, Donker, & van der Werf, 2014). The results mainly showed that programmes incorporating metacognitive processes and cooperative learning were the most effec tive strategies to improve reading comprehension in primary school. Furthermore, tutoring by an adult was found to be especially effective for students with low level of academic achievement (Swanson *et al.*, 2017).

Among the programmes focused on reading comprehension, Reciprocal Teaching (RT) – advanced by Palincsar and Brown (1984) – is one of the programmes that received particular attention in educational research. It can be considered a scaffolded and interactive discussion technique, based on cognitive strategies that a good reader should use to understand the information of a text and summarise its content (Oczkus, 2018). In detail it consists of the following four strategies:

1. *Predicting*, which requires students to make assumptions about the content that will be presented in the text.
2. *Questioning*, which requires students to generate questions at different levels such as recalling details of the text and making inferences.
3. *Clarifying*, which requires students to highlight words that may be difficult to understand such as unknown terms, new concepts, idioms.
4. *Summarising*, which requires students to identify, paraphrase and summarise relevant information of the text.

Firstly, the teacher introduces and carries out these four strategies through modelling and thinking aloud techniques, explaining his/her way of thinking until this meta-cognitive process is internalised by the students. After teacher modelling students take turns to apply the strategies and to work in small group to comprehend the text. Through RT guided practice students gradually develop understanding and autonomy in applying the four strategies and more in general in the reading process (Palincsar, 2013).

Three meta-analyses (i.e. systematic reviews that combine statistically results of experimental studies) studied the impact of RT on reading comprehension skills in primary school. The oldest review was conducted by Rosenshine and Meister (1994) showed a high level of impact of RT, although differentiated on the basis of the type of measure used. The authors distinguished between standardised measures of academic achievement and measures created by the researchers for the specific study. Cheung and Slavin (2016) found that on average researcher-made measures affect the magnitude of the intervention effect.

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-123-0

Galloway (2003) included 22 studies, where 18 were RCTs (Randomised Controlled Trials) or quasi-experimental studies and 4 used a pre- to post-test design without a control group. The overall effect size was +0.74 with a higher mean effect size when measures made by the researchers were used (ES = 0.92) compared to norm-referenced tests (ES = 0.56). Galloway (2003) found an increased level of effectiveness when RT was delivered through modelling and practice than when conducted through direct instruction. No differences in effects were found between grade levels – in late elementary (grades 4-6) RT was equally effective than in early secondary (grades 7-9) – nor between student achievement levels. A small effect size was found on far-transfer tests of reading comprehension and the authors concluded that more research is needed in order to study the transfer impact of RT. Lee and Tsai (2017) carried out a meta-analysis on interventions aiming at enhancing the reading skills of students with specific poor comprehension, namely those who showed difficulties in the specific domain of comprehension of the text at the transition between the third and fourth grade (Chall & Jacobs, 2003). Among the interventions RT demonstrated the largest impact. Students who used RT achieved significantly better learning outcomes than students in the business-as-usual group (ES = 0.86, with 95% CI [.56, 1.15]).

Although previous reviews detected a large impact for RT, significant limitations could have affected the results. Among them three methodological features can be pointed out (de Boer *et al.*, 2014; Cheung & Slavin, 2016): (i) the inclusion of results derived from studies with different research designs, (ii) the small sample sizes, and (iii) the adoption of researcher-made measures. Galloway included experimental studies and pre- to post-test design without a control group. As affirmed by Lipsey and Wilson (2001), mixing the effects of studies with different research designs (e.g. correlational and experimental studies) would not produce comparable results since the two designs do not share similar objectives. In this case Galloway included experimental studies with pre- to post-test design without a control group that could have led to overestimating the impact of RT. Lee and Tsai (2017) included the six studies on RT with a small sample size ($18 < n < 147$) that may have affected the magnitude of the impact in terms of effect size. Cheung and Slavin (2016) found a negative relationship between effect size and sample size, with larger effects for studies with a small sample ($n < 250$). These studies also used measures developed by the researchers and included their results with the ones gathered from independent measures. As mentioned above, a substantial difference in effect sizes were found in previous studies between researcher-made measures and independent measures (de Boer *et al.*, 2014; Pellegrini *et*

al., 2019; Wolf, 2021). Researcher-made measures had almost three times the effect size of independent measure and this value remained significant when controlling for other factors of potential impact (Pellegrini *et al.*, 2019). Therefore, the methodological quality of the studies could have led to overestimating the mean effect size. For this reason and because of the small number of recent studies published on RT, new experimental studies are needed in order to evaluate the level of effectiveness of RT and its components.

2. The intervention

With the aim of supporting teaching and learning for improving reading and summarising skills in Italian primary schools, the RC-RT programme (Reading Comprehension-Reciprocal Teaching) was developed (Calvani & Chiappetta Cajola, 2019). The programme is based on the theoretical framework and strategies of the RT but some modifications have been made on the operational level to reduce the risks of dispersion, loss of time, and distraction related to cooperative activities (Johnson & Johnson, 1999) as well as on the theoretical level, to establish a closer connection with the instances highlighted by the most recent international research (PIRLS, PISA). On the operational level in the RC-RT programme:

1. The work through the programme has the objective of making students to be able to write a good synthesis in a limited number of words. Summarising is therefore the key strategy.
2. The cooperative work has been revised. The first three RT strategies are applied individually by the students, while summarising is done in pairs. Each student works in pairs after an individual reflection on the text. In case of disagreement between the students, the pair tries to explain their opinion.
3. The clarifying strategy is kept under control by using simple texts, avoiding the influence of unknown terms or contexts on student understanding.

On the theoretical level, in relation to level 2 of PIRLS (Make straightforward inferences) and level 3 of PISA (Developing an interpretation), the «activation of inferential processes» strategy has been added in the second part of the programme. Teachers ask students to «look beyond the text» with «semantic» or «elaborative» inferences (e.g., «Why did the character do this?», «Why did the author write this?») (Tressoldi & Zamperlin, 2007).

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-123-0

The programme RC-RT consists of 25 hours of activities in a primary school fourth grade class, with biweekly one and a half hour lessons, for a total of about three months of work.

It is divided into two parts of progressively increasing length and difficulty. The first part includes ten texts and is dedicated to literal text comprehension, and the second one includes twenty-four texts and is dedicated to inferential text comprehension.

The programme begins with a working example by the teacher who, after reading the text, uses thinking aloud strategies to answer questions related to the content of the text. After teacher modelling, the students are requested to think aloud and answer the questions as well as the teacher had done. Later, students are invited to analyse the texts themselves according to the scheme indicated in the work notebook (*Fig. 1*) [3].

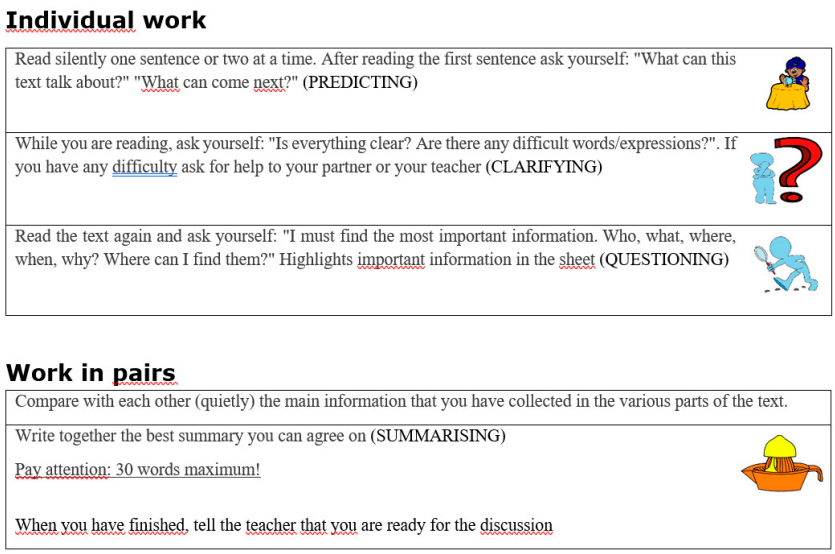

Figure 1. – Student book: sequence of activities related to individual work and work in pairs in the first part of the programme.

Figure 1 shows an example of the activities that students, after teacher modelling, carry out autonomously on each text. The activity sequence

[3] For a detailed description of the intervention see https://www.sapie.it/wp/wp-content/uploads/2020/01/RC-RT-presentazione-completa.pdf. See also Rizzo, Traversetti, & Pellegrini, 2023.

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-123-0

implies an initial part of individual reading and a second collaborative part in pairs aimed at writing a summary.

Considering the full inclusion model that characterizes the Italian school system and the importance of promoting the participation of pupils with special educational needs, a specific version of the programme was developed for pupils with moderate intellectual disabilities (Rizzo & Traversetti, 2021). This material was drawn up by task facilitation techniques and it was made easier in relation to linguistic (lexical and syntactic) and graphic aspects. For example, the information was pertinent to the pupils' daily experience and the text was segmented into paragraphs corresponding to the individual narrative sequences.

3. Method

3.1. *Objective and design*

To evaluate the effectiveness of the RC-RT programme a quasi-experimental design with a pre- to post-test control group was adopted. Quasi-experimental design – also called non-equivalent control group study – compares an experimental group with a control group to test hypotheses about the effects of a treatment but lacks the process of random assignment that occurs in true experiments (Cook *et al.*, 2002).

Students in the experimental group received the RC-RT programme while students in the control group continued with the usual practice offered by their teachers. Currently in Italian schools, reading is generally taught using textbooks characterised by exercises of a predominantly linguistic or stylistic nature. When questions with cognitive value are occasionally included in the textbooks, they do not aim to strengthen specific metacognitive strategies in a systematic way.

The study was conducted according to the ethics guidelines of the Declaration of Helsinki (2013) and approved by the Ethics Committee of the participating schools. Students and their parents were asked for an informed consent about the evaluation, and it was obtained for each participant.

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-123-0

3.2. *Participants*

The participants were 995 students in 46 fourth grade classes across 33 primary schools throughout Italy. These schools were located in urban and suburban areas of 6 different Regions of Northern (n = 1), Southern (n = 4) and Central (n = 1) Italy.

At the beginning of the study 24 classes and 571 students were assigned to the experimental group and 22 classes and 424 students to the control group. Although the assignment could not be random, students at the baseline were similar for most of the characteristics considered. Of these students 324 had no pre- or post-test available scores (n = 174 in the control classes; n = 150 in the intervention classes) and were deleted from the study. The overall student attrition (i.e. percentage of students who left the study) was 32.6%, and the student differential attrition (i.e. difference in percentage between students who left the study in the experimental and control group) was 14.8%. After the attrition that occurred at the individual level – no whole classes left the study – the analytical sample consisted of 250 students in the control group, and 421 students in the experimental group (*Tab. 1*).

Students with physical or specific learning disabilities or low socio-economic status (SES) were included in the sample [4]. Students with medium and severe intellectual disabilities were provided with different materials to include them in the class activities and their scores were not included in the sample (n = 6 in the experimental group; n = 5 in the control group).

Table 1. – Characteristics of the experimental and control groups.

Characteristic	Experimental (n = 421)	Control (n = 250)
Age (Mean)	8.89	8.80
Female	49%	53%
Low SES	2.85%	2.80%
Specific Learning Disabilities	1.90%	2.80%
Physical disabilities	0.24%	–

[4] Students' data were reported by schools.

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-123-0

3.3. *Measures*

Two measures were administered at the pre and post-test, also in order to gauge the effects on reading comprehension and summarising skills, together with a lexical test administered only at the pre-test.

Test of Verbal Meaning (TVM). This test aims to meet the need for controlling the possible differences between experimental and control groups at the baseline, which could affect the results. As a matter of fact, amongst the possible confounding factors, the vocabulary evaluation appeared the most significant. Many studies in the literature have addressed the impact of vocabulary on text comprehension and have analysed the relation between these two abilities even if the extent of this relationship remains controversial.

In order to assess the initial lexical skills of the students, a revised version of the subtest Test of Verbal Meaning of Primary Mental Abilities (PMA, Thurstone & Thurstone, 1965) was administered at the pre-test. The test was modified specifically for this research, including items with a higher grade of difficulty compared to the standard version (Thurstone & Thurstone, 1965). It consists of 30 items, of which 15 were modified from the previous version. The time available to complete the test is 7 minutes. Children must identify the correct synonym of a word among four possible alternatives (Montesano, 2020). The score is calculated as follows: one point for each correct answer and zero points for each incorrect or omitted answer. In the presence of a double answer, half a point is attributed provided that one of them is correct. The total score is obtained by the sum of each point. The psychometric characteristics of this measure were studied through a sample of 1177 fourth graders and the reliability coefficient was found to be Cronbach's $\alpha = .85$.

Summarising Test (ST). This measure developed by the research team aims to evaluate students' reading comprehension and summarising skills (Menichetti, 2018). The students are provided with four texts and three closed-ended questions on titles, summaries, and key words with six answer options for each text. Students have to choose the three most suitable options for all the questions. For each text, the maximum score is nine points, for a maximum of 36 points for the whole measure. The measure was split into two equivalent versions (STa and STb) through an item analysis administered at the pre-test and the post-test. The full version of the measure has a good reliability coefficient (Cronbach's $\alpha = .85$), as do the two split versions (STa Cronbach's $\alpha = .83$; STb Cronbach's $\alpha = .80$).

Summary Qualitative Assessment (SQA). This qualitative measure was developed by the research team with the aim of evaluating students' skills

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-123-0

in writing a summary of a text (Pečjak & Pirc, 2018; Menichetti & Bartolini, 2019). It requires children to write a short summary based on a read text. The text is presented to students divided into three parts and for each of them the students are asked to produce a short summary (max. 20 words). To score the summaries written by the students, the research team considered two criteria: length, that is the quantity of words used to make the three summaries (max. 6 points), and content, that is the congruence of the answer with a list of words that define the semantic field of important ideas, previously delineated by experts in the reading field (max. 14 points). Test validation was carried out through a repeated process based on random samples until two independent evaluators achieved a concordance level higher than 90%. The test was split into two equivalent versions (SQAa and SQAb) administered at the pre-test and the post-test according to the scheme shown in *Figure 2*.

Other measures. In addition to these tests a sub-group analysis was conducted on 106 subjects randomly selected from the main sample: 56 from the experimental group and 50 from the control group. This analysis consisted in the administration of specific tests, standardised for the Italian language (De Beni *et al.*, 2003; Tressoldi & Zamperlin, 2007), aimed at evaluating the impact of the programme on the lexical and semantic inference. Lexical inference means finding the meaning of an unknown word from context, semantics means extracting information implicit in the text (Tressoldi & Zamperlin, 2007).

A final questionnaire for customer satisfaction addressing both teachers and students was administrated.

3.4. *Procedure*

Pre-test measures were administered in January 2019, classroom intervention lasted from February to April 2019, post-test measures were administered between late April and May 2019.

The implementation of the programme consisted of the following phases.

- *Phase 1 – Pre-test administration*. Trained research assistants administered pre-test measures (TVM, ST, and SQA) to students in the experimental and control classes. As mentioned in the Measures section two equivalent versions of ST and SQA were developed (version A and B). According to the scheme shown in *Figure 2* half of the participants in each condition were presented with set A at the pre-test and with set B at the post-test, while the opposite was true for the other half.

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-123-0

- *Phase 2 –Teacher training*. Experimental teachers were provided with professional development by the research group. Teachers received a half-day (4 hours) training that included the presentation of the materials for students and teachers and a workshop that used video modelling to show the implementation of the programme and proposed a reflection with teachers on the most important strategies and phases of the intervention: modelling, thinking aloud, working in pairs, class feedback. Teachers were also provided with five short videos (about 6 minutes each) to be watched prior to or during the intervention. Teachers received a guide to support their work during the implementation of the programme.
- *Phase 3 – Intervention implementation*. The intervention was delivered over the course of seventeen learning units (twice a week) of one and a half hours each, for a total of twenty-five hours across three months. During the intervention in the experimental group, the control group continued with the regular practice that the teachers used to follow. Since the intervention was delivered during regular school days, the control group benefited from the same amount of time working on reading comprehension than the experimental group. To ensure the fidelity of implementation the procedure was analytically described in a guide which was presented to teachers at training meetings. The first meetings in the classes were led by researchers from the universities that carried out the study.
- *Phase 4 – Post-test administration*. After the end of the intervention post-test measures were administered to the experimental and control groups in the same condition of pre-test. Students presented with set A at the pre-test were presented with set B at the post-test, while the opposite was true for the other half (*Fig. 2*).

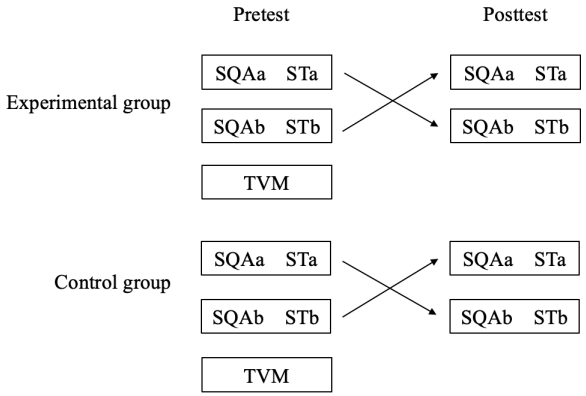

Figure 2. – Test administration scheme.

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-123-0

3.5. *Data analysis*

We first analysed the baseline equivalence of the experimental and the control conditions (*Tab. 2*).

Table 2. – Results obtained at pretest by the experimental and control group.

	Experimental group		Control group		t	*p*	Effect-size d_{Cohen}
	N	M (SD)	N	M (SD)			
Pre-test							
ST measure	421	22.83 (4.78)	250	22.72 (4.57)	-.286	.775	.02
SQA measure	421	10.58 (3.52)	250	10.60 (3.66)	.080	.936	.01
Vocabulary	406	23.15 (4.59)	238	22.71 (5.10)	-1.128	.260	.09

Note: M = Mean; SD = Standard Deviation; t = Student's t-test.

For all the measures used at the pre-test we found no significant differences between students' scores in the two groups: t(669) = -.286, p = .775, Cohen's *d* = .02 for ST; t(669) = .080, p = .936, Cohen's *d* = .01 for SQA; t(642) = -1.128, p = .260, Cohen's *d* .09 for vocabulary. Standardised mean differences were less than 0.25 SD for each measure, indicating baseline equivalence between the two groups, as required by the What Works Clearinghouse (2020).

ANCOVA with cluster-adjusted standard errors to take into account the cluster assignment was used to determine whether differences detected in post-tests between the experimental group and the control group were statistically significant. The model included the centered pre-test assignment variable as a covariate at the student level and cluster-adjusted standard error. Since we administered three outcome measures, the Bonferroni correction for multiple comparisons was used. The packages sandwich (Hothron *et al.*, 2015) and lmtest (Zeileis *et al.*, 2020) were used to estimate the models in the R statistical software (R Core Team, 2020).

One-way ANCOVA was used to test for any statistically significant difference in results in the experimental group between students with different initial levels of lexical skills. The students were divided into three groups based on the mean score achieved in the TVM measure at the pre-test: low proficiency (M < 20; n = 85), medium proficiency (21 < M < 26; n = 224), and high proficiency (M > 26; n = 98).

The assumptions for the ANCOVA use were checked and the only one violated was the normality of the distribution. Levy (1980) stated that

for large samples with unequal group sizes – when homogeneity of regression slopes is not violated as in this case – ANCOVA appears to be robust and could be used even if normality is violated. We tested the homogeneity of variances of the dependent variables because of the unequal size of the sample in the two groups. Levene's test for equality of variances indicated no significant differences for the ST measure [F(1, 669) = .153, p = .696], the SQA measure [F(1, 669) = .032, p = .858].

Effect size measure for pre-test to post-test control group designs was used in order to evaluate the magnitude of the impact of the intervention on dependent variables. The Morris (2008) procedure was used with the d_{ppc2} effect size:

$$d_{ppc2} = c_p \left[\frac{\left(M_{EG,post} - M_{EG,pre}\right) - \left(M_{CG,post} - M_{CG,pre}\right)}{SD_{pre}} \right]$$

M is the mean of the experimental and control group and at pre-test and post-test; SD_{pre} is the pooled standard deviation of the pre-test, and C_p is the correction factor for small sample sizes.

4. Results

Table 3 shows the unadjusted and adjusted post-test scores for baseline differences in the experimental and control group.

Table 3. – Results obtained at post-test by the experimental and control group (post-test unadjusted means and standard deviations and post-test ANCOVA-adjusted means and standard deviations).

	Experimental group			Control group		
	N	M (SD) Unadjusted	M (SD) Adjusted	N	M (SD) Unadjusted	M (SD) Adjusted
Post-test						
ST measure	421	25.17 (5.05)	25.15 (0.22)	250	23.56 (4.79)	25.59 (0.28)
SQA measure	421	13.12 (2.86)	13.12 (0.14)	250	11.22 (3.18)	11.22 (0.19)

Note: M = Mean; SD = Standard Deviation.

The results of the impact analysis showed a statistically significant difference between the two conditions on outcome measures (*Tab. 4*). Effect sizes showed that the experimental group outperformed the control group

with a larger effect size for SQA than for ST (d_{ppc2} = 0.32 for ST; d_{ppc2} = 0.54 for SQA).

Table 4. – Parameter estimates for impacts on ST and SQA scores.

	ST score	SQA score
Intercept	23.63 (0.31)	11.30 (0.37)
Pre-test	0.65 (0.04)	0.17 (0.04)
Treatment effect	1.50 (0.48)	1.77 (0.48)
Effect Size (d_{ppc2})	0.32	0.54

Note: Significant parameter estimates p < .05 are marked in bold type.

A one-way ANCOVA was used to measure if a difference in intervention effectiveness existed between experimental students with different levels of vocabulary proficiency. As mentioned above, three categories (high, medium, low) were formed based on scores obtained by the students at the vocabulary measure. Results of the analysis revealed no statistically significant difference in reading skills by level of proficiency in vocabulary, whilst controlling for pre-test scores: F(3, 416) = 1.41, p = .237 for ST measure; F(3, 415) = .964, p = .410.

As regards the sub-group analysis on semantic and lexical inference, the results obtained are significant for semantic inference (d_{ppc2} = 0.48), while they are not on lexical inference.

Only few students with intellectual disabilities (n = 11) attended the classes involved in the study and completed a simplified version of the intervention. Due to this small number, it was not possible to make any analysis on this sub-group of students.

The final questionnaire was taken by 29 teachers who indicated that they were satisfied with the programme implementation. On programme reproducibility, the answer is positive, but they suggested some modifications. The most common changes that were requested were to have more time available and to distribute the programme throughout the year. The final questionnaire on the students' interest in the programme was taken by the experimental group. On the rating scale used, 29% responded very high, 40% high, 35% enough, 3% little. In the open-ended question, 20% of pupils said they found the task difficult or they experienced boredom; 7% found problems related to cooperation.

5. Discussion

This study investigated the effects of the RC-RT programme on reading comprehension and summarising skills of fourth-grade students. After three months of intervention implementation results showed larger outcomes for students in the experimental condition compared to students that continued with regular teacher practice. The magnitude of the impact was equal to 0.32 for students' reading comprehension and summarising skills (ST measure) and 0.54 for qualitative analysis of skills in writing a summary of a text (SQA measure). The starting lexical level does not affect the final result.

The interpretation of the effect sizes and its practical significance is one of the most discussed issues in educational research (Cohen, 1988; Shaver, 1993; Kirk, 1996). In order to provide researchers and practitioners with elements for their interpretation, it is useful to compare them with appropriate criterion values (Lipsey *et al.*, 2012). Regarding this point one of the most cited proposals in the literature is the one advanced by Cohen (1988) for the area of behavioral science, whose benchmark values are widely used today: 0.20 small; 0.50 medium; and 0.80 large. However Cohen himself warned about their use, stressing the fact that these parameters were «recommended for use only when no better basis for estimating the ES index is available» (Cohen, 1988, p. 25). Recently Kraft (2020) proposed a more structured schema with new empirical benchmarks for interpreting effects from causal studies of preK-12 education interventions with standardised achievement outcomes: < 0.05 small, from 0.05 to < 0.20 medium, and ≥ 0.20 large. Based on Kraft's model the impact of the intervention on reading comprehension and summarising skills can be considered large.

There are some major limitations that should be considered in interpreting these results and for future research. Firstly, regarding the research design, it is worth noting that it was not possible to adopt a random allocation due to the constraints determined by the organization of instructional activities in the participating schools that did not allow the randomization of classes and children. Secondly, outcome measures were not standardised reading tests, but measures made by the RC-RT programme developers, and this may affect the magnitude of the intervention effect; the researcher-made measures may be aligned in content and form with the treatment implemented in the experimental group, favoring positive results in this group compared to the control one (Cheung & Slavin, 2016; Pellegrini *et al.*, 2019; Wolf, 2021). For this reason, future research on the intervention should include an independent measurement of reading achievement.

Thirdly, the intervention for fourth graders mainly focused on summarising skills, which the programme developers considered one of the most important dimensions to improve reading comprehension. In a broader sense, recent models of reading comprehension (e.g. the one adopted by the OECD PISA) suggest taking into account other strategies, such as the activation of inferential processes. The intervention also included sessions dedicated to these inferential reading skills. Outcome measures to evaluate them were not possible to administer to the entire sample, but a standardised test (Tressoldi & Zamperlin, 2007) was administered to a subgroup of 106 students (50 control group and 56 treatment group). The results are promising (d_{ppc2} = 0.48), but they still deserve further confirmation.

Finally, post-testing happened immediately after the intervention's conclusion, providing the opportunity to measure the effectiveness of the programme in developing students' reading skills, but not the longevity of its impact. Several studies in reading, especially involving struggling readers, have found strong evidence of effectiveness in the immediate post-test, but no long-term benefits after one year or more from the intervention's conclusion (Neitzel *et al.*, 2022).

Although the limitations of this study should be considered carefully, the results indicate that the intervention improved fourth graders' reading comprehension and summarising skills. Considering the large sample size of the study, the sustainable nature of the intervention (which took place in an authentic situation, with whole classes with their teacher and only a few hours of teachers training), this study offers adequate evidence and reasonable grounds for the decision to proceed with a large-scale implementation of the RC-RT programme.

References

Calvani, A., & Chiappetta Cajola, L. (Eds.). (2019). *Strategie efficaci per la comprensione del testo. Il Reciprocal Teaching*. Firenze: S.Ap.I.E. Scientifica.

Chall, J. S., & Jacobs, V. A. (2003). The classic study on poor children's fourth-grade slump. *American Educator*, *27*(1), 14-15.

Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283-292.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*, Vol. 1195. Boston, MA: Houghton Mifflin.

De Beni, R., Cornoldi, C., Carretti, B., & Meneghetti, C. (2003). *Nuova guida alla comprensione del testo*, Vol. 1: *Introduzione teorica generale al programma: le prove criteriali livello A e B*. Trento: Erickson.

de Boer, H., Donker, A., & van der Werf, M. (2014). Effects of the attributes of educational interventions on students' academic performance: A meta-analysis. *Review of Educational Research*, *84*(4), 509-545.

Galloway, A. M. (2003). *Improving reading comprehension through metacognitive strategy instruction: Evaluating the evidence for the effectiveness of the reciprocal teaching procedure*. Doctoral dissertation, University of Nebraska-Lincoln, NE.

Hothorn, T., & Zeileis, A. (2015). Partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, *16*(1), 3905-3909.

Johnson, D. W., & Johnson, R. T. (1999). Making cooperative learning work. *Theory into Practice*, *38*(2), 67-73.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*(5), 746-759.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241-253.

Lee, S. H., & Tsai, S. F. (2017). Experimental intervention research on students with specific poor comprehension: A systematic review of treatment outcomes. *Reading and Writing*, *30*(4), 917-943.

Levy, K. J. (1980). A Monte Carlo study of analysis of covariance under violations of the assumptions of normality and equal regression slopes. *Educational and Psychological Measurement*, *40*(4), 835-840.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. Washington, DC: National Center for Special Education Research.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. London: Sage.

Menichetti, L. (2018). Valutare la capacità di riassumere. Il Summarising Test, uno strumento per la scuola primaria. *Journal of Educational, Cultural and Psychological Studies (ECPS)*, *18*, 369-396.

Menichetti, L., & Bertolini, C. (2019). Prova qualitativa per la valutazione della capacità di riassunto. Il Summary Qualitative Assessment (SQA). In A. Calvani & L. Chiappetta Cajola (a cura di), *Strategie efficaci per la comprensione del testo. Il Reciprocal Teaching* (pp. 431-462). Firenze: S.Ap.I.E. Scientifca.

Montesano, L. (2020). *Vocabolario e comprensione del testo. Uno strumento per la valutazione del lessico nella scuola Primaria*. Firenze: S.Ap.I.E. Scientifca.

Online ISSN 2037-7924 - Print ISSN 2037-7932 - ISBN 978-88-5513-123-0

NAEP – National Assessment of Educational Progress (2019). *NAEP report card: 2019 NAEP reading assessment*. https://www.nationsreportcard.gov/highlights/reading/2019/

National Center for Educational Statistics (2019). *National assessment of educational progress*. Washington, DC: National Center for Educational Statistics.

Neitzel, A. J., Lake, C., Pellegrini, M., & Slavin, R. E. (2022). A synthesis of quantitative research on programmes for struggling readers in elementary schools. *Reading Research Quarterly*, *57*(1), 149-179.

NICHD – National Institute of Child Health and Human Development (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction (NIH Publication No. 00-4769)*. Washington, DC: U.S. Government Printing Office.

Oczkus, L. D. (2018). *Reciprocal teaching at work: Powerful strategies and lessons for improving reading comprehension*. ASCD.

Palincsar, A. S. (2013). Reciprocal teaching. In J. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 369-371). London - New York: Routledge.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension fostering and comprehension-monitoring activities. *Cognition and Instruction*, *1*(2), 117-175.

Pečjak, S., & Pirc, T. (2018). Developing summarizing skills in 4th grade students: Intervention programme effects. *International Electronic Journal of Elementary Education*, *19*(5), 571-581. doi: 10.26822/iejee.2018541306.

Pellegrini, M., Inns, A., Lake, C., & Slavin, R. E. (2019). Effects of researcher-made vs. independent measures on outcomes of experiments in education. Paper presented at the *Annual Meeting of the Society for Research on Educational Effectiveness*, Washington, DC.

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Rizzo, A. L., & Traversetti, M. (2021). *Il programma RC-RT per la comprensione della lettura. Percorso didattico evidence based per la scuola primaria. Guida per gli insegnanti*. Firenze: S.Ap.I.E.

Rizzo, A. L., Traversetti, M., & Pellegrini, M. (2023). *Potenziare la comprensione del testo*. Roma: Carocci.

Rosenshine, B., & Meister, C. (1994). Reciprocal teaching: A review of the research. *Review of Educational Research*, *64*(4), 479-530.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education*, *61*(4), 293-316.

Swanson, E., Stevens, E. A., Scammacca, N. K., Capin, P., Stewart, A. A., & Austin, C. R. (2017). The impact of tier 1 reading instruction on reading

outcomes for students in grades 4-12: A meta-analysis. *Reading and Writing*, *30*(8), 1639-1665.

Thurstone, T. G., & Thurstone, L. L. (1965). *P.M.A. Primary Mental Abilities*. Science Research Associates.

Tressoldi, P. E., & Zamperlin, C. (2007). La valutazione della comprensione del testo. Proposta di una batteria di approfondimento. *Psicologia clinica dello sviluppo*, *11*(2), 271-290.

van Dijk, T. A., & Kintsch, W. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*, 363-394.

Wolf, R. (2021). *Average differences in effect sizes by outcome measure type (WWC 2021)*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse.

Zeileis, A., Köll, S., & Graham, N. (2020). Various versatile variances: An object-oriented implementation of clustered covariances in R. *Journal of Statistical Software*, *95*, 1-36.

Riassunto

L'articolo presenta i risultati della valutazione di un programma volto a migliorare la comprensione della lettura e le capacità di sintesi di allievi della quarta classe della scuola primaria italiana attraverso uno studio quasi sperimentale condotto su 671 partecipanti (421 nel gruppo sperimentale e 250 nel gruppo di controllo). Gli allievi assegnati al gruppo sperimentale hanno ricevuto tre mesi di intervento, mentre gli allievi del gruppo di controllo hanno continuato a seguire le lezioni tradizionali. I risultati hanno mostrato una differenza statisticamente significativa tra le due condizioni per quanto riguarda la comprensione della lettura e la capacità di sintesi. Gli studenti del gruppo sperimentale hanno superato quelli del gruppo di controllo in entrambe le misure utilizzate (d_{ppc2} = 0,32 nel Summarising Test e d_{ppc2} = 0,54 nel Summary Qualitative Assessment). Non sono state riscontrate differenze tra allievi con diversi livelli di competenza lessicale nel gruppo sperimentale. Le implicazioni per la pratica della ricerca e i limiti dello studio sono discussi.

Parole chiave: Capacità di sintesi; Comprensione del testo; Disegno quasi-sperimentale; Scuola primaria.

How to cite this paper: Calvani, A., Marzano, A., Montesano, L., Pellegrini, M., Rizzo, A. L., Traversetti, M., & Vivanet, G. (2023). Improving reading comprehension and summarising skills in primary school: A quasi-experimental study [Migliorare la comprensione del testo e le capacità di sintesi nella scuola primaria: uno studio quasi-sperimentale]. *Journal of Educational, Cultural and Psychological Studies (ECPS)*, *28*, 81-100. https://doi.org/10.7358/ecps-2023-028-calv