Abstract Title Page
Not included in page count.

Title: The effects of within class grouping on reading achievement: A meta-analytic synthesis

Author(s): Kelly Puzio; Glenn Colby

For questions, please contact Kelly Puzio @ kelly.puzio@vanderbilt.edu

Abstract Body

Limit 5 pages single spaced.

Background/context:

Description of prior research, its intellectual context and its policy context.

Although some literacy researchers consider grouping students for reading instruction to be a proven educational practice, the support for this belief is lacking from a research synthesis perspective. With this idea in mind, Slavin comments in the middle of his review on the effects of grouping: "there is not enough research on within-class ability grouping in reading to permit any conclusions" (Slavin, 1987, p. 320). Because of this, the question of whether *and* how to group students is often cast and answered ideologically rather than empirically. This review attempts to see if the empirical research available can answers either or both of these questions.

Seven reviews have been conducted on the effects of grouping on student achievement in the last twenty years (Elbaum, 1999; Johnson & Johnson, 1989; Kulik, 1992; Kulik & Kulik, 1987; Lou, Abrami, Spence, Poulsen, Chambers, & d'Apollonia, 1996; Slavin, 1987, 1990). In these previous reviews, two distinct bodies of research have emerged: *between-class* grouping, which is related to the institutional organization of the school, and *within-class* grouping, which is concerned with classroom organization. Overwhelmingly, previous reviews have focused on between-class grouping; when they have considered within-class grouping, they focused on math outcomes and have neglected reading outcomes. This review will contribute to the literacy field by focusing specifically on reading outcomes for classroom teachers, who instruct a wide variety of students.

Kulik and Kulik (1987) considered nineteen within-class grouping studies. For four programs specifically designed for gifted and talented students, they found a statistically significant average overall effect size of +0.62 on all outcomes. For the nine programs designed for all students, they found an average effect size of +0.17, but this difference was not statistically significant. Slavin's (1987) meta-analysis considered eight within-class grouping studies, but only one of these studies measured a reading outcome. Citing mean effect sizes of +0.65, +0.27, and +.41 for low-, average-, and high-achievers, Slavin supported the practice of within-class ability grouping in the upper elementary mathematics. He also stated that there was not enough research on within-class ability grouping in reading to permit any conclusions. Citing 578 studies, Johnson and Johnson (1986) conducted a systematic meta-analysis on cooperative and competitive group learning, yet only 21 of their studies reported on any reading outcome and less than 5 reported on text comprehension. With a focus on the reading development of students with disabilities, Elbaum (1999) reported a mean weighted effect size of +0.58 for intraclass grouping. The majority of their studies, however, were cross-age tutoring dyads in special education resource pull-out programs rather than traditional within-class groups *per se*.

With a specific focus on within-class grouping, Lou et. al's meta-analysis (1996) reported an overall mean weighted effect size of +0.17 in favor of within-class grouping. Citing only four studies that compared heterogeneous to homogenous groups in reading, Lou et. al reported a mean effect size of +0.36 in favor of homogeneous ability reading groups. Overall, however, they also noted that low-achieving students performed, on average, worse in homogeneous ability groups (ES = -0.60). In a later analysis, Lou, Abrami, and Spence (2000) used multiple

regression and a hierarchical regression models to explore effect size moderators; they found outcome source, teacher training, grouping specificity, type of small-group instruction, grade level, and relative ability level to be statistically significant predictors of the mean weighted effect size. Like other previous meta-analyses, however, their overall sample of studies was composed mainly of mathematics achievement outcomes.

Purpose / objective / research question / focus of study:
Description of what the research focused on and why.

Informed by previous research on within-class grouping, the following three research questions guide the present study: 1) To what extent does within-class grouping impact student achievement in reading? 2) For which grade(s) or which students is within-class grouping most or least beneficial? 3) Do any moderators, especially those identified by previous research (measurement source, teacher development, and grouping type), help explain this effect?

Setting:
Description of where the research took place.

Two setting criteria were used in this study. First, to be included in the meta-analysis, the study must have been published after 1980. Although we initially intended to include older studies, it seemed inappropriate to compare some of the earliest research (1920-1950) with today's research – not because the practice of grouping has changed dramatically but because the comparison condition has changed dramatically. Pre-1950 comparison groups often had entire classes reading the exact same "grade-appropriate" book. Notably, some of the earliest studies (Evans, 1927) showed extremely large effect sizes (above +1.0 for some grades). Second, studies from any region in any country were eligible; only studies published in English were included.

Population / Participants / Subjects:
Description of participants in the study: who (or what) how many, key features (or characteristics).

This review includes studies of interventions delivered to Grade 2-10 school-aged children (or equivalent ages in international settings) in regular classroom settings during school hours. Special education classrooms and alternative schools were eligible school settings, but studies (like most reciprocal teaching interventions) that pulled students out of their regular classroom for special instruction were not eligible. Only studies conducted with full, intact classrooms were eligible. Studies of after-school programs were also not eligible. In the event of mixed-aged classrooms, partnering students must have been within one year of each other. No studies of college or adult students were included.

Intervention / Program / Practice:
Description of the intervention, program or practice, including details of administration and duration.

All interventions reported on the effects of within-classroom grouping. The within-classroom grouping intervention was often combined with another intervention like strategy instruction or cooperative learning. Between-classroom grouping studies—often known as tracking or streaming—that placed children in different rooms for a semester or a year were not eligible. Any intervention that specifically trained students in one-to-one peer tutoring was also not

eligible. The outcome of primary interest was reading achievement. Standardized reading assessments approximate global "Reading" and "Language" ability by measuring Vocabulary, Comprehension, Word Analysis, Language Mechanics and Expression, and sometimes other subcategories (depending on the test and forms are delivered). For the standardized assessments, the effect size for each study cohort was calculated from the "Total Reading" score. Researcher and teacher developed assessments of reading were also allowed.

Research Design:
Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

This review allowed only experimental studies or quasi-experimental studies that utilized a pretest, posttest, and comparison group design. Many studies utilized post hoc matching, but evidence of pretest equivalence on reading outcomes was required. Due to the limited resources for this review, only English-language studies were considered. Two studies conducted outside the United States were included. To the extent that this was possible, the control groups received "ungrouped" instruction. Studies without control or comparison groups were not eligible.

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.

For this review, studies on group interventions were identified using several methods. First, the reviewers searched Cambridge Scientific Abstracts (CSA) using the following databases: ERIC (Education Resources Information Center), IBBS (International Bibliography of the Social Sciences), PsycARTICLES, and PsycINFO. Secondly, bibliographies of relevant meta-analyses and literature reviews were inspected. Thirdly, Proquest Digital Dissertations was used to search for unpublished dissertations. Fourthly, bibliographies of retrieved studies were examined for potentially eligible research reports. In total, over 4600 articles were screened.

A pre-established coding form was used to code each study. Including effect size statistics, a total of 51 variables were coded on each study. Study coding and data management was initially done using Excel software; after all coding was complete, the data was transferred to SPSS and CMA (Comprehensive Meta-analysis) for statistical analyses. Studies were coded independently by two coders and then checked for reliability. Reliability coding was conducted on all studies. Overall, the coding reliability was satisfactory. The mean kappa coefficient on categorical moderators was 0.86 with a minimum value of 0.41 and a maximum value of 1.0. In cases of disagreement, the coders discussed each item to resolution. The coding form was updated (with more detailed definitions or different categories) when items had reliability below 0.7. Pearson's correlation for continuous variables (including effect size data) was 0.98 with a minimum value of 0.91 and a maximum value of 1.0. Again, in cases of disagreement, the coders discussed each item to resolution. Importantly, the effect size statistic was calculated exactly the same for both coders.

We used Hedges' unbiased estimate (Hedges's g) of the standardized mean difference effect size statistic (the difference between the treatment and control group means on an outcome variable divided by the pooled standard deviations for the posttest measure). Because every intervention implemented their treatment to a naturally occurring cluster of students, observed sample sizes

were adjusted to effective sample sizes. In order to calculate the effective sample sizes and, therefore, the correct standard errors for our effect size estimates, an Excel spreadsheet created by Kathleen McHugh (2004) and modified by Mark Lipsey was used. This spreadsheet was developed based on the work of Hedges (2004a; 2004b). This calculation requires an intraclass correlation; because individual studies did not report intraclass correlations, an imputed default estimate of 0.2 was used (Hedges & Hedberg, 2007) as an approximation for K-12 reading outcomes. Also, each effect size was weighted by its inverse variance in all computations so that its contribution was proportionate to its reliability. A random effects statistical model was used to analyze effect sizes throughout this meta-analysis because significant variability across effect sizes was expected and because we did not want to restrict generalization of the findings only to the specific studies located for the analysis.

Findings / Results:

Description of main findings with specific details.

In the end, the search yielded 15 unique articles, reports, or conference papers that met our inclusion criteria (See Figure 1). Among these 15 studies, we identified 28 unique "study cohorts", as some reports focused on multiple cohorts and analyzed them separately (normally by grade level). The evidence base described here relies upon an observed sample of 5,410 study participants, 2,424 of whom were in the treatment groups and 2,986 of whom were in the control groups. The overall weighted random effects mean was 0.22 (p=0.002), indicating that subjects in the intervention groups had significantly higher reading achievement scores than comparison subjects after participating in a within-class grouping intervention.

Figure 2 shows the forest plot for the effect size distribution, using random effects methods. The effect sizes range from -0.26 to 0.73. This figure shows that the majority of effects (over 75%) are positive. The 95% confidence interval around this weighted mean ($0.08 < \mu < 0.349$) does not include zero and reveals the relative precision of the estimate of the mean random effects size of the population of studies from which these 28 effects are presumably drawn. The standard error around this mean is 0.069, and the variance is less than 0.005. The homogeneity analysis test determines if variations in the effect sizes are due to sampling error or other factors (such as treatment intensity, study design, or publication type). A test of the homogeneity of the effect sizes was conducted using the Q-statistic. The analyses revealed that there was not significant variability across different studies in the effect sizes to reject the null hypothesis of homogeneity ($Q_{27} = 9.93$). That is, the variability across effect sizes did not exceed what would be expected for sampling error alone. Lastly, a publication bias analysis was also conducted (Figure 3). Egger's regression method for the 28 effect sizes produced an intercept (B0) of 0.06 and a 95% confidence interval that includes zero (-0.69, 0.82) with t(26)=0.17. The one-tailed p-value is 0.43; the two-tailed value is 0.86. The Egger's regression interception technique does not support the conclusion of publication bias in this sample.

Conclusions:
Description of conclusions and recommendations based on findings and overall study.

The overall mean weighted effect size of +0.22 suggests that within-class grouping is effective at improving reading achievement. This is slightly larger than the effect size (0.17) reported by Lou et. al (1996) for within-class grouping overall. There was not significant heterogeneity in outcomes across the 28 studies we reviewed; therefore, moderator analysis was not warranted (research questions 2 and 3 could not be answered) and would have capitalized on chance.

What is the practical significance of an effect size of 0.22? Is this effect size big or small? If we use Cohen's index for effect sizes in social science, we might consider this to be a small effect. But, if we consider this in relation to normative yearly reading growth on standardized achievement outcomes we might interpret this effect size differently. Hill, Bloom, Black, and Lipsey (2008) analyzed the annual reading gains on 7 standardized tests and found that the mean effect size for reading growth for students in Grade 2 through 6 was 0.6, 0.36, 0.4, and 0.32 respectively. Thus, when we consider that the mean Grade for the students in this sample was Grade 4, we can see that an effect size of 0.22 translates into approximately an extra half of a year's growth in reading.

School administrators and teachers have a wide variety of pedagogical choices and are interested in making effective educational decisions. Within-class grouping is a common classroom practice and this review has shown it to be effective at improving reading achievement. It is perhaps, however, a diminishing practice; recent studies (Chorzempa and Graham, 2006) have shown that between 56 and 60% of teachers group students for reading instruction within their classrooms. This review recommends that teachers and schools continue to use grouping to improve reading instruction. More research needs to be conducted using high quality designs in order to answer the questions related to the type of grouping that is most beneficial and how to support teachers' implementations of these practices.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Austin, M. C., & Morrison, C. (1963). *The first R: The Harvard report on reading in elementary schools.* New York: Macmillan.

Baker, L., Dreher, J., & Guthrie, J. (Eds.). (2000). *Engaging young readers: Promoting achievement and motivation.* New York: Guilford Press.

Baumann, J. F., Hoffman, J. V., Duffy-Hester, A. M., & Moon Ro, J. (2000). The first R yesterday and today: U.S. elementary reading instruction practices reported by teachers and administrators. *Reading Research Quarterly*, 35, 338–377.

Chorzempa, B. F., & Graham, S. (2006). Primary-grade teachers' use of within-class ability grouping in reading. *Journal of Educational Psychology*, 98(3), 529-541.

Elbaum B., Vaughn, S., Hughes, M., & Moody, S. W. (1999). Grouping practices and reading outcomes for students with disabilities. *Exceptional Children*, 65(3), 399-415.

Englert, C. S., & Mariage, T. V. (1991). Making students partners in the comprehension process: Organizing the reading "posse." *Learning Disability Quarterly*, 14, 123-138.

Evans, Mark M. "The Effect of Variable Grouping on Reading Achievement." Unpublished Dissertation, University of Pittsburgh, 1942.

Fuchs, Douglas, Fuchs, Lynn S., Mathes, Patricia G., Simmons, Deborah C. (1997). Peer-Assisted Learning Strategies: Making Classrooms More Responsive to Diversity *American Educational Research Journal*, 34: 174-206.

Hedges, L. V. (2004a). Correcting significance tests for clustering. Unpublished manuscript.

Hedges, L. V. (2004b). Effect sizes in multi-site designs using assignment by cluster. Unpublished manuscript.

Hedges, L. (2007). Effect Sizes in Cluster-Randomized Designs. *Journal of Educational and Behavioral Statistics*, 32 (4), 341–370.

Hill, C., Bloom, H., Black, A., Lipsey, M. (2008). Empirical Benchmarks for Interpreting Effect Sizes in Research. *Child Development Perspectives*, 2, 172-177.

Johnson, D. W. & Johnson, R. W.(1986). *Cooperation and Competition: Theory and Research*. Edina: Interaction Books.

Kulik, J. A. (1992). An analysis of research on ability grouping: Historical and contemporary perspectives. Research-based Decision-Making Series. Storrs, CT: University of Connecticut, National Research Center on the Gifted and Talented (ERIC Document Reproduction Service No. ED 350 777).

Kulik, J. A., & Kulik, C.-L. C. (1987). Effects of ability grouping on student achievement. *Equity and Excellence*, 23, 22–30.

Lipsey, M. W., & Wilson, D.B. (2000). *Practical Meta-analysis*. Thousand Oaks: Sage Publications.

Lou, Y., Abrami, P. C., & Spence, J. C. (2000). Effects of within-class grouping on student achievement: An exploratory model. Journal *of Educational Research*, 94(2), 101–112.

Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., & d'Apollonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research*, 66(4), 423–458.

Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research*, 57, 293–336.
Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research*, 60(3), 417–499.
Stevens, R. J., Slavin, R. E. (1995). The Cooperative Elementary School: Effects on Students' Achievement, Attitudes, and Social Relations. American *Educational Research Journal*, 32, 321-351.

Appendix B. Tables and Figures

Not included in page count.

Figure 1

Characteristic	N	%	Characteristic	N	%
Publication Year			**Race/Ethnicity** *(Predominant)*		
1980s	5	33%	White	2	13%
1990s	8	53%	Black	1	7%
2000s	2	13%	Hispanic	3	20%
			Mixed	4	27%
Form of Publication			Lebanese	1	7%
Non Peer Review Journal	2	13%	Cannot Tell	4	27%
Peer Review Journal	13	87%			
			Grouping Type		
Country of Study			Heterogeneous Grouping	14	93%
USA	13	87%	Homogeneous Grouping	0	0%
Other	2	13%	Unable to Determine	1	7%
Quasi-experimental Design			**Group Size**		
Cluster Randomized	3	20%	Dyads	1	7%
Matched Comparison	12	80%	Mixed (Pairs and groups)	8	53%
			4 to 5	3	20%
Pretest Adjustment			Unable to Determine	3	20%
Yes	11	73%			
No	4	27%	**Intervention Instructor**		
			Teacher	14	93%
Observed Sample Size			Researcher	1	7%
Under 50	1	7%			
51-100	1	7%	**Training Provided for Instructors (hours)**		
101-250	5	33%	< 10	4	27%
251-500	6	40%	10 to 25	1	7%
501 +	2	13%	25 to 40	5	33%
			Unable to Determine (or NA)	5	33%
Sample Grade					
Grade 2 - 6	13	87%	**Length**		
Grade 7 - 10	2	13%	< 10 weeks	3	20%
			10 - 20 weeks	4	27%
Outcome Measures			20 - 40 weeks	7	47%
Standardized Tests	11	73%	> 40	1	7%
Researcher Developed Tests	4	27%			
			Intensity of Treatment		
Location			< 30 minutes	1	7%
Urban	6	40%	30 - 60 minutes	4	27%
Suburban	6	40%	60 - 90 minutes	3	20%
Rural	1	7%	> 90 minutes	4	27%
Unable to Determine	2	13%	Unable to Determine	3	20%

Figure 2

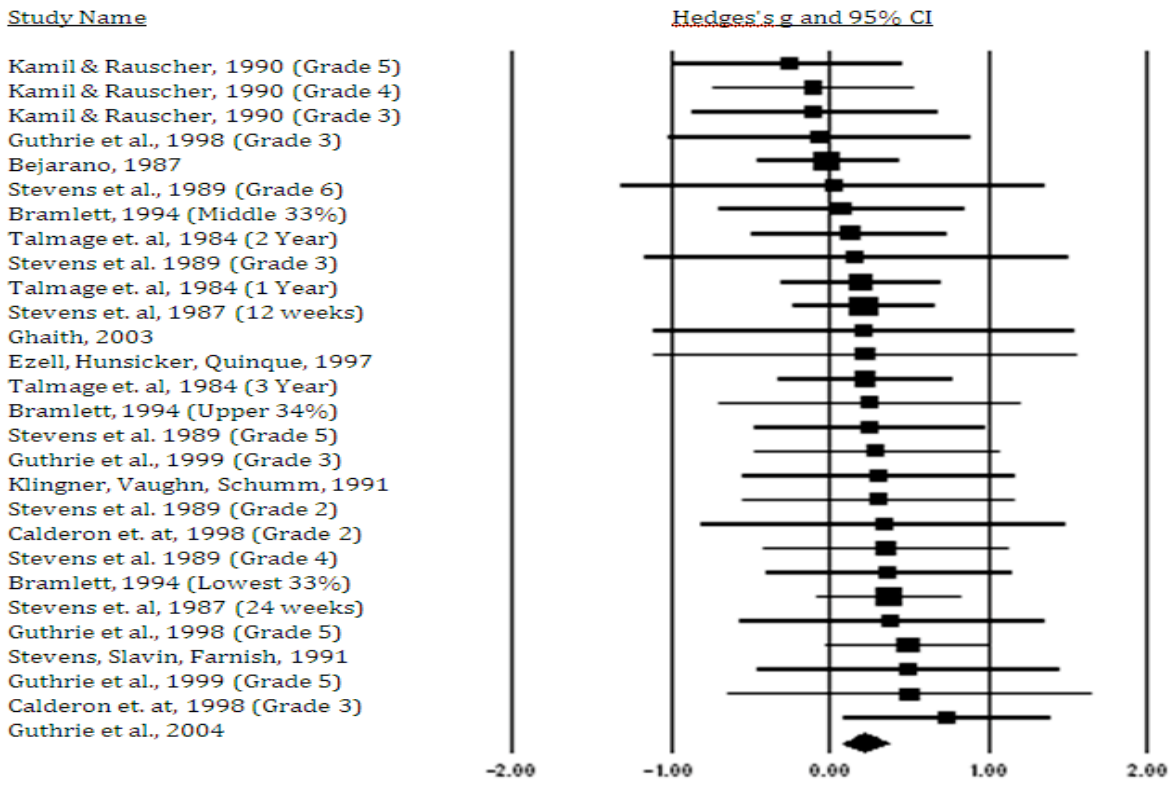

Figure 3