



Published in final edited form as:

*J Educ Psychol.* 2010 February 1; 102(1): 32–42. doi:10.1037/a0017288.

## “Teacher Effects” in Early Literacy Development: Evidence from a Study of Twins

**Brian Byrne** and **William L. Coventry**

University of New England

**Richard K. Olson** and **Sally J. Wadsworth**

University of Colorado

**Stefan Samuelsson**

Linköping University and Stavanger University

**Stephen A. Petrill**

Ohio State University

**Erik G. Willcutt** and **Robin Corley**

University of Colorado

### Abstract

It is often assumed that differences in teacher characteristics are a major source of variability in children’s educational achievements. We examine this assumption for early literacy achievement by calculating the correlations between pairs of twin children who either share or do not share a teacher in kindergarten, Grade 1, and Grade 2. Teacher effects, or more strictly classroom effects, would show up as higher correlations for same- than different-class twin pairs. Same-class correlations were generally higher than different-class correlations, though not significantly so on most occasions. On the basis of the results we estimate that the maximum variance accounted for by being assigned to same or different classrooms is 8%. This is an upper-bound figure for a teacher effect because factors other than teachers may contribute to variation attributable to classroom assignment. We discuss the limitations of the study and draw out some of its educational implications.

---

Factors influencing the educational achievement of students are the subject of sustained journalistic, political and scientific debate (Darling-Hammond, 2000; Darling-Hammond & Youngs, 2002). Among those factors are “teacher effects,” differences in effectiveness of individual classroom teachers, whose presumed importance is suggested by the following quotes from the popular press:

“...anyone with an ounce of brains knows what must be done... It’s time to move from identifying failing schools to identifying failing teachers.” Jonathan Alter, *Newsweek*, February 12, 2007.

“Studies claim that 40 per cent of the variation in student performance is the result of teacher quality.” John Della Bosca, NSW Minister for Education, *Sydney Morning*

---

Correspondence concerning this article should be addressed to Brian Byrne, School of Behavioural, Cognitive and Social Sciences, University of New England, Armidale NSW 2351, Australia. bbyrne@une.edu.au.  
Brian Byrne and William L. Coventry, School of Behavioural, Cognitive and Social Sciences, University of New England; Richard K. Olson, Sally J. Wadsworth, Erik G. Willcutt and Robin Corley, Institute for Behavioral Genetics, University of Colorado; Stefan Samuelsson, Department of Behavioral Sciences, Linköping University; Stephen A. Petrill, Department of Human Development and Family Science, Ohio State University.

*Herald*, April 18, 2008, in the context of an article about performance-based pay for teachers.

Estimating and understanding teacher effects are important not only for gaining insight into individual students' performance but for guiding investments in teacher preparation and certification (Darling-Hammond, 2000) and evaluating teachers for career advancement and remuneration (McCaffrey, Lockwood, Koretz, & Hamilton, 2003). In this report we focus on possible teacher effects on early literacy development, a skill which plays a fundamental role in all educational achievement. In reviewing the literature, we make the case that additional sources of information on this question would be of value.

We use the phrase *possible teacher effects* because most research, including ours, is in reality dealing with *classroom effects*. As we discuss below, this is particularly true of statistical approaches that strive to identify between-classroom variance in academic attainment that cannot readily be explained by characteristics of students, schools, neighborhoods, and so on. The idea that variance lying between classrooms is due to the teacher is mostly an assumption. Indeed, as we also document below, classroom-level factors that can influence class performance other than teacher have been identified. For these reasons, we will refer from here on to *classroom effects*.

The literature on classroom effects is vast. For example, as early as 1992 Hattie was able to summarize the results of 9 meta-analyses of teacher "background" and "style" (school subjects unspecified but presumably a reasonable number including literacy), which in turn summarized a total of 329 studies and 1097 effects, with an average effect size of 0.60 for background and 0.42 for style. Thus, a review of the literature is beyond the scope of this article, with its limited goal of introducing a novel method of estimating these effects and describing the results when we use that method. What we do need to do is provide a justification for embarking on this study at all, and that involves identifying some of the challenges to the validity of inferences about classroom effects that currently exist. We do that in what follows.

## Classroom effects are real

There is no doubt that there are genuine classroom effects for early literacy, as demonstrated in experimental studies. For instance, Connor, Morrison, Fishman, Schatschneider, and Underwood (2007) developed an algorithm-guided individualized instruction program for implementation by teachers using lesson-planning software. They reported stronger reading growth in classrooms implementing these methods than in control classes, and that greater growth accompanied greater fidelity to the methods. Blachman, Tangel, Ball, Black, and McGraw (1999), using regular teachers and teaching assistants, showed greater gains in basic reading skills for kindergarten and first-grade students of teachers who had received training to emphasize phonemic awareness and letter knowledge compared to teachers who followed the normal curriculum. These and other intervention studies for early literacy (see Bus & van IJzendoorn, 1999, for a survey) amount to a proof-of-principle; that scientifically-guided, special interventions managed by individual teachers can affect students' achievements. However, this research does not directly address what is occurring in the normal running of actual school systems, which is the arena in which classroom effects are generally evaluated.

## Findings and issues

The ability to identify classroom effects is made complex by the large number of factors that could and do affect children's literacy growth. For example, the substantial variability in children's reading and spelling skills is partly a product of variability in genetic endowment, which, for example, accounts for between 50 and 80% of individual differences at the end of first grade in the U.S., Australia, and Scandinavia (Byrne et al., 2006; Byrne et al., 2007; Petrill,

Deater-Deckard, Thompson, DeThorne, & Schatschneider, 2006; Petrill et al., 2007). Other influences include practices in the home (Petrill, Deater-Deckard, Schatschneider, & Davis, 2005), socio-economic level and ethnicity (McCoach, O'Connell, Reis, & Levitt, 2006), and, in the case of reading fluency at least, peer influences within the classroom (Foorman, York, Santi, & Francis, 2008: see Papaioannou, Marsh, & Theodorakis, 2004 and Ryan, 2000, for a broader discussion of peer influences).

Thus, interpretation of classroom-effect studies depends on how successfully they control for other influences on achievement. Consider an example from Darling-Hammond (2000). Using state-aggregated data, she showed that there is a substantial relationship between the proportion of well-qualified teachers in a state and the state's level of results in the National Assessment of Educational Progress (NAEP). The controls, again at state level, were the percentage of students in poverty and a measure of non-English language status. As the author acknowledges, however, adding other measures of student background, such as parent education levels, might make a difference to the overall picture.

Researchers agree that among the background factors that are useful is prior achievement in the domain because it is assumed to fold into a single number the effects of background variables such as genetically-driven aptitude for a subject and any effects of the home environment up to that point in time (Rowan, Correnti, & Miller, 2002; Wayne & Youngs, 2003). Achievement models of this type are referred to as *value-added* because they assume that the classroom (or in most researchers' view, the teacher) to which a child is assigned adds to (or subtracts from) progress over the period of assignment relative to other equivalent classrooms in the school, district, state or nation (McCaffrey et al., 2003).

Behavior-genetic analyses of individual differences in early literacy suggest, however, that there is an assumption in this model that may not hold true. A child's genetic endowment for literacy can change across development. For instance, new genetic influences on word reading skill come into play as children move from kindergarten to Grade 1 (Byrne et al., 2007; Samuelsson et al., 2008). Further, reading comprehension becomes more genetically complex with age; early in schooling its genetic correlation with word reading is very high, with virtually no genetic input that is independent of that shared with word reading (Byrne et al., 2007). But at a later stage genes that separately affect "higher level" language variables such as vocabulary also become important in influencing scores on tests of reading comprehension (Byrne et al., 2009; Keenan, Betjemann, Wadsworth, DeFries, & Olson, 2006). Thus, at least some changes in literacy levels across time, at both word and text levels, will be caused by changing genetic influences and not by classroom-derived factors. The influence of such changes in particular children or their aggregated effect in a class cannot be known in the absence of markers for the genes in question (which we do not have), but in none of the literature that we have seen on classroom effects do they appear to be acknowledged as a source of uncertainty. In the method of comparing twins who either share or do not share a teacher that we outline in detail later, genetic influences are controlled at least for monozygotic twins, such that any greater difference between members of such pairs who are in separate classes cannot be due to changing genetic endowment.

This caveat aside, evidence from value-added models appears to support classroom effects for literacy. Studies might calculate a  $\Delta R^2$ , the gain in variance explained by including the classroom as a dummy variable on top of whatever other background variables are included, or classroom effects might be estimated directly in a hierarchical linear model analysis (Rowan et al., 2002). Nye, Konstantopoulos and Hedges (2004) summarized a selection of studies using these approaches. The teacher effect, as they prefer to style it, ranged from .03 to .16 for reading and vocabulary.

McCaffrey et al. (2003) provide another summary and analysis of value-added models, some of which are published in peer-reviewed outlets and include measures of reading and language; Rowan et al. (2002) and Wright, Horn, and Sanders (1997). Rowan et al. estimate that classroom (teacher) accounts for between 4 and 16% of variance in reading scores adjusted for student characteristics and prior scores. They also claim that classroom accounts for around 60% of what they term “reliable variance” in reading growth. McCaffrey et al. raise some questions about the definition and computation of reliable variance, and report that when an alternative definition is used the classroom effect is 13 to 14%. They also comment on a variety of other issues in the Rowan et al. analysis, such as the large amount of missing data and the omission of covariates. They conclude that Rowan et al. “provide convincing evidence of likely teacher effects, although the exact magnitude is less well established. Their results should be interpreted cautiously because the effect sizes are relative to only part of the variability in scores (or growth) rather than to the total variability” (p. 30). McCaffrey et al. reached a similar kind of conclusion with respect to the paper by Wright et al.: “we believe that [Wright et al.] provide evidence of residual classroom variance predicting gains, although we cannot evaluate the absolute or relative size of the effect given the information in the...paper” (p. 24). Based on these and other considerations, they concluded that “the research base is currently insufficient to support the use of VAM (their acronym for value-added models) for high-stakes decisions. We have identified numerous possible sources of error in teacher effects and any attempt to use VAM estimates for high-stakes decisions must be informed by an understanding of these potential errors” (p. xx). These conclusions underline the value of a novel approach to estimating classroom effect sizes.

Some studies of classroom effects attempt to identify qualities of teachers that are critical in determining whether students prosper or falter. Research in this tradition has yielded mixed results. To take a particular example, Connor, Son, Hindman, and Morrison (2005) reported that teachers’ years of education related positively to Grade 1 vocabulary but *negatively* to letter-word (literacy) scores. A recent survey is by Wayne and Youngs (2003). They canvass a number of considerations that condition the interpretation of results, such as the possibility of “lagged” effects, whereby the effects of a previous teacher shows up in a particular period and is wrongly attributed to the current teacher, and the possibility that a measured variable, such as holding a masters degree or not, is confounded with another, such as years of experience. In the face of these and other sources of possible bias, they declined to discuss observed effect sizes even when the surveyed articles reported them. Wayne and Youngs also declined to include studies that attempt to assess the effects of teacher experience. They considered that such studies are too difficult to interpret because, for example, experience measures would capture any differences between teachers who leave the profession and those who remain, and they also capture “the effect of whether teachers were hired during a shortage or a surplus” (p. 106).

The primary positive findings from the Wayne and Youngs (2003) survey are evidence that the teachers college that teachers attended and the test scores that they received while studying had a positive relation to student gains. Other aspects of certification such as the nature of the qualification and the coursework it contained did not, except for mathematics.

Another approach to classroom-level effectiveness is to track the practices of “effective” teachers (for a summary, see Alvermann, Fitzgerald, and Simpson, 2006). The search for these practices should draw a blank if there are no teacher effects. A representative study with young children is from Pressley et al. (2001). They selected 15 “most-effective-for-locale” and 15 least-effective-for-locale first-grade teachers as identified by school officials on criteria deliberately left unspecified by the researchers. The effective teachers were superior in classroom management, balance of teaching, scaffolding of instruction to match student ability, encouragement of self-regulation by students, and connections made across the curriculum.

As with other approaches, interpretation depends on the effectiveness of control measures, in particular in this case how the students were originally assigned to the two types of teachers' classes and the bases on which the school officials made their judgments. If the most effective teachers are assigned more able students or are designated as effective *because* their students are achieving better results, interpretation remain ambiguous. The same issue arises, though perhaps less forcefully, even when effectiveness of teacher is defined by student growth during the period of the study (Taylor, Pearson, Peterson, & Rodriguez, 2003). If effective teachers have been assigned students who show higher promise, the results also remain ambiguous. As Taylor, Pearson, Clark, and Walpole (2000) comment in an earlier study: "When all is said and done, we are examining natural correlations between program and teaching factors on the one hand and student performance on the other. These correlations may be useful in planning more definitive research...however, they cannot be used to identify causes for improvements (or decrements) in student achievement. For that, more systematic experimentation is needed, including control groups, randomization, and careful analyses of growth over time" (p. 160).

Nye et al. (2004) added to the list of factors that can be difficult to control by pointing out that schools may have information about individual students not available to researchers that influence the way students are allocated to teachers. On the negative side, impending divorce, parental unemployment, and the like, and on the positive side, parental involvement in the school, improvements in motivation, and so on, may influence classroom placement. Random assignment of children to teachers and teachers to children would finesse these problems of interpretation, reducing pre-existing correlations of levels of teaching skill with levels of class achievement to chance values. Nye et al. had such a research resource available as a consequence of the earlier Tennessee Class Size Experiment (Nye, Hedges, & Konstantopoulos, 2000). Over a four-year span, kindergarten to Grade 3, children and teachers were randomly assigned to classes in 79 schools. With prior achievement as a covariate, the amount of variance between classrooms in student gains in reading ranged between .066 and .074 in Grades 1-3. In kindergarten, where only achievement, not gains, could be measured, the value was .100. All these estimates were significant beyond .05. The classroom effect was more substantial in low-SES schools than high-SES schools, .079-.140 versus .038-.049. Neither teacher education nor teacher experience explained very much of these classroom effects.

Although the study was conducted in a broad cross-section of schools in a diverse US state and represents one of the most scientifically sound approaches to the question of classroom effects, it is one of a kind and it was limited to a single US state. Because it involved state-wide cooperation in randomizing thousands of students and hundreds of teachers at a cost of around \$12 million, it may be difficult to replicate. It may also have had its own limitations. It was, for instance, subject to substantial attrition, a problem in value-added models that McCaffrey et al. (2003) identify. As well, age differences between classes within grades approached significance for two of the four years of the study. Therefore studies with differing methodologies that converge on the conclusions would be of benefit, if only on the grounds of generalization of the findings.

There is one other issue that adds to the uncertainty around the interpretation of classroom effects: Classroom effects may change with time. For instance, if in an educational jurisdiction the general level of student achievement rises, classroom effects may fade. This may have happened in Tennessee, where a classroom-delivered regimen known as peer-assisted-learning generated average advances in kindergarten reading in 1997 and several subsequent years. But by 2005 the effect had dissipated due to a rise in the level of reading in control classes to match the experimental group following changes in kindergarten practices (Fuchs, Saenz, McMaster, Yen, Taylor, Lemons et al., 2008). Thus, evidence about classroom effects can be conditioned by when and where the data are collected, and may not apply in perpetuity and in other regions.

This observation underlines a more general point, that classroom-level effects may depend on the context in which the data are collected. For those effects that are attributable to teachers, for example, a high degree of uniformity in teacher preparation for the educational jurisdiction in question would tend to lower the variance that teachers account for. In contrast, highly diverse teacher training such that, for example, a substantial proportion of the teacher workforce was delivering suboptimal instruction and an equal proportion delivering optimal instruction would result in substantial classroom effects. We return to this issue in the Discussion.

In summary, although there is evidence in favor of a classroom effect in early literacy development, there are several sources of uncertainty that cloud the interpretation and generalization of research; control of background variables, including those of students and teachers, technical details of analyses, methods of teacher assignment that may be based on confidential information, and locale and time of the study. Genuinely experimental studies can overcome some of these difficulties, but such studies are rare, and may indeed have their own limitations. None of the approaches so far used are geared to account for the possible effects of genetic changes that can alter an individual child's developmental trajectory. Newer methods of investigating classroom effects are therefore of value in that, whatever their own limitations, they may be different ones than those that apply with other methods.

### Data from Twins as a Novel Method of Assessing Classroom Effects

If there are detectable classroom effects on achievement, they should be evident from the comparison of twin children who either share or do not share a teacher. In fact, the size of the classroom effect can be estimated directly from the data, as we explain: Twin data can be partitioned into variance due to genetic influences, generally symbolized as  $A$  in variance component models, and variance due to environmental influence. The environmental variance can in turn be partitioned into shared environmental influences that stem from twins living in the same family, symbolized as  $C$ , and environmental influences that are unique to each child, symbolized as  $E$  (Plomin, DeFries, McClearn, & McGuffin, 2008). Additive genetic effects can be estimated by doubling the difference between the monozygotic (MZ) and dizygotic (DZ) twin correlations,  $r_{MZ}$  and  $r_{DZ}$  respectively. Thus,  $A = 2(r_{MZ} - r_{DZ})$ . If MZ twins are more alike than can be accounted for by  $A$ , the source of the additional similarity is, by definition,  $C$ , so  $C = r_{MZ} - A$ . Insofar as MZ twins are not alike, the source of the dissimilarity is, by definition,  $E$ , so  $E = 1 - r_{MZ}$ . For twin children in the same classrooms, a classroom effect would increase shared environmental influence, simultaneously decreasing unique environment influence. Given this, for MZ and DZ twins, the difference between same- and different-class correlations can be used to estimate a classroom effect. For example, if the correlation for reading in a sample of MZ twins where both members of a pair share a classroom is 0.8, so that  $E = 0.2$ , and in a sample where both members of a pair are in separate classrooms is 0.6, so that  $E = 0.4$ , then the classroom effect is 0.2, the addition to  $E$  from being in different classrooms.

One twin study has already examined classroom effects. Using teacher ratings of school performance and couching the results in terms of changes in *shared* environment, the Twins Early Development Study (TEDS) in the UK identified non-significantly higher shared environment effects in same- versus different-class twins assessed in reading, other aspects of English, mathematics, and science at ages 7, 9, and 10 (Kovas, Haworth, Dale, & Plomin, 2007). Importantly, the differences, statistical nonsignificance notwithstanding, mostly disappeared when objective test results replaced teacher ratings in reading, mathematics, and general cognitive ability, suggesting that rater bias might explain the rating-based observations; children in the same class also have the same rater, those in different classes, different raters. (The reading data alone continued to show a trend, not statistically significant, for higher shared

environment effects in same-class twin pairs, .17, versus .07 for different-class pairs, when objective testing replaced teacher ratings.)

Our twin data are based exclusively on objective testing and represent a relatively dense sampling, with assessments of different aspects of literacy across each of the first three years of school, ages 5 and up, when the skill is first being acquired. This offers ample opportunity to check for classroom effects in this important domain during the early school years, when they might be expected to be most visible because of the spurt in reading development that occurs in those years, particularly Grade 1 (McCoach et al., 2006).

## Method

### Participants

Our data come from an ongoing longitudinal study of young twins, the International Longitudinal Twin Study (ILTS) with samples from Colorado, USA, and New South Wales, Australia (Byrne et al., 2002,2005,2006,2007,2008; Samuelsson et al., 2005,2007,2008; Wilcutt et al., 2007). (The ILTS also has a Scandinavian sample, but only a few of the twin pairs are placed with different teachers there; hence that sample cannot inform the classroom-effect issue.) The maximum sample comprised 711 twin pairs, 483 recruited from the Colorado Twin Registry in the U.S. and 228 recruited via the National Health and Medical Research Council's Australian Twin Registry. There were 355 MZ pairs and 356 same-sex DZ pairs (zygosity by country is given in the tables in Results). No payment was given for participation in Australia, but the parents in the U.S. sample received a payment of \$100 for participation. Parents of the Colorado twins were approached by mail or phone, and 88% of the 60% of families who could be contacted when the children were 4 agreed to participate. The twins' parents in Australia were all approached by mail with a participation rate of 62%. All twins were in their final preschool year at initial contact with ages ranging from 54 to 71 months (mean 58.8) in the U.S. and 47 to 68 months (mean 57.8) in Australia. In kindergarten the children were about 18 months older than in preschool, and in Grades 1 and 2 12 months older again on each occasion. Zygosity was determined in most cases (81%) from DNA collected via cheek swabs, and in the other cases from selected items from the Nichols and Bilbro (1966) questionnaire concerning, for example, hair color and texture, eye color, facial appearance and complexion, and birth weight. The questionnaire has 95% accuracy when compared with blood samples.

A requirement for participation was that the child had English as the first language, necessitated by the nature of the tests and the young age at which we first assessed the children. In the US, one likely result of this stipulation was underrepresentation of families identifying as Hispanic, at 7%. Its effects in Australia are not clear; the Australian Bureau of Statistics (2006) reports that 17% of the population over 5 years of age speaks a language other than English at home, and that less than 1% speaks no English at all. The picture for younger children, a substantial proportion of our participants at initial recruitment, is unclear. However, it is safe to assume that the language requirement excluded some families. In both samples, no children receiving ESL education were part of the study.

This is a longitudinal study in which the two most recently recruited waves in Australia have yet to be tested in all grades. This accounts for the fact that the sample sizes decline across the school grades (see tables in Results) because at the time of writing not all children had completed their assessments at all grades. Actual attrition is close to zero; typically, only families who leave the geographical areas drop out of the study.

## Materials and Procedure

Reliabilities derived from our sample are helpful in evaluating the difference between same- and different-class twin pairs' correlations; higher reliabilities mean that modest but significant differences can be treated as meaningful with a degree of confidence. For two of the tests we can report test-retest correlations. For others, we can treat the correlations for same-class MZ twins as lower-bound estimates of reliability because the deviation from unity is caused by nonshared environment factors that include measurement error (but not classroom). These values will be underestimates to the degree that environmental influences of biological (e.g., birthing complications) or educational origin (e.g., exposure to educational television) affect members of a pair differently (Samuelsson et al., 2005). Table 2 contains these correlations for all variables in both countries and for the combined sample, and we repeat the combined sample values below. As can be seen, the MZ correlations for the individual tests range from .72 to .89, and for the combined literacy variable (sum of all the individual tests) they range from .84 to .90 (mean .883). Overall, therefore, the tests we have used have acceptable reliabilities.

**Test of Word Reading Efficiency (TOWRE)**—In this test (Torgesen, Wagner, & Rashotte, 1999), administered at all three school grades, children read a list of words (*Sight Word Efficiency*) and a list of nonwords (*Phonemic Decoding Efficiency*) as quickly as possible, with the score being the number correctly read in 45 seconds. There are two equivalent forms of the test, Forms A and B, and we administered both to optimize the reliability of the scores. Sample correlations between forms are as follows: kindergarten and Grade 1 Sight Word Efficiency, .97 and .95 respectively; kindergarten and Grade 1 Phonemic Decoding Efficiency, .94 and .94. respectively

**Woodcock Passage Comprehension**—This test, from the Woodcock Reading Mastery Test-Revised (Woodcock, 1989), uses a cloze procedure in which the child orally fills a blank in a passage that they are reading to assess the ability to understand passages of connected text. (Same-class MZ correlations, Grade 1 = .77, Grade 2 = .73)

**Spelling**—In kindergarten, the spelling test consists of 10 real words (examples *man*, *come*, *went*) and 4 nonwords (examples *sut*, *ig*). The scoring system honors phonological as well as orthographic accuracy, so that, for example, *kum* for *come* earns next-to-maximum points. The test has been used in studies of an intervention focusing on phonological awareness for preschoolers, including a group at familial risk for developing reading difficulties (Byrne & Fielding-Barnsley, 1993; Hindson et al. 2005). (Same-class MZ correlation = .80)

In Grades 1 and 2 we used the Wide Range Achievement Test Spelling subtest (Jastak & Wilkinson, 1984). Children spell words ranging from simple ones like *go* to complex ones like *belligerent* until they make ten consecutive errors. Score is total number orthographically correct. (Same-class MZ correlations, Grade 1 = .82, Grade 2 = .81)

Each member of a twin pair was assessed by a separate tester at the same time, in the home during the summer for the majority of the US sample and in school during the final three months of the school year in Australia (a minority of Australian pairs was assessed at home). Each test session ran approximately one hour.

## Results

In Table 1 we present means and standard deviations for those tests where national (US) norms are available as a guide to the representativeness of the samples. The US sample was close to the national norm of mean 100 and standard deviation 15 throughout, although variability was somewhat restricted in the kindergarten TOWRE. The Australian children generally performed at a higher level than their US counterparts, a feature of these data from the start (Byrne et al.,



2002). We have speculated that this difference might be due ascertainment bias: the Australian twins were recruited through a voluntary database, the Australian Twin Registry, and the US twins through birth records (Samuelsson et al., 2005). We cannot be sure how representative of the population the Australian sample is because the norms are derived from a US sample, but it is probably safe to assume that these children are a relatively high functioning group on average. Visual inspection of the distributions indicated normality in most cases, though for both the US and Australia there was positive skew in kindergarten for the TOWRE, with a small group of accomplished kindergartners in both samples, and a negative skew for spelling, with a group of children with zero or near-zero spelling ability.

Prior to further analyses, and following standard practice, the scores were age- and gender-adjusted, truncated to  $\pm 3SD$ , and standardized within country. In Table 2 we present the intraclass correlations (correlations between Twin 1 and Twin 2 for MZ and for DZ pairs) for same- and different-class twins for individual tests at each grade. The table also shows these correlations for the average of the individual tests for each child (*Total literacy*) and for the total sample (*Total*). Combining the country samples is justified by the similarity of the overall MZ and DZ twin correlations as well as of the differences between same- and different-class correlations. Averaging the measures for each child is justified by Principal Components Analyses of the individual measures, which at each grade indicated just a single factor, accounting for over 80% of variance.

For both MZ and DZ twins, pairs in the same class were more highly correlated than pairs in separate classes for most measures at each grade. For the individual tests in each country sample, the same-class-pair correlations were numerically higher than those for different-class pairs in a total of 21 out of 22 instances in MZ twins and 19/22 in DZ twins, a result which indicates that the trend held across countries, even though only a few of these contrasts reached statistical significance. For totals summed across the two country samples, the contrast reached statistical significance (one-tailed at  $p < .01$ , to adjust for multiple comparisons) on just 2 occasions out of 22 contrasts for individual tests, kindergarten PDE for MZ pairs and Grade 1 Spelling for MZ pairs, and just once for *total literacy* out of six opportunities, for kindergarten MZ pairs. The general trend of the results, individual contrast significance aside, is for higher correlations among same-class pairs. The mean differences based on the individual contrasts by each test and each country, weighted by sample sizes, are .066 for MZ twins and .096 for DZ twins, and overall rounded mean difference of .08, with a 95% confidence interval of .05 - .11.

In Table 3 we report the correlations for change from one school grade to the next as a function of same or different class assignment in the second year, based on the total sample. We residualized the second year's performance on the first's to control for pre-existing differences, and to enable us to include spelling, which was assessed with different lists in kindergarten and Grade 1.

Only one difference, for the change from kindergarten to Grade 1 spelling for MZ twins, was significant. Nine of the 14 correlations were higher for same-class pairings and 5 went the other way. The overall mean difference and 95% confidence intervals were .04 and  $-.01 - .05$  respectively. These data, therefore, do not provide convincing support for a classroom effect on change trajectories. The generally higher MZ compared to DZ correlations indicate that MZ twins stayed more in step with each other over time than did DZ twins, consistent with other evidence that stability in literacy growth is partly of genetic origin (Byrne et al., 2006,2007).

The higher within-year correlations between twins in the same class versus those in different classes, already evident in kindergarten, could not be attributed to pre-existing greater dissimilarities for twin pairs assigned to different classes, as far as we can determine. This

explanation would go through if there was a tendency for parents and/or schools to separate twins who prior to school appeared to be on divergent trajectories for literacy. Our check on this was to compute intraclass correlations for preschool scores on variables known to predict later literacy growth, with the children retrospectively classified by same and different kindergarten classes. We did this for the total sample. If pre-existing differences influenced class assignment, we should see lower intraclass correlations for those twins who were separated in kindergarten. This was not the pattern of results. To take some salient examples: MZ twins who were subsequently assigned to different kindergarten classes correlated .83 on preschool print knowledge, those assigned to the same class correlated .82. For dizygotic pairs, the respective values were .61 (different classes) and .65 (same class). For phonological awareness, the four values were .78 (MZ different), .78 (MZ same), .38 (DZ different), .49 (DZ same); in this latter case because the difference was fairly substantial and in the direction supporting the “pre-existing difference” hypothesis, we report the substantially overlapping confidence intervals of .20 - .54 and .37 - .58, respectively. For rapid naming of colors and objects the correlations were, in order, .64, .63, .39, .36. For nonverbal IQ (Block Design) they were .60, .66, .47, .38. and for vocabulary were .75, .73, .61, .65. We also checked inattention as it is known to relate to literacy levels and to preliteracy variables in preschool (Willcutt et al., 2007). We calculated the correlations based on parent ratings of the symptoms of inattention from the Disruptive Behavior Rating Scale (DBRS; Barkley & Murphy, 1998). The four values were .73, .71, .45, .37 (the latter two not approaching a significant difference). We conducted the same set of analyses based retrospectively on Grade 1 assignment and found a similar pattern. There was no hint of Grade 1 class decisions being based on the degree of preschool divergence between members of twin pairs.

## Discussion

In this study we assessed the contribution to variance in early word and nonword reading, reading comprehension, and spelling contributed by the classroom by comparing the intraclass correlations for MZ and DZ twin children assigned to the same or different classrooms. Although most of the differences in the correlations failed to reach significance, some did, and in the majority the same-class correlations were higher than the different-class correlations, with an average difference of approximately .08.

We also compared the intraclass correlations for change from one school year to the next in the literacy variables. Here, there was not as clear a distinction between the same- and different-class correlations, with an average of .04 and a confidence interval containing zero. The single significant difference was for spelling in the movement from kindergarten to Grade 1, suggesting that change in spelling ability may be more sensitive to classroom effects than reading. Interestingly, Grade 1 spelling was one of only two individual literacy measures that showed a significantly higher same- than different-class correlation in the within-year comparisons (Table 2). On the whole, however, two MZ twins who move from a shared classroom to separate ones from one year to the next do not diverge more in growth (in reading at least) than two who remain in the same classroom. This result is somewhat surprising in view of the claim that the growth is a particularly sensitive way to assess classroom effects (Rowan et al., 2002).

In the only experimental study of classroom effects (Nye et al, 2004), the proportion of variance in early literacy attributable to classroom was approximately .07, close to our estimate of .08. Both values are towards the lower end of the estimate for a classroom effect on reading in the studies summarized by Nye et al. Our other point estimate, a nonsignificant .04 for a classroom effect on change from one year to the next, is lower, and was generally less consistent than the single-year comparisons of same- versus different-class twins. Overall, our data are not consistent with the strong claims of classroom effects seen in the popular press, there uniformly

described as teacher effects, but are consistent with lower-end estimates seen in the research literature.

An effect of the size we have identified is not trivial. Nye et al. frame their interpretation assuming that teachers are the effective element of classroom—level effects, and point out that “the difference in achievement gains between having a 25<sup>th</sup> percentile teacher (a not so effective teacher) and a 75<sup>th</sup> percentile teacher (an effective teacher) is over one third of a standard deviation (0.35) in reading...” (p. 253). There would be a similar difference between an average teacher (50<sup>th</sup> percentile) and a highly effective teacher (90<sup>th</sup> percentile). Thus, while 8% of variance and differences of the order of one third of a standard deviation would generally be regarded as small effects, any increase in literacy for an individual child is to be welcomed.

### Classroom or teacher effects?

In view of this, identifying the source of classroom effects is important. As we have noted, most researchers treat classroom-derived variance as equivalent to teacher-derived variance. But there are reasons to wonder whether classroom effects are purely, or even mainly, teacher effects. First, children who are in different classes are likely to have social networks less similar than those of children in the same class, and peers outside the classroom could influence early literacy if they engage in activities that either support or compete with literacy development. Second, there is evidence that peers within the classroom might play a role in psychological well-being and academic progress (Rutter & Maughan, 2002) and in reading achievement (Foorman et al., 2008). More generally, researchers have identified a construct termed *classroom climate*, exemplified by the perceptions by students in a class of the class’s attitudes to learning (Papaioannou et al, 2004). Classroom climate can be independent of particular teachers, and can affect the motivation of individual students (Marsh, Martin, & Cheng, 2008; Rutter & Maughan, 2002). Third, in our educational systems non-teacher adults are often present to assist with literacy in classrooms, and the numbers (and presumably the competence) of these paraprofessionals can vary (Pressley et al., 2001). These and other factors may have a hand in classroom effects. Consequently, we propose that our figure is an upper-bound estimate of genuine teacher effects.

The analysis, however, may be more complicated than the way we have presented it. We have been discussing teacher effects and classroom effects as if they are nested, such that there may be more classroom-level processes at work than can be attributed to the teacher. This would have to be logically the case, but the distinction may be a matter of definition. For example, if a teacher implements an effective, new curriculum, is that a teacher or curriculum effect? If children are influenced by their peers, could not a competent teacher understand this and use it to good effect by simple actions such as seating arrangements? Insofar as it is good teachers who understand and exploit such factors, and accepting for argument that our estimate of 8% for classroom effects in total is accurate, it may be partly a matter of definition as to how close to that figure the teacher effect comes.

### Generalizing from twin studies

The degree to which results of twin studies can be generalized to the broader population is an important question (Plomin, et al., 2008; Stromswold, 2006). In the current context, it might be argued that twin data will underestimate a classroom effect; twins who begin to diverge in literacy because they are in different classrooms may converge again, perhaps as the result of parental actions designed to close the gap between the twins, perhaps as the result of activities the twins engage in whose effects keep them in step, or perhaps as a result of communication between teachers and their sensitivity to family dynamics which put a premium on raising the weaker child’s literacy to the level of the stronger one’s. Note, however, that convergence in behavior is not a uniform characteristic of twins; researchers acknowledge competition as

occurring as well (Neale & Maes, 2003). More directly relevant to these possibilities, however, is that shared environment effects are zero or near zero in Grades 1 and 2 literacy scores in our samples (Byrne et al., 2007, 2008). A tendency for members of a twin pair to read and spell at similar levels because of outside or within-pair pressures would result in lowered within-family variance, that is, higher within-pair correlations, and this in turn would emerge as a detectable, possibly substantial, shared environment effect. This aspect of the data therefore does not support convergence as undermining the generalization of our results to the broader population of young children.

Clustering of teachers within schools may result in an underestimate of a classroom effect from twin data. If teachers cluster within a school according to some metric of effectiveness, then twins who are in different classes may nevertheless be receiving similarly effective instruction, more similar than had they been in separate schools. This would result in a higher intraclass correlation for different-class twins than would be the case if teachers did not cluster in this way, and almost all members of twin pairs in our sample attend the same school as each other when they are in different classes. There is some evidence that clustering of this kind can happen, as when schools that happen to serve higher proportions of at-risk students tend to have less experienced and less credentialed teachers (McCaffrey et al., 2003). However, we can again appeal to our estimates of shared environment to evaluate this possible source of underestimation. To the degree that teachers account for significant variance in student achievement *and* cluster within schools, twins in the same school should be more similar to each other irrespective of zygosity. This is the situation that generates significant levels of shared environment in variance component modeling, and it is precisely the shared environment component that is low or nonexistent in our data (Byrne et al., 2007, 2008). Thus, either teachers do not account for sufficient variance in early literacy to produce significant shared environment effects or, more likely perhaps, the characteristics on which they cluster (experience and credentials) are not the variables that determine teacher effectiveness. In either case, we believe that the clustering phenomenon, if true of the schools in our samples, does not undermine our basic conclusions.

## Limitations

The schools included in our sample were not selected to be representative of either the USA or Australia, drawn as they are from Colorado and the Sydney metropolitan area respectively. So generalization to either country may be unsound. Having said that, it may be of interest that in New South Wales, Australia, where Sydney is situated, there is a state-wide curriculum for early literacy that mandates that 35% of the school week be devoted to literacy and includes recommendations for milestones to be achieved and the content of instruction. In Colorado, there is no analogous set of guidelines, with teachers apparently freer to determine when, what, and how they teach. Despite this contrast, the correlations that we report are similar across country, and so it appears that a more centralized curriculum, which carries the potential for less variability in instruction, has not resulted in less of a classroom effect in Australia. This affords some reassurance that generalization across the dimension of curriculum centralization is possible.

Having said that, we again emphasize that other aspects of the educational landscape may affect estimates of classroom effects. Thus, although Colorado and New South Wales organize early literacy instruction somewhat differently, there may be other features of the two regions that result in relative uniformity. For instance, in both countries there have been widely-publicized reviews of the teaching of literacy, both of which have recommended the inclusion of “phonics” as an important part of instruction (National Inquiry into the Teaching of Literacy, 2005; National Reading Panel, 2000). This, in turn, may have affected the content of teacher training in ways that bring instruction into conformity in both countries. If in other nations there have

been no analogous developments, teachers may vary more in what as well as how they teach. As an in-principle proof of the influence of educational context on early literacy development, we have shown that the heritability of reading and spelling in kindergarten is lower in Scandinavia than it is in Australia, with the US as an intermediate case (Samuelsson et al., 2008). We attributed the contrast to the less constrained literacy instruction that occurs in Norway and Sweden, with teachers free to offer only minimal instruction. In general, contextual factors may influence estimates of classroom effects.

The stipulation that the children had native proficiency in English is also a limitation in this study. Not only may this have distorted the structure of the sample relative to population characteristics, any classroom effects that apply uniquely to ESL classrooms or children are outside the range of our conclusions. We do note, however, that the descriptive data in Table 1 indicates that the US sample at least was a representative one judging by the sample means and standard deviations for those tests that have national norms.

Our conclusions are also limited to the literacy tests that we administered. We have no measures of writing, none of motivation to read, and our test of comprehension is known to be more dependent on word-level decoding skills than some other comprehension tests (Keenan, Betjemann, & Olson, 2008). The relative high correlations that exist among differing literacy measures reported in most studies that use multiple measures suggest that the picture would not change substantially had we elected to use different tests, but we are on less secure ground when it comes to aspects of literacy that are rarely measured or that present particular problems in assessment, such as the degree and quality of planning for writing that students exhibit (De La Paz & Graham, 2002).

We have not ventured into other domains, such as mathematics, and so our results are silent of the broader question of classroom effects. However, the general agreement between our results and those of the TEDS project cited earlier, which does cover a wider range of school subjects, includes older children, and is set in a different country (Kovas et al., 2007), suggests that ours are not outliers. Although in the analyses by Kovas et al. the shared environment effect of .10 for reading (.17 - .07 for same- and different-class twins, respectively) was not significant, the classroom effect that it represents was approximately of the same magnitude as ours.

The assignment of children in a twin pair to the same or different classrooms is not random, but depends on a combination of school facilities (one class per grade or more than one), school policy (some schools do mandate that parents adopt one practice or the other, or apply pressure to do so) and parental preferences. There may also be an interaction of factors, such as where a rural area has a small school with only one class per grade and where people of a predominant socio-economic level tend to reside. We do not know how these factors played out in decisions that the many parents in our samples made as they first sent their children to school, although the increasing proportion of pairs split into separate classes with each higher school grade (Table 2) indicates that overall pressures and/or preferences changed as the children developed. The failure to find greater differences prior to the start of school within pairs subsequently placed with separate teachers (see final paragraph of Results) suggests that the within-pair differences that did emerge at school did not pre-exist, but the same cautions that are called for in generalizing results from non-experimental methods could be in order. This study does not represent the equivalent of random assignment of twins to same and different teachers.

## Coda

Behavior-genetic research examines what is, not what could be. Although we did not find classroom effects as large as many expect, there is nothing in our data that is incompatible with the proposition that better instruction and better-prepared teaching staff will lead to higher

levels of literacy. But the data are incompatible with the strident claims sometimes seen from journalists and echoed by politicians that much of the blame for poor literacy levels in some young children can be largely attributed to defective teaching. Not only do such claims fly in the face of evidence for substantial genetic influence on literacy development, but they are contradicted by the relatively modest classroom effect sizes that we have reported. To the degree that classroom effects are not due to teachers, the case for basing high-stakes decisions about teacher accountability on them must remain less secure. Future research should find ways to identify the components of classroom-based variance.

## Acknowledgments

Funding was provided by the Australian Research Council (DP0663498 and DP0770805) and the National Institute for Child Health and Human Development (HD27802 and HD38526). We thank the Australian Twin Registry, our testers, and the twins, parents, and teachers involved.

## References

- Alvermann, DE.; Fitzgerald, J.; Simpson, M. Teaching and learning in reading. In: Alexander, P.; Winne, P., editors. *Handbook of Educational Psychology*. 2nd ed. Erlbaum; Mahwah, NJ: 2006.
- Australian Bureau of Statistics. A picture of the nation: the statistician's report on the 2006 census: Cultural diversity overview. [Retrieved March 8, 2009]. 2006 from [http://www.ausstats.abs.gov.au/ausstats/subscriber.nsf/0/C724250359785DC6CA25754C0013DC0A/\\$File/20700\\_cultural\\_overview.pdf](http://www.ausstats.abs.gov.au/ausstats/subscriber.nsf/0/C724250359785DC6CA25754C0013DC0A/$File/20700_cultural_overview.pdf)
- Barkley, RA.; Murphy, K. *Attention-deficit hyperactivity disorder: A clinical workbook*. 2nd ed. Guilford Press; New York: 1998.
- Blachman BA, Tangel DM, Ball EW, Black R, McGraw CK. Developing phonological awareness and word recognition skills: A two-year intervention with low-income, inner-city children. *Reading and Writing: An Interdisciplinary Journal* 1999;11:239–273.
- Bus AG, van IJzendoorn MH. Phonological awareness and early reading: A meta-analysis of experimental training studies. *Journal of Educational Psychology* 1999;91:403–414.
- Byrne B, Delaland C, Fielding-Barnsley R, Quain P, Samuelsson S, Høien T, et al. Longitudinal twin study of early reading development in three countries: Preliminary results. *Annals of Dyslexia* 2002;52:49–73.
- Byrne B, Fielding-Barnsley R. Evaluation of a program to teach phonemic awareness to young children: A 1-year follow-up. *Journal of Educational Psychology* 1993a;85:104–111.
- Byrne B, Olson RK, Hulslander J, Samuelsson, Wadsworth S, DeFries JC, et al. A behavior-genetic analysis of orthographic learning, spelling, and decoding. *Journal of Research in Reading* 2008;31:8–21.
- Byrne, B.; Olson, RK.; Keenan, JM.; Coventry, WL.; Byrne, B. Reading comprehension at Grade 4: Data from a twin study. Behavior-genetic perspectives on reading comprehension; Symposium conducted at the Seventeenth Annual Pacific Coast Research Conference; Coronado, CA, USA. 2009.
- Byrne B, Olson RK, Samuelsson S, Wadsworth S, Corley R, DeFries JC, et al. Genetic and environmental influences on early literacy. *Journal of Research in Reading* 2006;29:33–49.
- Byrne B, Samuelsson S, Wadsworth S, Hulslander J, Corley R, DeFries JC, et al. Longitudinal twin study of early literacy development: Preschool through Grade 1. *Reading and Writing: An Interdisciplinary Journal* 2007;20:77–102.
- Byrne B, Wadsworth S, Corley R, Samuelsson S, Quain P, DeFries JC, et al. Longitudinal twin study of early literacy development: Preschool and kindergarten phases. *Scientific Studies of Reading* 2005;9:219–235.
- Connor CM, Morrison FJ, Fishman BJ, Schatschneider C, Underwood P. Algorithm-guided individualized reading instruction. *Science* 2007;315:464–465. [PubMed: 17255498]
- Connor CM, Son S-H, Hindman AH, Morrison FJ. Teacher qualifications, classroom practices, family characteristics, and preschool experience: Complex effects on first graders' vocabulary and early reading outcomes. *Journal of School Psychology* 2005;43:343–375.

- Darling-Hammond, L. Teacher quality and student achievement: A review of state policy evidence. *Educational Policy Analysis Archives*. [Retrieved 2 April, 2007]. 2000 from <http://epaa.asu.edu/epaa/v8n1/>
- Darling-Hammond L, Youngs P. Defining “highly qualified teachers”: What does “scientifically-based research” actually tell us? *Educational Researcher* 2002;31:13–25.
- De La Paz S, Graham S. Explicitly teaching strategies, skills, and knowledge: Writing instruction in middle school classrooms. *Journal of Educational Psychology* 2002;94:687–698.
- Foorman BR, York M, Santi KL, Francis D. Contextual effects on predicting risk for reading difficulties in first and second grade. *Reading and Writing: An Interdisciplinary Journal* 2008;21:371–394.
- Fuchs, D.; Saenz, L.; McMaster, K.; Yen, L.; Taylor, K.; Lemons, C., et al. Scaling-up peer-assisted learning strategies: A multi-site randomized controlled study; Paper presented at the 16th Annual Pacific Coast Research Conference; San Diego, CA. Feb. 2008
- Hattie JA. Towards a model of schooling: A synthesis of meta-analyses. *Australian Journal of Education* 1992;36:5–13.
- Hindson BA, Byrne B, Fielding-Barnsley R, Newman C, Hine D, Shankweiler D. Assessment and early instruction of preschool children at risk for reading disability. *Journal of Educational Psychology* 2005;94:687–704.
- Jastak, S.; Wilkinson, GS. *The Wide Range Achievement Test-Revised: Administration manual*. Jastak Associates; Wilmington, DE: 1984.
- Keenan JM, Betjemann RS, Olson RK. Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading* 2008;12:281–300.
- Keenan JM, Betjemann RS, Wadsworth SJ, DeFries JC, Olson RK. Genetic and environmental influences on reading and listening comprehension. *Journal of Research in Reading* 2006;29:79–91.
- Kovas Y, Haworth CMA, Dale PS, Plomin R. The genetic and environmental origins of learning abilities and disabilities in the early school years. *Monographs of the Society for Research in Child Development* 2007;72:1–144. whole number. [PubMed: 17661895]
- Marsh HW, Martin AJ, Cheng JHS. A multi-level perspective on gender in classroom motivation and climate: Potential benefits of male teachers for boys? *Journal of Educational Psychology* 2008;100:78–95.
- McCaffrey, DF.; Lockwood, JR.; Koretz, DM.; Hamilton, LS. *Evaluating value-added models for teacher accountability*. Rand Corporation; Santa Monica, CA: 2003.
- McCoach DB, O’Connell AA, Reis SM, Levitt HA. Growing readers: A hierarchical linear model of children’s reading growth during the first 2 years of school. *Journal of Educational Psychology* 2006;98:14–28.
- National Inquiry into the Teaching of Literacy. *Teaching reading: Report and recommendations*. Australian Government Department of Education, Science and Training; Canberra, Australia: 2005.
- National Reading Panel. *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. National Institute for Child Health and Human Development; Washington, DC: 2000.
- Neale, MC.; Maes, HHM. *Methodology for genetic studies of twins and families*. Kluwer; Dordrecht, The Netherlands: 2003.
- Nichols RC, Bilbro WC. The diagnosis of twin zygosity. *Acta Genetica* 1966;16:265–275.
- Nye B, Hedges LV, Konstantopoulos S. The effects of small classes on achievement: The results of the Tennessee class size experiment. *American Educational Research Journal* 2000;37:123–151.
- Nye, B.; Konstantopoulos, S.; Hedges, LV. How large are teacher effects?; *Educational Evaluation and Policy Analysis*. [Retrieved 27 June, 2008]. 2004 p. 237-257. from <http://www.jstor.org/stable/2699577>
- Papaioannou A, Marsh HW, Theodorakis Y. A multi-level approach to motivational climate in physical education and sport settings: An individual or a group level construct? *Journal of Sport & Exercise Psychology* 2004;26:90–118.
- Petrill SA, Deater-Deckard K, Schatschneider C, Davis C. Measured environmental influences on early reading: Evidence from an adoption study. *Scientific Studies of Reading* 2005;9:237–259.

- Petrill SA, Deater-Deckard K, Thompson LA, DeThorne LS, Schatschneider C. Genetic and environmental effects of serial naming and phonological awareness on early reading outcomes. *Journal of Educational Psychology* 2006;98:112–121. [PubMed: 19444324]
- Petrill SA, Deater-Deckard K, Thompson LA, Schatschneider C, DeThorne LS, Vandenberg DJ. Longitudinal genetic analyses of early reading: The Western Reserve Reading Project. *Reading and Writing: An Interdisciplinary Journal* 2007;20:127–146.
- Plomin, R.; DeFries, J.C.; McClearn, G.E.; McGuffin, P. Behavioral genetics. 5th ed. Worth; New York: 2008.
- Pressley M, Wharton-McDonald R, Allington R, Block CC, Morrow L, Tracey D, et al. *Scientific Studies of Reading* 2001;5:35–58.
- Rowan B, Correnti R, Miller RJ. What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record* 2002;104:1525–1567.
- Rutter M, Maughan B. School effectiveness findings 1979–2002. *Journal of School Psychology* 2002;40:451–475.
- Ryan AM. Peer groups as a context for the socialization of adolescents' motivation, engagement and achievement in school. *Educational Psychologist* 2000;35:101–111.
- Samuelsson S, Byrne B, Olson RK, Hulslander J, Wadsworth S, Corley R, et al. Response to early literacy instruction in the United States, Australia, and Scandinavia: A behavioral-genetic analysis. *Learning and Individual Differences* 2008;18:289–295. [PubMed: 19122888]
- Samuelsson S, Byrne B, Quain P, Wadsworth S, Corley R, DeFries JC, et al. Environmental and genetic influences on prereading skills in Australia, Scandinavia, and the United States. *Journal of Educational Psychology* 2005;97:705–722.
- Samuelsson S, Byrne B, Wadsworth S, Corley R, DeFries JC, Willcutt E, et al. Genetic and environmental influences on prereading skills and early reading and spelling development in the United States, Australia, and Scandinavia. *Reading and Writing: An Interdisciplinary Journal* 2007;20:51–75.
- Stromswold K. Why aren't identical twins linguistically identical? Genetic, prenatal and postnatal factors. *Cognition* 2006;101:333–384. [PubMed: 16797523]
- Taylor BM, Pearson PD, Clark K, Walpole S. Effective school and accomplished teachers: Lessons about primary-grade reading instruction in low-income schools. *The Elementary School Journal* 2000;101:121–165.
- Taylor BM, Pearson PD, Peterson DS, Rodriguez MC. Reading growth in high-poverty classrooms: The influence of teacher practices that encourage cognitive engagement in literacy learning. *The Elementary School Journal* 2003;104:3–28.
- Torgesen, J.; Wagner, R.; Rashotte, CA. *A Test of Word Reading Efficiency (TOWRE)*. PRO-ED; Austin, Texas: 1999.
- Wayne AJ, Youngs P. Teacher characteristics and student achievement gains: A review. *Review of Educational Research* 2003;73:89–122.
- Willcutt E, Betjemann R, Wadsworth S, Samuelsson S, Corley R, DeFries JC, et al. Preschool twin study of the relation between attention-deficit/hyperactivity disorder and prereading skills. *Reading and Writing: An Interdisciplinary Journal* 2007;20:103–125.
- Woodcock, RW. *Woodcock Reading Mastery Tests*. American Guidance Service; Circle Pines, MN: 1989.
- Wright S, Horn S, Sanders W. Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education* 1997;11:57–67.



**Table 1**

Means and standard deviations for normed tests for separate country samples

Grades	Test	US		Australia	
		Mean	SD	Mean	SD
Kindergarten	TOWRE SWE	97.0	10.5	104.0	10.3
	TOWRE PDE	102.1	8.4	107.3	8.8
Grade 1	TOWRE SWE	102.3	14.0	110.0	12.8
	TOWRE PDE	100.8	12.8	109.5	12.7
Grade 2	WRAT Spelling	100.2	15.4	113.1	15.5
	TOWRE PDE	100.8	12.8	109.5	12.7
	TOWRE SWE	102.5	14.4	110.7	13.2
	TOWRE PDE	100.0	13.4	110.3	14.4
	WRAT Spelling	96.8	16.0	107.0	5.9
	WPC	100.0	12.1	109.4	10.3

Note: TOWRE SWE = Test of Word Recognition Efficiency, Sight Word Efficiency Form A; TOWRE PDE = Test of Word Recognition Efficiency, Phonemic Decoding Efficiency Form A; WRAT = Wide Range Achievement Test; WPC = Woodcock Reading Mastery Test (Revised), Passage Comprehension.

**Table 2**  
 Intraclass Correlations for Same- and Different-Class Twins Pairs (N Pairs in Parentheses; Contrasts Significant at .01 in Bold)

Grade	Test	MZ pairs						DZ pairs					
		US		Australia		Total		US		Australia		Total	
Kindergarten	TOWRE SWE	.89 (171, 53)	.84	.85	.82	.88	.83	.64	.51	.51	.41	.61	.50
	TOWRE PDE	.76	.61	.73	.54	<b>.75</b>	<b>.59</b>	.55	.38	.38	.34	.51	.52
	Spelling	.78	.75	.83	.73	.80	.75	.59	.46	.46	.55	.56	.55
	Total literacy	.88	.79	<b>.90</b>	<b>.76</b>	<b>.89</b>	<b>.78</b>	.59	.57	.54	.42	.58	.54
Grade 1	TOWRE SWE	.88	.86	.84	.76	.86	.83	.49	.36	.36	.32	.46	.45
	TOWRE PDE	.83	.73	.83	.80	.81	.76	.34	.44	.44	.44	.41	.41
	WRAT Spelling	.82	.73	<b>.84</b>	<b>.60</b>	<b>.82</b>	<b>.69</b>	.48	.23	.23	.33	.41	.39
	WPC	.74	.82	.82	.66	.77	.77	.58	.40	.51	.29	.56	.37
Grade 2	Total literacy	.88	.86	<b>.90</b>	<b>.74</b>	.89	.82	.55	.46	.36	.34	.50	.44
	TOWRE SWE	.88	.79	.79	.78	.84	.79	.55	.41	.46	.24	.52	.38
	TOWRE PDE	.85	.75	.80	.83	.83	.78	.50	.46	.35	.20	.45	.42
	WRAT Spelling	.84	.77	.78	.77	.81	.76	.54	.35	.35	-.05	.48	.29
Total literacy	WPC	.72	.70	.76	.66	.73	.68	.56	.38	.52	.16	.54	.33
	Total literacy	.90	.84	.84	.76	.88	.81	.60	.47	.40	.24	.55	.41

Note: S = Same class, D = Different class; TOWRE SWE = Test of Word Recognition Efficiency, Sight Word Efficiency; TOWRE PDE = Test of Word Recognition Efficiency, Phonemic Decoding Efficiency; WRAT Spelling = Wide Range Achievement Test, Spelling; WPC = Woodcock Reading Mastery Test (Revised), Passage Comprehension. Total literacy in each grade is based on mean of all tests at that grade for each child.

**Table 3**  
 Intraclass Correlations for Change Scores for Same- and Different-Class Twins Pairs (N Pairs in Parentheses; Contrasts Significant at .01 in Bold)

Grades	Test	MZ pairs		DZ pairs	
		S	D	S	D
K - 1	Spelling <sup>d</sup>	<b>.70</b>	<b>.54</b>	.45	.50
	TOWRE SWE	.72	.69	.46	.51
	TOWRE PDE	.65	.64	.45	.40
1 - 2	TOWRE SWE	.53	.57	.39	.29
	TOWRE PDE	.46	.43	.29	.30
	WRAT Spelling	.38	.45	.27	.20
	WPC	.38	.31	.32	.07

Note: S = Same class, D = Different class, TOWRE SWE = Test of Word Recognition Efficiency, Sight Word Efficiency; TOWRE PDE = Test of Word Recognition Efficiency, Phonemic Decoding Efficiency; WRAT Spelling = Wide Range Achievement Test, Spelling; WPC = Woodcock Reading Mastery Test (Revised), Passage Comprehension.

<sup>d</sup>In kindergarten the children were assessed with our own spelling test and in Grade 1 with the WRAT Spelling.