American Psychological Association

Journal of Educational Psychology

© 2019 American Psychological Association
ISSN: 0022-0663

http://dx.doi.org/10.1037/edu0000418

Estimating Classroom-Level Influences on Literacy and Numeracy: A Twin Study

Katrina L. Grasby
QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia

Callie W. Little, Brian Byrne, and William L. Coventry
University of New England

Richard K. Olson
University of Colorado Boulder

Sally Larsen
University of New England

Stefan Samuelsson
Linköping University

Classroom-level influences on literacy skills in kindergarten through Grade 2, and on literacy and numeracy skills in Grades 3, 5, 7, and 9, were examined by comparing the similarity of twins who shared or did not share classrooms with each other. We analyzed two samples using structural equation modeling adapted for twin data. The first, Study 1, was of Australia-wide tests of literacy and numeracy, with 1,098; 1,080; 790, and 812 complete twin pairs contributing data for Grades 3, 5, 7, and 9, respectively. The second, Study 2, was of literacy tests from 753 twin pairs from kindergarten through Grade 2, which included a sample of United States and Australian students and was a reanalysis and extension of Byrne et al. (2010). Classroom effects were mostly nonsignificant; they accounted for only 2–3% of variance in achievement when averaged over tests and grades. Although the averaged effects may represent a lower-bound figure for classroom effects, and the design cannot detect classroom influences limited to individual students, the results are at odds with claims in public discourse of substantial classroom-level influences, which are mostly portrayed as teacher effects.

Educational Impact and Implications Statement

Our paper addresses the importance of classroom-level processes, which include teacher practices, in influencing between-student variation in achievement in literacy and numeracy from kindergarten to Grade 9. We took advantage of large twin samples in Australia and the United States to compare the similarity of twins who either shared or did not share a classroom with their cotwin. We found that twins in separate classrooms were almost as similar in achievement as those who were placed together. Our best estimate is that just 2% to 3% of total student variability in literacy and numeracy is attributable to classroom-level processes (although our design is not sensitive to individual teacher-student interactions that may help or hinder a particular student independently of the whole class). Our estimate falls well short of commonly accepted values of up to 30%, and suggests that educational policies that are based on those high estimates, such as teacher hiring and advancement depending on class progress, are misplaced.

Keywords: classroom effects, teacher effects, twins, literacy, numeracy

Supplemental materials: http://dx.doi.org/10.1037/edu0000418.supp

Katrina L. Grasby, Psychiatric Genetics, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia; Callie W. Little, Brian Byrne, and William L. Coventry, Department of Psychology, University of New England; Richard K. Olson, Department of Psychology, University of Colorado Boulder; Sally Larsen, Department of Psychology, University of New England; Stefan Samuelsson, Department of Behavioural Sciences and Learning, Linköping University.

This research was supported by an Australian Research Council Grant (DP150102441). The Australian Twin Registry is supported by an enabling Grant (628911) from the National Health and Medical Research Council. The authors declare that they have no conflict of interest. Informed consent was obtained from all individual participants included in the study. All procedures were performed in accordance with the ethical standards of the University of New England (HE12-150). We thank the Australian Twin Registry, and all of the twins, triplets, and parents involved.

Correspondence concerning this article should be addressed to Callie W. Little, Department of Psychology, University of New England, S07 Library Road, Armidale, NSW 2351, Australia. E-mail: calliew01@gmail.com

Most children first learn to read from school instruction beginning in kindergarten or first grade, but they vary in how well and how quickly they learn across the first and subsequent years of formal reading instruction. Similarly, differences in numeracy emerge early in schooling and continue throughout. The reasons for individual differences in reading and numeracy development have often been attributed to environmental factors. This environmental focus has been reflected in United States federal and state laws that assume variation in teacher quality is a major reason for individual differences in student success, and that teachers should be held accountable when their students fall behind (Every Student Succeeds Act; https://www.ed.gov/essa). Despite this high-stakes focus on teacher quality, the empirical evidence for classroom influences (including teacher effectiveness) is inconclusive.

Value-added estimates of teacher effectiveness are increasingly used, but can be biased in how they attempt to account for student background characteristics, leading to uncertainty of the estimates (Chetty, Friedman, & Rockoff, 2014). Value-added methods assume classroom influences either increase children's expected achievement (add value), have no influence, or subtract from expected achievement. Previously, Rowan, Correnti, and Miller (2002) estimated that classroom environment accounts for between 4 and 16% of variance in reading scores, and in a highly cited paper Nye, Konstantopoulos, and Hedges (2004) summarized several value-added studies that examined classroom influences, and found a range of estimates (3–16%) for literacy skills. Classroom influences on numeracy outcomes have also been summarized, with findings in a similar range to that for literacy (6–13%; Rowan et al., 2002). However, in a widely endorsed Australian summary of teacher effectiveness, Hattie (2003) claims that teacher quality accounts for 30% of variance in school performance. The range of estimates from nongenetically sensitive studies of classroom influences on literacy and numeracy skills suggests that classroom influences could provide anywhere between small to moderate effects on student achievement.

In contrast, research with large samples of identical and fraternal twins selected from many different homes, schools, and classrooms across several countries with universal education have suggested that on average children's genetic differences are the main reason for individual differences in reading ability and numeracy (Kovas et al., 2013; Olson, Keenan, Byrne & Samuelsson, 2014). Taylor, Roehrig, Soden Hensler, Connor, and Schatschneider (2010) provide a median estimate of $\approx$ 65 as the percent variance accounted for in children's reading by genes. The substantial size of the estimates of genetic influence suggests that, on average, differences in teacher quality are not the primary reason for individual differences in student performance. Still, the mean estimate of 65% of the variance in children's reading development due to genes leaves 35% of the variance due to the environment, which could potentially include strong classroom effects.

Research employing school-age twins can do more than quantify overall genetic and environmental influences on academic performance; it can help identify individual sources of variation. By comparing the similarity of twins in pairs taught in the same class versus twins taught in different classes, the influence of classroom on performance can be estimated. Using this method, Byrne et al. (2010) published estimates of so-called teacher effects on literacy development from kindergarten to Grade 2 in the United States and Australia. In the current paper we resume and extend the examination of these effects using two separate samples. In Study 1 we examined classroom influences on Australian twins who have completed nation-wide tests in aspects of literacy, language and numeracy in Grades 3, 5, 7, and 9. The tests are known as the National Assessment Plan—Literacy and Numeracy (NAPLAN). In Study 2, we reexamine classroom influences on twins from the International Longitudinal Study of Twins (ILTS) which included twins from the United States and Australia who completed several literacy assessments in kindergarten through Grade 2 (Byrne et al., 2010). The reanalysis of the ILTS data was included since it uses a larger dataset than that of Byrne et al. (2010) by including the final waves of data collected. Further, rather than an analysis that compared the correlations of same and different classes, here we use structural equation models, which still do this but also (a) estimate a single source of classroom variance, not separate estimates for monozygotic (MZ) twins and dizygotic (DZ) twins; (b) use maximum likelihood estimation for more precise estimates; (c) concurrently estimate sources of genetic and shared environmental variance; and (d) ensure the estimates of genetic variance are constant across the same and different classes, as is expected given classroom effects represent, in theory at least, a trade-off between shared and unshared sources of environmental variation. Critically, while we do not anticipate different results in the reanalysis, reporting them here fills what would otherwise be a missing step for any contrast of the results of Byrne et al. (2010) and this NAPLAN sample.

Before proceeding, we note that we will follow the practice of Byrne et al. (2010) and use the term *classroom effects* in preference to the commonly used teacher effects. This is because our data do not allow us to separate the teacher from other classroom-level processes when analyzing the contribution of classroom assignment to variation in student achievement. Researchers in education often refer to those additional processes as classroom climate (Fraser, 2012), factors that can be independent of particular teachers (Marsh, Martin, & Cheng, 2008). Elements include, for example, peers (Rutter & Maughan, 2002) and the class's attitudes to the value of learning (Papaioannou, Marsh, & Theodorakis, 2004).

Previous Results and Their Limitations

Byrne et al. (2010) assessed word reading, nonword reading, and spelling in kindergarten, and those plus reading comprehension in Grades 1 and 2, for a total of 11 measures. They then compared the correlations between twin children (maximum N = 711 pairs, half identical [MZ], half fraternal [DZ]) where both members of a pair were in the same classroom with those where the members of a pair were in different classrooms. A classroom effect would show up as lower correlations in the case of different-class twins than same-class twins. The size of the difference in correlation coefficients can be interpreted as the percentage of variance attributable to classroom factors (see Byrne et al., 2010 for details of this calculation). Averaged over the 44 comparisons (two countries, 11 literacy measures and two twin types), the mean difference in the size of the correlations was .08, with same-class twins being more highly correlated than different-class twins. Although the differences in correlations between same- and different-class pairs were significant in only a minority of the 44 comparisons, the same-class twin correlations were numerically higher than the different-class twin correlations in 40 of the 44. The pattern of results was very similar in the two countries.

Correlations for MZ twins over all 11 measures averaged over the two countries ranged from .73 to .89 for those in the same classes, and from .59 to .83 for those in different classes; the analogous figures for DZ twins were .41 to .61 and .29 to .55. While not central to this paper, the higher values for MZ twins than for DZ twins is evidence for the substantial heritability for early literacy in these countries; see Olson et al. (2014) for a summary of evidence from the twin study from which the classroom effect data were derived and from other studies using the classic twin design.

Byrne et al. (2010) interpreted their findings as indicating that around 8% of the variance in early literacy skills is accounted for by classroom-level factors, which will include any differential influence of teachers as well as any factors contributing to classroom climate that affect mean classroom achievement and that are independent of teacher. Even if the teacher effect comprised the entire 8% variance, itself unlikely, this estimate is considerably lower than ones proposed by other researchers (e.g., Hattie, 2003; Nye et al., 2004).

Byrne et al. (2010) qualified their conclusions in several ways. One was the fact that the samples were drawn from restricted regions, the "front range" area of Colorado around Denver and Boulder in the case of the United States and the Sydney metropolitan area in the case of Australia. A second limitation was that the tests were restricted to literacy; other academic domains such as writing, numeracy, and the physical and social sciences may show different patterns of results. Third, the grade and age range of the twins was limited, from kindergarten to Grade 2 and from approximately 5 to 8 years; classroom effects may change in higher school grades and with increasing age. In the current study, by using data from the NAPLAN Twin Study we are able to address some of the limitations of Byrne et al. (2010). The NAPLAN tests are given nationwide in Australia rather than in restricted regions. They include additional linguistic skills (spelling, writing and aspects of grammar and punctuation), and cover numeracy as well (details are given in the Method section). The twins are spread across a broader grade range, 3–9, and are older, approximately 7 to 15 years. By comparing estimates of classroom influences in the ILTS and NAPLAN samples, we are able to not only address some limitations of the former study, but also provide evidence as to whether these limitations are or are not associated with any potential differences in level of estimated effects.

The Current Study: Novel Aspects and Some Expectations

As a preliminary comment, and as mentioned earlier, the heritability and the extent of environmental influence on each domain is not of central interest in this article, but it is worth mentioning that genetic factors accounted for between half and three quarters of the variance across most domains and grades. The "shared environment," the potential environmental influences that twins in a family would share, such as socioeconomic status, parental attitudes to the value of literacy and numeracy, school attended, and sometimes the same teacher, accounted for a small and often nonsignificant portion of the remaining variance. Environmental influences potentially unique to individual members of twin pairs, such as illnesses, peers, measurement error, and sometimes different teachers, exerted more substantial influence, at around a fifth to a half of the total variance, always significant (Grasby, Coventry, Byrne, Olson, & Medland, 2016). These results broadly fit the pattern found in other twin studies of academic achievement (e.g., Kovas, Haworth, Dale, & Plomin, 2007), offering some assurance about the representativeness of this sample.

Range of Academic Domains

Although reading and numeracy have been the subjects of research into classroom-level influences, we know of no studies that specifically focus on classroom effects on writing and grammar and punctuation, so this is a novel aspect of the current study. It is not clear what to expect with these academic domains, though it may be of interest that writing proved to be the least heritable of the five NAPLAN test domains, with around 50% of the variance accounted for by genes (Grasby et al., 2016). In turn, writing had the highest influence from factors unique to each twin, also close to 50% of total variance. This appears to give more scope for influences from individual experiences in school, such as the degree to which a teacher may encourage writing and the instructional style that the teacher brings to writing tasks.

Elementary and High School Differences

Another feature of the current study is that results are available for both primary (elementary in some educational jurisdictions) school and high school. The level of academic content is, obviously, more advanced in high school, and consequently the amount of training that teachers have had might begin to matter more. In turn, if this training is variable we might expect more teacher influence, with some better prepared than others to lead students through the curriculum. In Australia, there are recognized shortages of teachers in some subjects, particularly in mathematics and the physical sciences. According to the Productivity Commission (2012) *Schools Workforce* report of the Australian government, three quarters of high school principals report difficulty in recruiting qualified mathematics teachers. The 2009 Program for International Student Assessment from the Organisation for Economic Co-operation and Development (OECD) reports that around 30% of 15-year-old Australian students were enrolled in schools whose leaders reported that a lack of qualified mathematics teachers was hindering instruction. This compares to 18% for other OECD countries. International comparisons aside, these figures suggest that twins in separate high school classes may be in the hands of teachers with considerably different levels of qualifications and therefore may show more of a classroom effect that in primary school and compared to other high school academic domains where qualified teacher shortages are not so marked.

Apart from subject-specific factors like more variation in teacher qualifications in high school, the transition from primary to high school can be challenging to children in ways that might emerge as higher levels of classroom effects. Children are moving in many cases from a small school to a larger one, and from a stable set of teachers and peers to a more varied one. There are more choices of school subjects, and children are subject to variation in how connected they feel to the new school environment. In these challenging circumstances there is evidence that peer

(Ganeson & Ehrich, 2009) and teacher support (Bru, Stornes, Munthe, & Thuen, 2010) matter in how well a child settles into the school, with flow-on effects on academic performance (for a summary, see Hanewald, 2013). A full review of these processes is beyond the scope of this article, but given that teachers and peers can matter for the transition, it stands to reason that twins in the same class will be subject to more similar degrees of influence from these sources, positive or negative, than twins in separate classes. So it stands to reason, too, that classroom effects, measured by the correlation between twins in same versus different classroom contexts, may be more marked in high school than primary, and perhaps most marked in the first high school year (Grade 7 in our sample), decreasing once the transition phase is complete.

Academic Streaming

One issue that requires attention in the kind of study we conduct here is academic streaming. It is the practice in some schools to assign students to different classes based on their emerging academic skills. According to Clarke (2014):

> in Australia there is no informed, explicit and coherent policy approach to ability grouping. There is in fact a federal and state government policy silence in relation to the issue. That has not stopped systems, schools and teachers from grouping students according to their perceived ability. (p. 1)

Johnston and Wildy (2016) also document the use of streaming in Australia and its likely effects.

The reason why this matters for the design we adopt is that classroom separation could be confounded with preexisting academic performance differences, making it less clear whether any classroom effect is genuine (attributable to factors operating after classrooms are allocated) or spurious (attributable to streaming, operating before class allocation). We also assume that, given the documented genetic influence on school performance, it is more likely that dizygotic twins will exhibit achievement differences and therefore be more likely to be subject to streaming. But monozygotic twins whose academic skills have been differentially influenced by unique environment factors could also be streamed (Larsen et al., 2019). With these considerations in mind, we made an attempt to identify reasons why twins were kept together or separated into different classrooms (see the Method section) with a view to accounting for possible academic streaming in our analyses.

Reanalysis of Previous Data

In this paper we also take the opportunity to reanalyze the data from Byrne et al. (2010). We did this because we have adopted a more sophisticated method of analysis for the currently available data (see the Method section), and we wished to build an integrated picture of classroom effects spanning kindergarten to Grade 9 using a uniform analytic procedure. By including both samples, we are able to compare the magnitude of classroom influences across a range of different measures of literacy skills and to explore any potential differences in classroom influences from literacy to numeracy skills. Furthermore, with the inclusion of the ILTS we can compare the magnitude of potential classroom influences in early primary school with those in later primary school and into secondary school. Similar results across the two samples will provide more robust evidence for the size and extent of classroom-level influences on achievement, whereas divergent results could pinpoint potential moderators such as achievement domain, age, or nationality.

Tentative Hypotheses

We propose some tentative hypotheses about classroom effects on achievement across the subject domains and grades we had available. First, we hypothesize that the magnitude of classroom influences will be similar across the ILTS and NAPLAN studies. Next, we hypothesize larger effects in high school as students negotiate a more challenging educational environment where peers and teachers can be important to well-being, perhaps during the first year in particular; additionally we hypothesize larger effects in numeracy in high school in the face of a teacher shortage and consequent variable qualifications in the teaching workforce; we also hypothesize that writing might be a subject that leaves more room for classroom effects, given that it shows the highest unique environment influence of the suite of test subjects.

Method

Participants

Study 1. Participants were a part of the Australian Twin-Study of the NAPLAN (Grasby & Coventry, 2016). Those included in this study were twins with NAPLAN results who had attended the same school and sat the tests in the same calendar year as their cotwin, and were concordant in their class allocation in the test year and previous year. Of the initial 5,136 twins with NAPLAN results, 75% met these criteria and data were available for 3,850 twins, 51.4% female. There were 2,196 twins with Grade 3 results (1,098 pairs); 2,160 with Grade 5 results (1,080 pairs); 1,580 with Grade 7 results (790 pairs); and 1,624 with Grade 9 results (812 pairs). Number of twins and pairs by subject and grade and zygosity are detailed in Table 1.

For sensitivity analyses we further restricted the sample to exclude participants allocated to classrooms due to academic ability. For these analyses the sample was reduced both because it required a parent report on the reason for class allocation and that this reason was not attributed to academic ability for the current or the previous year. Parent report on the reason for class allocation was available on only 78%, 74%, 60%, and 54% of participants in Grades 3, 5, 7, and 9, respectively. Of those with a report, 6% were removed for academic reasons in Grade 3, 7% in Grade 5, 20% in Grade 7, 32% for literacy in Grade 9, and 40% for numeracy in Grade 9.

Since 2008, NAPLAN has been administered to students in Grades 3, 5, 7, and 9; therefore, the longitudinal data available on participants is staggered, such that not all students in latter grades sat tests in earlier grades and not all students in earlier grades have yet sat tests in later grades. Due to the staggered nature of the data, across all grades data was only available on 124 pairs for literacy and 116 pairs for numeracy, so for this study we have implemented cross-sectional analyses.

Table 1
Descriptive Statistics by Class Allocation for Each Grade and Domain in Study 1

NAPLAN grade	Reading		Spelling		Grammar and punctuation		Writing		Numeracy	
	Same	Different	Same	Different	Same	Different	Same	Different	Same	Different
Grade 3										
M (*SD*)	450.9 (85.2)	458.3 (85.7)	428.1 (76.3)	429.5 (76.6)	460.3 (93.4)	464.3 (90.2)	436.7 (55.8)	437.0 (58.8)	423.4 (73.1)	429.3 (71.2)
N	917	1269	915	1270	908	1267	911	1264	913	1261
MZ pairs	239	286	238	284	236	283	235	283	236	281
DZ pairs	218	346	219	349	216	346	218	343	218	345
Grade 5										
M (*SD*)	531.4 (77.8)	530.0 (77.8)	514.9 (65.5)	509.1 (68.3)	534.5 (80.4)	531.7 (82.2)	504.5 (60.2)	500.6 (58.9)	515.1 (72.5)	516.0 (69.2)
N	726	1417	729	1422	725	1411	727	1408	724	1413
MZ pairs	187	306	188	307	188	305	188	306	186	304
DZ pairs	174	401	176	402	174	396	174	392	173	399
Grade 7										
M (*SD*)	580.1 (66.4)	576.6 (66.7)	564.0 (64.4)	560.3 (64.1)	574.1 (75.6)	569.8 (74.2)	548.5 (70.3)	545.7 (66.1)	579.6 (74.3)	576.6 (67.5)
N	487	1010	484	1013	488	1009	487	1004	472	998
MZ pairs	131	232	131	232	133	233	132	231	127	226
DZ pairs	112	270	109	272	109	267	109	265	107	271
Grade 9										
M (*SD*)	631.8 (66.0)	617.2 (63.3)	618.7 (67.1)	599.6 (64.7)	619.0 (74.3)	603.7 (73.9)	603.0 (78.0)	590.8 (76.9)	639.1 (73.1)	616.0 (64.6)
N	341	1140	339	1140	342	1134	341	1131	401	997
MZ pairs	91	255	90	251	90	250	90	250	116	209
DZ pairs	78	311	78	316	79	313	79	310	84	286

Note. *N* = number of individuals; MZ pairs = number of complete monozygotic pairs; DZ pairs = number of complete dizygotic pairs.

At the time of testing, average age was 8.6 (0.39) years in Grade 3, 10.6 (0.40) years in Grade 5, 12.6 (0.43) years in Grade 7, and 14.6 (0.44) years in Grade 9. Zygosity was determined by parent report of a DNA test, or with a short questionnaire (Lykken, Bouchard, McGue, & Tellegen, 1990), which classified a subsample of twins in this study with 95% accuracy (Grasby & Coventry, 2016).

The families in this study are reasonably representative of Australia, with relative participation by states and territories within 5% of the expected, excepting a higher rate of participation from Western Australia. Parent education scored on a 9-point scale, from 1 (*some high school but did not finish*) to 9 (*a doctoral degree*), was proportional to the Australian population (aged 25–54; Australian Bureau of Statistics, 2014), with average scores of 5.0 (*SD* = 1.8) for mothers and 4.6 (*SD* = 2.0) for fathers (where a score of 5 = a 3-year university degree). In a similar vein, parental occupation covered a broad range, with average International Socio-Economic Index of Occupational Status (ISEI) scores of 53.4 (*SD* = 13.97, range 10–89) for mothers and 54.8 (*SD* = 15.51, range 16–89) for fathers (ISEI calculated according to Ganzeboom & Treiman, 1996 using Ganzeboom & Treiman, 2010 for score conversion). Of those reporting ancestry, 95% reported their ancestry as Australian or of European descent. One caveat to the representativeness of the sample, as previously reported in Grasby, Coventry, Byrne, & Olson, 2019, is that the schools attended by participants in this study are, on average, more advantaged than the general population.

Study 2. Participants from the ILTS came from an ongoing longitudinal study of twins in the United States, Australia, and Scandinavia (although the Scandinavian sample was excluded from the present analyses because only a few of the twin pairs were in different classrooms). There was a maximum of 753 twin pairs (previously 711; Byrne et al., 2010) recruited in Colorado (489 pairs) and the Sydney metropolitan area (264). Participants were recruited in preschool and followed to Grade 2, ranging in age from 4.9 (.23) years in preschool, 6.4 (.41) in kindergarten, 7.4 (.42) years in Grade 1, and 8.4 (.44) years in Grade 2. 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) similarity questionnaire. Females represented 49.9% of the sample.

Materials

Study 1.

Achievement.

National Assessment for Literacy and Numeracy (NAPLAN). The NAPLAN is an Australia-wide assessment of students in Grades 3, 5, 7, and 9 on literacy and numeracy. There are five standardized tests: reading comprehension, writing, grammar and punctuation, spelling, and numeracy. Scores are scaled and range from 0–1,000 across all grades. Technical papers and example tests and writing prompts are available online (www.nap.edu.au).

Reading. Students read 7–8 passages extracted from books, newspaper articles, poems, or posters. Extracts vary in length from a brief single paragraph to several paragraphs. For each extract students answer several (~5–8) questions, most are multiple choice format. Tests are to be completed in 45–65 min (depending on the grade).

Spelling. Students are required to identify and correct misspelled words in sentences. Most questions are constructed-response format, with 25–30 questions (depending on grade and year). Although scored separately, this measure is administered with the Grammar and Punctuation measure as a part of a single Language Conventions test, for which students are allowed 40–45 min to complete (depending on year).

Grammar and punctuation. Students are required to identify or provide correct tense, pronouns, conjunctions, punctuation, and verb forms. In later grades relative pronounces, clauses, and comparative adjectives are also assessed. Questions are multiple choice or constructed-response format, with 25–28 questions.

Writing. Students are provided with an idea or topic and given 40 min to write a response passage in a specified style (e.g., narrative or persuasive). For example, "It is cruel to keep animals in cages. What do you think? Do you agree or disagree? Perhaps you can think of ideas for both sides of this topic." Writing is assessed against 10 criteria: audience, text structure, ideas, vocabulary, cohesions, paragraphing, sentence structure, punctuation, spelling, and character and setting (for narrative style) or persuasive devices (for persuasive style).

Numeracy. The numeracy test assesses students on various aspects of mathematics, including problem solving, reasoning, interpretation, measurement, geometry, computation, algebra, statistics, and probability. In Grades 3 and 5 students take a single numeracy test, while in Grades 7 and 9 students sit one numeracy test that allows calculator use and one that does not; yet one numeracy score is provided for Grades 7 and 9. In Grades 3 and 5 there were 35–42 questions to be completed in 45–50 min (depending on grade and year). In Grades 7 and 9 there were 48–66 questions to be completed in 60–80 min (depending on the year).

Parent questionnaire. Parents of the NAPLAN participants were administered a questionnaire about their children in which they were asked if the twins were in the same or different classes in each year of school. In high school, classes are frequently different by subject, so from Grade 7 to 9 parents were asked about class allocation for both English and mathematics. In order to identify academic ability as a reason for class allocation, in the first round of questionnaires, parents were asked to comment on the reason for the allocation of twins to same or different classes. Parent comments that were categorized as being due to academic ability included comments like: "Classes were graded so they weren't together"; "Graded classes"; "Twin 1 was in an advanced stream for academic students, Twin 2 was in a lower stream"; "First born in extension class in year 7 and 8, both in extension in year 9"; and "Initially they were in different classes based on academic performance. Twin 2 was promoted to the same class as Twin 1 in Year 9 based on academic performance." From the comments a list of reasons was created, which included "different academic ability." Subsequent questionnaires provided this list of reasons for parents to select from, with the comment field retained for alternate reasons to be provided if required. From these reasons, twins who were placed in the same or different classes due to academic ability were identified.

Study 2

Achievement.

Test of Word Reading Efficiency. 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 s. 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.

Spelling. In kindergarten, the spelling test consists of 10 real words (examples man, come, went) and four 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). 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 10 consecutive errors. Score is total number orthographically correct.

Parent questionnaire. ILTS parents provided information on class status, together or separate, for each pair in each year of school, kindergarten to Grade 2.

Procedure

Study 1. NAPLAN tests are administered over three consecutive days each year in the second full week of May (approximately 3.5 months into the school year). On the first day students sit the language conventions test (comprising the spelling and grammar and punctuation domains) and later in the same day they sit the writing test. On the second day students sit the reading test, and on third day they sit the numeracy tests. After consent to participate was provided by both parents and twins, the NAPLAN results were obtained from state departments of education.

The parent questionnaires were either posted or administered online via Qualtrics. Although NAPLAN tests have been administered each year since 2008, data collection for this project began in 2013. Thus, the first wave of data collection included twins who had sat NAPLAN tests in the years 2008–2013. From 2014 the parental questionnaires have been sent out in the calendar year in which the twins sit a NAPLAN test.

Study 2. 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 United States 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.

Analyses

For Study 1, we conducted a class-allocation model for each of the five NAPLAN tests in each of the four grades, and for Study 2, we conducted a class-allocation model for three literacy tests in kindergarten and four literacy tests in Grades 1 and 2. Prior to running the class-allocation model with NAPLAN data, univariate outliers (±3.5 *SD*) were removed; sex, age, age-squared, Age × Sex, and cohort effects were regressed out of the scaled scores; and twin pair outliers (Mahalanobis distance >13.82) were removed.

The ILTS data were prepared using the procedures described in Byrne et al. (2010); scores were age- and gender adjusted, truncated to ±3 *SD*, and standardized within country.

The class-allocation model is a variation on the classic twin model. The classic twin model partitions variance into additive (A), shared environmental (C), and unique environmental (E) influences. Given that MZ twins share all and DZ twins share (on average) half of their segregating genetic variants, the covariance of A within MZ pairs is fixed to 1 and within DZ pairs is fixed to 0.5. Regardless of zygosity, the covariance of C is fixed to 1 and the covariance of E is fixed to 0. This model is extended in the class-allocation model by estimating a fourth variance component, the class variance (CL). Each zygosity group is split into pairs that shared a classroom, where CL is fixed to 1, and pairs that were in different classes, where CL is fixed to 0. Structural equation models fitted to raw data and estimated using full information maximum likelihood in OpenMx (Boker et al., 2011) were used to decompose the variance into A, C, E, and CL using the equations (Figure 1):

$$\text{MZ same class covariance} = A + C + CL$$

$$\text{MZ different class covariance} = A + C$$

$$\text{DZ same class covariance} = \frac{1}{2}A + C + CL$$

$$\text{DZ different class covariance} = \frac{1}{2}A + C$$

$$\text{Variance} = A + C + CL + E$$

The NAPLAN tests are conducted just 3.5 months into the school year, too early perhaps to show full effects of any classroom status and therefore possibly underestimating effects of classroom separation when the twins had shared a classroom the previous year. Thus, our allocation of twin pairs to "same" and "different" class in Study 1 required concordance in class allocation across the test year and the previous year. This requirement was not necessary for the ILTS data because testing was conducted near the end of each school year or immediately following in summer.

An assumption of the classic twin model is the equal environments assumption (EEA), which requires that MZ and DZ twins be equally exposed to environments that affect covariation. In the scenario that academic streaming does influence class allocation, we would expect MZ twins to share classes more than DZ twins, given their greater genetic similarity and evidence that academic performance is heritable. In this scenario the EEA is an issue for a classic twin ACE model if class allocation also contributes to covariation, because the greater covariation of MZ twins compared to DZ twins would be due to both greater genetic similarity and shared classes. This bias is avoided in the class-allocation model, because it does not conflate class allocation covariation within zygosity. However, in the class-allocation model academic streaming would mean that academic similarity could contribute to greater covariation between twins in the same class compared to those in different classes, and this would inflate the estimate of classroom variance. To check for this, in Study 1, sensitivity analyses were run after excluding twins with academic ability as their reason for class allocation. We note that any residual streaming, due to parents perhaps not being aware that streaming was occurring and reporting an alternative reason, would continue to inflate the classroom estimates. Data from Study 2 did not include the reason for class allocation; therefore, sensitivity analyses could not be conducted.

Based on heritability estimates of 60% and shared environmental variance of 10% (which was the average across grade and domain in the initial analysis of these NAPLAN data in Grasby et al., 2016), and given the number of twin pairs by zygosity allocated to same and different classes, there was over 90% power to detect a classroom effect of 8% for Grades 3 and 5, 80% power for Grade 7, and 70% for Grade 9.

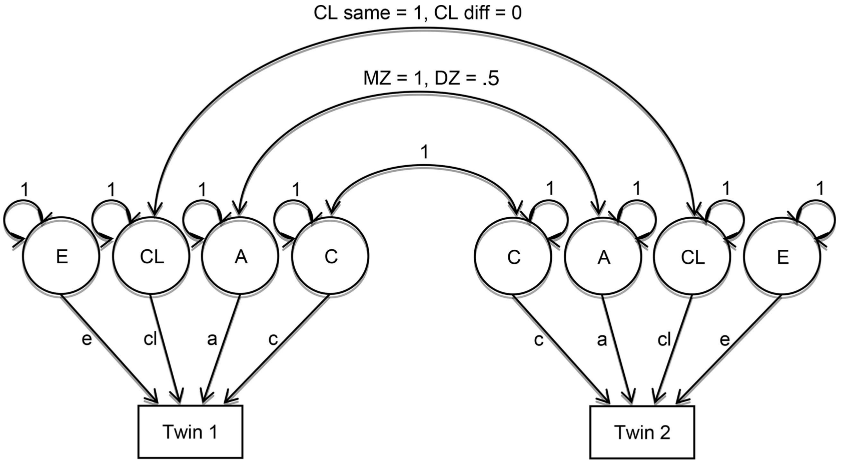

Figure 1. Path diagram of the classroom model. Variance due to classroom (CL), genes (A), shared environmental factors (C), and unique environmental factors (E) are estimated from monozygotic (MZ) and dizygotic (DZ) twins who either shared a class (CL same) or were in different classes (CL diff). The same class correlation is fixed to 1 for twins who shared a class and fixed to 0 for twins who were in different classes. The genetic correlation for MZ twins is fixed to 1 and for DZ twins is fixed to .5, while the shared environment correlation is fixed to 1 for all twins.

Results

Study 1

Within each grade and NAPLAN domain less than 1% of scores were removed for being outliers. Table 1 details means, standard deviations, and number of twins and number of complete pairs by zygosity in each grade for each domain for same and different class allocation.

Prior to running the class-allocation model, we tested if means and variances could be equated across zygosity and class allocation, and tested if class allocation was associated with sex or zygosity. Chi-square tests showed that females were more likely to be in the same class than males in Grade 5 and for mathematics in Grade 7. There was no difference in covariance between female and male twin pairs in any subject in Grade 5 or in Grade 7 numeracy. Males did have a greater variance than females in numeracy in Grades 5 and 7, but this greater variance for males did not translate into greater variance for twins in different classes, indicating that the class-allocation model would not be confounded with class-allocation by sex effects. Chi-square tests also showed that MZ twins were more likely to share a class than DZ twins (26–46% of MZ twins in the same class and 20–38% of DZ twins in the same class). Except for Grade 9 numeracy, this effect was no longer significant when removing participants where the reason reported for their class allocation was due to academic ability. Chi-square group sizes and test results are reported in the online supplemental materials (Table 1). Assumption testing of the class-allocation model found that means for class allocation could not always be equated in Grade 9 without a significant loss of fit to the model. This difference in means was not significant when the streamed participants were removed. However, for consistency across the models and greater precision in the variance decomposition, means were estimated separately for same and different classes in all models. Variances could be equated in all subjects and grades.

Estimates of classroom variance from the class-allocation model ranged from 0–7% and were not significantly different from zero except for spelling in Grade 3 (7%; Table 2; Supplemental Table 2, available online, details the variance components after dropping the class variance component). After removing participants with academic reasons for class allocation, the estimate of variance due to classroom for Grade 3 spelling dropped to 5% and was no longer significantly different from zero (Table 3). No other classroom estimate in Table 3 was significant either. Generally, the results with and without participants with academic reasons for class allocation were very similar, indicating that streaming of students into same or different classes due to their past academic performance was not inflating estimates of classroom influence in the full model. However, the variance due to classroom was less than the 8% estimated previously by Byrne et al. (2010), and the sample was underpowered. To detect a classroom influence of 5% we had only 50% power in Grades 3 and 5, 40% in Grade 7, and 30% in Grade 9.

Given the small estimates for classroom variance, we tested a more general classroom effect at each grade by allowing all five domains to load onto a latent NAPLAN performance variable and

Table 2
Study 1 Sample Size, Correlations, and Standardized Variance Components for NAPLAN Domains in Grades 3, 5, 7, and 9

	Number of complete pairs				Correlations				Class-allocation model variance components			
NAPLAN subscale by grade	MZs	MZd	DZs	DZd	MZs	MZd	DZs	DZd	A	C	E	CL
Grade 3												
Reading	239	286	218	346	.72	.76	.50	.34	.73 [.59, .79]	.03 [0, .15]	.24 [.21, .28]	0 [0, .06]
Spelling	238	284	219	349	.76	.76	.55	.24	.73 [.64, .78]	0 [0, .08]	.20 [.17, .24]	.07 [.01, .12]
Grammar and punctuation	236	283	216	346	.73	.71	.56	.35	.60 [.46, .74]	.11 [0, .23]	.25 [.21, .29]	.05 [0, .11]
Writing	235	283	218	343	.47	.47	.43	.32	.28 [.10, .47]	.18 [.03, .32]	.47 [.40, .56]	.06 [0, .15]
Numeracy	236	281	218	345	.74	.77	.54	.38	.64 [.51, .78]	.12 [0, .25]	.24 [.21, .27]	0 [0, .05]
Grade 5												
Reading	187	306	174	401	.71	.71	.47	.45	.54 [.40, .68]	.18 [.05, .30]	.28 [.23, .33]	.01 [0, .07]
Spelling	188	307	176	402	.79	.81	.58	.42	.73 [.61, .84]	.08 [0, .20]	.17 [.14, .21]	.01 [0, .05]
Grammar and punctuation	188	305	174	396	.73	.66	.48	.42	.52 [.38, .67]	.15 [.01, .27]	.27 [.22, .33]	.06 [0, .13]
Writing	188	306	174	392	.59	.55	.39	.34	.44 [.27, .60]	.11 [0, .26]	.40 [.34, .48]	.04 [0, .13]
Numeracy	186	304	173	399	.75	.76	.54	.41	.63 [.50, .77]	.14 [.01, .26]	.23 [.19, .27]	0 [0, .06]
Grade 7												
Reading	131	232	112	270	.76	.70	.52	.49	.45 [.29, .61]	.26 [.11, .40]	.24 [.19, .30]	.05 [0, .12]
Spelling	131	232	109	272	.77	.71	.54	.31	.77 [.68, .81]	0 [0, .07]	.21 [.17, .26]	.02 [0, .08]
Grammar and punctuation	133	233	109	267	.73	.67	.53	.35	.62 [.44, .74]	.07 [0, .22]	.26 [.21, .33]	.05 [0, .13]
Writing	132	231	109	265	.59	.55	.40	.29	.52 [.31, .64]	.05 [0, .23]	.42 [.35, .49]	0 [0, .10]
Numeracy	127	226	107	271	.87	.80	.58	.40	.77 [.62, .85]	.05 [0, .19]	.15 [.12, .19]	.03 [0, .08]
Grade 9												
Reading	91	255	78	311	.77	.73	.62	.43	.61 [.45, .77]	.15 [0, .29]	.22 [.17, .27]	.03 [0, .09]
Spelling	90	251	78	316	.83	.76	.31	.37	.81 [.74, .84]	0 [0, .06]	.19 [.14, .23]	0 [0, .05]
Grammar and punctuation	90	250	79	313	.71	.64	.51	.41	.51 [.33, .70]	.15 [0, .30]	.29 [.22, .37]	.05 [0, .14]
Writing	90	250	79	310	.51	.52	.52	.29	.34 [.10, .56]	.16 [0, .34]	.47 [.38, .57]	.03 [0, .15]
Numeracy	116	209	84	286	.85	.78	.68	.32	.70 [.53, .81]	.07 [0, .23]	.17 [.13, .21]	.06 [0, .12]

Note. MZs = monozygotic twins in the same class; MZd = monozygotic twins in different classes; DZs = dizygotic twins in the same class; DZd = dizygotic twins in different classes; A = additive genetic; C = shared environment; E = unique environment; CL = classroom. Confidence interval 95% in square brackets.

Table 3
Study 1 Sample Size, Correlations, and Standardized Variance Components for NAPLAN Domains in Grades 3, 5, 7, and 9 Excluding Participants With Academic Ability as a Reason for Class Allocation

	Number of complete pairs				Correlations				Class-allocation model variance components			
NAPLAN subscale by grade	MZs	MZd	DZs	DZd	MZs	MZd	DZs	DZd	A	C	E	CL
Grade 3												
Reading	163	241	141	254	.71	.76	.45	.39	.69 [.53, .78]	.06 [0, .21]	.25 [.21, .29]	0 [0, .04]
Spelling	164	239	141	254	.77	.75	.49	.26	.74 [.62, .78]	0 [0, .11]	.22 [.17, .27]	.05 [0, .11]
Grammar and punctuation	161	238	139	252	.74	.70	.54	.33	.59 [.42, .74]	.10 [0, .25]	.26 [.22, .32]	.05 [0, .12]
Writing	161	239	141	249	.43	.44	.31	.34	.28 [.06, .50]	.17 [0, .35]	.55 [.45, .63]	0 [0, .12]
Numeracy	163	236	142	251	.74	.78	.50	.43	.62 [.47, .78]	.14 [0, .28]	.24 [.21, .28]	0 [0, .03]
Grade 5												
Reading	118	238	107	271	.71	.71	.39	.46	.57 [.40, .74]	.15 [0, .30]	.28 [.23, .33]	0 [0, .05]
Spelling	119	239	108	272	.78	.82	.52	.47	.71 [.57, .84]	.11 [0, .25]	.18 [.15, .21]	0 [0, .03]
Grammar and punctuation	119	237	107	268	.75	.64	.36	.42	.57 [.40, .72]	.10 [0, .26]	.27 [.21, .35]	.06 [0, .14]
Writing	119	238	107	263	.57	.55	.36	.37	.42 [.21, .61]	.14 [0, .31]	.41 [.33, .50]	.03 [0, .13]
Numeracy	117	235	106	270	.74	.77	.52	.45	.61 [.46, .78]	.16 [0, .30]	.23 [.19, .27]	0 [0, .04]
Grade 7												
Reading	57	136	51	143	.74	.69	.50	.55	.41 [.21, .63]	.29 [.11, .47]	.23 [.15, .33]	.07 [0, .16]
Spelling	56	137	50	143	.76	.78	.52	.37	.80 [.64, .84]	0 [0, .15]	.20 [.16, .25]	0 [0, .05]
Grammar and punctuation	56	137	50	140	.65	.71	.45	.36	.71 [.47, .77]	0 [0, .21]	.24 [.17, .34]	.05 [0, .15]
Writing	56	136	50	141	.58	.58	.45	.38	.42 [.16, .66]	.17 [0, .39]	.38 [.28, .50]	.03 [0, .16]
Numeracy	51	129	45	140	.88	.85	.51	.55	.70 [.52, .88]	.16 [0, .34]	.13 [.10, .17]	0 [0, .05]
Grade 9												
Reading	24	119	26	131	.74	.72	.62	.57	.41 [.20, .64]	.35 [.13, .52]	.24 [.15, .32]	.01 [0, .12]
Spelling	23	120	26	133	.86	.76	.28	.49	.82 [.59, .86]	0 [0, .21]	.18 [.10, .24]	0 [0, .09]
Grammar and punctuation	23	120	26	131	.69	.61	.49	.45	.47 [.19, .72]	.19 [0, .41]	.28 [.17, .42]	.07 [0, .22]
Writing	23	118	26	129	.66	.48	.26	.35	.40 [.05, .61]	.10 [0, .38]	.40 [.23, .61]	.09 [0, .28]
Numeracy	33	90	22	106	.84	.81	.75	.43	.63 [.38, .84]	.17 [0, .39]	.17 [.11, .25]	.03 [0, .13]

Note. NAPLAN = National Assessment Plan—Literacy and Numeracy; MZs = monozygotic twins in the same class; MZd = monozygotic twins in different classes; DZs = dizygotic twins in the same class; DZd = dizygotic twins in different classes; A = additive genetic; C = shared environment; E = unique environment; CL = classroom. Confidence interval 95% in square brackets.

partitioning this latent variable into A, C, E, and CL variance components (Figure 2). The latent NAPLAN performance factor accounted for 61.8%, 60.9%, 63.2%, and 65.1% of the total variance in all five tests in Grades 3, 5, 7, and 9, respectively. In this model, given a heritability estimate of 80% and a shared environmental estimate of 10%, there was 80% power to detect a classroom influence of 5%. Estimates of classroom influence ranged from 1–5% of the variance of this latent NAPLAN factor, and were not significant for any grade (Table 4; Supplemental Table 3, available online, details the factor and path loadings for the latent models). The 95% confidence intervals were tighter, giving further support to the finding that the influence of classroom variability on variation in NAPLAN performance is small.

Based on correlations, dizygotic twins appeared to show larger classroom effects than monozygotic twins did. The within-pair correlations for DZ twins sharing a classroom were larger than the ones in separate classes on 19 out of 20 occasions, as against 11 out of 20 occasions for MZ pairs (see Table 2). The difference in the weighted average correlation for twins in the same versus different classes was .14 for DZ twins and only .02 for MZ twins. However, this DZ-specific effect mostly disappeared once twins who had been streamed for academic reasons had been removed, with larger correlation differences in same-class DZ pairs dropping from 19/20 to 12/20 (see Table 3). Similarly, the difference in the weighted average correlation dropped to .04 for DZ twins. This suggests that the DZ effect was spurious, occasioned by preexisting performance differences in DZ pairs. These differences may in part reflect genetic differences within DZ pairs, meaning that genetic endowment is being conflated with classroom status. (Removing similarly streamed MZ twins had virtually no effect on the pattern of same- vs. different-class correlations, with 12/20 larger for same-class pairs for the sample with streamed pairs removed and a difference of .01 in the weighted average correlation for MZ twins in the same vs. different class; Table 3).

Study 2

Estimates of the classroom effect also ranged from 0 to .07, with only one, Grade 1 spelling, reaching significance (Table 5). Supplemental Table 4, available online, includes the variance components estimated without the class variance component. Thus the pattern of results closely mirrored those from Study 1. In Byrne et al. (2010), a model was formulated for estimating classroom influences based on MZ and DZ correlations within the ILTS sample. The value of the present study is in the use of structural equation modeling to estimate the classroom influences in both the ILTS and the NAPLAN data. This method generates more precise values than the previous one and provides confidence intervals. The change in analysis explains the small differences in magnitude between the current and previous estimates for ILTS data.

Discussion

In this project we have estimated the proportion of variance in school achievement in literacy and numeracy that can be assigned to classroom-level factors. We have done this by modeling the

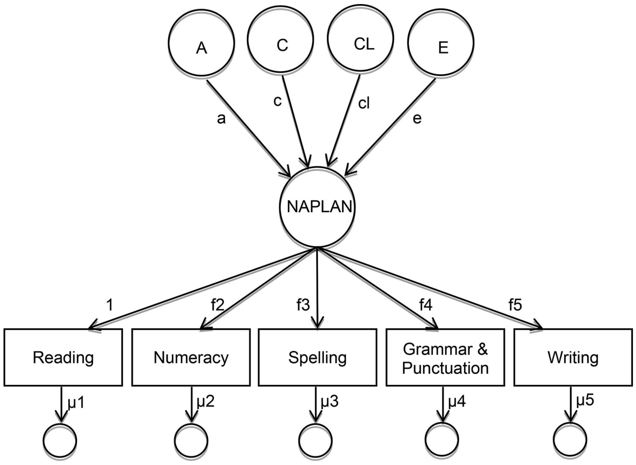

Figure 2. Path diagram representing the latent factor model. Lower case paths are estimated and the variance common to the five National Assessment Plan—Literacy and Numeracy (NAPLAN) domains is decomposed into variance due to classroom (CL), genes (A), shared environmental factors (C), and unique environmental factors (E) using the covariation between monozygotic and dizygotic twins who either shared a class or were in different classes. Only one twin is represented. The residual variances (μ) were allowed to correlate between pairs, and these correlations were estimated separately for MZ and DZ pairs.

consequences of assigning members of twin pairs to the same or different classrooms. If classroom factors are in play, it is expected that twins who are separated will be less similar than twins who share a classroom, detectable as a classroom factor alongside genetic, shared environment, and unique environment factors. The students involved in Study 1 were Australian, in Grades 3, 5, 7, and 9, and the tests were countrywide assessments of reading, spelling, grammar and punctuation, writing, and numeracy administered in May of each year, around three and a half months into the Australian school year. Because of the limited duration spent in each grade prior to assessment, we required the twins to have been together or separate in the previous school year as well as the test year. We modeled the data for the full sample and for a substantial subsample that excluded pairs that had been separately streamed on academic grounds to ensure that any classroom effects reflected current placement. The students in Study 2 were derived from students from Australia and the United States. The analyses were conducted with the same models as for the Study 2, and somewhat differently from those in Byrne et al. (2010), which used just raw correlations.

Overall, the analyses showed that classroom effects were negligible. In all but one of the models in each study, the classroom factor was not significant. This was spelling in Grade 3 in Study 1 (this effect of 7% dropped to a nonsignificant 5% in the subsample in which participants who were streamed for academic reasons were removed) and spelling in Grade 1 in Study 2. Significance aside, the mean classroom effect across all tests and grades of Study 1 was 3% for both the full participant sample and the subsample (calculated from Tables 2 and 3, respectively). It was 2% when the NAPLAN test results were loaded onto a latent factor common to all test domains (see Table 4). Similarly, Study 2 showed negligible classroom effects, with the mean classroom effect across grades and tests also of 2% (see Table 5).

The results from the current study largely converge with those from other groups' research, some employing twins and some not, with our results at the lower end of others' estimates. One twin study, from Kovas et al. (2007), used results from the Twins Early Development Study conducted in England and Wales and which recruited around 12,000 sets of twins. The authors couched their findings in terms of the influence of twins' shared environment on school performance for twins sharing a classroom versus those in separate classrooms. Kovas et al. reported higher C for same-class pairs, though the increase in C compared with different-class twins did not reach significance in most cases. Further, when objective tests that were available for some of the students were used instead of teacher ratings, the same-class effect of higher C mostly disappeared. This suggests that idiosyncratic use of the scales by individual teachers may have been responsible for the initial finding—twins in the same class have the same rater, those in different classes, different raters. It was only reading that continued to show (nonsignificant) higher C in same-class pairs, a value of .17 compared with .10 for different-class pairs. If taken a face value, this translates to a classroom effect of .07.

Another twin study that was based primarily on teacher ratings is by White et al. (2018). The authors reported on ratings for reading, writing, and mathematics, averaged into an overall achievement score, in Canada (Quebec) at Ages 7, 9, 10, and 12. In the United Kingdom., they used ratings for English and mathematics, again averaged, at Ages 7, 9, 10, 12, and 14, along with results from the public examination known as General Certificate of Secondary Education (GCSE) at Age 16. White et al. used mean within-pair differences in achievement for twins taught by the

Table 4
Sample Size and Standardized Variance Components From the Latent Variable Model of Classroom Effects in Study 1

NAPLAN grade	Number of complete pairs				Standardized latent variance components			
	MZs	MZd	DZs	DZd	A	C	E	CL
Grade 3	240	286	220	352	.80 [.67, .92]	.10 [0, .22]	.08 [.06, .11]	.02 [0, .06]
Grade 5	189	310	177	404	.70 [.58, .83]	.21 [.08, .33]	.08 [.05, .10]	.01 [0, .05]
Grade 7	121	216	104	256	.81 [.65, .94]	.11 [0, .26]	.07 [.04, .10]	.01 [0, .06]
Grade 9	81	185	62	260	.81 [.63, .91]	.07 [0, .24]	.07 [.04, .12]	.05 [0, .10]

Note. MZs = monozygotic twins in the same class; MZd = monozygotic twins in different classes; DZs = dizygotic twins in the same class; DZd = dizygotic twins in different classes; A = additive genetic; C = shared environment; E = unique environment; CL = classroom. Confidence interval 95% in square brackets.

Table 5
Standardised Variance Components for Study 2 Tests in Kindergarten, Grade 1, and Grade 2

	Variance components			
ILTS achievement by grade	A	C	E	CL
Kindergarten				
TOWRE SWE	.79 [.65, .90]	.08 [0, .23]	.11 [.09, .13]	.02 [0, .07]
TOWRE PDE	.72 [.55, .81]	.04 [0, .22]	.22 [.18, .26]	.02 [0, .10]
Spelling	.40 [.28, .54]	.40 [.27, .52]	.19 [.17, .23]	.00 [0, 0]
Grade 1				
TOWRE SWE	.80 [.66, .86]	.04 [0, .17]	.15 [.12, .18]	.02 [0, .06]
TOWRE PDE	.73 [.59, .82]	.05 [0, .19]	.19 [.16, .23]	.02 [0, .08]
Spelling	.72 [.61, .77]	0 [0, .11]	.21 [.17, .25]	.07 [.02, .17]
Passage comprehension	.67 [.53, .79]	.09 [0, .23]	.23 [.20, .27]	0 [0, .06]
Grade 2				
TOWRE SWE	.80 [.67, .84]	0 [0, 0]	.17 [.14, .21]	.02 [0, .07]
TOWRE PDE	.81 [.67, .85]	0 [0, .15]	.16 [.13, .19]	.02 [0, .07]
Spelling	.78 [.67, .81]	0 [0, .10]	.22 [.19, .26]	0 [0, .03]
Passage comprehension	.63 [.47, .77]	.10 [0, .24]	.26 [.21, .31]	.01 [0, .08]

Note. TOWRE = Test of Word Recognition, Sight Word Efficiency; TOWRE PDE = Test of Word Recognition, Phonemic Decoding Efficiency; A = additive genetic; C = shared environment; E = unique environment; CL = classroom. Confidence interval 95% in square brackets.

same and different teachers as one index of similarity. They showed that although twins taught together tended to have more similar ratings than those taught apart, only a few effects were significant, and with small effect sizes (2.7% for Age 12 in Quebec and 3.4% for the United Kingdom GCSE).

Classroom effects research using nontwin populations often rely on value-added models. These models utilize the gains or losses students in a class make over the levels achieved in the previous year. The assumption is that the previous year's achievement folds into a single number the effects of a large number of variables, including genetic endowment, home support, homework habits, and so on. Hence when the current year's trajectory is compared to other classes in the same school or district it forms an index of classroom effectiveness, or in the view of most researchers, teacher effectiveness. Our estimates are at the lower bound of estimates of "teacher effects" on literacy and numeracy skills, for example, ranging from 3–16% (Nye et al., 2004; Rowan et al., 2002) that have employed value-added methods. McCaffrey, Lockwood, Koretz, and Hamilton (2003) reviewed several value-added studies and pointed to a variety of shortcomings, including the omission of covariates. They concluded that value-added research was prone to numerous errors and cautioned about its use in high-stakes decisions.

Another research technique, rarely achievable, would be to assign children to classes and teachers randomly in a genuine experiment. This reduces to chance levels the role of factors other than classroom effects including any due to teachers. The Tennessee Class Size Experiment (Nye, Hedges, and Konstantopoulos, 2000) is one project that was able to achieve the status of an experiment; children and teachers were randomly assigned to classes in 79 schools from kindergarten to Grade 3. Classroom effects on reading ranged from .066 to .110 (but see Byrne et al., 2010 for some possible problems with the study which may mean that these are overestimates).

The current data cover a relatively wide spread of language, literacy, and numeracy skills, and are derived from a different country than those just cited. Thus, the case for very modest classroom-level effects for the core domains of language, literacy, and numeracy holding in Western educational systems is strengthened.

One factor that may contribute to the very low classroom effect within Australia is a set of nation-wide expectations as to what students should be taught. The Australian Curriculum and Assessment Authority (ACARA) states that "the Australian Curriculum sets the expectations for what all young Australians should be taught, regardless of where they live in Australia or their background" (acara.edu.au). These teaching objectives have been developed for several domains, including English and mathematics, up to Grade 10. They constitute a detailed list of objectives for each grade in each subject. Environmental uniformity, to which a national curriculum will contribute, will, in turn, contribute to lower environmental influence (and higher genetic influence) than will be the case with wide environmental variability. In nations where there are no country-wide stipulations like Australia's, teachers presumably have greater leeway in what they teach, which may, in turn, lead to larger classroom-level influences on achievement.

None of our tentative hypotheses were confirmed. One was that writing may be a domain more vulnerable to classroom effects because it was more subject to environmental influences than the other domains (Grasby et al., 2016), possibly therefore including the influence of particular teachers. However, the writing test in the NAPLAN is likely more subject to measurement error because scoring is subjective to a higher degree than for the other tests (see also Caldwell & White, 2017), and this, rather than classroom-based factors, may explain the higher E component in our data. In any case, to the extent that the unique environment component of variance represents nonerror factors, they do not appear to include ones that operate on a classroom-wide basis.

We had also expected that numeracy may be more susceptible to classroom effects because of a scarcity of fully trained mathematics teachers. This expectation was not fulfilled either. We should say, however, and foreshadowing the Limitations section, that a classroom effect may emerge in higher school grades (up to 12 in

Australian schools) as the syllabus becomes more demanding and therefore possibly more dependent on teacher qualifications.

The third expectation that we entertained was that classroom effects may be more pronounced in high school than elementary school because of variation in needed peer and teacher support in the more challenging educational environment of high school. We saw no evidence of this in our data. At least, if classroom variation does influence how well a student settles into high school, it does not appear to impinge on the academic domains that the NAPLAN assesses. However, we acknowledge that, just as in the case of numeracy, where the timing of the tests may have missed a (later) classroom effect, so timing may matter for the high school-elementary school comparison. The Grade 7 assessments are administered just over 3 months into the school year, the first in high school, too early perhaps to be affected by variation in the hypothesized support factors. By Grade 9, the next NAPLAN assessments, these factors may have worked their way through students' sense of place in school and no longer affect academic achievement. Admittedly, these are post hoc suppositions, but at least if the peer and teacher support processes do play a role in academics, they may be fleeting.

Limitations

These data cover a larger range of school grades and test domains than any previous set of which we are aware—respectively, kindergarten to Grade 9, and spelling, writing and grammar/punctuation as well as the commonly tested reading and numeracy. However, classroom effects may emerge later in high school than Grade 9 because more advanced content may be subject to more substantial influence from individual teachers, for instance, as already mentioned in connection with mathematics. Our data are also silent on school subjects other than those assessed in the ILTS and NAPLAN, such as languages, art, music, social sciences, and sciences.

The NAPLAN data are restricted to one country, and we have alluded to possible differences that might emerge in nations that do not enforce a uniform curriculum. Similarly, differences across countries in terms of the uniformity of teacher training, class size, and other factors that would be expressed at the classroom level may generate larger classroom effects than we present here, although the similar results found within the cross-national ILTS sample and between the ILTS and NAPLAN suggest these differences are not present between the United States and Australia.

While twin samples are largely considered representative of the general population for most traits, including achievement, we acknowledge that there may be twin-specific factors which could influence the classroom experience such as twin-to-twin sharing of information or collaborative studying. Additionally, there may be classroom effects that, in principle, remain undetectable using the kind of data that we had available, For example, an individual teacher or classroom peers may have a profound academic influence, positive or negative, on an individual student, influences that do not spread to others in the class. Even MZ twins can experience the same classroom differently, and those differences have been shown to relate to differences in mathematics and science achievement (Asbury, Almeida, Hibel, Harlaar, & Plomin, 2008). Thus, there can be individual rather than group experiences in the classroom that affect academic performance, and indeed these may be more prevalent than classroom-wide experiences.

Finally, and referring to a possible teacher influence as part of a classroom effect, if teaching quality matters and if it clusters within schools, data like ours will underestimate the classroom effect. This is because even twins in separate classes will both be in the hands of higher or lower quality teachers, depending on the quality of the teachers with which their school is staffed. Clustering of this kind may occur on a regional basis (e.g., metropolitan vs. remote schools), on the basis of school type (e.g., private vs. public), or on some other basis. There is some evidence from the United States that schools that happen to serve a higher proportion of at-risk students tend to have less credentialed and less experienced teachers (McCaffrey et al., 2003). Thus it is prudent to treat our estimate of a classroom effect as lower bound in case systematic factors like clustering are obscuring a larger effect.

Conclusion

Classroom-level influences on students' variability in literacy and numeracy were almost all nonsignificant in the data we report, which were structured to reflect any average effects of twins sharing or not sharing classrooms. If statistical significance is ignored, we reach an estimate of around 3% of variance in school achievement attributable to classroom influences. This figure is much lower than ones in popular discourse, which can range up to 40% (Byrne et al., 2010). We concede that our results may represent a lower-bound estimate of classroom effects (see Limitations), and that we cannot detect individual student-classroom interactions, but it is unlikely that the "real" effect approaches the publicly canvassed ones. Thus legislative actions and educational policies that are based on those high estimates are unsound. Instead, our results and other similar ones indicate that for the educational jurisdictions in which they are generated teaching practices and aspects of classroom climate are sufficiently uniform to allow students' own potentials to substantially influence their academic achievements. We are of the view that this situation, rather than one in which substantial influences stem from variation in teacher skill and other classroom-level processes, is one feature of a well-tempered educational system.

References

Asbury, K., Almeida, D., Hibel, J., Harlaar, N., & Plomin, R. (2008). Clones in the classroom: A daily diary study of the nonshared environmental relationship between monozygotic twin differences in school experience and achievement. *Twin Research and Human Genetics, 11,* 586–595. http://dx.doi.org/10.1375/twin.11.6.586

Australian Bureau of Statistics. (2014). *Education and work, Australia*. Retrieved from http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/6227.0Main+Features1May2014?OpenDocument

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., . . . Fox, J. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika, 76,* 306–317. http://dx.doi.org/10.1007/s11336-010-9200-6

Bru, E., Stornes, T., Munthe, E., & Thuen, E. (2010). Students' perceptions of teacher support across the transition from primary to secondary school. *Scandinavian Journal of Educational Research, 54,* 519–533. http://dx.doi.org/10.1080/00313831.2010.522842

Byrne, B., Coventry, W. L., Olson, R. K., Wadsworth, S. J., Samuelsson, S., Petrill, S. A., . . . Corley, R. (2010). "Teacher effects" in early literacy development: Evidence from a study of twins. *Journal of Educational Psychology, 102,* 32–42. http://dx.doi.org/10.1037/a0017288

Byrne, B., & Fielding-Barnsley, R. (1993). Evaluation of a program to teach phonemic awareness to young children: A 1-year follow-up. *Journal of Educational Psychology, 85*(1), 104–111.

Caldwell, D., & White, P. R. (2017). That's not a narrative; this is a narrative: NAPLAN and pedagogies of storytelling. *Australian Journal of Language and Literacy, 40,* 16–27.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *The American Economic Review, 104,* 2593–2632. http://dx.doi.org/10.1257/aer.104.9.2593

Clarke, M. (2014, July 27). Novice teachers challenged by ability grouping contrary to evidence. *EduResearch Matters, Australian Association for Research in Education*. Retrieved from http://www.eduresearchmatters.devave.com/category/education-policy/

Fraser, B. J. (2012). *Classroom climate*. Milton Park, UK: Routledge.

Ganeson, K., & Ehrich, L. C. (2009). Transition into high school: A phenomenological study. *Educational Psychology and Theory, 41,* 60–78. http://dx.doi.org/10.1111/j.1469-5812.2008.00476.x

Ganzeboom, H. B. G., & Treiman, D. J. (1996). Internationally comparable measures of occupational status for the 1988 International Standard Classification of Occupations. *Social Science Research, 25,* 201–239. http://dx.doi.org/10.1006/ssre.1996.0010

Ganzeboom, H. B. G., & Treiman, D. J. (2010, July 27). *International Stratification and Mobility File: Conversion tools*. Amsterdam, the Netherlands: Department of Social Research Methodology. Retrieved from http://www.harryganzeboom.nl/ismf/index.htm

Grasby, K. L., & Coventry, W. L. (2016). Longitudinal stability and growth in literacy and numeracy in Australian school students. *Behavior Genetics, 46,* 649–664. http://dx.doi.org/10.1007/s10519-016-9796-0

Grasby, K. L., Coventry, W. L., Byrne, B., & Olson, R. K. (2019). Little evidence that socioeconomic status modifies heritability of literacy and numeracy in Australia. *Child Development, 90,* 623–637. http://dx.doi.org/10.1111/cdev.12920

Grasby, K. L., Coventry, W. L., Byrne, B., Olson, R. K., & Medland, S. E. (2016). Genetic and environmental influences on literacy and numeracy performance in Australian school children in Grades 3, 5, 7, and 9. *Behavior Genetics, 46,* 627–648. http://dx.doi.org/10.1007/s10519-016-9797-z

Hanewald, R. (2013). Transition between primary and secondary school: Why it is important and how it can be supported. *The Australian Journal of Teacher Education, 38,* 62–74. http://dx.doi.org/10.14221/ajte.2013v38n1.7

Hattie, J. A. (2003, October). *Teachers make a difference: What is the research evidence?* Paper presented at the Building Teacher Quality: What Does the Research Tell Us. ACER Research Conference, Melbourne, Australia. Retrieved from http://research.acer.edu.au/research_conference_2003/4/

Hindson, B., Byrne, B., Fielding-Barnsley, R., Newman, C., Hine, D. W., & Shankweiler, D. (2005). Assessment and Early Instruction of Preschool Children at Risk for Reading Disability. *Journal of Educational Psychology, 97,* 687–704.

Jastak, S., & Wilkinson, G. S. (1984). *Wide Range Achievement Test-Revised*. Wilmington, DE: Jastak Associates.

Johnston, O., & Wildy, H. (2016). The effects of streaming in the secondary school on learning outcomes for Australian students—A review of the international literature. *Australian Journal of Education, 60,* 42–59. http://dx.doi.org/10.1177/0004944115626522

Kovas, Y., Haworth, C. M. A., Dale, P. S., & Plomin, R. (2007). The genetic and environmental origins of learning abilities and disabilities in the early school years. *Monographs of the Society for Research in Child Development, 72,* 1–144.

Kovas, Y., Voronin, I., Kaydalov, A., Malykh, S. B., Dale, P. S., & Plomin, R. (2013). Literacy and numeracy are more heritable than intelligence in primary school. *Psychological Science, 24,* 2048–2056. http://dx.doi.org/10.1177/0956797613486982

Larsen, S. A., Byrne, B., Little, C. W., Coventry, W. L., Ho, C. S., Olson, R. K., & Stevenson, A. (2019). Identical genes, unique environment: A qualitative exploration of persistent monozygotic-twin discordance in literacy and numeracy. *Frontiers in Education, 4,* 21. http://dx.doi.org/10.3389/feduc.2019.00021

Lykken, D. T., Bouchard, T., Jr., McGue, M., & Tellegen, A. (1990). The minnesota twin family registry: Some initial findings. *Acta Geneticae Medicae et Gemellologiae, 39*(1), 35.

Marsh, H. W., Martin, A. J., & Cheng, J. H. S. (2008). A multi-level perspective on gender in classroom motivation and climate: Potential benefits of male teachers for boys? *Journal of Educational Psychology, 100,* 78–95. http://dx.doi.org/10.1037/0022-0663.100.1.78

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: Rand. http://dx.doi.org/10.1037/e658712010-001

Nichols, R. C., & Bilbro, W. C., Jr. (1966). The diagnosis of twin zygosity. *Human Heredity, 16,* 265–275.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2000). The effects of small classes on achievement: The results of the Tennessee class size experiment. *American Educational Research Journal, 37,* 123–151. http://dx.doi.org/10.3102/00028312037001123

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26,* 237–257. http://dx.doi.org/10.3102/01623737026003237

Olson, R. K., Keenan, J. M., Byrne, B., & Samuelsson, S. (2014). Why do children differ in their development of reading and related skills? *Scientific Studies of Reading, 18,* 38–54. http://dx.doi.org/10.1080/10888438.2013.800521

Papaioannou, A., Marsh, H. W., & Theodorakis, Y. (2004). 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, 26,* 90–118. http://dx.doi.org/10.1123/jsep.26.1.90

Productivity Commission. (2012). *Schools workforce* (Research Report). Canberra, Australia: Australian Government. Retrieved from https://www.pc.gov.au/inquiries/completed/education-workforce-schools/report

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the "Prospects" study of elementary schools. *Teachers College Record, 104,* 1525–1567. http://dx.doi.org/10.1111/1467-9620.00212

Rutter, M., & Maughan, B. (2002). School effectiveness findings 1979–2002. *Journal of School Psychology, 40,* 451–475. http://dx.doi.org/10.1016/S0022-4405(02)00124-3

Taylor, J., Roehrig, A. D., Soden Hensler, B., Connor, C. M., & Schatschneider, C. (2010). Teacher quality moderates the genetic effects on early reading. *Science, 328,* 512–514. http://dx.doi.org/10.1126/science.1186149

Torgesen, J., Wagner, R., & Rashotte, C. A. (1999). *A Test of Word Reading Efficiency (TOWRE)*. Austin, TX: PRO-ED.

White, E. K., Garon-Carrier, G., Tosto, M. G., Malykh, S. B., Li, X., Kiddle, B., . . . Kovas, Y. (2018). Twin classroom dilemma: To study together or separately? *Developmental Psychology, 54,* 1244–1254. http://dx.doi.org/10.1037/dev0000519

Woodcock, R. W. (1989). *Woodcock Reading Mastery Tests*. Circle Pines, MN: American Guidance Service.

Received February 6, 2019
Revision received August 1, 2019
Accepted August 23, 2019 ■