Published in final edited form as:

Aust J Educ. 2015 April ; 59(1): 5–21. doi:10.1177/0004944114563775.

Validity of large-scale reading tests: A phenotypic and behaviour-genetic analysis

Katrina L Grasby,

Discipline of Psychology, School of Behavioural, Cognitive and Social Sciences, University of New England, Australia

Brian Byrne, and

Australian Research Council Centre of Excellence for Cognition and its Disorders, and Discipline of Psychology, School of Behavioural, Cognitive and Social Sciences, University of New England, Australia

Richard K Olson

Director of Colorado Learning Disabilities Research Center, and Department of Psychology, University of Colorado, USA

Abstract

Each year, all Australian students in grades 3, 5, 7, and 9 sit nationwide large-scale tests in literacy and numeracy, which have their validity frequently questioned. We compared the performance of Grade 3 twins on these large-scale reading tests with their performance on three individually administered literacy tests in comprehension, word reading and vocabulary within a genetically sensitive design. Comprehension, word reading, and vocabulary accounted for a substantial amount of the variance in school reading tests. Performance on large-scale reading tests and individually administered tests was moderately to substantially heritable and the same genes contributed to performance in both types of test. These results confirm that large-scale school reading tests measure, at least in part, the literacy skills tapped by individual tests that are frequently considered to be the “gold-standard” in testing.

Keywords

Reading ability; literacy skills; academic achievement; twins; genetics; environmental influences

Introduction

In this paper, we intend to contribute to the debate about the validity of so-called high-stakes testing in Australian schools by reporting on how achievement in those tests relates to performance in alternative measures of literacy. We have a database of the test results of 250 Grade 3 students in New South Wales who completed the literacy assessments of the Basic

Corresponding author: Katrina Grasby, Psychology, University of New England, NSW 2351, Australia kgrasby2@une.edu.au.

Declaration of conflicting interests

None declared.

Skills Test (BST) or its successor, the National Assessment Program: Literacy and Numeracy (NAPLAN). In addition to these assessment results, we have a number of well-accepted individually administered (IA) tests of literacy skills for each child that can serve as a validity check for the large-scale, group-administered school results. Further, the children are all members of twin pairs, which allows us to estimate the relative influences of genes and aspects of the environment on individual differences in performance on both individual and large-scale assessments. In brief, we address issues of the validity of large-scale tests and their “behaviour-genetics,” meaning we explore the relative influence of genes and the environment on variation in performance (as we explain further below).

These twins were part of a large international study on literacy, for which they were administered the labour-intensive individual literacy tests. Although there are sufficient participants from each of the BST and NAPLAN years for phenotypic analyses (analyses of the physically expressed trait), the number of participants within each test type is small for genetic analyses. Hence we used both tests in our measure of large-scale tests for some of the analyses (see the method section for further details regarding the combining of these scores). We acknowledge that it would have been more desirable to use data from just one type of large-scale test. However, the availability of these twin-study data represents a unique opportunity to explore the relationship between large-scale reading tests and individual literacy tests within a genetically sensitive design.

In 2008, nationwide assessment in literacy and numeracy began in Australia (Australian Curriculum Assessment and Reporting Authority [ACARA], 2008). The NAPLAN includes standardised tests in reading, writing, language conventions (spelling, grammar and punctuation), and numeracy. Each year, students in Grades 3, 5, 7, and 9 from government and non-government schools take these tests. For each achievement domain, performance is measured on a common scale from Grade 3 to Grade 9; this common scale allows for comparisons between cohorts and provides a measure of growth within a cohort over time. The tests are developed with reference to the National Statements of Learning in English and Mathematics, and state and territory curricula (ACARA, 2010a).

Prior to the NAPLAN, statewide testing in literacy and numeracy was implemented in New South Wales (NSW) in 1989 in the form of the BST (Masters et al., 1990). Though initially the BST was administered to Grade 6 students, by 1996 students in Grades 3 and 5 were tested and performance at each grade level measured on a common scale (Wasson, 2009). Many hallmarks of the current NAPLAN were already present in the BST. These included comparing performance of a school with a similar school as measured on remoteness, socioeconomic status, and percentage of Aboriginal and Torres Strait Islander student enrolments, and providing schools with data and software to link performance and test items with the curriculum and teaching strategies (Wasson, 2009). These characteristics of the assessment and feedback process were intended to identify individuals, classes, or schools that were performing below a minimum standard and to assist with teaching resources and support (Masters et al., 1990; Smith, 2005; Wasson, 2009).

Even though NSW now has a 25-year history of large-scale assessment in literacy and numeracy, the NAPLAN remains a controversial topic, particularly in relation to the

interpretation of results. These tests have relatively few items, and though reliable information can be obtained when the cohort is sufficiently large, results from individuals and small groups, such as classrooms, are easily misinterpreted or over-interpreted (Wu, 2010, 2011). In addition, complaints have been raised of undue pressure on students, teachers, and schools, perhaps fuelled by poorly informed responses to the publicly available results online (Athanasou, 2010; Wu, 2011). Criticisms have been made of some test items and questions raised about whether the items actually test the skills intended. For example, the NAPLAN language conventions test requires students to identify and correct spelling, punctuation and tense errors, but Williams (2009) suggests that this does not assess the ability to apply grammatical knowledge in any depth. Willet and Gardiner (2009) analysed the spelling errors of students from both the NAPLAN and dictation tests and found that students scored far better on dictation tests. In NAPLAN some spelling items present a misspelt word and students are required to write the word correctly. In contrast to the general pattern of scoring better for dictation, Willet and Gardiner found these constructed errors sometimes facilitated the correct spelling of a word. This occurred when the constructed error was not in the part of a word that students were most likely to misspell.

Similarly, question format can assist in solving numeracy problems. In a critique of the Grade 9 numeracy tests in 2008, Norton (2009) noted that 57% of Queensland students correctly answered an algebraic problem in multiple-choice format, which was almost double the 29% of correct answers when a similar algebraic problem required a written solution. Continuing to identify and investigate validity errors and misinterpreted results is important to appropriately report on performance and improve future tests. These item-level flaws, assuming that is what they are, open the question of what these large-scale tests measure. Our research does not address issues of individual items, but it does explore whether the results from the large-scale tests converge with well-accepted, individual forms of assessment.

The National Statements of Learning in English describe the reading skill level of Grade 3 students as follows:

When students read and view texts, they identify the main topic or key information, some directly-stated supporting information, and the order of events. They can draw inferences from directly-stated descriptions and actions (e.g. *infer a character's feelings*) and talk about how people, characters and events could have been portrayed differently (e.g. *more fairly*). They relate their interpretations to their own experiences.

(Curriculum Corporation, 2005, p. 5)

Mastery of this level of reading comprehension requires the execution of a range of literacy skills. A proficient reader must have acquired the *alphabetic principle*, essentially that words are composed of letters that systematically represent phonemes (Byrne, 1998). The ability of a reader to decode regular words must be expanded to include the acquisition of orthographic rules and recognition of irregular words (Ehri, 2005). Skilled readers will execute these basic skills with accuracy and, eventually, fluency. Beyond this, the text must be related to meaning, a process that relies on higher-level language skills such as

vocabulary and syntax. Deficiencies at any of these levels will undermine reading comprehension, as captured by the *Simple View* of reading (Hoover & Gough, 1990). Therefore the validity of the school literacy tests could be approached by assessing the relationship of performance on the large-scale tests with performance on well-established tests of literacy skills. As further evidence of validity we would expect the relationship between literacy skills and school reading to be stronger than the relationship between literacy skills and school numeracy. This is the approach we adopt here.

The data in the current analyses are part of the International Longitudinal Twin Study (ILTS), a large study that followed twins from preschool to Grade 2 in four countries (Australia, Norway, Sweden, and United States). In the year prior to the start of formal schooling the twins were tested on preliteracy skills and several aspects of language and cognition including: phonological awareness, print knowledge, naming fluency, vocabulary, grammar, morphology, verbal learning and memory, and nonverbal IQ (Byrne et al., 2002; Samuelsson et al., 2005). The twins were tested in subsequent years on measures including phonological awareness, word and nonword reading, reading comprehension, spelling, vocabulary, and orthographic learning (Byrne et al., 2008; Byrne et al., 2009; Byrne et al., 2006; Byrne et al., 2007; Byrne et al., 2005). Along with these literacy measures, the BST data were collected from the ILTS students in NSW from 2003 to 2007 followed by the collection of NAPLAN data from 2008 to 2010, when data collection for the project ceased in Australia.

The BST and NAPLAN each include tests of reading, language conventions, writing, and numeracy. We have chosen to focus on the reading and numeracy measures.

These performance data span numerous years, and data for the language conventions and writing tests were not combined across the years for a number of reasons. Firstly, from 2003 to 2010, the number of questions in the language conventions tests ranged from 26 to 50, and the number of marks possible in the writing tests ranged from 47 to 70. Secondly, the marking criteria for the writing tests changed considerably from BST to NAPLAN (see ACARA, 2010b; NSW Department of Education, 2006). Hence, while it was inappropriate to combine the language conventions and writing tests into measures of large-scale school tests that spanned from 2003 to 2010 we explain the appropriateness of combining of the other measures in the method section).

One aim of the ILTS was to examine the genetic and environmental influences on the development of literacy from before formal instruction through the early years of school (Byrne et al., 2002). The inclusion of several countries allowed the pattern of genetic and environmental influence on literacy skills to be compared across countries (Byrne et al., 2009; Samuelsson et al., 2008; Samuelsson et al., 2007). However, for the purpose of comparing achievement on large-scale tests with individually administered tests of literacy, only the Australian students are reported on in this article. Although much information about the relationship of school assessment and other tests of literacy skills can be obtained from a sample of unrelated children, an advantage of analysing data from twins is the ability to estimate how much of the phenotypic relationship would be due to shared genes or shared and unique environments across skills and over time.

The classical twin design compares the relative similarity of identical (monozygotic; MZ) and fraternal (dizygotic; DZ) twins on a trait. The known genetic and environmental relationships between MZ and DZ twins can be used to estimate the amount of variation among individuals that is due to genes (A), the shared or common environment (C), and the unique environment (E). In brief, MZ twins share all their genes and DZ twins share, on average, half of their segregating genes. Within each twin pair that is raised within the same family there are some environmental influences that are shared, such as family income, common schools and teachers, and shared friends, and some that are unique to each twin, such as individual illnesses and accidents, different teachers (when twins are separated at school) and separate friends. By comparing the similarity of MZ and DZ twins it is possible to derive quantitative estimates for the influence of genes, the shared environment, and the unique environment on traits. The correlation of MZ twins on a trait is due to both their shared genes and their shared environment, while the correlation of DZ twins is due to a genetic influence that is half that of MZ twins and a shared environment that is the equivalent of the MZ twins.

The genetic influence on a trait can be calculated as double the difference in the MZ and DZ correlations (Falconer & Mackay, 1996). The degree to which the DZ correlation is greater than half of the MZ correlation reflects the shared environment. The unique environment is the difference between the MZ correlation and unity (see Plomin, DeFries, Knopik, & Neiderhiser, 2013, for an introduction to the classic twin design methodology). The genetic analyses from the ILTS have shown that by the time of testing in Grade 2, genes explain 67–82% of the variation in performance for most of the literacy measures (Byrne et al., 2009). For the most part, the unique environment (which also includes any error in measurement) explains the remaining variation in performance, with very little influence from sources of shared environment (Byrne et al., 2009). The exception is vocabulary, where genes explain 44% of the variation in performance, with 36% due to the shared environment (Byrne et al., 2009). Given the substantial heritability of these literacy skills that contribute to mastery of reading, a substantial influence of genes on test performance in the school reading tests can be hypothesised.

Using multivariate extensions to the classic twin design, phenotypic correlations can be broken down into how much of the covariation between two traits is due to genes that affect both traits and how much is due to environmental factors (shared and nonshared) that affect both traits. The extent of genetic and environmental overlap is expressed as genetic and environmental correlations. Genetic correlations are independent of univariate heritability; two traits that are each highly heritable can be influenced by different genes and thus would show no genetic correlation. However when related skills are measured an overlap in the genetic and environmental factors would be expected. For example, the genetic correlations between word reading and comprehension in the Grade 2 sample in the ILTS is .88, and the genetic correlations of those skills with vocabulary are .36 and .46 respectively (Byrne et al., 2009). These moderate to high correlations suggest that much of the genetic influence on literacy is shared among the component skills. The fact that these correlations are lower than one also suggests some independence in the genes that influence specific literacy skills (Byrne et al., 2009). In this example, the high genetic correlation between word reading and comprehension compared to that of word reading and vocabulary means that if a specific

gene was identified as influencing word reading then that gene is highly likely to also influence comprehension, but is somewhat less likely to influence vocabulary. As such, if the large-scale reading tests require students to employ the same literacy skills as are tested by the individually administered tests, then we would expect substantial overlapping genetic influences among the literacy measures and school reading tests. The analyses reported in this article aim at testing this hypothesis.

To summarise, the first purpose of this article is to explore if the large-scale reading tests of the BST and NAPLAN show convergent validity with individually administered and well-established tests of literacy skills. We approach this first by analysing if literacy skills as measured by the individually administered tests are substantial predictors of performance on the large-scale reading tests, and we compare these results with the relationship between the test results from the individually administered literacy tests and the results on the large-scale numeracy tests. The second purpose is to examine whether the genes that substantially influence performance on the individually administered tests of literacy skill also influence performance on the large-scale tests, and we compare if this genetic overlap is greater for literacy than numeracy performance.

Method

Participants

The participants were 250 children who sat the BST or NAPLAN in one of the years from 2003 to 2010; the respective numbers from each year were 22, 16, 22, 30, 54, 40, 34, and 32. The 125 twin pairs were all same-sex, with 67 monozygotic (37 male, 30 female) and 58 dizygotic (32 male, 26 female) twins. The twins were recruited through the Australian Twin Registry in the year before beginning school. Parents were approached by mail, with a 60% participation rate, of which 93% (246 twin pairs) continued with the study to the end of Grade 2 when the skills we analyse here were tested. Of these 125 pairs, 51% provided Grade 3 school test results. Those who returned results of large-scale tests had no missing data on any of the IA literacy tests. Seven participants returned either a large-scale reading or numeracy test but did not return both. A MANOVA showed that those who returned results of large-scale tests performed significantly better on the IA literacy tests at the end of Grade 2 than those who did not return results from large-scale tests ($\lambda = .98$, $F(3,488) = 3.33$, $p = .019$, $\eta^2 = .02$). Although statistically significant, the size of this effect was very small; performance on the literacy tests accounted for only 2% of the variation in returning school tests.

The means on each IA literacy skill test for students who did and did not return results of large-scale tests are presented in Table 1. The mean age (*SD*) in months at the time of the IA literacy tests was 95.5 (4.2). While the BST and NAPLAN tests were sat during different months, no significant difference in the age of participants who sat the BST and the NAPLAN could be observed (mean age (*SD*) in months 105.5 (3.8) and 104.1 (4.0) respectively). Zygosity was determined by DNA analysis from cheek swab collection or, in a minority of cases, by selected items from the Nichols and Bilbro questionnaire (1966).

Materials

Individually administered test of word reading—The Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999) was used to test word reading efficiency and phonemic decoding. A list of words and a list of nonwords constitute two subtests. The score on each subtest is the number of correctly read items in 45 seconds. Each subtest has two versions, and both versions of both subtests were administered to each child. Scores were standardised ($M = 100$, $SD = 15$). The test manual reports test-retest reliability for children aged 6–9 years of .97 for word and .90 for nonword standardised scores. The average from these four lists was used to obtain a single score for word reading.

Individually administered test of vocabulary—The Boston Naming Test (Kaplan, Goodglass, & Weintraub, 2001) was used to assess vocabulary. This test requires the viewing of sixty concrete pictures that range from common (bed) to rare (abacus) with the score being the number of items correctly named, with a maximum possible score of 60 and an internal reliability of .84 in the whole ILTS sample.

Individually administered test of comprehension—The Woodcock Passage Comprehension from the Woodcock Reading Mastery Test-Revised (Woodcock, 1989) was used to assess reading comprehension. This test is a modified cloze procedure where a short passage with a missing word is read. The child supplies a single word that is suitable to replace the blank in the sentence. The test includes 43 items and scores were standardised ($M = 100$, $SD = 15$). The test manual reports a split-half reliability for first grade students of .94.

Large-scale reading tests—For both the BST and NAPLAN, students were required to read a range of specially prepared texts and answer questions that asked them to identify or interpret information contained in the texts. Responses were predominantly multiple choice with the occasional short answer. The reported internal reliability of the Grade 3 NAPLAN tests in reading from 2008–2010 was .88–.89 (ACARA, 2013). For the BST and NAPLAN, the number of questions and maximum raw scores varied from 35 to 38 over the years that these results were obtained. The scaled scores for the NAPLAN were not used, as they are not comparable with the BST scores.

Large-scale numeracy test—The numeracy test of the BST and NAPLAN include questions assessing the aspects of number, measurement, and space. Responses were predominantly multiple choice with the occasional short answer. The reported internal reliability of the Grade 3 NAPLAN tests in numeracy from 2008–2010 was reported to be .87–.92 (ACARA, 2013). The number of questions was the same each year, with a maximum raw score of 35.

Procedure

The individual literacy skills tests were administered in the school or the home of the twins in a session that lasted about an hour. Two test administrators assessed each twin pair at the same time, one test administrator per child (Byrne et al., 2009).

The large-scale tests were administered in class in line with prescribed, standardised procedures. The BST tests were sat during the first week of August in Grade 3, and the NAPLAN tests were sat during the second week of May in Grade 3. The test results for each student were forwarded to the research team by parents of the participants.

Data Analysis

Our data has eight cohorts of participants, and although the IA literacy tests do not vary from year-to-year, the large-scale tests do. For those participants who returned the large-scale tests, we have standardised the scores on both the IA literacy tests and the large-scale tests within each year. This was done to control for any cohort effects in the IA literacy tests and to control for any cohort effects or inadvertent differences in test difficulty in the large-scale assessments. This creates a cohort specific, relative score in standard deviation units for each test to be used in the correlation and behaviour genetic analyses, where it is the relationship between variables for individuals or for twin pairs that is of interest and not the mean scores. Scores were adjusted for age and sex; outliers were truncated at ± 3 .

To account for non-independence within twin pairs, for the phenotypic correlations the Griffin and Gonzalez's (1995) formula was implemented to adjust the degrees of freedom when evaluating significance. The regression analyses involved multilevel modelling with restricted maximum likelihood to derive the estimates. Multilevel models with dyads have the individual at level 1 and the dyad at level 2 (Kenny, Kashy, & Cook, 2006). With only two units at level 1 the slopes are constrained to be equal across dyads while the intercepts can vary; the non-independence within dyads is modelled in the variation of these intercepts and the variation of the slopes is part of the error variance. This leaves two random effects to be estimated, the dyad covariance (s_{dd}) and the error variance (s_e^2). To determine the variance explained, s_{dd} and s_e^2 are calculated for the model with the predictors included and with the predictors excluded. From these variances, pseudo R^2 can be calculated as

$$R^2 = 1 - \frac{s_{dd} + s_e^2}{s_{dd}' + s_e'^2}$$

where the prime indicates the estimate is from the model without predictors. Using this calculation pseudo R^2 gives the proportion of variance in the model that is explained by the predictors (for more detail see Kenny et al., 2006).

To have sufficient data for some basic genetic parameter estimates we wanted to combine the results from the BST and the NAPLAN and use the combined score as the school result for reading and numeracy. Although the BST and the NAPLAN are of similar length (35–38 questions), have similar types of questions (both literal and inferential), and require similar response formats (predominantly multiple choice), the tests have been shown to vary from year-to-year. For example, the proportion of questions requiring the interpretation of images has differed over the years (Unsworth & Chan, 2009). Moreover, Keenan, Betjemann, and Olson (2008) have shown that reading comprehension tests vary in the extent to which they tap underlying literacy skills. While standardising within year will control for any variation

in difficulty, before combining these measures of school reading we needed to assess if each type of test was measuring reading comprehension in a similar way. To explore if the BST and NAPLAN were similar in the nature of their assessment of comprehension, we examined whether or not the correlations between each of the BST and the NAPLAN results and the IA literacy tests were equivalent. The correlations are reported in Table 2.

As can be seen, the correlations between the BST reading scores and the IA literacy scores were within .07 of the NAPLAN reading scores. The correlations between the BST numeracy scores and the IA literacy scores were within .09 of the NAPLAN numeracy scores. The difference in BST and NAPLAN correlations were tested using *Z*-difference scores (as per Field, 2013) with the degrees of freedom adjusted for non-independence between twins (see Table 2.). None of the correlations were significantly different with an alpha of .05. As well as correlations we used multilevel models to check if there was much discrepancy in how much variation was explained by the three IA literacy tests in BST and NAPLAN reading. These three IA literacy tests predicted 50% and 46% of the variance in BST and NAPLAN reading, respectively. With numeracy, the three IA literacy tests predicted 26% and 30% of the variance in BST and NAPLAN, respectively. As such, the relationship between the IA literacy tests and the BST closely resembled that of those same literacy tests with the NAPLAN; therefore, the scores (standardised within year) from the BST and NAPLAN were combined into a single score for each of the domains of reading and numeracy. At this point, any variance in scores that is systematically due to the type of test (i.e. BST or NAPLAN) will contribute to error variance in our genetic analyses. To remove this error variance from the genetic analyses, test type was covaried out.

Results

Phenotypic analyses

All of the measures shared significant positive correlations ($p < .001$; see Table 3). The correlations among the three IA literacy tests ranged widely from .26 between word reading and vocabulary to .70 between word reading and comprehension. The correlations between the large-scale tests and the IA literacy tests were moderate to high, with the numeracy being slightly lower than reading.

Analyses of multilevel models were conducted to examine how much of the variation in student performance on large-scale reading tests was accounted for by the IA literacy tests. The variables were standardised prior to entry, so the coefficients reported are equivalent to standardised beta weights. The pseudo R^2 indicates the proportion of variance explained in the outcome by the predictor variables. The IA literacy tests accounted for 49% of the variation in performance on the large-scale reading test (see Table 4). The coefficients indicated each of the literacy skills explained a unique portion of the variation in the school reading test, with each literacy skill assessed by an IA test remaining a significant predictor of large-sale reading results after controlling for the effects of the other predictors. With each standard deviation increase in IA comprehension, performance on the large-scale reading test increased by .33 of a standard deviation. Similarly, a standard deviation increase in IA word reading was related to an increase in performance on the large-scale reading test by .31 of a standard deviation, and a standard deviation increase in IA vocabulary related to

an increase in reading performance as measured by the large-scale test of .23 of a standard deviation.

Although still substantial at 28%, the IA literacy tests accounted for considerably less variation in numeracy as assessed by the large-scale tests. Interestingly, the unique contributions of IA word reading and IA vocabulary to predicting differences in large-scale numeracy performance were as substantial as their contribution to predicting performance in reading on the large-scale test.

Behaviour-genetic analyses

The correlations within zygosity type are reported as intraclass correlations in Table 5. The MZ correlations are consistently higher than the DZ correlations, indicating a genetic influence on all of the tests. The proportion of variance in each measure that was attributable to either genetic (A), shared environment (C), or unique environment (E) was estimated using OpenMx (Boker et al., 2011), and are also reported in Table 5. Due to our use of an opportune data set, the sample size is small for behaviour-genetic analyses and the confidence intervals are consequently wide, as such the parameter estimates should be interpreted cautiously and with this low power in mind. These results suggest that genes and unique environment both substantially influence performance in large-scale assessments of reading and numeracy, with little evidence of shared environmental influence.

Genetic and environmental correlations represent the amount of overlap there is in the genes (or environment) that are influencing two traits. In Table 6 we have reported the genetic and environmental correlations between the IA literacy tests and the school administered reading and numeracy tests. The shared environment correlations are unreliable as there was insufficient power to obtain a reliable estimate of the shared environment for any of the measures. Nevertheless, the model used to estimate these parameters was the full ACE model to avoid biasing our heritability estimates (Coventry & Keller, 2005; Keller & Coventry, 2005).

The genetic correlations are high between performance on the large-scale reading tests and on the IA tests that assess the literacy skills. The genetic correlation between large-scale reading and IA comprehension is very high, being reported here at unity, essentially meaning that the genes that influence comprehension as assessed by the IA test also influence reading as assessed by the large-scale tests. The rest of the genetic correlations are high but not all-inclusive, indicating there are genes that contribute to performance on large-scale reading tests that do not also contribute to the individually assessed literacy skills of word reading or vocabulary.

The unique environmental correlations between the IA literacy skills and large-scale reading performance are considerably smaller than the genetic correlations. The unique environment variance estimates include any measurement error, which, by definition, does not correlate across measures. Therefore, unique environment correlations are evidence that the environmental variance is not solely due to measurement error. That is, the unique environmental correlations indicate specific factors in the unique environment that contribute

to both performance on IA literacy tests and large-scale reading tests. However, the data analysed in this article do not allow the identification of those factors.

Similar to performance on large-scale reading tests, performance on large-scale numeracy tests also shared moderate to high genetic correlations with the IA literacy tests, indicating substantial overlap in the genes that influence performance on large-scale numeracy tests and on these IA tests of specific literacy skills. Again, some overlap in the environmental factors contributed to performance on large-scale numeracy tests and IA word reading and large-scale numeracy performance and IA vocabulary.

Discussion

Our principal aim in this article was to contribute to the dialogue on the validity of large-scale school assessment tests in Australia. We assessed if student performance on reading tests administered on a large scale converged with student performance on individually administered tests of literacy skills. We chose to explore this relationship with both phenotypic and behavioural genetic analyses.

Performance on the IA literacy skills tests accounted for 49 per cent of the variation in performance on large-scale reading tests. Although the remaining variation in performance on large-scale reading tests needs to be explained, it is worth noting that our estimates are comparable in size to the amount of variance explained by individually administered word reading and listening comprehension in several widely used reading comprehension tests (Keenan, et al., 2008). Furthermore, the correlations among different measures of reading comprehension are often quite modest (Cain, Oakhill, & Bryant, 2004; Keenan et al., 2008). As such, our findings are consistent in that they show that reading as assessed in large-scale tests taps into some of the same literacy skills that are measured by the individually administered tests. Note also that this relationship between performance on large-scale reading tests and the IA literacy skills tests is not driven exclusively by any one skill. Each of the literacy skills measured (i.e. comprehension, word reading, and vocabulary) contributes uniquely to the variation in performance on the large-scale reading tests, suggesting that performance in reading tests is best predicted by a composite of the literacy skills as measured by the IA tests.

In contrast with school reading, performance on the IA literacy tests account for less variation in school numeracy. Literacy skills still do account for a substantial amount of variation in numeracy, almost 28 per cent. A certain amount of overlap between literacy skills and numeracy is a realistic finding given that the numeracy test requires reading of the test items and that there is extensive support in the literature for covariation in performance across academic domains (Hart, Petrill, Thompson, & Plomin, 2009; Helwig, Rozek-Tedesco, Tindal, Heath, & Almond, 1999; Kovas, Harlaar, Petrill, & Plomin, 2005. Taken together, these phenotypic results provide some evidence that large-scale reading tests preferentially assess the literacy skills that are thought to underpin reading comprehension.

The heritability estimates of performance on large-scale tests of reading and numeracy are moderate and consistent with the estimates obtained from studies conducted in the United

Kingdom and the United States (Harlaar, Dale, & Plomin, 2007; Harlaar, Hayiou-Thomas, & Plomin, 2005; Hart et al., 2013; Oliver et al., 2004; Petrill et al., 2012). The heritability estimates of the literacy skills are also in line with larger studies, with a lower estimate for vocabulary than for comprehension and word reading (Harlaar et al., 2010). Moreover, the estimates of this sub-sample of Australian students who provided information on their large-scale test results were very close to the heritability estimates from the full ILTS sample at Grade 2 for word reading (.82), vocabulary (.44), and comprehension (.67; Byrne et al., 2009).

Given that our purpose is to explore if there is convergence in performance on the individually administered and large-scale tests, we need to break down the heritability and environmental variance estimates further. Specifically, if the large-scale reading test is tapping into and assessing skills that are also measured by the IA literacy tests then this would be evidenced, in part, by common influences of genes and environmental factors. The high genetic correlations between performance on the large-scale reading test and the IA literacy tests indicate that common genes influence performance on both types of tests. This particularly applies to the correlation between the large-scale reading and IA comprehension tests. However, common genes can affect multiple academic domains. There is evidence for substantial overlap in genes across academic domains (Kovas et al., 2005; Plomin & Kovas, 2005; Plomin, Kovas, & Haworth, 2007), which was supported by the moderate to high genetic correlations among performance on the large-scale numeracy test and the IA literacy skills that emerged in our analyses.

Of particular interest in our data are the unique environmental correlations among the IA literacy tests and school reading. If the unique environment is genuinely influencing performance in the IA literacy tests and those same literacy skills are being tapped by school reading, then we would expect some positive unique environmental correlations. The unique environmental correlations between reading and the literacy skills indicate that there are environmental factors that contribute to both school reading and the underlying literacy skills. The substantial genetic and unique environmental correlations between school reading and the literacy skills support the phenotypic findings that a portion of performance in school reading is capturing those component literacy skills. Furthermore, this relationship between the component literacy skills is considerably stronger with school reading than school numeracy.

Limitations

The greatest limitation with our study is a consequence of using a sub-sample of a larger study. The sample size for our genetic analyses is small, evident in the large confidence intervals around many of the parameter estimates. However, the parameter estimates are very similar to those found in much larger studies on similar measures, such the Twins Early Development Study in the United Kingdom (Harlaar et al., 2007; Haworth, Kovas, Petrill, & Plomin, 2007), which gives us more confidence in the estimates of this small sample. The sample is also limited in its representativeness for a number of reasons. Firstly, the students were all sampled from the Sydney metropolitan area and are from families who have registered to participate in twin studies. As such they are unlikely to be representative of

Australian school students. Indeed, when compared to the national reports on the NAPLAN at least three quarters of our NAPLAN students were performing above the NSW average. Secondly, only those participants in the full ILTS study who sent in results of large-scale assessments were included in this paper. Those from the Grade 2 ITLS assessment who returned this information and who are included in this paper performed, on average, slightly better in the individually administered tests than those who did not return results from large-scale tests, although given the small differences between groups, it seems unlikely that the lack of participation is driven by very poor results in the large-scale tests. Moreover, the heritability estimates at Grade 2 in the full ILTS sample were very close to the estimates in this sub-sample of Australian students who returned results of the large-scale tests. This lends support to the sub-sample in these analyses being representative of the ILTS sample as a whole.

Conclusions

One key finding was that a few individually administered tests of component literacy skills accounted for a reasonable amount of variation in performance on large-scale reading tests. The effectiveness of these literacy skills at predicting performance on large-scale tests was greater for reading than for numeracy. These results indicate that large-scale reading tests are preferentially testing, to some extent, a student's reading skills as assessed by well-accepted individually administered tests of component skills. In contrast to some of the reports in newspapers, these large-scale tests are not simply measures of student anxiety regarding exams, or their ability to colour in multiple-choice bubbles (Coulson, 2011).

Still, the results reported in this article are not evidence that the results from large-scale tests are sufficiently accurate for all the purposes for which parents, schools, or politicians attempt to use them. As is the case with any test the validity of large-scale tests does not rest solely on their measurement characteristics, but also on an accurate and appropriate interpretation of their results.

The other key finding is that performance on these Australian large-scale reading and numeracy tests are heritable to about the same degree as other measures of literacy and numeracy, whether individually administered (e.g. Byrne et al., 2009) or based on teacher assessment (e.g. Harlaar et al., 2007), from a variety of studies and countries. This evidence further supports the validity of the BST and NAPLAN.

Acknowledgments

Funding

This work was supported by the Australian Research Council [grant numbers DP0663498, DP0770805].

The Australian Twin Registry is supported by an enabling grant from the National Health and Medical Research Council. We thank the Australian Twin Registry, our testers, and the twins and parents involved.

References

Athanasou J. NAPLAN and MySchool survey. Independent Education Union of Australia. 2010

- Australian Curriculum Assessment and Reporting Authority. NAPLAN achievement in reading, writing, language conventions and numeracy. Sydney: Author; 2008. Retrieved from <http://www.nap.edu.au/results-and-reports/national-reports.html>
- Australian Curriculum Assessment and Reporting Authority. Submission to the Senate Education, Employment and Workplace Relations Committee: Inquiry into the administration and reporting of NAPLAN testing [Appendix C: NAPLAN technical paper]. Canberra: Parliament of Australia; 2010a. (Submission 261)Retrieved from http://www.aph.gov.au/Parliamentary_Business/Committees/Senate_Committees?url=eet_ctte/completed_inquiries/2010-13/naplan/submissions.htm
- Australian Curriculum Assessment and Reporting Authority. Writing: Narrative marking guide. Sydney: Author; 2010b.
- Australian Curriculum Assessment and Reporting Authority. Cronbach alpha formula and reliabilities - NAPLAN 2008 - 2012. ACARA; Sydney: 2013 Jun 28.
- Boker S, Neale M, Maes H, Wilde M, Spiegel M, Brick T, ... Fox J. OpenMx: An open source extended structural equation modeling framework. *Psychometrika*. 2011; 76:306–317. [PubMed: 23258944]
- Byrne, B. The foundation of literacy: The child's acquisition of the alphabetic principle. Hove, UK: Psychology Press; 1998.
- Byrne B, Coventry WL, Olson RK, Hulslander J, Wadsworth S, DeFries JC, ... Samuelsson S. A behaviour-genetic analysis of orthographic learning, spelling and decoding. *Journal of Research in Reading*. 2008; 31:8–21.
- Byrne B, Coventry WL, Olson RK, Samuelsson S, Corley R, Willcutt EG, ... DeFries JC. Genetic and environmental influences on aspects of literacy and language in early childhood: Continuity and change from preschool to grade 2. *Journal of Neurolinguistics*. 2009; 22:219–236. [PubMed: 20161176]
- Byrne B, Delaland C, Fielding-Barnsley R, Quain P, Samuelsson S, Høien T, ... Olson R. Longitudinal twin study of early reading development in three countries: Preliminary results. *Annals of Dyslexia*. 2002; 52:47–73.
- Byrne B, Olson RK, Samuelsson S, Wadsworth S, Corley R, DeFries JC, Willcutt E. 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 J, ... Olson R. Longitudinal twin study of early literacy development: Preschool through Grade 1. *Reading and Writing*. 2007; 20:77–102.
- Byrne B, Wadsworth S, Corley R, Samuelsson S, Quain P, DeFries JC, ... Olson RK. Longitudinal twin study of early literacy development: Preschool and kindergarten phases. *Scientific Studies of Reading*. 2005; 9:219–235.
- Cain K, Oakhill J, Bryant P. Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology*. 2004; 96:31–42.
- Coulson, J. Why my kids won't sit the NAPLAN. *Sydney Morning Herald*. 2011 May 9. Retrieved from <http://www.smh.com.au/federal-politics/society-and-culture/why-my-kids-wont-sit-the-naplan-20110508-1ee67.html>
- Coventry WL, Keller MC. Estimating the extent of parameter bias in the classical twin design: A comparison of parameter estimates from extended twin-family and classical twin designs. *Twin Research and Human Genetics*. 2005; 8(3):214–233. [PubMed: 15989749]
- Curriculum Corporation. Statements of Learning for English. Carlton South: Author; 2005. Retrieved from http://www.mceecdya.edu.au/verve/_resources/SOL_English_Copyright_update2008_file.pdf
- Ehri, LC. Development of sight word reading: Phases and findings. In: Snowling, MJ.; Hulme, C., editors. *The science of reading: A handbook*. Oxford, UK: Blackwell Publishing Ltd; 2005. p. 135-154.
- Falconer, DS.; Mackay, TFC. *Introduction to quantitative genetics*. 4. Harlow, UK: Longman; 1996.
- Field, A. *Discovering statistics using IBM SPSS Statistics*. 4. London: Sage Publications; 2013.

- Griffin D, Gonzalez R. Correlational analysis of dyad-level data in the exchangeable case. *Psychological Bulletin*. 1995; 118:430–439.
- Harlaar N, Cutting L, Deater-Deckard K, DeThorne L, Justice L, Schatschneider C, ... Petrill S. Predicting individual differences in reading comprehension: A twin study. *Annals of Dyslexia*. 2010; 60:265–288. [PubMed: 20814768]
- Harlaar N, Dale PS, Plomin R. From learning to read to reading to learn: Substantial and stable genetic influence. *Child Development*. 2007; 78:116–131. [PubMed: 17328696]
- Harlaar N, Hayiou-Thomas ME, Plomin R. Reading and general cognitive ability: A multivariate analysis of 7-year-old twins. *Scientific Studies of Reading*. 2005; 9:197–218.
- Hart SA, Logan JAR, Soden-Hensler B, Kershaw S, Taylor J, Schatschneider C. Exploring how nature and nurture affect the development of reading: An analysis of the Florida twin project on reading. *Developmental Psychology*. 2013
- Hart SA, Petrill SA, Thompson LA, Plomin R. The ABCs of math: A genetic analysis of mathematics and its links with reading ability and general cognitive ability. *Journal of Educational Psychology*. 2009; 101:388–402. [PubMed: 20157630]
- Haworth CMA, Kovas Y, Petrill SA, Plomin R. Developmental origins of low mathematics performance and normal variation in twins from 7 to 9 years. *Twin Research and Human Genetics*. 2007; 10:106–117. [PubMed: 17539370]
- Helwig R, Rozek-Tedesco MA, Tindal G, Heath B, Almond PJ. Reading as an access to mathematics problem solving on multiple-choice tests for sixth-grade students. *The Journal of Educational Research*. 1999; 93:113–125.
- Hoover WA, Gough PB. The simple view of reading. *Reading and Writing*. 1990; 2:127–160.
- Kaplan, E.; Goodglass, H.; Weintraub, S. Boston naming test. 2. Baltimore, MD: Lippincott Williams & Wilkins; 2001.
- 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.
- Keller MC, Coventry WL. Quantifying and addressing parameter indeterminacy in the classical twin design. *Twin Research and Human Genetics*. 2005; 8(3):201–213. [PubMed: 15989748]
- Kenny, DA.; Kashy, DA.; Cook, WL. Dyadic data analysis. New York, NY: Guildford Press; 2006.
- Kovas Y, Harlaar N, Petrill S, Plomin R. “Generalist genes” and mathematics in 7-year-old twins. *Intelligence*. 2005; 33:473–489. [PubMed: 19319204]
- Masters, G.; Lokan, J.; Doig, B.; Khoo, ST.; Lindsey, J.; Robinson, L.; Zammit, S. Profiles of learning: The basic skills testing program in New South Wales 1989. Hawthorn: Australian Council for Educational Research; 1990.
- Nichols RC, Bilbro WC. The diagnosis of twin zygosity. *Acta Genetica*. 1966; 16:265–275.
- Norton S. The responses of one school to the 2008 Year 9 NAPLAN numeracy test. *Australian Mathematics Teacher*. 2009; 65:26–37.
- NSW Department of Education, E. M. a. S. A. D. Writing tasks marking procedures. Sydney: Author; 2006.
- Oliver B, Harlaar N, Hayiou Thomas ME, Kovas Y, Walker SO, Petrill SA, ... Plomin R. A twin study of teacher-reported mathematics performance and low performance in 7-year-olds. *Journal of Educational Psychology*. 2004; 96:504–517.
- Petrill S, Logan J, Hart S, Vincent P, Thompson L, Kovas Y, Plomin R. Math fluency is etiologically distinct from untimed math performance, decoding fluency, and untimed reading performance: Evidence from a twin study. *Journal of Learning Disabilities*. 2012; 45:371. [PubMed: 21890908]
- Plomin, R.; DeFries, JC.; Knopik, VS.; Neiderhiser, JM. Behavioral genetics. 6. New York: Worth; 2013.
- Plomin R, Kovas Y. Generalist genes and learning disabilities. *Psychological Bulletin*. 2005; 131:592–617. [PubMed: 16060804]
- Plomin R, Kovas Y, Haworth CMA. Generalist genes: Genetic links between brain, mind, and education. *Mind, Brain, and Education*. 2007; 1:11–19.

- Samuelsson S, Byrne B, Olson RK, Hulslander J, Wadsworth S, Corley R, ... DeFries JC. 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, ... Olson R. Environmental and genetic influences on prereading skills in Australia, Scandinavia, and the United States. *Journal of Educational Psychology*. 2005; 97:705–722.
- Samuelsson S, Olson R, Wadsworth S, Corley R, DeFries J, Willcutt E, ... Byrne B. Genetic and environmental influences on prereading skills and early reading and spelling development in the United States, Australia, and Scandinavia. *Reading and Writing*. 2007; 20:51–75.
- Smith, M. Getting SMART with data in schools: Lessons from NSW. Paper presented at the Australian Council for Educational Research Conference on Using Data to Support Learning; Melbourne, VIC. 2005 Aug.
- Torgesen, JK.; Wagner, RK.; Rashotte, CA. Test of word reading efficiency. Austin, TX: PRO-ED; 1999.
- Unsworth L, Chand E. Bridging multimodal literacies and national assessment programs in literacy. *Australian Journal of Language & Literacy*. 2009; 32:245–257.
- Wasson, D. Large cohort testing – How can we use assessment data to effect school and system improvement?. Paper presented at the Australian Council for Educational Research Conference on Assessment and Student Learning: Collecting, Interpreting and Using Data to Inform Teaching; Perth, WA. 2009 Aug. Retrieved from http://research.acer.edu.au/research_conference/RC2009/18august/4
- Willett, L.; Gardiner, A. Testing Spelling—Exploring NAPLAN. Paper presented at the AATE /ALEA National Conference on Bridging Divides: Ensuring Access, Equity and Quality in Literacy and English Education; Hobart, TAS. 2009 Jul.
- Williams L. Down and dirty with grammar. *Literacy Learning: The Middle Years*. 2009; 17:11–20.
- Woodcock, RW. Woodcock reading mastery tests. Circle Pines, MN: American Guidance Service; 1989.
- Wu M. Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*. 2010; 29:15–27.
- Wu M. The use of NAPLAN data for english language teaching. *Idiom*. 2011; 47:38–41.

Table 1

Means and standard deviations of the individually administered (IA) tests for participants who did and did not return results of large-scale tests.

Measure	<u>Returned large- scale test results</u>		<u>Did not return large-scale test results</u>	
	Mean	(SD)	Mean	(SD)
IA vocabulary	35.0	(5.6)	33.9	(6.1)
IA word reading	112.3	(13.1)	108.6	(14.6)
IA comprehension	110.1	(10.4)	107.8	(10.5)

IA: individually administered.

Table 2
Pearson’s correlations between BST, NAPLAN, and the individually administered literacy (IA) tests.

School domain	Test type	IA comprehension			IA word reading			IA vocabulary		
		<i>r</i>	<i>Z_{diff}</i>	<i>p</i>	<i>r</i>	<i>Z_{diff}</i>	<i>p</i>	<i>r</i>	<i>Z_{diff}</i>	<i>p</i>
LS reading	BST	.64	−0.01	.991	.61	0.37	.714	.46	0.60	.549
	NAPLAN	.64			.57			.39		
LS numeracy	BST	.41	−0.71	.477	.45	0.01	.993	.37	−0.02	.982
	NAPLAN	.50			.45			.38		

IA: individually administered test; LS: large-scale test; *Z_{diff}*: *Z*-difference score.

Phenotypic correlations among the individually administered literacy (IA) tests and the large-scale (LS) tests.

Table 3

Measure	IA vocabulary	IA word reading	IA comprehension	LS reading	LS numeracy
IA vocabulary	1	.26	.38	.43	.37
IA word reading		1	.70	.59	.45
IA comprehension			1	.64	.45
LS reading				1	.62
LS numeracy					1

IA: individually administered test; LS: large-scale tests.

p < .001 for all correlations.

Table 4

Multilevel regression model predicting performance on large-scale (LS) tests from individually administered (IA) literacy tests.

Variable	Coefficient	df	t	p	pseudo R^2
LS reading					.49
IA comprehension	.33	237	5.07	< .001	
IA word reading	.31	238	4.81	< .001	
IA vocabulary	.23	230	4.46	< .001	
LS numeracy					.28
IA comprehension	.14	224	1.80	.073	
IA word reading	.31	236	4.14	< .001	
IA vocabulary	.25	236	4.15	< .001	

IA: individually administered test; LS: Large-scale tests. Pseudo R^2 was calculated as the proportion of variance explained by the predictors in the model.

Table 5

Intraclass twin correlations and proportion of variance due to genes, shared environment, and unique environment for individually administered (IA) literacy tests and large-scale (LS) reading and numeracy tests.

Measure	MZ	DZ	A	C	E
LS reading	.51	.38	.45 [0, .71]	.12 [0, .51]	.43 [.29, .63]
LS numeracy	.62	.32	.64 [.21, .76]	0 [0, .35]	.36 [.24, .52]
IA comprehension	.65	.36	.68 [.25, .78]	0 [0, .37]	.32 [.22, .47]
IA word reading	.77	.32	.81 [.59, .87]	0 [0, .20]	.19 [.13, .29]
IA vocabulary	.65	.40	.42 [0, .74]	.21 [0, .59]	.37 [.26, .52]

IA: individually administered test; LS: Large-scale tests; MZ: monozygotic; DZ: dizygotic.
95% confidence intervals are in parentheses.

Table 6

Genetic and unique environmental correlations between individually administered (IA) literacy tests and large-scale (LS) reading and numeracy tests.

	Genetic correlation	Shared environment correlation	Unique environment correlation
LS reading			
IA comprehension	1 [.68, 1]	−1 [−1, 1]	.25 [.04, .45]
IA word reading	.83 [.56, 1]	1 [−1, 1]	.35 [.13, .54]
IA vocabulary	.45 [−1, 1]	.73 [−1, 1]	.28 [.04, .49]
LS numeracy			
IA comprehension	.68 [.43, 1]	−.20 [−1, 1]	.04 [−.19, .27]
IA word reading	.64 [.39, 1]	−1 [−1, 1]	.21 [−.01, .45]
IA vocabulary	.41 [−1, 1]	1 [−1, 1]	.31 [.07, .51]

IA: individually administered; LS: Large-scale tests. 95% confidence intervals are in parentheses.