The Etiology of Individual Differences in Second Language Acquisition in Australian School Students: A Behavior-Genetic Study

William Coventry
University of New England

Inés Antón-Méndez
University of New England

Elizabeth M. Ellis
University of New England

Christina Levisen
University of New England

Brian Byrne
University of New England and Linköping University

Victor H. P. van Daal
University of Stavanger

Nick C. Ellis
University of Michigan

We present one of the first behavior-genetic studies of individual differences in school students' levels of achievement in instructed second language acquisition (ISLA). We

Correspondence concerning this article should be addressed to William Coventry, Discipline of Psychology, University of New England, Armidale, NSW 2351, Australia. Internet: coventrywill@gmail.com

1

assessed these language abilities in Australian twin pairs (maximum N pairs $= 251$) by means of teacher ratings, class rankings, and self-ratings of proficiency, and used the classic twin design to estimate the relative influences of genes, shared (family/school) environment, and unique environment. Achievement in ISLA was more influenced by additive genetic effects (72%, 68%, and 38% for teacher ratings, class rankings, and twin self-ratings, respectively) than by shared environment effects, which were generally not substantial (20%, 07%, and 13%). Genetic effects distinct to speaking and listening, on the one hand, and reading and writing, on the other, were evident for the twin self-ratings. We discuss the limitations and implications of these findings and point to research questions that could profitably be addressed in future studies.

Introduction

There is now a considerable amount of research which explores the influence of genetic endowment on levels of attainment in school subjects (for summaries, see Byrne, Khlentzos, Olson, & Samuelsson, 2010; Grigorenko, 2007; Kovas & Plomin, 2006; Plomin, DeFries, McClearn, & McGuffin, 2008; Plomin, Kovas, & Haworth, 2007). Findings from this research have practical as well as scientific implications. Scientifically, they form part of the "nature-nurture" debate, a topic of abiding interest, particularly as it concerns language (Christiansen & Chater, 2008; Elman et al., 1996; Pinker, 1994). Practically, they can inform and guide educational policies, practices, and research. For instance, if heritability of a subject is low, teachers would be constrained in the degree to which they can attribute success and failure to some sort of natural ability, and research would instead need to focus on identifying how the environment (school policies, teaching practices, family support, and so on) affects progress and ultimate levels of success in that subject. If the heritability of a subject is high, as it is in the case of reading ability when literacy is taught intensively from an early age (Samuelsson et al., 2008), it can be expected that some children will struggle with it much more than others, and policies and practices designed to compensate these children would deserve close attention. (For an extended essay on what behavior-genetic research means and does not mean for education and how it can inform educational practices and policies, see Byrne, Khlentzos et al., 2010. For examples of how genetic and environmental factors may interact to determine educational attainment, see Samuelsson et al., 2008, and Taylor, Roehrig, Soden Hensler, Connor, & Schatschneider, 2010.)

In this article, we report on the heritability of instructed second (and subsequent) language acquisition (henceforth ISLA) in adolescent students in Australia, using pairs of twin students where both are studying a second language (L2). We highlight the issue of instruction in this acronym to make it explicit that we know of no evidence on the degree to which sources of individual differences within learners, the main focus of this study, hold equally across natural and educational settings for additional language learning, and hence we make no assumption that our findings generalize beyond the educational (instructed) settings that hold for Australian school students. As Ellis (2008) notes, in SLA there is "a general assumption that the learning that takes place in natural and educational settings is very different in nature" (p. 288), and establishing the nature of the interface between instructed and less formally acquired knowledge of a second language is a continuing challenge in SLA research (Ellis, 2005).

The study reported here aimed at investigating the extent to which ability to learn a second language in school is heritable, in other words, the degree to which observed individual variability among students in formal language learning success might be attributable to genes vis-à-vis environmental influence from factors such as family, friends, school, and so on. A secondary goal was to inspect the evidence for partially independent genetic influence on spoken versus written ISLA factors, given existing prior evidence that spoken (speaking and listening) and written (reading and writing) first language abilities are subject to differing behavior-genetic principles.

Theoretical and Methodological Background to the Behavior-Genetic Framework

The use of twins to separate genetic and environmental factors governing individual differences relies on the fact that monozygotic (MZ) twins share all of their genes whereas dizygotic (DZ) twins share, on average, half of their segregating genes (i.e., genes that make people different from each other). It also relies on an assumption that within families both types of twins share equally similar environments, known as the equal environments assumption (EEA), which allows researchers to distinguish family-based (i.e., shared or common) environment influences from those that affect one twin but not the other (i.e., nonshared or unique environment). In the case of school achievement, the shared environment could include the school, if both twins attend the same one, and the teacher, if both are in the same class. The nonshared environment could include the teacher, if the twins have different teachers, or a stronger

affinity in just one twin in the pair for the learning of a particular language. Any measurement error is incorporated into the nonshared environment because measurement error is assumed to be independent from one person to the next. The results from twin studies are generally expressed in terms of the proportion of total variance in a trait attributable to genetic (A estimates expressed as a^2), shared environment (C estimates expressed as c^2), and nonshared environment (E estimates expressed as e^2) sources. These parameters sum to unity (i.e., 100%).

For characteristics that are fully determined by genes, MZ twins will be identical and DZ twins 50% alike, on average. For characteristics that are fully determined by shared environment, both types of twins will be identical. For characteristics that are fully determined by nonshared environment, twins will be no more alike that randomly selected individuals of the same age and gender. These are idealized situations, but departures from these ideals can be used to estimate the mix of genetic and environmental (shared and nonshared) factors affecting the characteristic of interest. Heritability can be estimated by doubling the difference between the correlations for monozygotic and dizygotic twins; shared environment influence can be estimated from the difference between the MZ correlation and the heritability value; and nonshared environment can be estimated from the difference between the MZ correlation and unity (Falconer, 1960). The current study also employs multivariate analysis in a comparison of spoken and written language in ISLA (see Plomin et al., 2008, for an introduction to twin methodology, including quantitative estimation techniques).

It may occur to readers unfamiliar with twin methodology that monozygotic twins are not only more alike genetically but are also treated more similarly by others because they look alike. In the present situation, that might play out as a bias for teachers to rate MZ twins as more similar, over and above any real similarity, in which case part of the difference in any correlations between MZ and DZ twins would be spurious. This would amount to a violation of the EEA, and hence a threat to the accuracy of estimates of genetic and environmental influence by overestimation of heritability. The EEA has been the focus of considerable research (Derks, Dolan, & Boomsma, 2006; Kendler, Neale, Kessler, Heath, & Eaves, 1994; Plomin et al., 2008; Richardson & Norgate, 2005; Slutske et al., 1997) and has been found to hold in many situations, along with occasional violations. In our work, we have presented evidence on literacy achievement that is relevant to the issue (Byrne, Coventry et al., 2010). We showed that estimates of the heritability of literacy achievement were unaffected by whether twin pairs each had the same teacher, which could

in principle involve a violation of the EEA because a teacher might instruct an MZ pair more similarly than a DZ pair, or a different teacher, which could not involve a violation of the EEA. Therefore, although possible violations of the EEA cannot be discounted, they need not be assumed to be in place, and often are not. Note, too, that as far as possible rating biases are concerned, EEA is not an issue when the twins are rating themselves, as is the case with one of our measures and as was the case in the research of Vinkhuyzen, van der Sluis, Posthuma, and Boomsma (2009), as described below.

Contributions of Genetics to School Achievement and First Language Literacy

All aspects of twin school achievement studied to date are subject to moderate-to-substantial genetic influence. For example, Oliver et al. (2004) reported data from 7-year-old twins in the Twins Early Development Study (TEDS) in the United Kingdom, which employed UK National Curriculum (NC) teacher ratings of various aspects of mathematics. Genes accounted for around two-thirds of the variability, shared environment for a small proportion, and unique environment for a quarter to a third. Wainwright, Wright, Luciano, Geffen, and Martin (2005) analyzed data on calculations and mathematical problem solving from 17-year-olds in Queensland, Australia, reporting a total heritability of 73%, a marginally significant shared environment effect of 11%, and a unique environment effect of 19%. A preliminary report from the TEDS project on heritability of science achievement (Plomin et al., 2007) showed that at age 9 teacher ratings using the NC criteria yielded a heritability estimate of 63%, and at age 10 of 47%.

Literacy abilities have been extensively studied with the behavior-genetic framework. Using the twin design described above, reading and spelling abilities have been shown to be heritable from the earliest school years, with an estimate of 70% in kindergarten in a combined US/Australian/Scandinavian sample (Byrne et al., 2005), rising to 82% in Grade 2 (Byrne et al., 2009). The shared environment influence on literacy has been shown to be relatively low, 3% in Grade 2, a value that was not statistically significant.

Scope of the Present Study

In the present study we also have the opportunity to explore possible differences between spoken and written language learning *within* ISLA because we separately evaluate ISLA for spoken (speaking and listening) and written language

(reading and writing). There are good reasons to believe that, when it comes to first language (L1) acquisition, these two forms of language are subject to differing principles (Byrne, 2011; Byrne, Khlentzos et al., 2010; Byrne & Liberman, 1999; Liberman & Liberman, 1992). In brief, these reasons are: all human communities have speech but only a minority have developed writing independently rather than adopting it from other societies; speech is presumably as old as the species, but writing only emerged a few thousand years ago; the normal child need only be among mature competent speakers to develop speech whereas reading almost always requires explicit instruction; and children with reading difficulties are, for the most part, competent speakers. It is an open question, however, whether this distinction holds for ISLA, and we take the opportunity to explore this question here.

At the time this project began, the only published large-scale twin study of second language learning was from Vinkhuyzen et al. (2009), in which self-report data were collected from 1,685 pairs of Dutch twins, aged 12 to 26 years. It is not clear what mix of naturalistic and educational contexts held as these twins were acquiring other languages. Participants were asked to rate their ability to speak one or more foreign languages in relation to the general population (i.e., less competent, average competence, more competent, exceptional skill). Heritability was estimated at 71%, with the rest of the variance best attributed to nonshared environment. Our project lacks the large numbers of the Dutch study, but it includes more comprehensive assessments of ISLA. In a study by Ando (1992) there was evidence of genetic influence on English learning in Japanese fifth graders, but the small numbers of twin pairs (7 MZ, 4 DZ) made quantitative estimates inappropriate. Based on the data from Vinkhuyzen et al., however, SLA, instructed and otherwise, appears to be about as heritable as early literacy progress.

Subsequent to the initiation of the project, members of the group have published research that utilizes teacher report data on 604 pairs of 14-year-old twins studying a second language in UK schools (Dale, Harlaar, Haworth, & Plomin, 2010). The data were from the UK NC standards, which we also employed in the current study (see below). The heritability of ISLA was estimated at 67%, shared environment influence at 13%, and nonshared environment influence at 20%, a similar pattern to that reported by Vinkhuyzen et al. (2009) using a very different measure. Dale et al. also reported relatively high genetic correlations between ISLA and two measures of L1, although sources of shared environment influences appeared to differ. The current study is therefore in part a replication of Dale et al., but it includes two extra measures of ISLA as well as exploring both written and spoken language achievement; it also encompasses a wider

age range among participants, and it is situated in a different country. All of these aspects furnish tests of the generality of both the Vinkhuyzen et al. and Dale et al. findings.

Our particular research concerns ISLA in primary and secondary school classrooms in Australia, where the teaching and curricula vary across the eight states. In Victoria, for instance, study of one of eight languages other than English (known as LOTEs in the Australian school curriculum) is compulsory from kindergarten to Year 10 (where Year 12 is the final year of schooling), whereas in the Australian Capital Territory there is no mandated language study; rather, it is left to the individual school to decide which language(s) to offer to which grades. It is not possible, therefore, to generalize about the kind of instruction the twins in our study received, other than to recognize that differing curricula, teaching methods, textbooks, and resources ensured a heterogeneous teaching environment for ISLA.

In this research we have employed questionnaires to assess ISLA, which we detail in the Method section. Here we consider the merits of this approach in general because it can be argued that more direct measures, such as objective testing in the face-to-face situation, are preferable. First, although data gathered in that way would be of high value, doing so runs into difficulties because in studies of relative levels of achievement within and across twin pairs a common metric is required. Thus, a test for French and one for Indonesian are difficult to equate (and the twins we have available are studying a considerable variety of languages). This would be less of an impediment if twins within each pair were studying the same language because within-pair similarity, captured in intraclass MZ and DZ correlations, is the primary basis for estimating the effects of genes and environment on achievement. However, 20% of the pairs in our sample were studying different languages, and restricting analyses to the remaining 80% would in turn restrict statistical power. The teacher-based rating scale that we have used is applicable to all languages and thus offers more in the way of a common measure. Second, the twins in our sample are spread across all of Australia, making face-to-face testing prohibitively expensive. Internet-based assessment is a possibility as a cheaper alternative, and we hope to develop and exploit such measures in future research (although the language variety problem would still remain). Third, one of the types of assessments that we have used, the NC teacher ratings developed in the United Kingdom, has been used extensively in the TEDS to examine literacy, language, mathematics, and science performance in English and Welsh samples. Importantly, when the NC ratings have been compared with objective testing, either face-to-face or telephone administered, they have fared well, with a correlation of $r = .68$ in the

case of reading and a Cohen's kappa (Cohen, 1960) of $k = .73$, indicating good agreement, in the case of mathematics (Oliver et al., 2004). In a meta-analysis of teacher ratings versus standardized testing, Hoge and Colardaci (1989) report a median correlation of $r = .66$, despite considerable variation in methods of teacher assessment. Teacher ratings in general have the advantage of being based on observations over extended periods, and they have been shown to add to the prediction of future outcomes over and above objective assessments (e.g., Alvidrez & Weinstein, 1999; McCall, Evahn, & Kratzer, 1992).

Method

Participants

The twins (59% female) in this study ranged from grades 4 to 12. For each measure, the numbers of participants, split by sex, are presented in Table 1. (The numbers of MZ and DZ pairs with data for each of the three assessments collected in this study are presented later in Table 3.) Because sampling issues are complex in the behavior-genetic study of twins, we provide some detail for readers here.

Four thousand and thirty one families of twins from the Australian Twin Registry (ATR) were initially contacted and parents were asked: (1) if both their twins were studying an additional language and, if so, (2) if they would be willing to participate in our study. We also asked the parents if they would consent for us to contact the twins' teachers and, if so, we requested the teachers' contact details. We received responses from 388 families. We cannot accurately identify a response rate from these figures because we do not know in how many families both twins were studying a second language. However, in previous research using the ATR to recruit young twins via their families we have had a response rate of around 62% (Byrne et al., in press). We mailed questionnaires to the 388 families, and out of these 308 returned material, a response rate of 79%. In some cases, neither or only one of the twin children returned a self-assessment, although information from parents about the home background was included. These cases were retained in the preliminary correlations and *t*-tests but not in the behavior-genetic analyses (see the Results section). We also excluded from the behavior-genetic analyses those twins who were studying at school the language spoken at home (5%), leaving a maximum of 251 pairs for the behavior-genetic analyses.

We were supplied with permission to approach 469 teachers, with educational districts stipulating that we do so via school principals, who retained the authority to allow teacher involvement or not. We received responses from 258

Table 1 Means (standard deviations, Ns) of the second language acquisition measures separately by Zygosity and Sex, with t test results

Sex	Females	Males	t (df)	95% confidence interval on the mean difference
Twin self-rating	1.36 (0.86, 174)	1.49 (0.85, 126)	−1.39 (298)	−0.33 to 0.06
Teacher NC ratings	3.88 (1.62, 117)	3.23 (1.44, 79)	2.96 (194)**	0.21 to 1.10
Teacher class rankings	39.99 (22.89, 80)	46.02 (27.42, 53)	1.10 (131)	−14.71 to 2.66

Zygosity	MZ	DZ	t (df)	95% confidence interval on the mean difference
Twin self-rating	1.40 (0.82, 95)	1.45 (0.86, 180)	−0.48 (273)	−0.26 to 0.16
Teacher NC ratings	4.23 (1.76, 53)	3.54 (1.46, 99)	2.29 (150)	0.16 to 1.22
Teacher class rankings	38.11 (24.99, 39)	41.17 (25.17, 67)	0.66 (104)	−13.09 to 6.96

Note. After Bonferroni correction for multiple testing, ** $= p < .01$. These correlations use one twin selected at random from the pair.

teachers, a minimum response rate of 59% (minimum because we do not know how many principals neglected or declined to pass the questionnaires on to the teachers in that school, though we suspect this number was small). Where twins within a pair were taught by separate teachers, we could only use those responses where both teachers returned a questionnaire for the behavior-genetic analyses, which excluded responses from 30 teachers. Some teachers omitted one of the two assessments we requested (see details under Materials), and some omitted both on various grounds (such as having taught the children in very large classes or only having taught a child for a few months of the year).

Materials
Questionnaires were completed by each of the twins independently, by one of their parents, and by each of their second language teachers. Questionnaires comprised a variety of measures and we present here those that are relevant to the current article.

The twin questionnaire included a question for each twin to rate his or her own language learning skill, referred to as *twin self-rating* hereafter. They separately rated their ability at speaking, listening, reading, and writing in a second language on a 5-point scale of "very good" (1), "good" (2), "ok" (3), "poor" (4), and "very poor" (5).

The teachers were firstly asked to give a class ranking of each twin's second language ability (e.g., 3rd out of 15 pupils). This was then converted to a

percentage (e.g., if they were 3rd out of 15 pupils, their score was 20%), so a low percentage represents a high ranking (referred to hereafter as *rank*). Teachers then rated the skill level each twin had attained in the four key learning areas of the standardized UK NC assessment guidelines.[1] For each ability, they selected which of the nine levels of the NC guidelines best described the child's attainment. This score is hereafter referred to as the *NC rating*. The nine levels of the NC guidelines comprise increasing proficiency, with each level accompanied by a brief description of the type and range of performance expected of a pupil having reached that particular level (for full details, see note 1). These descriptions provide the basis for making judgments and are meant to ensure consistency in the evaluation process.

In addition to the standard demographic questions and questions assessing zygosity (Nichols & Bilbro, 1966), both the twins and their parents completed a number of questions assessing whether there were any languages that the twins either spoke or were regularly exposed to at home (these questions are available from the authors on request). From this, we ascertained if the languages that the twins learned at school were also part of the home background.

Zygosity

For 86% of the twin pairs, zygosity was determined by the Nichols and Bilbro (1966) questionnaire, based on physical characteristics such as hair and eye color and height and weight, as well as the frequency with which the twins are mistaken by family, friends, and teachers. Nichols and Bilbro report that the questionnaire is 95% accurate when compared to blood-based diagnosis of zygosity. Ambiguous cases were resolved via telephone follow-up that afforded greater clarity on the subjects' responses to the Nichols and Bilbro questionnaire. Zygosity ascertainment in the remaining 14% of pairs was based on DNA tests reported by the parents or either of the twins or the ATR on their behalf, and these were based on either blood or buccal swabs.

Results

Data Preparation

We initially performed a principal components factor analysis of the speaking, listening, reading, and writing items for each of the teacher NC ratings, teacher rank, and twin self-ratings. The goal was to determine if we could combine the four types of items into a single measure, a procedure that increases the reliability of the measure if the factor structure licenses combining items by showing that a single factor provides a parsimonious account of the data (see

Samulesson et al., 2005, for an example within a twin study). For twin self-ratings, teacher NC ratings, and teacher rankings, a single factor accounted for 74%, 88%, and 90% of the variance, respectively, thus supporting the generation of a combined measure for each variable, created by averaging across the four item types. For the teacher NC ratings and teacher ranking variables, no sensible structure emerged from forcing a two-factor structure, with the second factor explaining only 5% additional variance for both variables. However, for twin self-ratings, a two-factor solution with oblimin rotation neatly differentiated speaking and listening (loading .69 and 1.00 on Factor 1) and reading and writing (loading .97 and .85 on Factor 2), with this additional factor explaining 12% more variance and being correlated .67 with the first factor. Hence, there was some evidence to support a two-factor solution for the twin self-ratings. To explore whether the distinction between spoken and written language acquisition that holds for first language acquisition emerges in any way for ISLA, we analyzed the twin self-ratings both as a single factor and two factors in the genetic analyses. As stated, we computed the single-factor variables by averaging across the speaking, listening, reading, and writing items for twin self-ratings (Cronbach's $\alpha = .82$), teacher NC ratings ($\alpha = .92$), and teacher rankings ($\alpha = .96$). For the two-factor solution of twin self-ratings, we separately averaged (1) the speaking and listening items and (2) the reading and writing items. These scales also had high internal reliabilities, $\alpha = .82$ and .84, respectively.

In order to approximate normal distributions, we transformed the twin self-rating and teacher class ranking variables with square root transformations, and the teacher NC ratings with a logarithmic transformation. All variables were then standardized. We also covaried out the effects of (1) school year, (2) years spent studying an additional language, (3) age, and (4) sex, as such effects can artificially inflate the correlations of MZ and DZ twin pairs and thus artificially affect the estimates of genetic and environmental influences. For example, it is to be expected that the number of years studying a language will be reflected in the NC ratings, higher for more years of study, and in a sample of twins that includes a range of years members of twin pairs will look more similar to each other than if the sample had covered just, say, one year. The covariation procedure recalculates the NC ratings (and the two other language measures) by using regression to adjust for the effect of years of study and then analyzes those residuals. This statistically equates the measures for years of study (and for school year, age, and sex, in the other recalculations). We also checked that the factor structures reported earlier remained constant with the effects of these covariates removed and they did.

Table 2 Phenotypic correlations (below diagonal) and *N*s (above diagonal) of second language acquisition measures

	1	2	3
1. Twin self-rating		153	105
2. Teacher NC ratings	−.30**		122
3. Teacher class rankings	.40**	−.24*	

Note. After Bonferroni correction for multiple testing, * = $p < .05$; ** = $p < .01$. These correlations use one twin selected at random from the pair.

Phenotypic Analyses

In behavior genetics, phenotypic analyses are those that ignore the twin structure of the data, in a sense treating twins as unrelated individuals. Our phenotypic analyses comprised *t*-tests and Pearson correlations, and for these we used a Bonferroni correction of the three measures. We did not remove the effects of the covariates from these particular analyses because (1) covarying out the effects of one of the covariates, sex, will have removed any sex differences and (2) the results differed little when these analyses were run on the covaried variables. We randomly selected one twin from each pair to avoid the interdependence of including both members of a twin pair in such analyses, an interdependence that can compromise statistical procedures. (Of course, for behavior-genetic analyses, both members of twin pairs are required.)

We initially used *t* tests to assess whether males and females or monozygotic and dizygotic pairs differed on any of the three measures. There were no significant differences on zygosity, though for sex, females were significantly higher ($p < .01$) than their male counterparts on teacher NC ratings but not on the teacher ranking or twin self-rating variables (see Table 1). For all the twin pair analyses, we controlled for gender to adjust for this difference. As shown in Table 2, the phenotype correlations among the variables were moderate, but were statistically significant, as expected. The scoring of teacher NC ratings was such that greater ability was represented by higher scores. By contrast, for teacher class rankings and twin self-rating, greater ability was represented by lower scores. Hence, the correlations between the teacher NC ratings and both teacher class rankings and twin self-rating were negative, as expected, while the correlation between teacher rankings and twin self-rating was positive.

Behavior-Genetic Analyses

For the behavior-genetic analyses, in addition to using the transformed variables with the effects of the covariates removed, we first removed twin-pair outliers by

Table 3 Twin pair correlations (and Ns) of second language acquisition measures split by Zygosity, and estimates of a^2 (additive genetic), c^2 (common environment), and e^2 (unique environment) variance Components (and their 95% Confidence Intervals)

	MZ		DZ				
	r	N	r	N	a^2	c^2	e^2
Twin self-rating	.51	85	.32	166	.38 (−.03 to .76)	.13 (−.18 to .42)	.49 (.37 to .66)
Teacher NC ratings	.92	40	.56	84	.72 (.49 to 1.03)	.20 (−.11 to .42)	.08 (.05 to .13)
Teacher class rankings	.75	29	.41	53	.68 (.18 to 1.22)	.07 (−.44 to .46)	.25 (.15 to .44)

computing a Mahalanobis distance of both twins for each variable and excluding participants with scores on any one variable greater than a chi-square of 13.82, 2 *df, p* < .001, which is the appropriately conservative *p* value recommended for this test (Tabachnick & Fidell, 1996). Furthermore, we then also removed the 5% of twins whose family spoke at home the same additional language they were being taught at school. These two manipulations corrected what would otherwise be biases in the correlations between twin pairs.

Table 3 presents the correlations between the MZ and DZ pairs. From these, we computed the estimates of additive genes (A), common environment (C), and unique environment (E) using Mx (Neale, 2004). These, as will be remembered from the Background section, are the proportion of total variance in a trait attributable to genetic (A estimates, a^2), shared environment (C estimates, c^2), and nonshared environment (E estimates, e^2) sources, summing up to unity, that is, 100%. The additive genetic effects were substantial for the teacher measures (teachers NC ratings, 72%; class rankings, 68%). However, they were only moderate for twin self-rating (38%), which was influenced more by unique environment effects (49%) than the two teacher variables (teachers NC ratings, 8%; class rankings, 25%). Shared environment effects were quite small for all measures: 13%, 20%, and 07% for twin self-rating, teacher NC ratings, and teacher rankings, respectively.

Table 3 also includes the 95% confidence intervals. They are broad, relative to many behavior-genetic studies, but this is expected with our small sample size (discussed under Limitations below). Further, some of the confidence intervals fall outside the bounds of zero and one. We simply note it is impossible for the estimates of A, C, and E (cf. a^2, c^2, and e^2 results in Table 3) to fall outside these bounds, because these parameters have been squared (so cannot be negative) and are constrained to sum to unity. Yet the confidence intervals, as presented, do indicate the variability expected were different samples used to derive these same estimates. Finally, in the case of A estimates for twin self-ratings (as opposed to teacher NC ratings or teacher class rankings) these

confidence intervals include zero, so this estimate is nonsignificant. We are mindful however that tests of significance are vulnerable to sample size (Field, 2009), and emphasis should be placed on the effect sizes, which are the estimates of A, C, and E presented in Table 3. An effect size of 38%, the A estimate for the self-ratings, is generally regarded as moderate.

The majority of twin pairs were taught (and therefore rated and ranked) by one teacher, 79% in the case of ratings and 82% in the case of rankings. Kovas, Haworth, Dale, and Plomin (2007) showed that in reading, mathematics, and cognitive ability, same-teacher ratings lead to higher estimates of shared environment effects (and conversely lower nonshared environment effects) than do different-teacher ratings, but that they do not affect heritability, as discussed above in relation to early literacy (Byrne, Coventry et al., 2010). When objective tests replace teacher ratings, these differences mostly disappear, suggesting that same-teacher ratings as an assessment method may bias the results. Unfortunately, our numbers are too small to compare same- with different-teacher ratings, but the fact that most of our assessments are from the same teacher for each member of a twin pair leaves open the possibility that even our modest estimates of shared environment may be overestimates. Of course, the same considerations do not apply to twins' self-ratings.

The evidence for two factors of twin self-rating, speaking/listening and reading/writing, was not strong at the phenotypic level (i.e., where the twin structure of the data was ignored). We nevertheless analyzed them separately at the etiological level (that is, using the twin structure of the data to examine the factor structure of the genetic variance, with the environmental variability parsed out). In justification, we note that phenotypic relations, in this case relatively weak separation of the written and spoken language factors, can mask genetic relations (and environmental ones as well). For example, Byrne et al. (in press) have shown that in phenotypic analyses young children's knowledge of letter names and sounds loads with measures of spoken language such as vocabulary as much as it does with later measures of word reading. But genetically, letter knowledge loads strongly on the same factor as does reading and only weakly on the (genetic) spoken language factor. This apparent mismatch between phenotypic and etiological relations can come about because measured (phenotypic) values are the sum of genetic and environmental influences, which are separated in behavior-genetic analyses.

The magnitude of genetic and environmental effects were of a similar order for the two factors, 44% and 34% for genetic influences on speaking/listening and reading/writing respectively, 06% and 10% for shared environment effects

respectively, and 50% and 56% for unique environment effects, respectively. The low levels of shared environment influence for both factors remains as a contrast to the situation for vocabulary and grammatical control in first language acquisition, where it is relatively high, as discussed earlier, but the main interest in this analysis is the relation between speaking/listening and reading/writing *within* ISLA.

To advance our understanding of these relations, we conducted a Cholesky analysis of the two factors. This analysis is similar in principle to a hierarchical regression. Namely, the effects of genes (or the environment) on one variable are assessed after the genetic (or environmental) effects of other, correlated variables are taken into account. This allowed us to firstly assess the etiology of the $r = .67$ correlation between these two factors. This correlation comprised 45% additive genetic, 14% common environment, and 55% unique environment effects. The Cholesky also allowed us to assess whether the two factors were influenced by the same or different genetic and environmental effects. The first Cholesky, with reading/writing entered before speaking/listening, showed that 46% of the genetic variance of speaking and listening was specific to speaking and listening, the balance being shared with reading and writing. Similarly, in a second Cholesky where the order of the variables was reversed, 42% of the genetic variance of reading and writing was specific to reading and writing with the balance shared with speaking and listening. In short, just under half the genetic variance was specific to the reading/writing and speaking/listening factors rather than shared between them. The unique environment effects were similarly specific: 55% and 58% were specific to speaking/listening and reading/writing, respectively, depending on the order of the Choleskys. Hence, just over half the unique environment variance was specific rather than shared between these two factors.

Discussion

The central findings of this study, based on estimates from teachers, were that (1) success at learning a second language in school is heritable, with about two-thirds of the variability among students attributable to genes; (2) the environmental influence from factors that twins in a pair share, such as family, friends in common, and school (i.e., the common environment), was, by comparison with the genetic influence, relatively modest, at up to one fifth of the variability (in the case of NC ratings); and (3) the environmental influence from factors outside of family, shared friends and shared school (i.e., the unique environment), was also modest, up to a quarter of the variability (in the case

of teacher rankings). These results are similar to those reported by Dale et al. (2010), using a UK sample of 14-year-olds.

The estimates derived from twins' self-reports of ability differ in that heritability was lower at 38% and nonshared environmental influence was higher, at about a half of the variability. The reasons for the discrepancy between teachers' and twins' ability estimates are not clear, although it is possible, and perhaps plausible, that in assessing their own performance levels students are influenced by individual-level factors unrelated to academic proficiency and relatively less heritable, such as how much they enjoy a subject and how motivated they are. We employed the self-ratings to examine the relation between spoken and written language skills because there was some evidence that they formed partially separable factors. Both showed similar patterns of genetic and environmental etiology, though multivariate analyses indicated that there was only partial overlap in the sources of both genetic and unique environmental influence. In other words, and focusing just on genes, there appear to be "generalist" genes (i.e., those common to different aspects of a general domain of ability) that affect ISLA for written and for spoken language and also "specialist" genes (those specific to each aspect of ability). However, we acknowledge that larger samples are needed to confirm or refute this tentative conclusion.

It is clear that the idea of "ability" for ISLA is well founded, if by ability is meant inborn characteristics the learner brings to the task. Our data are in line with the results presented in two other studies that put heritability of ISLA at around two-thirds of the variance (Dale et al., 2010; Vinkhuyzen at al., 2009). ISLA thus aligns with other aspects of academic achievement such as literacy and mathematics in being affected by genetic endowment, with most estimates placing heritability as the single largest influence (see Byrne, Khlentzos et al., 2010, for a summary).

As we have emphasized, this study focused on instructed second language learning, rather than on naturalistic acquisition of a second language as might take place within multilingual societies or through family migration to another country. School instruction is often considered to be focused on explicit knowledge, which "involves an attempt to construct explicit rules or generalizations of which the learners will have a high level of awareness" (Ellis, 1997, p. 116). Insofar as this is the case, it could be that ISLA is particularly affected by executive processes of cognition, often collectively referred to as working memory (WM) and broadly implicated in explicit learning. Sawyer and Ranta (2001) summarized their review of the role of WM in L2 learning by stating that "it is likely that WM serves as an arena in which the effects of other components of aptitude are integrated" (p. 342; see also Juffs & Harrington, 2011;

Miyake & Friedman, 1998). There is now considerable evidence for substantial genetic influence on WM components, such as the phonological loop (Baddeley, Gathercole, & Papagno, 1998; Brkanac et al., 2008) and executive processes more generally (Friedman et al., 2008; Monaco, 2007). These processes may be one source of genetic influence on ISLA.

Given the literature reviewed earlier suggesting that spoken (speaking and listening) and written language (reading and writing) are subject to differing principles (Byrne, 2011; Byrne & Liberman, 1999; Liberman & Liberman, 1992), we assessed whether this distinction holds for ISLA. Although no separate factors emerged from the teacher ratings of ISLA, there was some support for partially independent spoken and written factors for the twin self-ratings of ability. When we subjected the subcomponents derived from these self-ratings (written and spoken language) to genetic analyses, we observed that the etiological patterns were similar, with around one-third of the variance attributable to genetic influence, about half to nonshared environment (including measurement error), and about one-tenth to shared environment. More interesting, perhaps, was that written and spoken language showed both overlap and independence in the genes that influenced them in that approximately half the genetic variance in spoken and written ISLA was specific to each and not shared with the other. This finding was mirrored for nonshared environment influence, where there was both overlap and a degree of independence between the two components. We believe that this is the first time a distinction between etiological factors for subcomponents of ISLA has been suggested.

Limitations and Future Research

Our findings should be interpreted in the context of the relatively small sample (for a behavior-genetic study). Small samples make quantitative estimates of genetic and environmental influences less secure, and should be supplemented by larger samples when possible. Small samples also restrict the range of questions that can be addressed. For example, we have been unable to explore the effects on ISLA of degrees of similarity between L1 and L2. Even within Australia, with English as the almost exclusive L1, there is enough variety in L2s taught in schools to examine this interesting issue, had we had sufficient numbers. We have also been unable to explore the effects of age on ISLA. Based on a sensitive period hypothesis, and given that for some students SLA instruction begins in elementary school, we could, with sufficient numbers, test the hypothesis that the etiological pattern differs as a function of the age at which ISLA begins. As a final example, we have not been able to check on an

interesting idea emerging from the research by Ando (1992) mentioned earlier that there is an interaction between method and intelligence in that among the students that he studied more formal types of instruction appeared to be more effective for students of higher intelligence and less formal (immersion-type) instruction was more effective for students of lower overall intelligence. More generally, we have not been able to examine the consequences of variation in teaching techniques for ISLA within our genetically sensitive design.

Despite the sample size limitation, it is worth pointing out that the estimates of genetic and environmental influence on the measures of ISLA that we have presented are in line with those from Dale et al. (2010) and fromVinkhuyzen et al. (2009). Replication is a valuable tool in behavior-genetic research, and can compensate for the lowered confidence one can have in individual studies because of limited sample sizes.

One might expect the family environment to be quite important in a child's ability to learn a second language. For example, the home environment of a non–English-speaking family constitutes a substantial advantage in learning the family's language, should the child study it at school. Such an effect would show up as common environmental influence in a twin study. We removed participants whose family spoke at home the same language they were learning at school, to leave us with a direct account of each child's individual ability to learn a second language to which they had not been regularly exposed. However, we have not adjusted for other forms of multilingualism. If having some degree of multilingualism supports further language acquisition, as suggested by Cenoz (2003) and Ellis (2004), this feature of the study may have contributed to the (modest) shared environment effects that we have noted.

We cannot, and do not, claim that the genetic influence documented here is specific to ISLA. Indeed, it is known that there are genetic factors that spread their influence over a variety of scholastic domains such as language, mathematics, and science (Davis et al., 2008; Hart, Petrill, Thompson, & Plomin, 2009; Kovas & Plomin, 2006). The question of specificity versus generality of genetic influence (and for that matter, for sources of environmental influence) can only be determined empirically, for example by including results in mathematics for the same twins and subjecting the data to multivariate procedures such as Cholesky decomposition. The same applies to the broad cognitive capacities captured under the construct of intelligence. It is likely that some of the genetic influence on ISLA is shared with tests that measure IQ, but that too can only be determined empirically.

Goals for future research include broadening assessment to include other school subjects and cognitive capacities and environmental factors likely to

underpin ISLA, and bolstering ISLA assessment itself with more refined measures. With much larger samples, such as those offered by the higher prevalence of additional language learners throughout European countries, a wider range of issues can be addressed, including those we have mentioned such as the effects of age, gender, teaching methods, and L1/L2 relations. As is well known, and as we have demonstrated (Samuelsson et al., 2008), estimates of genetic and environmental influence depend in part on the samples from which data are collected, and so international collaboration will help refine the total picture of how nature and nurture contribute to the acquisition of second and additional languages.

Revised version accepted 10 February 2011

Note

1 It was decided to use the UK NC rankings because Australia has no national LOTE standards which could be used to rank twins' attainment. The UK NC attainment targets can be found at:
http://curriculum.qcda.gov.uk/uploads/MFL%201999%20pr
ogramme%20of%20study_tcm8–12069.pdf?return=/key-stages-3-and-4/subjects/
modern-foreign-languages/index.aspx

References

Alvidrez, J., & Weinstein, R. S. (1999). Early teacher perceptions and later student achievement. *Journal of Educational Psychology*, *91*, 731–746.

Ando, J. (1992). The effects of two EFL (English as a foreign language) teaching approaches studied by the cotwin control method: A comparative study of the communicative and the grammatical approaches. *Acta Geneticae Medicae et Gemellologiae: Twin Research*, *41*, 335–352.

Baddeley, A., Gathercole, S. E., & Papagno, C. (1998). The phonological look as a language learning device. *Psychological Review*, *105*, 158–173.

Brkanac, Z., Chapman, N. H., Igo, R. P. J., Matsushita, M. M., Nielsen, K., Berninger, V. W., et al. (2008). Genome scan of a nonword repetition phenotype in families with dyslexia: Evidence for multiple loci. *Behavioral Genetics*, *38*, 462–475.

Byrne, B. (2011). Evaluating the role of phonological factors in early literacy development: Insights from experimental and behavior-genetic studies. In S. Brady, D. Braze, & C. Fowler (Eds.), *The phonological deficit hypothesis* (pp. 175–196). New York: Taylor and Francis.

Byrne, B., Coventry, W. L., Olson, R. K., Samuelsson, S., Corley, R., Willcutt, E. G., et al. (2009). Genetic and environmental influences on aspects of literacy and

language in early childhood: Continuity and change from preschool to Grade 2. *Journal of Neurolinguistics*, *22*, 219–236.

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

Byrne, B., Khlentzos, D., Olson, R. K., & Samuelsson, S. (2010). Evolutionary and genetic perspectives on educational attainment. In K. Littleton, C. Wood, & J. K. Staarman (Eds.), *International handbook of psychology in education* (pp. 3–34). Bingley, UK: Emerald Group Publishing.

Byrne, B., & Liberman, A. M. (1999). Meaninglessness, productivity and reading: Some observations about the relationship between the alphabet and speech. In J. Oakhill & R. Beard (Eds.), *Reading development and the teaching of reading: A psychological perspective* (pp. 157–173). Oxford, UK: Blackwell.

Byrne, B., Wadsworth, S., Boehme, K., Talk, A., Coventry, W., Olson, R. K., et al. (in press). How general are the genes influencing learning to read words? *Scientific Studies of Reading*.

Byrne, B., Wadsworth, S., Corley, R., Samuelsson, S., Quain, P., DeFries, J. C., et al. (2005). Longitudinal twin study of early literacy development: Preschool and kindergarten phases. *Scientific Studies of Reading*, *9*, 219–235.

Cenoz, J. (2003). The additive effect on third language acquisition: A review. *International Journal of Bilingualism*, *7*, 71–88.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*, 489–558.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Dale, P. S., Harlaar, N., Haworth, C. M. A., & Plomin, R. (2010). Two by two: A twin study of second-language acquisition. *Psychological Science*, *21*, 635–640.

Davis, O. S. P., Kovas, Y., Harlaar, N., Busfield, P., McMillan, A., Frances, J., et al. (2008). Generalist genes and the Internet generation: Etiology of learning abilities by web testing at age 10. *Genes, Brain and Behavior, 7*, 455–462.

Derks, E. M., Dolan, C. V., & Boomsma, D. I. (2006). A test of the equal environments assumption (EEA) in multivariate twin studies. *Twin Research and Human Genetics*, *9*, 403–411.

Ellis, E. M. (2004). The invisible multilingual teacher: The contribution of language background to Australian ESL teachers' professional knowledge and beliefs. *International Journal of Multilingualism*, *1*, 90–108.

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition*, *27*, 305–352.

Ellis, R. (1997). *SLA research and language teaching*. Oxford, UK: Oxford University Press.

Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). Oxford, UK: Oxford University Press.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Falconer, D. S. (1960). *Introduction to quantitative genetics*. New York: Ronald Press Company.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Los Angeles: Sage.

Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General*, *137*, 201–225.

Grigorenko, E. L. (2007). How can genomics inform education? *Mind, Brain, and Education*, *1*, 20–27.

Hart, S. A., Petrill, S. A., Thompson, L. A., & Plomin, R. (2009). The ABCs of math: A genetic analysis of mathematics and its links with reading ability and general cognitive ability. *Journal of Educational Psychology*, *101*, 388–402.

Hoge, R. D., & Colardaci, T. (1989). Teacher-based judgments of academic achievement: A review of the literature. *Review of Educational Research*, *59*, 297–313.

Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching*, *44*, 137–166.

Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C., & Eaves, L. J. (1994). Parental treatment and the equal environments assumption in twin studies of psychiatric illness. *Psychological Medicine*, *24*, 579–590.

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*, 1–160.

Kovas, Y., & Plomin, R. (2006). Generalist genes: Implications for the cognitive sciences. *Trends in Cognitive Sciences*, *10*, 198–203.

Liberman, I. Y., & Liberman, A. M. (1992). Whole language versus code emphasis: Underlying assumptions and their implications for reading instruction. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 343–366). Hillsdale, NJ: Erlbaum.

McCall, R. B., Evahn, C., & Kratzer, L. (1992). *High school underachievers: What do they achieve as adults?* Pittsburgh, PA: Sage.

Miyake, A., & Friedman, N. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. Healy & L. Bourne (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 339–364). Hillsdale, NJ: Erlbaum.

Monaco, A. P. (2007). Multivariate linkage analysis of Specific Language Impairment (SLI). *Annals of Human Genetics*, *71*, 660–673.

Neale, M. C. (2004). *Mx: Statistical modelling (Version 1.5.01)*. Richmond, VA: Department of Psychiatry, Medical College of Virginia.

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

Oliver, B., Harlaar, N., Hayiou, T., Marianna, E., Kovas, Y., Walker, S. O., et al. (2004). A twin study of teacher-reported mathematics performance and low performance in 7-year-olds. *Journal of Educational Psychology*, *96*, 504–517.

Pinker, S. (1994). *The language instinct*. London: Allen Lane.

Plomin, R., DeFries, J. C., McClearn, G. E., & McGuffin, P. (2008). *Behavioral genetics* (5th ed.). New York: Worth.

Plomin, R., Kovas, Y., & Haworth, C. M. A. (2007). Generalist genes: Genetic links between brain, mind, and education. *Mind, Brain, and Education*, *1*, 11–19.

Richardson, K., & Norgate, S. (2005). The equal environments assumption of classical twin studies may not hold. *British Journal of Educational Psychology*, *75*, 339–359.

Samuelsson, S., Byrne, B., Olson, R. K., Hulslander, J., Wadsworth, S., Corley, R., et al. (2008). Response to early literacy instruction in the United States, Australia, and Scandinavia: A behavioral-genetic analysis. *Learning and Individual Differences*, *18*, 289–295.

Samuelsson, S., Byrne, B., Quain, P., Wadsworth, S., Corley, R., DeFries, J. C., et al. (2005). Environmental and genetic influences on prereading skills in Australia, Scandinavia, and the United States. *Journal of Educational Psychology, 97*, 705–722.

Sawyer, M., & Ranta, L. (2001). Aptitude, individual differences, and instructional design. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 319–353). New York: Cambridge University Press.

Slutske, W. S., Heath, A. C., Dinwiddie, S. H., Madden, P. A. F., Bucholz, K. K., Dunne, M. P., et al. (1997). Modeling genetic and environmental influences in the etiology of conduct disorder: A study of 2,682 adult twin pairs. *Journal of Abnormal Psychology*, *106*, 266–279.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: HarperCollins.

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.

Wainwright, M. A., Wright, M., Luciano, M., Geffen, G., & Martin, N. (2005). Multivariate genetic analysis of academic skills of the Queensland Core Skills Test and IQ highlight the importance of genetic g. *Twin Research and Human Genetics*, *8*, 602–608.

Vinkhuyzen, A. A. E., van der Sluis, S., Posthuma, D., & Boomsma, D. I. (2009). The heritability of aptitude and expectional talent across different domains in adolescents and young adults. *Behavior Genetics*, *39*, 380–392.