Educational Research Journal《教育研究學報》, *Vol. 21, No. 2, Winter 2006*

A Statistical Approach to Second Language Grammar Knowledge: A Longitudinal Study of Hong Kong Students

Ming Ming Chiu, David Coniam, Eunice Tang
Faculty of Education
The Chinese University of Hong Kong

This study investigates students' grammar learning rates, with the aid of advanced statistical models. During a three-year study, 3,227 English as Second Language (ESL) students in 16 Hong Kong high schools (Grades 7–9) took four grammar tests. After calibrating the test questions with a three-parameter logistic model, scores were analyzed using multi-level analysis. The results indicate that students' grammar learning rates decrease over time and best fit a logarithmic curve. Furthermore, while ability grouping by school predicts grammar scores, it does not predict learning rates. Finally, higher achieving students in each school learned at slower rates, suggesting that instruction fell short of students' learning potential. Together, these results suggest that instruction might be developmentally inappropriate.

Key words: English as a second language, individual differences, ability grouping

---

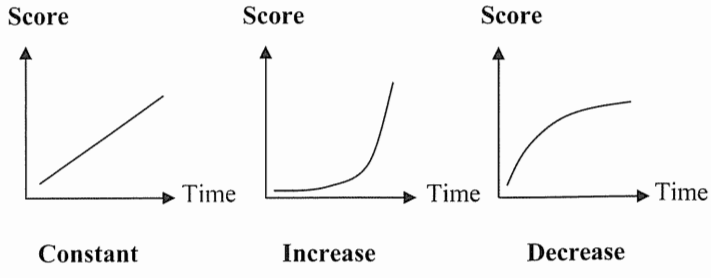
Correspondence concerning this article should be addressed to Ming Ming Chiu, Department of Educational Psychology, Faculty of Education, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong. E-mail: mingming@cuhk.edu.hk

In Hong Kong, secondary schools differ by student ability. Secondary schools are graded according to a school's student intake — generally through admitted students' achievement test scores at the end of primary school (at Grade 6 , aged 12). The ability grouping policy is a central government decision aimed toward improving the efficiency and effectiveness of teaching and learning. There are three broad bands of ability, with each band covering roughly 33% of the student ability range. Despite the school banding differences, most teachers use the same textbooks and instruction methods to teach students with different abilities and learning rates. A person's learning rate in a specific domain (e.g., English language) can remain the same, increase, or decrease over time (see Figure 1). If English grammar learning follows the principles of naturalistic acquisition, then the learning of each grammar item can be viewed as separate phenomena that require roughly the same amount of time, barring sharp differences in grammar item difficulties. As a result, naturalistic acquisition would tend to yield similar learning rates over time.

Instruction that corresponds to a learner's development is generally taken as building on knowledge of earlier learned items (proactive facilitation; Anderson, 1995; Dempster & Corkhill, 1999). This proactive facilitation would reduce the time needed to learn new items, thereby increasing the learning rate over time (Pienemann, 1984).

In contrast, a mismatch of grammar instruction and learner development can lead to falling learning rates over time. If students learn grammar

Figure 1 Possible Learning Rates: Constant (linear), Increasing (e.g., exponential), Decreasing (e.g., logarithmic)

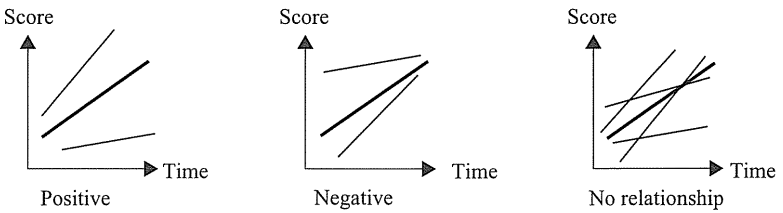


structures for which they are not developmentally ready, students may acquire a fragmented set of grammar structures. When learning a developmentally inappropriate grammar item that does not easily fit with their earlier grammar structures, they may spend more time learning it (proactive inhibition; Anderson, 1995; Dempster & Corkhill, 1999). In this case, the student's learning rate falls over time, lower than that of naturalistic acquisition.

Differences in students' learning curves can stem from individual differences or instructional differences or both. Studies have shown that these differences vary substantially across countries. For example, classroom and school differences account for over 50% of the differences among students in the Netherlands, but only 10% in Iceland (Organization for Economic Cooperation and Development, 1999, 2000).

Students with higher levels of achievement may have steeper learning curves if they can use their current knowledge to construct new information (e.g., using adjectives to learn related adverbs [quick, careful → quickly, carefully]). Graphically, high achieving students might have higher and steeper learning curves, while lower achievers have lower and flatter learning curves (see Figure 2). For example, Bahrick (1984) has shown that learners with high levels of grammar competence had less language attrition than learners of low competence levels. This being the case, low grammar competence learners might need more time to make up for their apparent language loss, with the ensuing learning gain slowing down or even decreasing.

Figure 2 Possible Line Graphs for Each Student



Note. Students with higher achievement might have higher learning gains (positive relationship), lower learning gains (negative relationship) or similar learning gains (no relationship) compared to lower achieving students.

Alternatively, high achieving students might have flatter learning curves. In this case, higher learning curves are flatter and lower ones are steeper. This can occur for at least two reasons. First, students' current knowledge can interfere with constructing new information. Yip (1995, p. 143) discusses the problems faced by Chinese students in grasping and correctly assimilating ergative verbs into their repertoire and in differentiating which verbs are used ergatively rather than passively. Second, high achieving students may reach a ceiling if the teaching/learning point is limited (e.g., forming comparative adjectives from the base form of the adjective).

Finally, high achieving students' learning rates may not differ from those of low achieving students. Figure 2 above illustrates the possible relationships (positive, negative and unrelated) between learners' previous achievement and learning gains. (The bold line shows the average initial grammar score and average learning gain.)

Ability grouping (or streaming) can also affect students' learning curves. Most studies show no significant overall effect of ability grouping (streaming) on academic achievement (Ireson, et al., 1999; Kulik & Kulik, 1992; Slavin, 1990). Although streaming does not affect high ability students' achievement, lower ability students perform better in mixed ability grouping (Newbold, 1977). Occasional significant effects (e.g., in Ireson et al., 1999, streaming improved mathematics scores) appear to stem from different opportunities to learn (e.g., through differences in the curriculum or the pacing of lessons). Differentiated curriculum materials have, in fact, shown the greatest effects (Kulik & Kulik, 1992). Kulik and Kulik's meta-analyses indicate that gifted pupils' achievement improved significantly given custom-designed programs to meet their needs. When groups proceed at the same pace and cover the same curriculum, learning outcomes do not differ significantly (Hallam & Toutounji, 1996; Ireson & Hallam, in press).

The Study

The present study began when students enrolled in secondary school (Grade

7), continuing for three years until the end of Grade 9. The students ranged from age 12 to 15. They had received six years of formal English instruction at the starting point of the study, i.e., the beginning of Grade 7.

In this study, the following research questions using advanced statistical methods (factor analyses and multi-level analysis of item response model test scores) are addressed.

1. What is the shape of Hong Kong students' English grammar learning (gain) rate?
2. Do ability differences affect grammar learning rates?
3. Does ability grouping across schools affect grammar learning rates?

Methodology

After designing the test items and choosing an appropriate subset to use on Hong Kong students, tests were administered to subjects four times within three years (resulting in both longitudinal and cross-sectional data, also known as panel data). Analyses of the data included examination of the structure of students' grammar competence, the quality of the test, the relative difficulty of test items and textbook items, estimation of student grammar scores, and the effects of years of schooling and school banding on student test scores.

Test Design

Fifty secondary schools were invited to participate in the study. Sixteen schools accepted, and fourteen schools completed the study. (Due to the small number of schools, the statistical power of this analysis to detect non-significant school level effects is limited.) Of the initial 3,227 participating students, 2,348 completed all tests; two of the mid-band schools withdrew at different stages. Using multilevel analysis (Goldstein, 1995) to minimize missing data distortions, we modeled all 3,227 students with all available test scores. (See the discussion of missing data in the analysis section below.)

These 16 schools were representative of Hong Kong schools' language of instruction, diverse bands, governing school boards and locations. In

thirteen of the schools, teachers primarily taught in Chinese. In the remaining three schools, teachers primarily used English. A total of 1,110 students from six high-band schools, 1,371 students from five middle-band schools, and 796 students from five low band schools participated. Of these, 950 high-band students, 841 middle-band students, and 557 low-band students completed the test.

The items in this study were drawn from an item bank constructed to create a common English language competence scale for the Hong Kong school system (described in Coniam, 1995). Three expert secondary school teachers (two heads of English departments and an English language teacher) and two tertiary English language professors judged these items to match the stage of development (Lightbown & Spada, 1993; Selinker, 1972) of the age and competence of the cohorts being tested. Of these 1,500 multiple-choice items, roughly 700 items were classified as tapping different features of English language grammar. These include 65% lexico-grammatical items (e.g., verb tense, number agreement, articles); 21% vocabulary items (e.g., word class, collocation, idioms, phrasal verbs); 9% usage items (e.g., appropriateness, register); and 5% syntax items (e.g., word order, sentence structure). For the purpose of this study, the term "grammar" is taken, denoting the above mentioned features. (See Batstone, 1994, for other interpretations of grammar.)

From these 700 items, four tests of 60 items each were created, linked by items common to pairs of tests to enable calibration of the different tests to a common standard. The 60 items in each test were tested before this study on a comparable (although different) cohort of students to ensure item stability and reliability. (Test 4 had one Test 1 item, three Test 2 items, and twenty-eight Test 3 items. Test 3 had ten Test 1 items and twenty-three Test 2 items. Test 2 had eleven Test 1 items.) Estimating learning gains across tests requires a common calibrated interval scale, created from the overlapping common items across different tests (e.g., the difference between levels 1 and 2 is the same as the difference between levels 3 and 4 on an interval scale). If this scale is arbitrarily decided (e.g., de Avila, 1997) or

includes learning gain assumptions (e.g., Plewis, 1996), then the shape of children's learning gains cannot be estimated.

Procedure

Hong Kong students regularly take tests, often multiple choice tests, and generally try their best at them (Biggs, 1996). The first test was administered at the beginning of the Grade 7 school year. The remaining tests were administered along with end-of-year tests. Answer sheets were carefully scrutinized after students completed each test. As test sheets with unusually regular patterns suggested careless test completion (e.g., all answers were choice "A"), these test sheets (less than 1% of the tests in actuality) were removed.

Analysis

Test properties were modeled to obtain more precise estimates of students' grammar competences with a factor analysis and an item response model. An explanatory model of students' grammar competences with multi-level analyses was created.

Structure of grammar competence. This factor analysis estimates the structure of students' grammar competence within and across (a) test questions and (b) the four types of language items. Each student response was scored 1 if correct or 0 if incorrect. To test the structure of students' understanding of grammar (single factor, hierarchical, separate chunks, or isolated elements), we used a tetrachoric correlation-based, principal factors, factor analysis ($y_i = a_{ij}f_j + e_i$; where y_i is each test item, f_j is one of j factors, with e_i error terms; Lord & Novick 1968) on student answers to each test question on each of the four tests. For binary variables (value = 0 or 1), tetrachoric correlations are unbiased, unlike the negatively-biased Pearson correlations.

Tests that are one-dimensional and yield data fitting an item response model well are ideal for assessing students. A well-designed test examines only one dimension of student competence (Lord & Novick, 1968). If the

factor analysis shows one dominant factor, then the student answers to the test questions likely reflect one competence dimension, most likely grammar. Evidence for a single dominant factor includes the following. First, the largest factor has an eigenvalue of 10 or more (Reckase, 1979). Second, the ratio of the largest and second largest factors' eigenvalues is large (Horn, 1965; Reckase, 1979). Third, aside from that of the largest factor, the eigenvalues of all adjacent factors sorted by size all show small differences (Tabachnick & Fidell, 1989). Fourth, the variance explained by the largest factor is 20% or more (Horn, 1965; Reckase, 1979), using the *Mplus* software package (Muthen & Muthen, 1998) to do the factor analysis. If the test has question items covering different dimensions, the test will be divided into subtests for each dimension with each subtest scored separately.

Test quality. After identifying the dimensional structure of the test items, the students' scores were estimated from the four linked English grammar tests. If a test question were to be considered ideal, all students below a specific competence level would have answered incorrectly, and all students at or above that competence level would have answered correctly. The degree to which test questions approximated this ideal was tested by fitting an item response (IR) model to each test question. A one-parameter logistic model (a Rasch model) allows test questions to have different levels of difficulty and calibrates different tests to the same standard.

$$P_i(\theta) = 1 / [1 + \exp(-1.7 * \bar{a}(\theta - b_i))] \quad (1)$$

In the above equation, θ is the estimated student competence, $P_i(\theta)$ is the probability that a student with competence θ will correctly answer test item i, and b_i is the difficulty of test item i. The Rasch model and all IR models therefore ensure that students with the same competence receive the same score even if the test item difficulties (b_i) differ. The Rasch model assumes that all test questions have the same precision for distinguishing among subjects with different abilities (discrimination; $\bar{a}$ in equation 1). In contrast, a two-parameter logistic (2PL) model allows items to have different discrimination powers.

$$P_i(\theta) = 1 / [1 + \exp(-1.7\, a_i(\theta - b_i))] \quad (2)$$

Instead of ā, discriminations can differ across test items (a_i). A three-parameter logistic (3PL) model further allows the possibility of subjects guessing successfully (c_i in equation 3).

$$P_i(\theta) = c_i + (1 - c_i)/ [1 + \exp (-1.7\ a_i(\theta - b_i))] \quad (3)$$

A 3PL model estimates an item's difficulty, discrimination, and guessing success rate. (See Hambleton & Swaminathan, 1985, for a detailed discussion of Rasch, 2PL, and 3PL models.)

To identify the best model for the test questions, each IR model was fitted to the student answers to each test question using Bayesian expected a posteriori (EAP) estimation (Bock & Mislevy, 1982; Mislevy & Bock, 1990). Log-likelihood difference χ^2 tests were then used to identify the model that best fits the data (Cohen & Cohen, 1983; Judge, Griffiths, Hill, Lutkepohl, & Lee, 1985). Goodness of fit χ^2 tests of each test question indicates how closely each test question resembles an ideal test question (Mislevy & Bock, 1990), and provides further criteria for choosing the appropriate item response model. After identifying the best model, EAP estimation also yielded students' English grammar competence scores from that model, using the *Bilog 3* software package (Mislevy & Bock, 1990). The best fitting item parameters were then taken to compute student competences for each test component: lexico-grammatical, vocabulary, usage and syntax. Finally, the means of the component scores for each school band for each year were then computed.

Item difficulties across tests and across textbooks. The mean item difficulties of each test were compared. With the expectation that the difficulty of the tests would increase, the later tests included more difficult items for students who would have learned more.

Test items to their first appearance in school textbooks were also matched. As with the tests, the grammatical concept assessed in higher difficulty items were expected to appear in textbooks for later grades.

Finally, multi-level analysis was used to model the panel data to capture differences across students and across time for each student (Goldstein, 1995;

also known as hierarchical linear modeling [HLM, Bryk & Raudenbush, 1992]).

$$y_{tij} = \beta_{000} + e_{tij} + f_{0ij} + g_{00j} + \beta_{wij}\mathbf{W}_{tij} + \beta_{00x}\mathbf{X}_{00j} \quad (4)$$

The test score at time t of student i from school j is y_{tij} with grand intercept β_{000}, time, student, and school residuals e_{tij}, f_{0ij}, and g_{00j}. $\mathbf{W}_{tij}$ is a vector of w time explanatory variables with w corresponding coefficients, β_{wij}. Likewise, $\mathbf{X}_{00j}$ is a vector of x school-level explanatory variables with x corresponding coefficients, β_{00x}. Time-series analyses (e.g., Mellow, Reeder, & Foster, 1996) can model longitudinal but not cross-sectional differences among students or among schools. In contrast, a multi-level analysis models panel data by capturing differences across students (i) and across time (t). This reduces the need for control variables that might otherwise introduce multi-collinearity into a model and thereby reduce the precision of explanatory predictors. Multi-level analyses also allow modeling of complex interaction effects across levels (e.g., do students in high band schools show higher learning gains over time compared to those in low band schools?) Altogether, this multilevel analysis has three levels: school (j), student (i), and time (t).

Students' grammar scores. Grammar score differences were estimated among schools, students and tests using sequential, multi-level analyses, with the sequential analyses allowing for the estimation of the additional variance explained by each added explanatory variable (Cohen & Cohen, 1983).

As the nested data consisted of multiple test scores per student and multiple students per school, *multi-level analysis* is needed to model these nested relationships. Analysis conducted using simple regressions (ordinary least squares) assumes that there are no school specific effects on students (i.e., students within a school differ from one another as much as students from different schools [$g_{00j} = 0$]). Likewise, it assumes that there are no student specific effects on an individual's test scores (tests completed by one student are as different from one another as tests completed by different students [$f_{0ij} = 0$]). Such assumptions were shown to be incorrect by

estimating the variance at each level (a variance components model, $y_{tij} = \beta_{000} + e_{tij} + f_{0ij} + g_{00j}$; Bryk & Raudenbush, 1992; Goldstein, 1995). The nested structure of tests within students within schools also handled missing data by allowing unequal numbers of students per test and unequal numbers of students per school. Analysis was also conducted with only the students who completed all four tests, the results were similar in both cases, only the results from the full data set are reported.

Dividing the variance at each level (school, student, test) by the total variance yields the percentage of the test score differences at each level. Such information helps in the selection of explanatory variables at the appropriate level: school, student or test. (For example, if most of the differences are at the student level, then school level variables will not explain much of the differences.) For this purpose, the *MLn* software package (Rasbash & Woodhouse, 1995) was used to estimate the variance components model and to do all the following analyses.

Learning curves. Different students learn at different rates, with different depictions of how learning progresses. Students may learn roughly the same amount each year (linear). Alternatively, their learning rate can increase in proportion to the years of instruction (quadratic) or by a similar percentage per year (exponential). The gains may also decrease in inverse proportion to the years of instruction (logarithmic). Finally, the gains may increase and then decrease over time as in an S-shaped function.

Using the best model of the grammar learning rate, how test scores differed across schools, across students and across time were examined. This was achieved by allowing the effect of years of instruction to vary at the school, student, and time levels (entering the best function of YEARS as a random parameter; YEARS ranged from 0 to 3). As multi-level analyses allow modeling of variations of effects within each level, it was used to test via slope-intercept effects (Goldstein, 1995) whether students with higher overall grammar scores show higher gain, and whether schools with higher overall grammar scores on average show higher gain.

Ability grouping. Next, test were conducted as to whether ability

grouping affects student grammar scores. To test for grammar score differences between high and medium band schools, a Wald test (for discussion, see Davidson & MacKinnon, 1993) was used to determine if effect sizes are significantly different.

Tests were also conducted as to whether school variables affected gains by adding interaction variables (HIGH_BAND*YEARS, MEDIUM_BAND*YEARS). The coefficients of the school banding interaction variables indicate how much learning gains in high and medium band schools differ from gains in low band schools.

While significant school-level effects in this explanatory model are valid, non-significant school-level results (taking an alpha level of .05) may not be reliable because of the small school sample size (Cohen & Cohen, 1983).

Results and Discussion

Single Structure of Grammar Competence

Hong Kong students' knowledge of grammar likely consists of one dominant underlying competence as the results indicate that a single factor model best fits the data of student answers to the test questions, according to the four criteria listed above. The eigenvalues were near or greater than 10 (9.8, 10.4, 12.4, 13.2, listed in chronological order of their tests). The ratio of the largest and second largest factors' eigenvalues all exceeded 6 (6.7, 8.0, 9.9, 9.8). The differences of the remaining adjacent eigenvalues were less than 0.3. Finally, the dominant factor accounted for roughly 20% of the variance (16.6%, 17.3%, 20.7%, and 22.1%). These students' knowledge of each grammar component is therefore very similar, so that most students who score high on one grammar component also score high on all the other components. Likewise, students who score low on one grammar component also score low on all the other components. This pictured consistency holds true for each annual test.

The quality of the test items differed substantially, with students also

often guessing successfully. Consequently, the 3PL model fit the data best (item difficulty: $M = -0.102$ [$SD = 0.517$]; true score: 0.451 [.202]) The precision with which test questions discriminated among students of different grammar competences differed as shown by the significant variation in discrimination ($M = 1.04$ [$SD = 0.355$]). Further evidence can be attributed to the fact that the 2PL model significantly fit the data better than the Rasch model (likelihood difference χ^2 [163] ratio = 6,515 [898,416 – 891,901], with 163 degrees of freedom, $p < .001$). As mentioned, guessing was often successful (guessing estimate: $M = 0.153$ [$SD = 0.056$]) and the 3PL model fit better than the 2PL model (difference likelihood χ^2 [163] ratio = 2,436 [891,901 –889,465], $p < .001$). In addition, the 3PL model performed better in goodness of fit χ^2 tests, fitting the data well for 82% of the test questions. The Rasch and 2PL models fit the data well for only 20% and 66% of the items, respectively. The 3PL model fit the data best.

The above results indicate that a Rasch model (or a 2PL model) does not necessarily fit the data well and can yield imprecise achievement estimates. Item discrimination and student guessing success using 2PL and 3PL models should be systematically tested for, as has been done in the current study.

Logarithmic Learning Curves

Results suggest that learning gains decrease over time, with a logarithmic curve best fitting students' scores (LL = 15,931; see Table 1, model #2; and Figure 3). The other models had higher log-likelihoods (all > 15,931), and did not fit the data as well: linear (17,028), quadratic (16,175), exponential (16,580) and logistic (16,562). Using the best fitting model (logarithmic), average gain was computed, emerging as 0.259/YEAR, namely 0.259 in the first year (0.259/1), 0.130 in the second year (0.259/2), and 0.086 in the third year (0.259/3). Component scores differed only a little for a given school band in a particular year (component results for each year are available from the authors).

Table 1 Significant, Unstandardized Parameter Coefficients of Sequential Set Multi-level Regressions Predicting Grammar Scores; Effects at Each Level; and Explained Variance (All standard errors are in parentheses)

	4 Multilevel regression models predicting grammar scores			
Predictor	Model 1	Model 2	Model 3	Model 4
Ln (Time)		0.259 **	0.246 **	0.252 **
		(0.003)	(0.028)	(0.028)
High band school				1.455 **
				(0.102)
Medium band school				0.834 **
				(0.101)
School level Grammar	0.333 *	0.334 *	0.317 *	0.031 *
Score Variation (3)	(0.118)	(0.118)	(0.112)	(0.011)
Relationship between			0.008	−0.009
Grammar Ability and Gain			(0.016)	(0.006)
Variation in Gain			0.011 *	0.011 *
at the school level			(0.004)	(0.004)
Student level Grammar	0.210 **	0.210 **	0.234 **	0.234 **
Score Variation (2)	(0.006)	(0.006)	(0.008)	(0.008)
Relationship between			−0.024 **	−0.023 **
Grammar Ability and Gain			(0.005)	(0.002)
Variation in Gain			0.020 **	0.020 **
at the student level			(0.005)	(0.004)
Test level Grammar	0.149 **	0.128 **	0.118 **	0.118 **
Score Variation (1)	(0.002)	(0.002)	(0.002)	(0.002)
Explained school variance	0%	0%	5%	91%
Explained student variance	0%	0%	0%	0%
Explained year variance	0%	14%	21%	21%
Total explained variance	0%	3%	7%	48%

Note. We included a constant term in each regression model.

*$p < .01$ **$p < .001$

Test Item Difficulty

In "developmentally-constructed" textbooks, the mean difficulty of test items in later Grades should be higher than those in earlier Grades. These did not differ significantly (0.35, 0.52, and 0.25 for Grades 7, 8, and 9 respectively). Some items in Grade 7 textbooks were likely too difficult, while certain items in Grade 9 textbooks were likely too easy—a finding which suggests that the textbooks themselves were not developmentally appropriate.

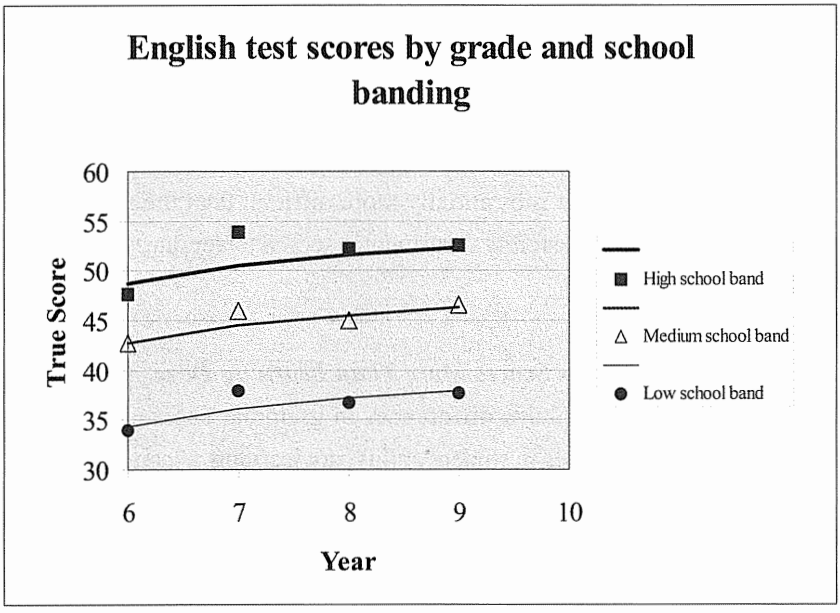
It might have argued that the low scores in Grades 8 and 9 might be due to substantially more difficult test items. The overlapping items permitted, however, for the calibration of all tests on the same scale so that a student taking an easy test and a difficult test would receive the same score. Furthermore, the mean difficulties of items on tests taken in Grades 6 through 9 were 0.10; 0.79; 0.48; and 0.37, respectively. This shows that the third and fourth tests were not substantially more difficult than the second test, thereby ruling out the possibility that students scored lower due to extremely difficult tests.

Student Ability Explains Scores More Than Learning Does

Student ability both explains differences in grammar scores and affects learning gains. Differences in student ability, not learning, accounts for most of the differences in grammar test scores as shown in the variance components model (see Table 1 above, model #1). The school and student level differences indicate differences in student ability while differences in annual test scores indicate student learning. Differences across schools accounted for 48% of the differences in grammar (48% = 0.333 / [0.333 + 0.210 + 0.149]). Student differences within the same school accounted for 30% (= 0.210 / [0.333 + 0.210 + 0.149]), and annual test score differences only accounted for 22% (= 0.149 / [0.333 + 0.210 + 0.149]). These results show, therefore, that student ability differences (cross-sectional) explain 78% of the differences in test scores over three years while learning (longitudinal differences) only explains 22%. This suggests that the learning of these Hong Kong students was unlikely to overcome differences in past achievement, an inference supported by Figure 3. The grammar competence of a typical Hong Kong student in a lower banding school with three years of schooling falls short of the grammar competence of a typical student in a higher banding school before secondary school instruction.

Student differences in learning rates within a school help explain the logarithmic shape of the learning curves. Surprisingly, students in the same school with higher overall grammar scores showed lower gains

Figure 3 Hong Kong Students' English Grammar Learning Gains Decrease Over Time (Grades 6 – 9)



(coefficient = –0.024, *SE* = 0.005, correlation = –0.35; see Table 1, model #3). This result supports the view that the pace of instruction was too slow for the high-achieving students. As low-achieving students learned more than high-achieving students in the same school, learning opportunities for these low-achieving students were likely greater than those for high-achieving students. This slow instruction can also account for the logarithmic shape of the learning curves if the slow instruction did not allow students to realize their potential.

It might be suspected that the students approached full grammar mastery (ceiling effect). That this is not the case, however, can be seen from the fact that the mean score of the last test was –0.349, equivalent to a true score of 49% of all the test questions. This suggests that students who completed Grade 9 understood less than half of the content covered by all the test questions. (See Hambleton & Swaminathan, 1985, for a discussion of the equivalence between achievement estimates and true scores.)

Ability Grouping Explains Scores, But Not Gains

Students in higher banding schools scored higher, but did not learn faster. As noted above, school differences accounted for 48% of the variance in grammar scores while within school differences only accounted for 30%. School banding can, therefore, be seen to accounts for most of the school level differences. Grammar scores in both high- and medium-band schools were higher than those in low-band schools (high: +1.46; medium: +0.83; see Table 1 above, model #4). Furthermore, students in high-band schools scored significantly higher than those in medium-band schools (Wald test, $\chi^2[1] = 46.4, p < .001$).

In contrast, school banding did not explain differences in learning gains as shown by the following three results. First, schools with higher overall grammar scores did not have significantly higher learning rates (coefficient = 0.008, $SE = 0.016$, $r = 0.15$; see Table 1, model #3). Second, both of the interaction variables (HIGH_BAND*YEARS, MEDIUM_BAND*YEARS) were not significant. Third, differences in learning rates across schools (variance = 0.011; see Table 1, model #3) were smaller than those within the same school (0.020). These results are also consistent with previous research, which shows that achievement grouping alone has no significant effect on student achievement (Ireson et al., 1999; Kulik & Kulik, 1992; Slavin, 1990).

Our full model (Table 1, model #4) explained much of the differences in students' cross-sectional and longitudinal grammar scores, accounting for about half of the total variance (48%) and included nearly all of the school level variance (91%). School banding accounts for a large portion of the explained variance. Differences in school bands accounted for at least 86% (= 91% – 5%) of the variation in grammar scores at the school level and at least 41% of the total variance (= 48% – 7%; see Table 1, models #3 and #4). In contrast, learning gain (logarithmic model) explained, at most, 21% of the grammar scores variation at the test year level (see Table 1, model #3) with at least 7% of the variation across years explained by the variation in gain at the school and student levels (21% – 14%; see Table 1, model #3).

Conclusion

From this study, a number of important observations can be made. First, Hong Kong students' grammar learning rates decrease over time. Second, within each school, students with higher achievement showed lower learning rates. Lastly, ability grouping showed no effects on learning.

Students' learning gains decrease over time, best fit with a logarithmic curve. As the difficulty of grammar items covered by textbooks for Grades 7–9 did not differ significantly, the textbooks examined are likely to be developmentally inappropriate. As teachers typically base their classroom instruction on these textbooks (Morris, 1995), the results suggest that their instruction could be developmentally inappropriate for their students and consequently might have lowered their learning curves. One corollary of this is that curriculum developers should consider reorganizing grammar items in Hong Kong textbooks to facilitate student learning (Pienemann, 1989).

High-achieving students within each school had lower learning rates. This result suggests that low-achieving students received more learning opportunities than high achieving students. To address this problem, teachers may use differentiated instructional materials to help meet students' individual needs. By so doing, they can promote greater learning gains among all students within a school. This result also implies that sequenced instruction of grammar should have sufficient flexibility to adapt to the needs of high-achieving students as well as those of low-achieving students.

Grouping students by ability does not improve student learning. It might be expected that students in higher banding schools with presumably higher ability would learn faster than students in lower banding schools. They do not. Such a practice would appear not to aid effective and efficient learning, possibly because most teachers use the same textbooks and instruction methods rather than adapting their instruction to students' needs (Lee, Sze, & Chung, 1998)

These results are generally consistent with other research results in

Europe and the United States (Ireson et al., 1999; Kulik & Kulik, 1992; Slavin, 1990). Indeed, studies have shown that labeling students as less capable harms their self-esteem (e.g., Boaler, William, & Brown, 2000) and reduces pro-achievement behaviors (such as perseverance, Eccles, et al., 1983), supporting the assertion that students should not be grouped by ability without adapting instruction to students' needs.

Acknowledgements

This project was supported by a grant from the Hong Kong Language Fund. We appreciate the research assistance of Yik-ting Choi and the comments on an earlier draft by Lawrence Khoo and Icy Lee.

References

Anderson, J. R. (1995). *Learning and memory*. New York: Wiley.

Bahrick, H. P. (1984). Fifty years of language attrition: Implications for problematic research. *Modern Language Journal, 68*(2), 105–118. (ERIC assessing number: EJ299802)

Batstone, R. (1994). *Grammar*. Oxford, U.K.: Oxford University Press.

Biggs, J. B. (1996). Review and recommendations. In J. B. Biggs (Ed.), *Testing: To educate or to select?* (pp. 298–325). Hong Kong: Hong Kong Educational.

Boaler, J., William, D., & Brown, M. (2000). Students' experiences of ability grouping. *British Educational Research Journal, 26*(5), 631–648.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431–444.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. London: Sage.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Coniam, D. (1995). Towards a common ability scale for Hong Kong English secondary school forms. *Language Testing,12*(2), 182–193.

Davidson, R., & MacKinnon, J. G. (1993). *Estimation and inference in econometrics*. New York: Oxford University Press.

De Avila, E. (1997). *Setting expected gains for non and limited English proficient students* (The National Clearinghouse for Bilingual Education — NCBE Resource Collection Series No. 8). Retrieved October 3, 1999, from http://www.ncbe.gwu.edu/ncbepubs/resource/setting/index.htm

Dempster, F., & Corkhill, A. (1999). Interference and inhibition in cognition and behavior. *Educational Psychology Review, 11*(1), 1–74.

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives* (pp. 75–146). San Francisco: W. H. Freemen.

Goldstein, H. (1995). *Multilevel statistical models*. Sydney: Edward Arnold.

Hallam, S., & Toutounji, I. (1996). *What do we know about the grouping of pupils by ability?* London: Institute of Education.

Hambleton, R., & Swaminathan, H. (1985). *Item response theory*. Boston, MA: Kluwer.

Horn, J. L. (1965). A rationale and a test for the number of factors in the factor analysis. *Psychometrika, 30*, 179–185.

Ireson, J., & Hallam, S. (in press). Raising standards. *Oxford Review of Education*.

Ireson, J., Hallam, S., Mortimore, P., Hack, S., Clark, H., & Plewis, I. (1999, September). *Ability grouping in the secondary school*. Paper presented at the British Educational Research Association Annual Conference, University of Sussex at Brighton.

Judge, G. G., Griffiths, W. E., Hill, R. C., Lutkepohl, H., & Lee, T. C. (1985). *The theory and practice of econometrics* (2nd ed.). New York: Wiley.

Kulik, J. A., & Kulik, C-L. C. (1992). Meta-analytic findings on grouping programs. *Gifted Child Quarterly, 36*(2), 73–77.

Lee, C. K., Sze, P., & Chung, C. K. W. (1998). Teachers' use and adaptation of TOC English textbooks. *Journal of Basic Education, 8*(1), 1–17.

Lightbown, P. M., & Spada, N. (1993). *How languages are learned* (2nd ed.). Oxford, U.K.: Oxford University Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mellow, J. D., Reeder, K., & Foster, E. (1996). Using time-series research designs to investigate the effects of instruction on SLA. *Studies in Second Language Acquisition, 18*, 325–349.

Mislevy, R. J., & Bock, R. D. (1990). *Bilog 3*. Mooresville, IN: Scientific Software.

Morris, P. (1995). *The Hong Kong school curriculum*. Hong Kong: Hong Kong University Press.

Muthen, L. K., & Muthen, B. O. (1998). *Mplus: The comprehensive program for applied researchers*. Los Angeles: Muthen & Muthen.

Newbold, D. (1977). *Ability grouping*. Slough, U.K.: NFER Publishing.

Organization for Economic Cooperation and Development [OECD]. (1999). *OECD in figures*. Paris: OECD.

Organization for Economic Cooperation and Development[OECD]. (2000). *Education at a glance* Paris: OECD.

Pienemann, M. (1984). Psychological constraints on the teachability of languages. *Studies in Second Language Acquisition, 6*(2), 186–214.

Pienemann, M. (1989). Is language teachable? *Applied Linguistics, 10*, 52–79.

Plewis, I. (1996). Reading Progress. In G. Woodhouse (Ed.), *Multilevel modeling applications* (pp. 103–129). London: University of London.

Rasbash, J., & Woodhouse, G. (1995). *MLn command reference*. London: Multilevel Models Project, Institute of Education.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests. *Journal of Educational Statistics, 4*, 207–230.

Selinker, L. (1972). Interlanguage. *International Journal of Applied Linguistics, 10*, 201–231.

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools. *Review of Educational Research, 60*, 471–490.

Tabachnick, B. G., & Fidell, L. S. (1989). *Using multivariate statistics* (2nd ed.). New York: Harper & Row.

Yip, V. (1995). *Interlanguage and learnability: From Chinese to English*. Amsterdam: John Benjamins.