Applied Linguistics 26/3: 317–342
doi:10.1093/applin/ami011

The Impact of Assessment Method on Foreign Language Proficiency Growth

STEVEN J. ROSS

Kwansei Gakuin University, Japan

Alternative assessment procedures have made consistent inroads into second and foreign language assessment practices over the last decade. The original impetus for alternative assessment methods has been predicated more on the ideological appeal this approach offers than on firm empirical evidence that alternative assessment approaches actually yield value-added outcomes for foreign and second language learners. The present study addresses the issue of differential language learning growth accruing from the use of formative assessment in direct comparison with more conventional summative assessment procedures in a longitudinal design. Eight cohorts of foreign language learners ($N = 2215$) participated in this eight-year longitudinal study. Four early cohorts in a 320-hour, four-semester EFL program were assessed with mainly conventional end-of-term summative assessments and tests. A sequence of sixteen EAP courses for these learners produced four time-varying grade point averages indexing stability and changes in achievement over the course of the program. Contrasted with these four cohorts were four latter cohorts of learners who engaged in considerably more formative assessment practices. The products of these formative assessments were also converted into manifest variables in the form of four time-varying grade point averages directly comparable to those generated by the four earlier cohorts. In addition to the series of grade point averages indicating achievement, the participants completed three time-varying EAP proficiency measures. Four research questions are addressed in the study: the comparative reliability of summative and formative assessment products; evidence of parallel changes in achievement differentially influencing proficiency growth; an examination of differential rates of growth in the two contrasted cohorts of learners; direct multivariate tests of differential growth in proficiency controlling for pre-instruction covariates. Analyses of growth curves, added growth ratios, and covariate-adjusted gains indicate that formative assessment practices yield substantive skill-specific effects on language proficiency growth.

The last decade has witnessed widespread change in language assessment concepts and methods. At the forefront of this change has been the increased experimentation with learner-centered 'alternative' assessment methods. From among different possible alternatives has emerged formative assessment, which, as its central premise, sees the goal of assessment as an index to learning processes, and by extension to growth in learner ability. In many

second and foreign language instruction contexts, assessment practices have increasingly moved away from objective mastery testing of instructional syllabus content to on-going assessment of the effort and contribution learners make to the process of learning. This trend may be seen as part of a wider *zeitgeist* in educational practice, which increasingly values the contribution of the learner to the processes of learning (Boston 2002; Chatteri 2003).

The appeal of formative assessment is motivated by more than its novelty. Black and Wiliam (1998), performing a meta-analysis of educational impact in 540 studies, found that formative assessment yielded tangible effects that apparently surpassed conventional teacher-dominated summative assessment methods. The current appeal of formative assessment thus is grounded in substantive empirical research, and has exerted an expanding radius of influence in educational assessment. Its long-term impact on language learning growth has not been examined empirically.

As recent contributions to the literature on second language assessment would suggest, conventional summative testing of language learning outcomes is gradually integrating formative modes of assessing language learning as an on-going process (Davison 2004). Measurement methods predicated on psychometric notions of reliability and validity are increasingly considered less crucial than formative assessment processes (Moss 1994; cf. Li 2003; Rea-Dickins 2001; Teasdale and Leung 2000), particularly in classroom assessment contexts where the assessment mandate may be different and where teacher judgment is central. The concern about the internal consistency of measurement products has shifted to focus on the way participants conceptualize their assessment practices. For instance, Leung and Mohan surmise:

> . . . student decision-making discourse is an important resource that could contribute to all subject areas. These matters do not fit well with the conventional standardised testing paradigm and require a systematic examination of the multi-participant nature of the discourse and of classroom interaction. (Leung and Mohan 2004: 338)

Their concern is centered on the processes involved in how participants arrive at formative decisions which may eventually get translated into a summative account of what has been learned.

Rationales for the increasing use of formative assessment in second language education vary in degree and focus. Huerta-Macias (1995), for instance, prioritized the direct face validity of alternatives to conventional achievement tests as sufficient justification for their use. This view also converges on the notion of learner and teacher empowerment (Shohamy 2001), especially in contexts reflecting a multicultural milieu. Shohamy, for instance, sees formative approaches as essentially more democratic than the conventional alternatives, especially when stakeholders such as the

learners, their parents, and teachers assume prominent roles in the assessment process. Other scholars (Davidson and Lynch 2002; Lynch 2001, 2003; McNamara 2001) have in general concurred by endorsing alternatives to conventional testing as a shift of the locus of control from centralized authority into the hands of classroom teachers and their charges. The enthusiastic reception that formative assessment has thus far received, however, needs to be tempered with limiting conditions and caveats; fair and accurate formative assessment depends on responsible and informed practice on the part of instructors, and on self-assessment experience for learners (Ross 1998).

A key appeal formative assessment provides for language educators is the autonomy given to learners. A benefit assumed to accrue from shifting the locus of control to learners more directly is in the potential for the enhancement of achievement motivation. Instead of playing a passive role, language learners use their own reckoning of improvement, effort, revision, and growth. Formative assessment is also thought to influence learner development through a widened sphere of feedback during engagement with learning tasks. Assessment episodes are not considered punctual summations of learning success or failure as much as an on-going formation of the cumulative confidence, awareness, and self-realization learners may gain in their collaborative engagement with tasks.

The move from objective measurement of learning outcomes to inter-subjective accounts of formative learning processes has raised a number of methodological issues. With less emphasis on conventional reliability and validity as guiding principles, for instance, questions of the ultimate accuracy and fairness have been raised (Brown and Hudson 1998). Studies of the actual practices observed in classroom-based assessment (Brindley 1994, 2001) have similarly pointed out issues that speak to dependability, consistency, and consequential validity. The consequences of process-oriented classroom-centered assessment practice have not become readily discernable, and remain on the formative assessment validation research agenda.

Much of the initial impetus for using formative assessment has been situated at the primary level in multicultural educational systems (e.g. Leung and Mohan 2004). The integration of formative assessment methods, however, has spread rapidly beyond the original primary-level ESL/EAL context to highly varied situations, now commonly involving foreign language education for adults. The ecological and systemic validity of formative assessment, with its incorporation of autonomous learner reflection and cooperative learning, has to date not been well documented in the increasingly varied contexts in which it is currently used. The influence of formative assessment now needs to be contrastively examined in how much it affects longitudinal growth in language learners' achievement and proficiency.

Formative assessment methods, especially those for second or foreign language learning adults, increasingly feature on-going self-assessment, peer-assessment, projects, and portfolios. While formative assessment processes can be seen as essentially growth-referenced in their orientation, questions remain as to how indicators of learner growth can be integrated into assessment conventions such as summative marks (Rea-Dickens 2001). The formative processes thought to motivate learning, in other words, may need to synthesize into tangible outcomes indicating both within and between-learner comparisons. The synthesis captures the distinction between summative and formative assessments as products. Summative assessments, as will be defined here, are comprised of criteria that are largely judged by instructors. In contrast, formative assessments, which are also tangible learning products, as well as learning processes, differ from summative assessments in that the language learners and their peers play a role in determining the importance of those products and processes as indicators of language learning achievement.

The trend towards formative assessment methods in the assessment of achievement has by now taken hold at all levels of second language education. At this stage of its evolution, empirical research is required on the impact of formative assessment in bolstering learner morale and on actual learning success. Of key interest is whether formative assessment manifests itself in observable changes in how learner achievement evolves over time and how putative changes in achievement spawned by innovations in assessment practices influence changes in language proficiency. Given that formative processes are dynamic, conventional experimental cross-sectional research methods are unlikely to detect changes in learning achievements and parallel changes in proficiency. Mainly for this reason, innovative research methods are called for in the examination of formative assessment impact.

RESEARCH QUESTIONS

The focus of the present research addresses various aspects of formative assessment applied to foreign language learning. We pursue four main research questions:

1 Are formative assessment practices that incorporate learner self-assessment and peer-assessment, once converted into indicators of achievement, less reliable than conventional summative assessment practices?
2 To what degree do changes in achievement co-vary with growth in language proficiency?
3 Does formative assessment actually lead to a more rapid growth in proficiency compared to more conventional summative assessment procedures?

4 Do language learners using formative assessment in the end gain more foreign language proficiency than learners who have mainly experienced summative assessments?

METHODS

To answer these research questions, a mixed mode approach was employed. Document analysis (Webb *et al.* 2000) was used initially to examine evidence of a shift in assessment practices within an English for academic purposes program situated in a foreign language environment. Once a pattern of shift appeared evident, the extent of the shift was quantified by converting the assessment criteria into percentages for direct comparison in time series mode. The first research question, concerning the comparative reliability, was addressed by examining the internal consistency of course achievements. The second research question was examined with the use of parallel growth models devised to provide comparative latent variable path analyses of changes in achievement and language proficiency. The third research question was examined with the use of a multiple group added growth model. The fourth research question was examined with the use of direct between-group comparisons of mean score differences on three repeated measures of EAP proficiency.

PARTICIPANTS

In this study, eight cohorts of Japanese undergraduates enrolled at a selective private university ($n = 2215$) participated in a multi-year longitudinal evaluation of an English for academic purposes program. Each cohort of students progressed through a two-year, sixteen-course English for academic purposes curriculum designed to prepare the undergraduates for English-medium upper-division content courses. The core curriculum featured courses in academic listening, academic reading, thematic content seminars, presentation skills, and sheltered (simplified) content courses in the humanities. Each cohort was made up of approximately equal numbers of males and females, all ranging from ages 18 through 20 years of age. All participants were members of an undergraduate humanities program leading to specializations in urban planning, international development, and human ecology in upper division courses.

Document analysis

Curriculum documents over the first eight years of the program provided archival evidence of the syllabus content and assessment practices in each of the sixteen courses in the core EAP curriculum. As part of each syllabus document, assessment criteria and relative weightings used in computing

grades were recorded. These documents became the basis for comparing a gradual shift in assessment practices from the first four cohorts to the latter four cohorts in the program. The shift suggested a gradual change in the assessment mandate (Davidson and Lynch 2002). The first four cohorts of learners were taught and tested in relation to an external mandate (policy) formulated by university administrators. In the first four years of the program, the EAP program staff was made up of veteran instructors—many with American university EAP program experience—where the usual direct mandate is to prepare language learners for university matriculation. The second four years of the program saw a nearly complete re-staffing of the program. The second wave of instructors, a more diverse group, many with more recent graduate degrees in TEFL, independently developed an 'internal' mandate to integrate formative assessment procedures into the summative products used for defining learner achievements. Their choice in doing so was apparently based on an emerging consensus among the instructors that learner involvement would be enhanced when more responsibility for achievement accountability was given to the language learners.

The refocusing of assessment criteria accelerated the use of formative assessment in the EAP program. The extent of assessment reform was considered substantive enough to motivate an evaluative comparison of its impact on patterns of achievement and proficiency growth in the program.

Syllabus documents revealed that for the first four cohorts ($n = 1113$), achievements were largely computed with summative information gathered from conventional instructor-graded homework, quizzes, assignments, report writing projects, and objective end of term tests sampling syllabus content. The latter four cohorts of learners ($n = 1102$), in contrast, used increasingly more self-assessment, peer-assessment, on-going portfolios, and cooperative learning projects, as well as conventional summative assessments. Learners in the latter cohorts thus had more direct input into formative assessment processes than their program predecessors, and received varying degrees of on-going training in the process of formative assessment. The archival data within the same program provides the basis for a comparative impact analysis of the shift in assessment practices in a single program where the curricular content remained essentially unchanged.

At this juncture it is important to stress that the comparisons of formative and summative assessment approaches are not devised as experiments. The two cohorts contrasted in this study were not formed by planned manipulations of the assessment processes as a usual independent variable would be. Rather, the summative and formative cohorts are defined by instructor-initiated changes in assessment practices. Tallies of the assessment weightings used in courses involving formative assessments that 'counted' in the achievement assessment of the students revealed a growing trend in the use of process-oriented formative assessment in the latter four cohorts of learners. These formative cohorts were in fact also assessed with the use

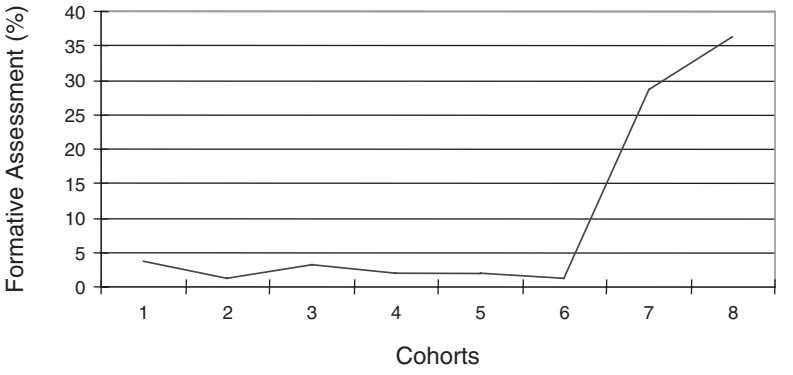

Figure 1: Average percentage use of formative assessment for achievement

of instructor-generated grades. The basis for the comparison is in the degree of formative assessment use. Figure 1 shows the trend[1] in the increased use of formative assessment, expressed in the percentage of each end of term summative grade involving formative assessment methods.

The reliability of achievement indicators

As is common in educational assessment, end-of-term grades are used to formally record learner achievement. In the sixteen-course sequence of EAP core courses, a grade point average (GPA) was computed as the average of each set of four EAP courses taken per semester. The content domain for the grade point average was linked directly to the syllabus document specifications detailing the criteria for assessment in each course. Although no course had specific criterion-referenced benchmarks for success, a university-wide standard based on a score of '60' yielded a minimum passing standard for credit-bearing courses. Credit was thus awarded for an average of at least '60' across the four EAP courses taken each semester. At the end of the two-year core curriculum, each learner in the program had four different grade point averages reflecting longitudinal achievement across the sixteen courses in the program.

A key unresolved issue in formative assessment is the possibility of weak reliability, internal consistency, or dependability because it involves several subjective observations of the interaction-in-context (Brindley 1994, 2000), which may in fact be recollected some time later by participants outside of the immediate context of the classroom (Rea-Dickins and Gardner 2000; Rea-Dickens 2001). This subjectivity, compounded by the influence of such possible learner personality factors as self-flattery, social popularity, social networks, accommodation to group normative behavior, and possible over-reliance on peers in cooperative learning ventures, may undermine the reliability of formative assessment when they are converted to summative

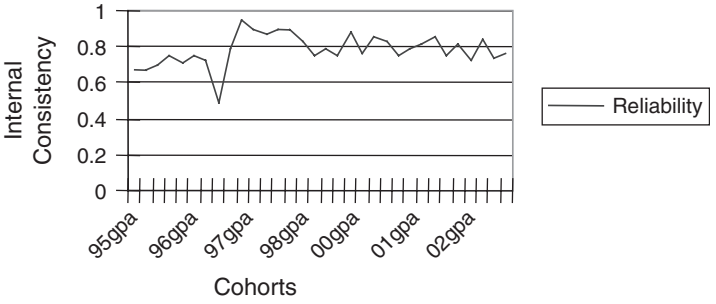

Figure 2: The reliability of achievement assessments

statements. Assertions of validity without evidence of reliability are still subject to interpretation as being less warranted than counter-assertions more firmly grounded in corroborating evidence (Phillips 2000). To date, little direct comparative evidence has been available to examine how much reliability is actually lost with the use of formative assessment relative to conventional summative assessment.

In the context of the present study, since each learner's term grade point average was computed from four core-course grades, each of which in turn was made up of an admixture of formative and summative criteria, the internal consistency of each grade point average could be readily computed.[2] The summative assessments used in cohorts 1–4 were based almost exclusively on instructor-scored objective criteria. If the instructor-determined assessments in cohorts 1–4 are in fact more internally consistent than the hybrid learner-plus-teacher-given assessments used to define achievement in cohorts 5–8, we would expect to find a notable drop in the internal consistency of the GPAs recorded in the last sixteen semesters of the program relative to those in the first sixteen semesters. Figure 2 plots the reliability estimate θ (Carmines and Zeller 1979; Zeller and Carmines 1980) which indicates the internal consistency of each grade point average across the thirty-two semester history of the program.

As Figure 2 suggests, the internal consistency among summative assessments used in the first sixteen semesters of the program (95gpa–98gpa) varies considerably. Since individual instructors would have been mainly responsible for scoring and recording objective criteria that would be used for the summative assessment, the variation in reliability may indicate differences among the classroom assessors, as well as variation in their agreement on standards. In contrast, and contrary to the expected influence of self-assessment and peer-assessment in particular, the formative assessment-based GPAs (99gpa–02gpa) appear to yield a more stable series of reliability estimates for the grade point averages reported in the latter sixteen semesters. Further, mean reliabilities[3] for the summative (.79) and formative (.80) cohorts suggest no difference in the internal consistency of the grade point average across the series of 32 semesters. A possible

interpretation of this phenomenon may be that for each language learner, the composite of the self–peer-instructor input to the assessment of achievement covaries enough to support the generalizability of even collaborative language learning tasks such as presentations, group projects, and portfolios when these are integrated into grade point averages.

Proficiency measures

In addition to monitoring learner achievement in the form of grade point averages, repeated measures of proficiency growth were made. Each learner had three opportunities to sit standardized proficiency examinations in the EAP domain. The reading and listening subtests of the Institutional TOEFL[4] were used initially as pre-instruction proficiency measures, and as a basis for streaming learners into three rough ability levels. At the end of the first academic year, and concurrent with the end of the second GPA achievement, a second proficiency measure was made in the form of the mid-program TOEFL administration. At the end of the second academic year, concurrent with the computation of the fourth GPA, the third and final TOEFL was administered. The post-test TOEFL scores are used in the program as auxiliary measures of overall cumulative program impact.

The four grade point averages index the achievements each learner made in the program. Arranged in sequential order, the grade point averages can be taken to indicate the stability of learner sustained achievement over the four semesters of the program. A growth in an individual's grade point average could suggest enhanced achievement motivation over time—or it could indicate a change in difficulty of assessment criteria. A decline in an individual's grade point average could indicate a loss of motivation to maintain an achievement level—or possibly an upward shift in the difficulty of the assessment standard. Given that there are different possible influences on changes in a learners' achievement manifested in the grade point average, the covariance of achievement and proficiency is of key interest.

The three measures of proficiency, equated on the same TOEFL scale, index the extent of proficiency growth for each learner in the program. Taken together, the dual longitudinal series of achievement and proficiency provides the basis for examining the influence of parallel change in a latent variable path analysis model. One object of interest in this study is how changes in the trajectory of achievement covary with concurrent growth or decline in language proficiency.

ANALYSES

Latent growth curve models

The major advantage of a longitudinal study of individual change is seen in the potential for examining concurrent changes. In the context of the

current study, changes in achievement over the 320-hour program potentially indicate learner engagement, motivation, participation, effort, and success in the EAP program. Measured in parallel are individual changes in each learner's proficiency. When changes in growth trajectory are of interest the focus moves from mean scores to growth curves that can be modeled when at least three repeated measures of the same variable are available for each participant. In the current study, achievement, with four GPA measures serving as indicators, and proficiency, with three TOEFL indicators, provide the longitudinal basis for assessing the impact of achievement on proficiency changes over a series of eight two-year panel studies.

Latent growth curve analysis has become an increasingly familiar method of longitudinal analysis in a number of social science disciplines (Curran and Bollen 2001; Duncan *et al*. 1999; Hox 2002; McArdle and Bell 2000; Muthen *et al*. 2003; Singer and Willett 2003). When cast as a covariance structure model,[5] individual and group change trajectories can be modeled and tested for linear and non-linear trends. Change trajectories can act as covariates of other changes such as proficiency growth, or as outcomes influenced by other static cross-sectional variables of interest. Most importantly for the present research goal, parallel change processes can be examined as time-varying predictors using latent variables, which represent the initial status in achievement and proficiency as well as individual differences in change over subsequent repeated measures indicating instructional effects.

Latent growth curve estimates can be compared across different groups in order to assess the generalizability of a structural equation model (Muthen and Curran 1997). In the context of the present study, four early cohorts experiencing mostly summative assessment defining their achievement outcomes are compared with four latter cohorts participating in relatively more formative assessment.[6] The comparative approach used here allows for an examination of the impact of formative assessment on achievement growth curves, as well as the consequential influence of achievement change on proficiency growth.

The model tested in this study uses seven indicators of growth on four latent variables. The four indicators of achievement GPA1–GPA4 are derived from individual case records ($n = 2215$). For these same learners, the three TOEFL administrations provide the basis for estimating the growth in EAP proficiency over the 320 hour program.

The two growth trajectories (achievement and proficiency) are modeled in parallel. In Figure 3, the four grade point averages (GPA 1–4) are indicators of the achievement changes for individual learners. Each of the four achievement indicator factor loadings is constrained to the achievement intercept (AI) latent variable. The achievement intercept indicates individual differences at the start of the longitudinal achievement series. Growth in achievement is estimated by changes of the trajectory from the intercept to the achievement slope (AS) indicator. Here, the first GPA is referenced to

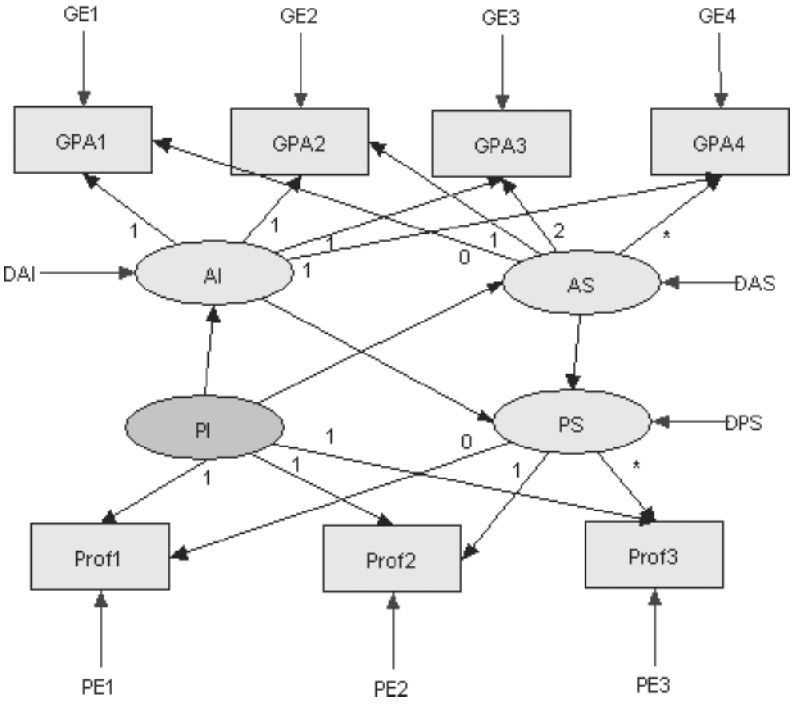

Figure 3: Parallel growth model for achievement and proficiency.
Note: Latent factor intercepts (PI and AI) are conceptualized as regressions on a constant equal to one. Paths from growth slopes (AS and PS) are set to zero on their first indicators, 1 on the second, etc, and are freely estimated (*) on the last indicator. The single headed arrows among the latent factors (ovals) represent hypothesized path coefficients. GE1–GE4 indicate errors (residuals) of the measured variables; DAI–DPS indicate the disturbances (residuals) of the latent variables.

the starting point (zero), while the second, third, and fourth GPAs are tested for a non-linear growth trajectory.[7]

The second growth model is indicated by the three proficiency measures. In Figure 3, the initial pre-program proficiency sub-test (Prof 1) is the baseline measure of proficiency. The proficiency intercept indicates individual differences in proficiency among the learners before the start of the 320 hour EAP program. Proficiency growth is also tested for non-linear growth by freely estimating the third proficiency indicator, Prof 3.

Once the shapes of the achievement and proficiency growth trajectories have been identified, the main focus, the latent variable path analysis, can be examined. In the present case, the path between initial individual differences in achievement (achievement intercept, AI) and initial proficiency (proficiency intercept, PI) is first tested for significance (PI → AI).

A significant path here would suggest that initial individual differences in proficiency influence individual differences in achievement by the end of the first academic term. Since both achievement intercept AI and proficiency intercept PI are initial states, a covariance between them would be unsurprising. EFL learners with more relative proficiency are likely by the first term to initially appear more capable to their instructors.

A second path from initial proficiency status to change in achievement (PI → AS) is also examined. Here, initial proficiency level is tested for its effect on the trajectory of changes in achievement over time during the four-semester, sixteen-course program. A positive path would indicate that higher proficiency learners progressively get higher grade point averages in the EAP courses. A negative path, in contrast, would indicate that the initial advantage of higher relative proficiency over time leads to a decline in EAP course achievement. A negative path here could also indirectly suggest motivational loss for the relatively more initially proficient learners in the program—though in this study no specific indicators of motivation are available to directly support such an inference.

The main object of interest in research question 2 is comparative change in proficiency over time. Covariances between initial achievement (AI), changes in achievement trajectory (AS), and growth in proficiency (PS) test the impact of course achievement as a causal influence on proficiency growth. A positive path from initial achievement (achievement intercept AI) to changes in proficiency (proficiency slope, PS), AI → PS, would indicate that individual differences at the end of the first term achievement outcome co-vary with eventual growth in proficiency. A substantive AI → PS path would suggest that the EAP program impact is limited to learners reaching high levels of achievement only at the beginning of the program.

The path of primary interest in this parallel growth analysis is the path from change in achievement (AS) to change in proficiency (PS), which directly tests the causal link between changes in achievement with change in proficiency. A significant positive path here would indicate that achievement growth serves to leverage the proficiency learning curve over the course of the program. The assessment system underlying the computation of learner achievements can also be assessed in this parallel change model. An examination of how the two assessment approaches compared here, formative or summative, differentially impact observed parallel changes in both achievement and proficiency, provides an opportunity to examine the second research question—that the formative assessment approach as it is used in this program results in substantive differences in the achievement-to-proficiency change relationship. By modeling parallel changes in achievement and proficiency, the effect of the two different assessment practices can be focused in a causal framework that hitherto could not be done effectively with cross-sectional analyses.

In order to test formative assessment impact, four latent path analyses employing the model in Figure 3 were conducted. Two sets of learner

cohorts, the first four groups using mainly summative assessment methods ($n = 1113$), and the latter four groups using more formative assessment ($n = 1102$) were compared on two measures of EAP proficiency. TOEFL Reading and Listening sub-tests were modeled separately.[9]

Imputation methods

Attrition has always been the bane of longitudinal research. Until recent innovations in simulation methodology employing Bayesian estimation, the only recourse for longitudinal analysis has been list-wise deletion of incomplete cases. Intermediate methods such as pair-wise deletion or replacement with a variable mean score have done little to solve the problem, and in some cases have even created others such as violations of distribution assumptions upon which many conventional effect estimation analyses rely. List-wise deletion omits possibly crucial data, while pair-wise deletion injects asymmetry into analyses that tends to bias outcomes (Bryne 2001; Little and Rubin 2002; Wothke 2000). Missing data in the context of educational assessment may inject particular kinds of bias into the analysis of outcomes and thereby complicate interpretation. It may be, for instance, that unsuccessful language learners are more likely to avoid proficiency tests. While some missing outcomes may be circumstantial and follow a random pattern across the ability continuum, others might hide systematic avoidance or planned omission. This phenomenon has made accurate language program evaluation problematic.

A current strategy for dealing with potentially-biased missing data in social science research is to use multiple imputation methods (Graham and Hofer 2000; Little and Rubin 2002; Schafer 1997, 2001; Singer and Willett 2003). Imputation serves to replace each missing datum with the most plausible substitute.[10] In the present study, missing data in each matrix of three proficiency measures and four achievement measures were arranged in chronological order before being input to ten imputation simulations per each of the four data sets. In each set,[11] imputed missing scores were saved after each 100 imputations, yielding ten sets of imputed data for each of the four matrices of three proficiency times four achievement longitudinal data arrays.

Parallel change analysis

Each of the 40 imputed data sets was tested in turn with the parallel growth model in Figure 3. For each EAP domain examined, listening and reading, the same model was tested on each of the ten data sets containing imputed scores. In this manner the summative cohorts and formative cohorts were tested directly against the same covariance structure model of parallel growth. After the tenth analysis of each imputed set, the combined effects were summarized according to methods outlined in Schafer (1997: 109).

The resulting estimates and their standard errors account for variance within and between the imputed data sets and provide the basis for testing the significance of the effects for each hypothesized influence of initial achievement and the parallel indicators of change.

Latent variable paths

Median path coefficients were computed for each of the four latent variable path relationships hypothesized in Figure 3. As the latent variable path analyses in Figure 4 and Figure 5 indicate, there are subtle changes in the sizes of path coefficients when the summative and formative cohort analyses are directly compared. In Figure 4, which shows the parallel influence of achievement changes on growth in TOEFL Reading sub-scores,

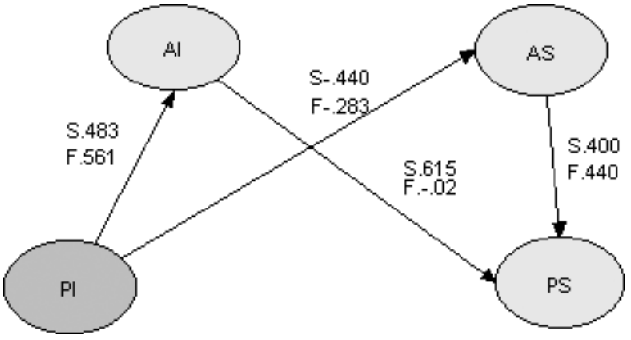

Figure 4: Parallel change analysis: Reading proficiency growth
Note: S = Summative cohort; F = Formative cohort. Single-headed arrows show paths testing effects on latent variables

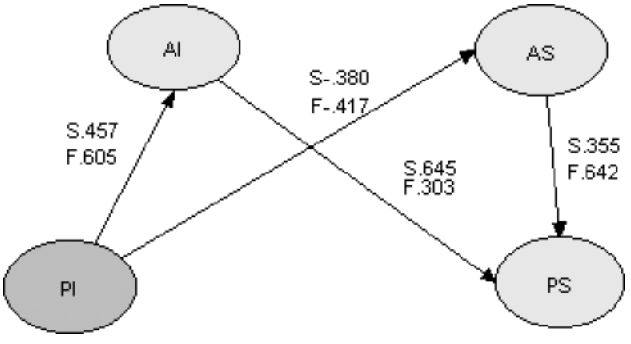

Figure 5: Parallel change analysis: Listening proficiency growth
Note: S = Summative cohort; F = Formative cohort. Single-headed arrows show paths testing effects on latent variables

it appears that, in the summative cohort, learners with relatively high achievement at the end of the first semester are the learners whose reading proficiency improves most. This path diminishes considerably for the formative cohorts; initial status shows no significant effect on TOEFL Reading growth. A possible interpretation is that the formative assessment approach shifts some of the causal influence from initial differences in achievement to a more dynamic parallel change process in the formative cohorts.

Evidence of growth can be inferred from the path from change in initial proficiency (PI) to changes in achievement (AS). Here, the formative cohort shows a smaller path than the summative cohort, suggesting that the decline in achievement by the initially proficient learners is less precipitous in the formative cohort. The parallel growth path from (AS) to growth in proficiency (PS) shows a smaller advantage for the formative cohort in academic reading proficiency growth.

Path size comparisons in Figure 5 show similar differences between the summative and formative cohorts on paths from initial achievement level to proficiency growth in TOEFL Listening, and on the substantial parallel slope covariance between achievement and proficiency. For listening proficiency growth, the path from initial achievement level for the summative cohort is approximately twice the size of that of the formative cohorts. This implies that the initial achievers in the summative cohort tended to increase their proficiency the most over time. The lower path from AI to PS for the formative cohort suggests that causal pathways have reversed. Initial achievement now has half the influence on listening proficiency gain. The path from change in achievement to listening proficiency changes also reflects a much larger impact in the formative cohorts. This path in particular suggests that formative assessment processes may serve to stimulate a gain in listening proficiency.

With the evidence thus far examined, it appears that formative assessment procedures change the relation between achievement and proficiency when compared to conventional summative assessment methods—at least for academic listening. A core thesis of formative assessment would thus get strong empirical evidence of systemic validity (Cohen 1996; Fredrickson and Collins 1989). By incorporation of formative assessment criteria into the award of course grades, which are the core components of the grade point averages, a new premium for achievement may be spawned.

Added growth models

The latent variable parallel growth models thus far suggest that formative assessment at least changes the relative influences on proficiency growth in academic English. Cross-cohort differences in developmental patterns of proficiency evolution do not, however, provide unimpeachable evidence that formative assessment creates a value-added outcome superior to the more conventional summative assessment methods. Further, Figure 4 and

Figure 5 suggest that the way proficiency evolves may differ in the formative cohort relative to their summative counterparts. These side-by-side comparisons of paths among the latent variables therefore do not index the size of actual growth trajectory difference between the cohorts. The third research question therefore addresses the issue of added growth accruing from formative assessment practices.

Muthen and Curran (1997) have formulated a latent growth model for testing multiple group differences directly by postulating a model to test if focus group change trajectories surpass those of a comparison group. For this research question, the influence of parallel changes in achievement is no longer the object of interest. The focus rather is on the comparative rate of change between the focus group and the comparison group. In the present context, the formative cohort takes the role of the focus group, and the summative cohort serves as the reference group. In this approach, the growth curve trajectory observed in the summative cohort is compared directly to that of the formative cohort, whose latent growth model contains an 'extra' or added growth latent factor. If the formative cohort growth trajectories surpass those seen in the summative cohort, the inference is that the added growth is most likely attributable to the formative assessment system. We note that the comparison featured in research question three is not in the mean and variance of the proficiency measures, but in the differences in rate of change over the 320 hours of program instruction.

In the added growth analyses, six pairs of summative and formative cohorts are compared.[12] The first pairing is based on the observed covariance matrix of three TOEFL results for each member of the respective cohorts. List-wise deletion was used in the generation of the observed data covariance matrices and descriptive statistics. Thereafter, the first through fifth imputed sets for TOEFL Reading and Listening were matched for the multiple group added growth curve comparisons. The unit of analysis was the value-added ratio[13] (Duncan *et al.* 1999), which shows the extent of extra growth observed in each paired comparison of a formative cohort with its summative cohort counterpart. Ratios larger than 1 indicate the growth observed in the mean slopes of the formative cohorts surpasses the mean growth slopes observed in the summative cohorts. Figure 6 plots the added growth outcomes, expressed as growth ratios, for TOEFL Listening and Reading.

As Figure 6 suggests, the observed data sets with list-wise deletion ('ObsSet'; $N = 1854$) parallel the five imputed data sets (Imp1–5; $N = 2215$). It appears also that the added growth benefiting the formative cohorts is limited to academic listening. Here, the average growth advantage is 36 per cent larger than the slope of the growth observed in the direct comparisons with matched summative cohorts (median t-ratio of 7.2, $p < .01$). The growth advantage for academic reading, in contrast, averages just 3.2 per cent (median t-ratio of .65, ns). The implication is that formative assessment

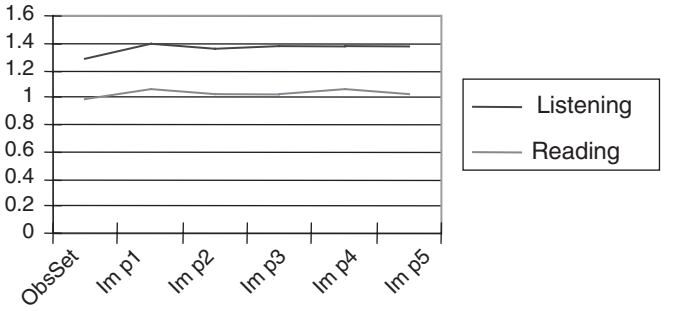

Figure 6: Added growth ratios

procedures show uneven, but possibly value-added influence on the growth of academic English for these young adult foreign language learners.

The emerging picture from the comparative parallel growth curve analyses and the added growth model comparison suggests that there are differences between the summative and formative cohorts in the way changes in achievement influence growth in proficiency. An inference made at this point might conclude that formative assessment yields benefits over conventional summative assessment. Such an inference might not in fact be correct unless there is also substantive evidence that the formative cohorts actually succeed in reaching and sustaining higher levels of proficiency than their summative counterparts. Research question 4 addresses this issue by focusing directly on the comparative gains observed on TOEFL Reading and Listening sub-scores.

Multivariate comparisons

The most direct method of examining proficiency level changes between the summative and formative cohorts is by comparing the means and variances of the two cohorts. Given the fact that this study was not devised as an experiment, a quasi-experimental design (Shadish *et al*. 2002) is used. Here, the pre-instruction measure of TOEFL listening and reading serves as the basis for between cohort baseline comparisons in proficiency. Considering the large samples in the study, even small actual differences in means and variances will suggest non-random differences between the cohorts. For this reason, the pre-instruction measures of reading and listening are used as covariates in the multivariate comparisons of the second and third measures of proficiency. Table 1 lists the results of the multivariate analysis of covariance for TOEFL Reading sub-scores.

After controlling for small but significant differences in reading proficiency prior to the start of the 320-hour program, the mean differences in reading proficiency gains indicate that there is an overall multivariate effect ($\lambda = .967$; $p < .001$) and two univariate effects favoring the formative cohort

Table 1: MANCOVA result for TOEFL reading gains

Variable	SS	DF	MS	F	P
Univariate tests					
RC2	862.44	1	862.44	52.99	.000
Error	35969.25	2210	16.276		
RC3	68.501	1	68.501	2.957	.086
Error	51202.124	2210	23.168		

Multivariate tests Wilk's $\lambda = .967$; $F = 37.351$; $DF = 2,2209$; $p = .000$

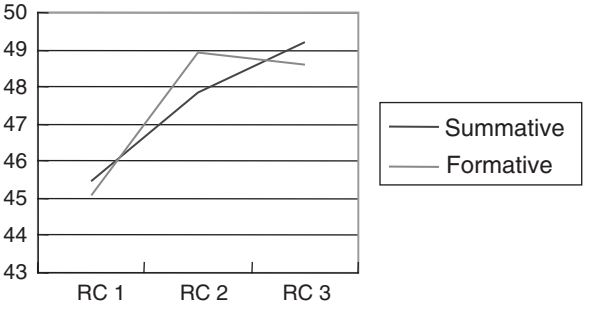

Figure 7: Mean reading proficiency comparisons

on the first post-test reading proficiency ($F = 52.99$; $p < .001$) given after the first two semesters of the EAP program. This effect diminishes by the third reading proficiency measure given at the end of the 320-hour program. Here the summative cohort seems to 'catch up' in academic reading ($F = 2.95$; $p = .086$), while the formative cohort reaches a plateau. The mean growth in reading proficiency, controlling for initial differences is shown in Figure 7.

The growth in listening proficiency differs from the pattern observed for TOEFL Reading sub-scores. For listening, there was a slightly more pronounced difference between the cohorts in the pre-instruction measure of proficiency ($t = 3.45$, $p < .001$) favoring the formative cohort. Here again, the pre-instruction measure of proficiency, LC1, was used as a covariate in a multivariate analysis of covariance. Table 2 shows the main effects analysis for the MANCOVA on TOEFL Listening sub-test gains over the course of the program.

The multivariate effect ($\lambda = .974$; $p < .001$) detects the overall difference between the two cohorts on the two post-tested measures of listening proficiency. Both the univariate tests of the effect of the cohort variable, controlling for initial proficiency differences (Appendices B1 and B2), suggest

Table 2: MANCOVA result for TOEFL listening gains

Variable	SS	DF	MS	F	P
Univariate tests					
LC2	386.179	1	386.179	23.9	.000
Error	35741.5	2212	16.158		
LC3	959.452	1	959.452	56.104	.000
Error	37828.298	2212	17.101		

Multivariate tests Wilk's $\lambda = .974$; $F = 29.267$; $DF = 2,2211$; $p = .000$

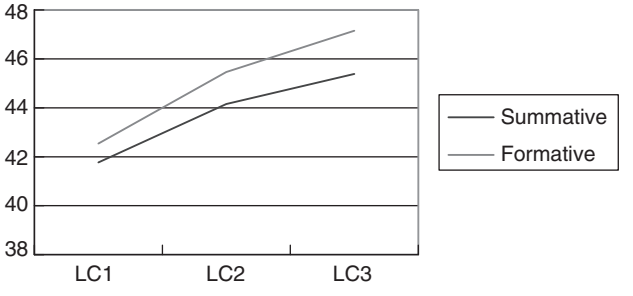

Figure 8: Mean listening proficiency comparisons

that there was a constant difference between the cohorts in proficiency gain in listening by the first post-test ($F = 23.9$; $p < .001$), which became more pronounced by the second post-test, LC3 ($F = 56.1$; $p < .001$). The mean score differences are represented graphically in Figure 8.

The mean comparisons of the summative and formative cohorts indicate that the formative cohort starts with a 0.79 TOEFL listening sub-test scaled score mean advantage, and thereafter the difference between the cohorts accelerates in a non-parallel manner[14] on the second (LC2) and third (LC3) measures of listening proficiency.

Taken together, the mean effects analyses of reading and listening corroborate the foregoing growth curve and added growth analyses. The consistent effect of the formative assessment approach appears limited to growth in listening comprehension. The apparent growth advantage for the formative cohort in reading is comparatively short-lived.

SUMMARY

The three analyses of achievement and proficiency growth reveal that the impact of the formative assessment approach is substantive but still

domain-dependent. The main effects analyses for the latent path analyses indicate that the consistent covariance between growth in proficiency and achievement change is limited to the academic listening domain. The comparatively small effects (R^2) observed in changes on the reading slopes for the formative group (Appendix A) are seen also in an AS → PS latent variable path coefficient comparable to that observed for the summative cohorts. The added growth analysis in Figure 6 corroborates this pattern with insignificant added effects for the formative cohorts *vis-à-vis* the summative cohorts in the academic reading domain. The multivariate analysis of covariance analysis indicates that the effect for the formative cohort in reading gains is tenuous, with the summative cohort eventually reaching the same level of reading proficiency by the end of the program.

The composite of results for academic listening growth show a more favorable outcome for the formative cohort. The latent growth model and added growth analyses concur in detecting that the pathways to greater listening gain are different for the two cohorts. In particular, the direct path from changes in achievement to gains in listening proficiency (AS → PS) suggest that positive changes in achievement more directly co-vary with proficiency gains for the formative group. The multivariate analyses, controlling for small initial differences between the cohorts, indicate that a gap in listening proficiency growth is slowly but consistently widening in favor of the formative cohort.

While the composite of latent variable paths, added-growth estimates, and mean comparisons tend to favor formative assessment, it should be noted that model fit is at best borderline in the formative cohort analyses.[15] This suggests that while the grade point averages for the formative cohort are at least as reliable as those for the summative cohort, there is evidently more noise in the formative cohort's latent indicators of achievement and proficiency intercepts and slopes. This fact may be a consequence of greater heterogeneity of criteria defining achievement in the formative cohort. Further research is needed to identify the sources of this relatively weaker model fit.

CONCLUSIONS

The overall emergent picture drawn in this longitudinal study suggests that formative assessment may produce its largest impact on learners' volitional stance likely to affect attention and participation in language learning activities leading at least to greater listening comprehension improvement. Growth in academic listening seems optimally conditioned by learner contribution to and direct participation in the definition of language learning

achievement. For academic reading growth, in contrast, it appears that there is less of a salient advantage for formative assessment. The premium for enhancing proficiency growth in reading may be more associated with coordinated management of the syllabus content through cyclical cross-referencing of reading materials in an integrated and coherent instructional framework (Ross 2003). Here again, more detailed research is needed to account for the differential influences of achievement on proficiency growth across skill domains.

Assessment procedures shifting more of the locus of control to the learners—through process oriented portfolios, self-assessment, peer-assessment, group projects, and co-operative learning tasks—may provide a domain-specific stimulant to enhanced learner engagement, especially when the formative assessments are recognized by learners as eventual inputs to summative criteria. Concerns that formative assessment procedures inject extraneous sources of variance into the assessment outcomes to the extent that such sources downgrade the reliability of the assessments are not borne out in macro-level analyses employed in this study. Formative assessment appears to offer these foreign language learners a larger share of direct control over the definition of 'achievement' and its consequential relation to proficiency growth. While the formative assessment may not be omnipotent in all academic skill domains, the results of this longitudinal study of comparative assessment approaches suggest that judicious use of formative assessment may well lead to tangible value-added outcomes.

Final version received May 2005

ACKNOWLEDGEMENTS

Thanks are due to Eric Rambo and Neil Matheson for their assistance in locating curriculum documents. Three anonymous reviewers and Gabriele Kasper provided critical feedback and suggestions on the original draft of the paper.

APPENDIX A. PROFICIENCY SLOPE EFFECT SIZES

Summative	PS R^2	Formative	PS R^2
Sumlc1	0.455	Frmlc1	0.408
Sumlc2	0.492	Frmlc2	0.442
Sumlc3	0.445	Frmlc3	0.371
Sumlc4	0.475	Frmlc4	0.412

APPENDIX A. CONT.

Summative	PS R^2	Formative	PS R^2
Sumlc5	0.45	Frmlc5	0.658
Sumlc6	0.422	Frmlc6	0.425
Sumlc7	0.473	Frmlc7	0.407
Sumlc8	0.498	Frmlc8	0.381
Sumlc9	0.448	Frmlc9	0.432
Sumlc10	0.463	Frmlc10	0.457
Medians	0.459		0.4185
Sumrc1	0.45	Frmrc1	0.13
Sumrc2	0.425	Frmrc2	0.218
Sumrc3	0.431	Frmrc3	0.179
Sumrc4	0.46	Frmrc4	0.234
Sumrc5	0.415	Frmrc5	0.193
Sumrc6	0.435	Frmrc6	0.242
Sumrc7	0.427	Frmrc7	0.185
Sumrc8	0.451	Frmrc8	0.167
Sumrc9	0.458	Frmrc9	0.22
Sumrc10	0.4	Frmrc10	0.198
Medians	0.433		0.1955

Note. Sum = Summative; Frm = Formative
lc = TOEFL Listening subtests; rc = TOEFL Reading subtests
PS = Proficiency Growth Curve Slope; R^2 = Coefficient of Determination

APPENDIX B1. DESCRIPTIVE STATISTICS FOR READING MEAN COMPARISONS

	Summative			Formative		
	RC1	RC2	RC3	RC1	RC2	RC3
No of cases	1111	1111	1111	1102	1102	1102
Minimum	24.000	27.000	26.000	22.000	26.000	26.000
Maximum	60.000	61.000	62.000	63.000	64.000	63.000
Mean	45.479	47.841	49.174	45.028	48.901	48.600
Standard deviation	4.931	4.532	5.216	5.849	4.724	5.769

APPENDIX B2. DESCRIPTIVE STATISTICS FOR LISTENING MEAN COMPARISONS

	Summative			Formative		
	LC1	LC2	LC3	LC1	LC2	LC3
No of cases	1113	1113	1113	1102	1102	1102
Minimum	28.000	26.000	30.000	28.000	31.000	28.000
Maximum	62.000	66.000	63.000	64.000	65.000	68.000
Mean	41.765	44.175	45.356	42.550	45.479	47.135
Standard deviation	4.593	4.588	4.756	6.123	5.658	5.630

NOTES

1 The 1999 curriculum document was incomplete, making confirmation of the exact weights used in the computation in all of the GPAs impossible. Word-of-mouth accounts indicate that by 1999 formative assessment was in fact used in some program courses. Figure 1 shows only the estimate for 1999 based on the incomplete records, and thus underestimates the percentage of formative assessment used that year.

2 An individual course grade was defined as the sum of weighted scores collected during the term. Individual instructors in each core course used the same criteria and the same weightings. An example for Core EAP Reading 3 (Summative Cohort, 1997) was computed as Grade = (Vocabulary $\times$10) + (ReadingJournal $\times$ 15) + (Midterm $\times$ 25) + (Final $\times$ 25) + (Participation $\times$ 25), all of which were teacher compiled and scored.

3 The mean reliabilities were based on Fisher Z transformations of the Theta (θ) estimates. The Theta estimates are derived from a principal components analysis of each matrix of grades (four per term). The largest extracted latent root (eigenvalue, lambda (λ) below) indicates the sum of the squared component loadings among the four GPA indicators. The θ is the upper bound estimate of Cronbach Alpha (α) $\theta = (k/(k-1))(1-(1/\lambda))$ where $k =$ number of grades used to compute the GPA.

4 The choice of TOEFL for program monitoring was made by university administrators prior to the launch of the EAP program as part of the external mandate. Academic reading and listening subtests are the focus of analysis in this study because they are most relevant to the instructional focus of the program. Other subtests such as TWE and TSE, though desirable for program evaluation, are not currently funded for use in the program.

5 The parallel growth curves were modeled with MPlus Version 2.14; Covariance structure diagrams were drawn with EQS 5.6 for Windows. The MANCOVAs were done with the MGLH module of SYSTAT 4.0 (DOS).

6 Over the first eight years of the EAP program, freshman admissions policies have varied little. The mean TOEFL of pre-tested freshmen indicates a small and variable downward trend of about 5 scale points in the last four years (including the last two cohorts comprising the formative cohort in this study). The EAP curriculum has changed little, with a core curriculum of 320 hours of academic reading, presentation skills, academic content seminar, academic listening.

7 The growth curve shapes for achievement suggest small downturns in each second term of the four-semester sequence. The resulting pattern is wave-like with slightly higher mean achievements in each first semester of each academic year.

8 The proficiency growth curves show a declining angle after the second semester. The shape of the proficiency curve is construction crane-shaped, with a slight decline from the angle of a direct linear growth.

9 Item level internal consistency estimates for each of the ITP administrations at the institution level were not made available by the ETS representative in Japan. However, θ estimates of the matrix of repeated sub-scores suggest ITP reliability in the .80 to .90 range.

10 Multiple imputation in this study was done with the use of NORM (Schafer 1997). More details can be found in the on-line version of this paper.

11 Program participants leaving the program permanently after the first year, or participants taking a leave of absence for a one-year study abroad program were omitted from the imputation and analysis.

12 Schafer (2001) notes that a range of 5–10 imputed sets usually suffices to correctly estimate variation across the parameters of interest in model-based simulations.

13 The added growth ratio is the added growth factor (PSadd) modeled in addition to the formative cohort slope, plus the slope factor for the summative cohort (PSsummative). This sum is divided by the summative cohort slope factor: (PSadd + PSsummative)/PSsummative.

14 Analysis of covariance imposes the assumption that regression slopes are parallel. In the case of listening gains, there is evidence of an interaction between the pre-instruction measure of listening proficiency (LC1), the covariate, and the cohort grouping variable (LC1 × Cohort). This interaction suggests that a subgroup of formative cohort members improves more or less rapidly than other formative cohort members.

15 Bryne (2001: 85) notes a Comparative Fit Index (CFI) > .90 or a Root Mean Squared Error of Approximation (RMSEA) fit range of .08 to 1.0 indicate 'mediocre' model fit.

REFERENCES

Black, P. and **D. Wiliam.** 1998. 'Assessment and classroom learning,' *Assessment in Education* March: 7–74.

Boston, C. 2002. 'The concept of formative assessment,' *Practical Assessment, Research, and Evaluation* 8/9: http://PAREonline.net/getn.asp?v = 8&n = 9

Brindley, G. 1994. 'Competency-based assessment in second language programs: some issues and questions,' *Prospect* 9: 41–85.

Brindley, G. 2000. 'Task difficulty and task generalizability in competency-based writing assessment' in G. Brindley (ed.): *Studies in Immigrant English Language Assessment, Vol. 1*. Sydney: NCELTR, pp. 45–80.

Brindley, G. 2001. 'Outcomes-based assessment in practice: some examples and emerging insights,' *Language Testing* 18/4: 393–407.

Brown, J. D. and **T. Hudson.** 1998. 'Alternatives in language assessment,' *TESOL Quarterly* 32: 653–75.

Bryne, B. M. 2001. *Structural Equation Modeling with AMOS*. Mahwah, NJ: Lawrence Erlbaum Associates.

Carmines, E. and **R. Zeller.** 1979. *Reliability and Validity Assessment.* Sage University Paper in Quantitative Methods in the Social Sciences. Newbury Park CA: Sage.

Chatteri, H. 2003. *Designing and Using Tools for Educational Assessment*. Boston, MA: Allyn and Bacon.

Cohen, A. 1996. *Assessing Language Ability in the Classroom*. Boston: Heinle and Heinle.

Curran, P. and **K. Bollen.** 2001. 'The best of both worlds: Combining autoregressive and latent curve models' in L. Collins and A. Sayer (eds): *New Methods for the Analysis of Change.* Washington, DC: American Psychological Association, pp. 105–36.

Davidson, F. and **B. Lynch.** 2002. *Testcraft.* New Haven: Yale.

Davison, C. 2004. 'The contradictory culture of teacher-based assessment: ESL assessment practices in Australian and Hong Kong secondary schools,' *Language Testing* 21: 305–34.

Duncan, T., S. Duncan, L. Strycker, F. Li, and **A. Alpert.** 1999. *An Introduction to Latent Variable Growth Curve Modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.

Dylan, W. and **P. Black.** 1996. 'Meanings and consequences: A basis for distinguishing formative and summative functions of assessment,' *British Educational Research Journal* 22: 537–48.

Fredrickson, J. and **A. Collins.** 1989. 'A systems approach to educational testing,' *Educational Researcher* 18: 27–31.

Graham, J. and **S. Hofer.** 2000. 'Multiple imputation in multivariate research' in T. Little, K. Schnabel, and J. Baumert (eds): *Modeling Longitudinal and Multilevel Data.* Mahwah, NJ: Lawrence Erlbaum Associates, pp. 201–18.

Hox, J. 2002. *Multilevel Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Huerta-Macias, A. 1995. 'Alternative assessment: Responses to commonly asked questions,' *TESOL Journal*, 5/1: 8–11.

Leung, C. and **B. Mohan.** 2004. 'Teacher formative assessment and talk in classroom contents: assessment as discourse and assessment of discourse,' *Language Testing* 21: 335–59.

Li, H. 2003. 'The resolution of some paradoxes related to reliability and validity,' *Journal of Educational and Behavioral Statistics* 28/2: 89–95.

Little, R. and **D. Rubin.** 2002. *Statistical Analysis with Missing Data.* Hoboken NJ: John Wiley & Sons.

Lynch, B. 2001. 'Rethinking assessment from a critical perspective,' *Language Testing* 18/4: 351–72.

Lynch, B. 2003. *Language Assessment and Program Evaluation*. New Haven: Yale.

McArdle, J. and **Bell, R.** 2000. 'An introduction to latent growth curve modeling for developmental data analysis' in T. Little, K. Schnabel, and J. Baumert (eds): *Modeling Longitudinal and Multilevel Data.* Mahwah, NJ: Lawrence Erlbaum Associates, pp. 69–108.

McNamara, T. 2001. 'Language assessment as social practice: Challenges for research,' *Language Testing* 18/4: 329–32.

Moss, P. 1994. 'Can there be validity without reliability?' *Educational Researcher* 23/2: 5–12.

Muthen, B. and **P. Curran.** 1997. 'General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation,' *Psychological Methods* 2: 371–402.

Muthen, B., S. T. Khoo, D. Francis, and **C. Boscardin.** 2003. 'Analysis of reading skills development from kindergarten through first grade: An application of growth mixture modeling to sequential processes' in S. Reise and N. Duan (eds): *Multilevel Modeling: Methodological Advances, Issues, and Applications*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 71–89.

Phillips, D. C. 2000. *The Expanded Social Scientist's Bestiary.* New York: Rowman & Littlefield.

Rea-Dickins, P. 2001. 'Mirror, mirror on the wall: identifying processes of classroom assessment,' *Language Testing* 18/4: 429–62.

Rea-Dickins, P. and **S. Gardner.** 2000. 'Snares and silver bullets: Disentangling the construct

of formative assessment,' *Language Testing* 17/2: 215–43.

Ross, S. 1998. 'Self-assessment in language testing: A meta-analysis and analysis of experiential factors,' *Language Testing* 15/1: 1–20.

Ross, S. 2003. 'A diachronic coherence model for language program evaluation,' *Language Learning* 53/1: 1–33.

Shadish, W., T. Cook, and **D. Campbell.** 2002. *Experimental and Quasi-Experimental Design for Generalized Causal Inference*. New York: Houghton Mifflin Co.

Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

Schafer, J. L. 2001. 'Multiple imputation with PAN' in L. Collins and A. Sayer (eds): *New Methods for the Analysis of Change*. Washington, DC: American Psychological Association, pp. 355–78.

Shohamy, E. 2001. *The Power of Tests*. London: Longman.

Singer, J. and **J. Willett.** 2003. *Applied Longitudinal Data Analysis*. New York: Oxford University Press.

Teasdale, A. and **C. Leung.** 2000. 'Teacher assessment and psychometric theory: A case of paradigm crossing?' *Language Testing* 17/2: 163–84.

Webb, E., D. Campbell, R. Schwartz, and **L. Sechrest.** 2000. *Unobtrusive Methods.* Thousand Oaks: Sage.

Wothke, W. 2000. 'Longitudinal and multi-group modeling with missing data' in T. Little, K. Schnabel, and J. Baumert (eds): *Modeling Longitudinal and Multilevel Data.* Mahwah, NJ: Lawrence Erlbaum Associates, pp. 219–40.

Zeller, R. and **E. Carmines.** 1980. *Measurement in the Social Sciences: The Link Between Theory and Data.* New York: Cambridge University Press.