Journal of Bunkyo Women's University
Vol.2, No.1, pp.165～175, 2000

A Path Analysis Approach to Course Grade Evaluation

Jeff Johnson*

Abstract

This study reports the use of path analysis to evaluate the effectiveness of English language courses. The first wave in the model uses pre-program proficiency measures of listening comprehension to predict success (operationalized as the end of year grade on a scale of 55 to 100), in an Oral English course, and reading comprehension to predict success in a Composition course and a Reading course. In the second wave the three course grades are used to predict post-program listening and reading proficiency, measured with an equivalent form of the pre-program test. A path model is analyzed first with full-length proficiency test scores, then again with proficiency scores made more reliable and valid by continually deleting items that have low discrimination values until the test reliability estimates peak. Results show that each of the courses has significant paths to or from the relevant proficiency measures, showing evidence of course effectiveness to varying degrees, and that the shortened proficiency measures more closely fit the model.

Key Words: language proficiency, path analysis, program evaluation, test item analysis, the SLEP test

和文要旨

この研究は，英語クラスの効果測定を，熟達度測定テストとの相関のパス分析を用いて行った結果

パス分析によるクラス成績結果の評価

＊Jeff Johnson
Correspondence Address: Faculty of Business Administration, Bunkyo Women's University, 1196 Kamekubo, Oimachi, Iruma-gun, Saitama 356-8533, Japan.

Accepted October 11, 2000. Published December 20, 2000.

の報告である。最初のモデル分析は，プログラム前に実施された聴解力と読解力の熟達度テストを，それぞれ聴解力テストはオーラルイングリッシュクラスの，そして読解力テストはコンポジションクラス及びリーディングクラスの成績（学年末の成績で，尺度は55～100）との相関を調べるために用いた。分析の結果，それぞれのクラスで成績と熟達度測定テストとの多様な強さでの相関が見られた。

The following basic idea is the genesis of this study: How can teachers or curriculum designers show that their English language courses are successful or effective ? How can they show parents, possible future employers, graduate school admissions officers, and other stakeholders in the education of their students that the grades they give in their Reading course, for example, are meaningful signs of English reading ability. Evidence for course effectiveness is not only something students and other stakeholders deserve to know, it is also advantageous for teachers or curriculum designers to be able to defend the results of their courses with objective data.

In this paper I explore a method to obtain these objective data by using path analysis to evaluate the effectiveness of English language courses at a Japanese women's college. Path analysis is a method used to test a theory of direct and indirect effects of variables thought to cause other variables. The theory I hypothesize here is that the English language proficiency students initially possess will strongly predict success in English language courses (in this study: Oral English, Composition, and Reading), and that success in English language courses should in turn be a strong predictor of proficiency measured after the course.

The general research question for this study is simply this: Are the English language courses in this curriculum effective ? In other words, does the pattern of course grades reflect ability in the skills taught? The specific research questions I address, each of which should be answered in the affirmative if my theory is correct, are as follows:

1) Is first-year Oral English course success predicted by initial listening comprehension proficiency significantly and more strongly than it is by initial reading comprehension proficiency ?

2) Is first-year Composition course success predicted by initial reading comprehension proficiency significantly and more strongly than it is by initial listening comprehension proficiency ?

3) Is second-year Reading course success predicted by initial reading comprehension proficiency significantly and more strongly than it is by initial listening comprehension proficiency ?

4) Is second-year Reading course success predicted by first-year Composition course success significantly and more strongly than it is by first-year Oral English success?

5) Does first-year Oral English course success predict listening comprehension proficiency measured after the course significantly and more strongly than first-year Composition course success and second-year Reading course success?

6) Does second-year Reading course success predict reading comprehension proficiency measured after the course significantly and more strongly than first-year Composition course success, and does first-year Composition course success predict reading comprehension proficiency measured after the course significantly and more strongly than first-year Oral English course success?

The variables, as in any statistical analysis, are assumed to be reliable and accurate measures. I assume that the course grades are strongly based on the results of achievement tests, and that the achievement tests test the domain indicated in the course title. In other words, I assume the grades from the Oral English course are largely an accurate reflection of student ability in English speaking and listening comprehension, grades from the Composition course are a measure of writing ability, and grades from the Reading class closely mirror the trait of reading comprehension. I suggest that if this is not the case, then the courses, as titled, can not be defended as effective, in that success in the courses does not correspond to student ability in the skills presumed to be taught.

The theoretical model I propose assumes that proficiency in one language skill is strongly correlated to achievement in the same skill. Therefore, for example, reading proficiency should correlate strongly with Reading course success. I also assume that there are stronger inter-channel correlations (the oral channel being listening and speaking skills, the written channel reading and writing skills) than there are cross-channel. In other words, reading proficiency will have a higher correlation than listening proficiency with Composition course grades, and listening proficiency will have a stronger correlation than reading proficiency with Oral English course grades.

Another assumption made is that the pre- and post-program proficiency tests are reliable and valid tests of listening comprehension and reading comprehension proficiency. For this study I use two sets of proficiency test scores, one with the original full-length scores from a commercially available proficiency test, and another with scores from the same tests shortened by removing items that were poor discriminators (see the Procedures section below). So a seventh research question is:

7) Will more reliable and valid versions of the proficiency measures lead to better results for research questions 1 to 6?

Method

Participants

The participants in this study were the entire year 2000 graduating class in the English Literature department at a women's junior college in Tokyo. The number of students in the program was initially 297, and 283 students graduated. All 297 students took both sections of the pre-program proficiency test, and all but one graduating student took both post-program proficiency tests. Some course grade data is missing because students withdrew from individual courses, or the college, and grades for one section of the Oral English course were not available. Students were divided into four groups of approximately 75 students each, and these groups were cut in half so that each of the English classes consisted of 38 or fewer students. All the participants were teenaged women Japanese nationals.

Materials

The materials used in this study were two forms of the listening and reading comprehension sections of the Secondary Level English Proficiency (SLEP) test, Form 1 administered pre-program and Form 2 administered post-program, and course grades for two first-year courses; Oral English and Composition, and one second-year course; Reading.

Proficiency. The SLEP test is divided into listening and reading comprehension sections of 75 multiple-choice questions each. The listening section has four subsections: photographs, dictation, map, and conversation. The photograph part presents the test-takers with 25 facsimiles of black-and-white photographs with no including text, and the test-taker hears a voice read four sentences, with the correct response being the one sentence which describes the content of the photograph. According to the SLEP order form (Educational Testing Service, 1998), this part tests "correct recognition of minimal pair contrasts, juncture, stress, sound clusters, tense, voice, prepositions and vocabulary" (pg. 10). The dictation portion has 20 or 19 questions (forms 1 and 2, respectively) and for each shows four similar sentences, one of which the test-taker must match with the sentence heard on the tape. The map section presents a road map with labeled streets and buildings and four cars traveling in different areas of the town. The test-taker listens to the two occupants in one of the cars discuss something dealing with a specific area in the town, and must take into account the location referred to (e.g., a specific building, such as a drug store), as well as the position and orientation of the cars, to choose the car in which the discussion originates. There are 12 map questions in test form 1, and 11 questions in test form 2. The final listening subsection is a conversation between three students and a staff member in a high school. The conversation is presented in sections followed by one or two oral questions for which the test-taker must choose one of four possible answers. The total time for the listening section, including instructions and example questions for each subsection, is approximately 40 minutes.

The reading comprehension section (which is also meant to measure grammar and vocabulary) is also divided into four subsections: comic, line drawings, cloze, and passage. There are 12 questions in the comic part. Test-takers see an example of the American comic strip Family Circus by Bill Keane, and for each question must match the written sentence with the appropriate character's thoughts, which appear as illustrations in a bubble above each character's head. The line drawing subsection has 15 questions in form 1, and 16 questions in form 2. For each question the test-taker matches a sentence with one of the four drawings. This section tests the "use of prepositions, pronouns, adverbs, and numbers" (Educational Testing Service, 1998, pg. 15). The next subsection includes three passages in which multiple-choice cloze questions are embedded, followed by comprehension questions. Each cloze question has four options of single words, two-word verbs, or prepositional phrases that the test-taker chooses from. Form 1 has 22 cloze questions and 18 comprehension questions, form 2 has 22 cloze questions and 17 comprehension questions. The passage subsection is a longer reading passage than the ones in the cloze subsection, and is followed by 8 multiple-choice questions.

The test was designed to assist high schools in America in admissions or placement decisions for their international students. Wilson (1993), however, found that over one third of the respondents to his survey used the test with college and university students. In a study with Japanese college and university students, Culligan & Gorsuch (1999) found that pruning the 84 items (56% of the test) that discriminated poorly (ID≤ .20), improved the reliability estimate and standard error of measurement for the test as a whole, with both listening and reading comprehension sections analyzed together. In another Japanese study, Johnson, Allen-Tamai, & Kasuya (1999) found using factor analysis that the listening and reading sections were generally measuring two separate factors, and correlation coefficients between subsections supported this. They also found that 54 items (36%) did not discriminate well, with a point biserial correlation of less than .30. In a study with subjects the test was designed for, non-English speaking high school students, Stansfield (1984) reported that 77% of the items discriminated well with a point biserial correlation of .30

or greater. He also showed the test to be very reliable for this group of 310 students, with reliability estimates of .94 for the listening section, .93 for the reading section, and .96 for the two sections combined. This compares to Culligan and Gorsuch estimates of .81 for the entire test, and .84 for the test shortened by deleting the poor items, and Johnson, et. al. estimates of .72 for listening and .82 for reading. (All the reliability estimates used the Kuder-Richardson formula 20)

Grades. Grades for the three courses were the other variables used in this paper. I do not attempt to ask why a course is successful or not in this study, so I do not describe the course materials. Rather, I simply ask if the courses are effective, or to what degree they agree with the theory I propose. The question of why a course works or does not is left for future qualitative investigation. Here the courses are described simply by their titles; Oral English, Composition, and Reading, and the course grade data are numerical, given on a 55 to 100 scale. Grades of “withdraw” are treated as missing data. Students who did not withdraw from a course or the college, but did not reach a score of 60, were given a “fail,” and to keep this data in the study I set this grade to a value of 55, five points lower than the minimum passing grade.

Procedures

The participants took the pre-program SLEP listening and reading comprehension tests in the second (listening section) and third (reading section) weeks during their first year at the school in their regularly scheduled Oral English classes. They were provided student booklets containing the questions, and were told not to make marks in the booklets but to fill in the answers on a separate answer sheet. Students who were absent from class on the test day were required to schedule a time to take the test proctored by school staff. Test scores were then hand-tallied by school staff, and answer sheets were stored. The answer sheets were made available to me, and I entered the data into a spreadsheet computer program.

Course grades were provided by each teacher at the end of the academic year, on a scale from 60 to 100, with students not achieving a mark of 60 given a “fail”. Each course was designed and coordinated by a two-teacher team, and each class used the same materials, followed the same lesson plan, and used identical midterm and final exams.

Post-program SLEP test scores were obtained during the final weeks of the students’ second and final year at school, and scored by computer scanner.

I analyzed the SLEP test items to delete poor items and make shorter, more reliable scores. To do this I calculated the discrimination of each item, using the point biserial correlation (rpb), and deleted, in waves, the items with the lowest discrimination while recalculating the test reliability estimates (Cronbach's alpha) at each stage. For example, the listening pre-test had a reliability estimate of .7706, and when I deleted the five items with a rpb of less than .10, the reliability went up to .7816. Then I eliminated the 8 items with a rpb <.15, and the reliability reached .7888. When the three items with a rpb <.16 were taken out, reliability went to .7903. Two items had a rpb of .16, but when these were deleted the reliability fell back to .7899, so for this test the 59 items with rpb of .15 or higher result in the shortened form that is most reliable.

Analyses

All analyses were performed on an IBM personal computer. SLEP test raw data were input into the spreadsheet program Quattro Pro 6.0 (1994), where listening and reading total-correct scores were tabulated. Course grades were then added to the spreadsheet and descriptive statistics (means, low values, high values, standard deviations, skewness, and kurtosis) were calculated with the SPSS 7.0 (1995) computer program. Cronbach's alpha estimates for test score reliability were also computed via SPSS 7.0, and the standard error of measurement values were then done by calculator. SLEP test item discrimination was measured as point biserial correlations calculated via Quest (Adams & Khoo, 1996). For this analysis items left blank in the SLEP test data were considered missing, not incorrect responses. Finally, the path analyses for predictive power of pre-proficiency on grades, and grades on post-proficiency, were calculated with the SPSS 7.0 regression program, with variables entered together in blocks (the enter method), and missing data excluded pair-wise.

Results

Table 1 lists the descriptive statistics for the SLEP test and course grades. Most of the variables appear to be fairly normally distributed, a requirement for the statistical analysis. All the scores distributions have subjects with scores both over and under two standard deviations from the mean, and no variable has a absolute skewness value of over 1.00. The post proficiency tests, however, do have kurtosis values over 1.0, suggesting a flattish bell-shaped curve, but this was not considered extreme, and statistical analysis was carried out without transformation of data.

The reliability estimates and standard error of measurement (SEM) statistics are shown in table 2. Reliability was estimated with the Cronbach's alpha statistic, and values for the full-length tests range from a low .66 for the reading proficiency pre-test to a high of .77 for the listening proficiency pre-test, low enough to question the use of this test with this population. The SEM is a value that can be considered a band size; approximately two thirds of students would get a score within plus or minus the SEM if they retake the test (without any significant raising or lowering of proficiency between tests), and one third of the students would likely score more than SEM points either higher or lower than their first score if they retake the test. A smaller SEM means a more reliable test. Table 2 shows that most students, if they retake the full-length tests, would receive a score within about three and one half points above or below their original score, which is less than 5% of the total score and considered reasonable.

When the tests were shortened by deleting poorly discriminating items, the reliability estimates of the listening tests went up by about .02 each, and the estimates for the reading pre-test improved considerably, up by more than .05, although this test, both in the full form and the shortened form, is the least reliable of all the proficiency measures. The reading post-test reliability improved the least, by about .01. The SEM values provide further evidence that the shortened tests are more reliable, with an average value of nearly one half a point lower than the SEMs for the full-length tests.

Answers to the first six research questions are found in table 3. The values in the table (and table 4) are standardized regression beta weights, which can be interpreted as the predictive power the independent variables in the first column have on the dependent variables across the top row of the table. For example, the pre-program listening comprehension proficiency variable (Pre-Listening) significantly predicts success in the Oral English course (Oral English) with a beta weight, or path value, of .225, while its path strength to Composition course success is .154, and it does not significantly predict Reading course success.

Research questions 1 to 6 ask first if the hypothesized relationships between the variables exist, and next whether these relational paths are stronger than paths with variables thought to have less predictive power. Answers to the first question are: yes, listening proficiency does significantly predict Oral English course success, but no, not more strongly than reading proficiency does. Answers to question 2 are: yes, reading proficiency predicts Composition success, and yes, more strongly than listening proficiency. For question 3 the answers are no, reading proficiency does not significantly predict Reading success, and no, the predictive power is not stronger than listening proficiency.

The fourth question considers the relationship between first-year courses and the second-

Table 1: SLEP Test and Course Grade Descriptive Statistics

Variable	k	N	Min	Max	Mean	S.D.	Skew	Kurtosis
Listening Pre-Test	75	297	25	71	47.73	7.15	.13	.68
Reading Pre-Test	75	297	25	63	45.10	5.90	-.20	.57
Listening Post-Test	75	283	30	73	44.98	6.68	.61	1.18
Reading Post-Test	75	282	12	71	50.52	6.23	-.75	4.89
Oral English Grade	100	261	55	92	74.44	7.42	-.61	.30
Composition Grade	100	295	55	94	73.41	9.07	-.29	.28
Reading Grade	100	286	55	93	76.06	7.18	-.46	.29
Short Listening Pre-Test	59	297	19	58	39.19	6.61	-.00	.19
Short Reading Pre-Test	44	297	9	38	25.01	4.99	-.13	-.14
Short Listening Post-Test	62	283	24	61	37.93	6.25	.61	1.18
Short Reading Post-Test	60	282	8	60	42.41	5.66	-.85	4.46

k=number of items or total possible score. N=number of students. Min=minimum score. Max=maximum score. S.D.=standard deviation. Skew=skewness.

Table 2: SLEP Test Reliability and Standard Error of Measurement Estimates

	Listening		Reading	
Time, Length of Test	alpha	SEM	alpha	SEM
Pre-Test Full Length	.7706	3.425	.6579	3.451
Pre-Test Shortened	.7903	3.027	.7119	2.678
Post-Test Full Length	.7243	3.507	.7249	3.268
Post-Test Shortened	.7439	3.163	.7377	2.899

alpha=Cronbach's alpha. SEM=standard error of measurement.

year Reading course. First-year Composition course success does significantly predict success in the Reading course, but, interestingly, there is a much stronger predictive power from the Oral English course. As for research questions 5 and 6, all answers are yes. Oral English course success is a significant predictor, and the strongest predictor, of post-program listening proficiency, and Reading course success is the strongest predictor of post-program reading proficiency, with Composition course success also a significant predictor, stronger than Oral English success.

The final research question considers improving the proficiency measures and reanalyzing the data. When the poor items were deleted form the tests, improving the reliability, the data better fit the hypothesized model (compare table 4 to table 3). Pre-program listening proficiency is now the strongest predictor of Oral English course success, however Composition course success is now predicted equally strongly by both listening and reading proficiency. There is no change in the course-to-course paths, as these variables were not altered, but post-program proficiency measures are now more strongly predicted by

Table 3: Path Analysis Results with the Full-Length Pre-and Post-Proficiency Tests

Effect on:	Oral English	Composition	Reading	Post-Listening	Post-Reading
of:					
Pre-Listening	.225	.154	n.s.	—	—
Pre-Reading	.301	.265	n.s.	—	—
Oral English	—	—	.503	.315	n.s.
Composition	—	—	.131	n.s.	.137
Reading	—	—	—	.148	.355

p≤.05. effects in standardized regression beta weights. n.s.=not significant. — = not applicable

Table 4: Path Analysis Results with the Shortened Pre-and Post-Proficiency Tests

Effect on:	Oral English	Composition	Reading	Post-Listening	Post-Reading
of:					
Pre-Listening	.320	.183	n.s.	—	—
Pre-Reading	.160	.182	n.s.	—	—
Oral English	—	—	.503	.390	n.s.
Composition	—	—	.131	n.s.	n.s.
Reading	—	—	—	n.s.	.401

p≤.05. effects in standardized regression beta weights. n.s.=not significant. — =not applicable

corresponding course success. Oral English course success is now the only significant predictor of listening proficiency, and Reading course success is the sole predictor of reading proficiency, while both path values are stronger than they were in the full-length test analysis. Thus, the path analysis results with the shortened test scores are closer to the expected outcome.

In this study I proposed a path analysis model to test the effectiveness English language courses using pre-program and post-program proficiency measures. More work is needed to study the effects of data that is not normally distributed, outliers, missing data, and the reliability and validity of pre- and post-proficiency measures. The next step for a program evaluation should include qualitative data and answer questions about the reasons a course does or does not effectively contribute to proficiency, but as an important first step I believe this model can be useful to collect objective data to evaluate and defend English language courses.

References

1. Educational Testing Service. (1998). *SLEP test order form.* Princeton, NJ: Educational

Testing Service.

2. Wilson, K. M. (1993). Uses of the secondary level English proficiency (SLEP) test: A survey of current practice. *TOEFL Research Report 93-9*. Princeton, NJ: Educational Testing Service.

3. Culligan, B., & Gorsuch, G. (1999). Using a commercially produced proficiency test in a one-year core EFL curriculum in Japan for placement purposes. *JALT Journal, 21*(1), 7-25.

4. Johnson, J., Allen-Tamai, M. & Kasuya, H. (1999). SLEP tesuto ni yoru eigonouryoku sokutei: bunkyo joshi daigaku ichinensei no bunseki [Measuring proficiency with the SLEP test: An analysis with first-year women's university students]. *Journal of Bunkyo Women's University, 1*(1), 141-162.

5. Stansfield, C. (1984). Reliability and validity of the secondary level English proficiency test. *System, 12*(1), 1-12.

6. *Quattro Pro Version 6.0.* (1994). Novell, Inc.

7. *SPSS Version 7.0.* (1995). SPSS Inc.

8. Adams, R. J., & Khoo, S-T. (1996). *Quest: The interactive test analysis system version 2.1.* [computer program]. Victoria, Australia: The Australian Council for Educational Research.