PAC3
at
JALT
2001
Conference
Proceedings

International
Conference
Centre
Kitakyushu
JAPAN
November
22–25, 2001

Results and Interpretation of Scores from a Program Designed for Rapid Oral Testing

David W. Dugas
Daejeon University

A rapid oral testing program for lower-level EFL conversation students, developed at Daejeon University in Korea, was part of a course in which homework, class work and tests were tightly integrated with syllabus objectives. Since the research classes differed in many ways, a group of 50 students was randomly selected from among all the classes each semester. Difference scores of sample groups showed highly significant improvement between the first and final tests for fall 1999, spring 2000, and fall 2000 (alpha = .01, using the Wilcoxon signed-ranks test). The oral test results occur as partial scores (production, errors, delay), a combined score, an efficiency index (p/e), and the test marking patterns. This work shows potential for practical achievement testing of lower level university conversation classes. The statistical properties of the test results, as well as the reliability and validity of the tests are under consideration.

Oral testing of lower level conversation students is problematic. Often such students are unable to sustain the "authentic dialog" required by modern speaking tests. Even assuming valid and reliable proficiency tests are used, they may not have the sensitivity needed to show progress for many students at this level after only a single semester. Also in our case, administration of proficiency tests to 2000 students

twice a year would require technical support, time and personnel that were not available. Clearly, there was need for a 'low-tech' speaking achievement test which could be adapted to the limited, specific objectives of classes for early learners. Such a test was developed over the last five years in Korea at Daejeon University as one component of a new design (Richards and Rodgers, p.20) for a lower level conversation course.

Parts of this program have been described (Dugas 2000, 2001a), and a detailed review of current practices is available in the form of a workshop manual (Dugas 2001b).

The Students and Classes

If others are to judge whether this program might be useful, they need to know the academic context in which the results occurred. During the research period, conversation classes at our university occurred in three formats. Two types of day classes met during the interval from 9 AM to 6 PM. One was two periods long, met once a week and was attended by non-English majors. Another met twice a week for a total of three periods, and was attended by freshmen English majors. Night classes occurred from 6:00PM to 10:30PM, and met once a week. Night classes contained older students with jobs as well as younger full-time and part-time students. The classes had 30% less content due to being about 10% shorter, and because students were usually in poorer condition after 6 PM.

Class sizes ranged from 11 to 38 students, with an overall average of 25 students (Table 1). Spring classes were much larger, on average about 33 students, than fall classes, with about 21 students. A trend in our classes toward increased mixing of students from different majors is best indicated by the numbers before and during fall 2000 in Table 1, when the average number of majors jumped from 4 per class to 6. With the exception of a few classes for new English majors (97-100% freshmen), Table 1 shows that more than 90% of the students in my research classes were sophomores, with an overall average of about 95 %.

Table 1: *Description of Classes in Fall 1999, Spring 2000, and Fall 2000*

Students per class	average	20	33	22
	range	11-29	21-38	14-29
Majors per class	average	4	4	6
	range	2-6	3-5	3-8
% of Sophomores*	average	91	98	95
	range	64-100	94-100	89-100

* Excluding first year classes for English majors, which ranged from 97 to 100% freshmen.

Classroom instruction offered during a semester (after removal of daily breaks, holidays and cancelled classes) ranged from 21 to 37 hours (Table 2). Twenty-four hours per semester were offered for typical day classes, 22 hours for night classes, and 36 hours for freshmen

English majors. (Instruction time varies from professor to professor, according to details of holidays, cancellations, sickness, etc.) Students attended an average of 84 % (range 72 - 89) of the time offered. Contrast this with the informed assertion that 150 - 180 hours/year of instruction are needed if students are to make enough progress to reap significant economic rewards (Nunan).

Table 2: *Hours* of Classroom Instruction Per Semester*

	Hours Offered		Hours Attended**		Percent Attended
	Average	Range	Average	Range	
Fall 1999					
Day Classes (4)	24	23-25	17	15-19	72
Night Classes (2)	22	21-23	17	17-18	81
For E Majors (1)	36	—	32	—	88
Spring 2000					
Day Classes (4)	23	22-23	20	19-21	86
Night Classes (2)	21	—	19	18-19	89
For E Majors (1)	35	—	30	—	86
Fall 2000					
Day Classes (3)	25	—	22	21-23	86
Night Classes (2)	22	21-23	19	—	87
For E Majors (1)	37	—	31	—	83

* Actual time in class after subtraction of time for holidays, canceled classes and scheduled breaks.

* *Calculated from averaged percent attendance of students in each class multiplied by hours offered.

A classroom survey taken during Fall 2000 provided other information about the background and study habits of students in those classes. Based on student estimates, 80% had studied English for six or more years before taking the course (Table 3). Most of this study time was presumably focused on reading, writing, vocabulary, and grammar, for 60% had spent 2 years or less learning to speak English. Students also estimated how much time they studied each week outside of class. Day students (non-majors) claimed an average of about 2.7 hours of study per week (range 2.3 - 3.5), while night students claimed an average of 3.7 hours (range 3.2 - 5.0).

Table 3: *Estimated Amount of English Study before Fall 2000*

Years of Study	General Study*	General Average	Oral Study*	Average for Conversation
0-2	8	0.8	49	0.8
2-4	3	3.2	24	3.1
4-6	6	5.8	6	5.1
6-8	58	7.4	2	6.4
8-10	11	9.9	0	—
	86 respondents		81 respondents	

* Number of students in each time interval.

Most often, both the English speaking ability and interest of the non-major students was marginal. By comparison to typical evaluation criteria, my impression was that most would have been rated from middle novice to low intermediate in oral proficiency interviews of the ACTFL type. Only a few students appeared to prepare for classes, or do any follow-up study outside of class, except when exams were imminent. This meant that the backwash effect (Hughes, p.1-2) from the exams was the primary learning process for most of my students, and prior to the new course design, there was no way to benefit from this situation.

The New Oral Test Design

My response to this unproductive situation was to restrict the course to a well-defined body of practical content appropriate to the usual level of English ability, to greatly increase the role of in-class repetition, and to focus on oral practice and oral testing. Each semester, use of two types of exams provided a progressive challenge (high scores: Type 1 exam = 40, Type 2 = 60, Type 3 = 90). Note that the series provided only potential for, not a guarantee of, a higher score. For example, a student who got a score of 35 on a Type 1 exam, but made no further effort at preparation, would likely score about 35 regardless of the type of exam. The scoring method for all test types was the same, so that higher scores showed real improvement in use of the target processes and content. The question became whether students would get higher scores as they worked through the course, and became familiar with the test process.

Before we answer that question, it must be clear what the scoring form looks like and how it works. Figure 1A (see Appendix 1; modified from Dugas 2001b, p.36) has labels on the functional units of the form. The central column contains the test items. The left and right sides (adjacent to the central column) are near mirror images and serve to record the marks for the two students during their performance. The columns labeled AAA show the weighting of each test item and serve to record the row totals for calculation. The blocks labeled BBB record the total positive points awarded to each student during the test (hereafter called 'production').

During production, a variety of mistakes might be marked and scored as negative, penalty points (hereafter called 'errors'). Unacceptable delays during production also result in penalty points (hereafter called 'delay'). Tests might also contain administrative penalties for unacceptable behavior during exams. The totals of negative points for each student are recorded in the red blocks labeled CCC. Subtotals are calculated and placed in the blocks labeled DDD. The base score (15 points to keep results positive even after poor performances) is added to complete the 'combined score', which is then recorded in the blocks labeled EEE.

Test Results

There are several useful aspects of results from these tests. There is, of course, the information in the partial scores (production, errors, delay). While interpretation of delay and production is straightforward, errors are more complicated. The absolute number of errors generally rises as students do more talking on the later tests. Just looking at errors would give an incomplete impression, and use of an efficiency index (p/e) is one way to avoid this. It is important to see errors in relation to changes in production. For this reason, changes in this index provide a useful reference to the quality of work. The combined score incorporates all the results. Finally, the pattern of marks on the scoring form provides information useful in explaining student strengths or weaknesses. Following are examples taken from real exams to illustrate this point.

Figure 1B (see Appendix 1; modified from Dugas 2001b, p.53) shows results from an extended answer (Type 2) test. Student A produced well with only one mistake, to finish with a score above the nominal maximum. Student B produced poorly and made mistakes, made even worse by delays. Figure 1C (see Appendix 1) shows results from an interactive chain (Type 3) test. Student C has accomplished the test goals, used many complex sentences, and made only a few mistakes to score well above the nominal maximum. The production of Student D was only a little off, but many errors and many delays resulted in a poor score. Students B and C each made one serious pronunciation error, as noted. Student D made an error by reversing the normal word order, as noted. Each round of effort (1-5) corresponds to one topic on the test card. For a Type 2 test, a good round has one question and three answers. A type 3 test requires three questions and three answers in each round. Once a teacher is familiar with the test results, a glance at these patterns shows where the student strengths and weaknesses are, regarding the test tasks.

To find out whether students were getting positive results after a semester of study, a sample of fifty conversation students was randomly selected each semester. The difference scores for each student selected, created by comparing 'production' and 'combined score' on the first and last tests, were always from two different exam types (i.e. exam types 1 and 2, or exam types 2 and 3). Results from first and last exams of each selected student were considered two related measures from a single sample. The data were considered interval level measures with students acting as their own controls.

Two operative definitions for describing positive results were needed to provide a basis for the necessary hypotheses. Groups of students which showed significant increases in combined scores were defined as showing clear Improvement. Groups of students which showed significant increases in production, regardless of any other problems, were defined as showing Progress.

The logical framework of the analysis is shown in the experimental design in Table 4 (adapted from Wiersma, p.107-108).

Table 4: *Experimental Design for Two Related Measures from One Sample*

Group	Pretest	Treatment	Post-test
RG1	$OTi_{(p,cs)}$	X	$OTf_{(p,cs)}$
RG2	$OTi_{(p,cs)}$	X	$OTf_{(p,cs)}$
RG3	$OTi_{(p,cs)}$	X	$OTf_{(p,cs)}$

R = randomly selected OTi = initial oral test; OTf = final oral test, p = production, cs = combined score Treatment = sum of English study, practice and testing experience prior to exam 4.

The hypotheses necessary for testing were formulated as shown below. H_o stands for the null hypothesis; H_A stands for the alternate hypothesis.

Production hypotheses: To infer whether groups made progress as defined by P.

$H_o : P_f = P_i$

(production score on final test = score on initial test) ($\alpha = .01$)

$H_A : P_f > P_i$

(production score on final test > score on initial test)

Production Decision: If H_o is rejected, then H_A is true. We would infer that students had significantly higher scores in production on their final exams, and thus showed progress as defined.

Combined score hypotheses: To infer whether groups improved as defined by CS.

$H_o : CS_f = CS_i$

(combined score on final test = score on initial test) ($\alpha = .01$)

$H_A : CS_f > CS_i$

(combined score on final test > score on initial test)

Combined Score Decision: If H_o is rejected, then H_A is true. We would infer that students had significantly higher combined scores on their final exams, and thus showed improvement as defined.

The results of testing the difference scores using the Wilcoxon signed-ranks test (Siegel and Castellan, p.87-88; 91-95) are presented in Table 5. If z_p is less than or equal to alpha (α), the null hypothesis (H_o) is rejected. The number represented by z_p is the probability that the null hypothesis is true.

Table 5: *Rapid Testing Results for Three Semesters of 1999 and 2000*

Semester	Value for z	P (α)	P (z_p)	Rej H_o	Value for z	CS (α)	CS (z_p)	Rej H_o	Test Results*
Fa1999	6.154	.01	<.000005	Yes	6.125	.01	<.000005	Yes	SP and SI
Sp2000	6.154	.01	<.000005	Yes	5.999	.01	<.000005	Yes	SP and SI
Fa2000	6.154	.01	<.000005	Yes	6.038	.01	<.000005	Yes	SP and SI

*Showed Progress (SP = significantly > P)
Showed Improvement (SI = significantly > CS)

Conclusions

These results show that, within the definitions provided, and for the tasks which they were trained to do, students were able to demonstrate improvement in speaking tasks each semester using this program. This important result was necessary to justify further work to identity and rectify remaining weaknesses. For similar students in required English classes at other universities, this program might provide a viable new approach to teaching and testing lower level EFL conversation.

Whether it ever does, depends on several unresolved issues. There is much published support for the notion that difference scores (or gain scores) are inherently unreliable (for example, Cronbach and Furby). That this is not always true has been pointed out and supported by Zimmerman and Williams (1998, 1982) provided that the original scores are reliable. Since determination of the statistical properties of the test results, as well as the reliability and validity of these exams, is the next phase of work, this program may be modified by those results. Also, it is not known whether significant gains on these achievement tests relate to gains in English ability. Another goal of future research is to find out if the tests may serve as an index to English proficiency.

Regardless of the outcome of future work, the logical approach, integrated syllabus objectives, and strong efforts at objectivity through the use of formal protocols make this program superior to anything I have used before. For this reason, I urge that teachers formally evaluate the benefits and liabilities of their current testing programs. The professional growth which occurs during such an evaluation will certainly improve conversation testing in Asia.

References & Citations

Cronbach, L. J. & Furby, L. (1970). How we should measure change - or should we? *Psychological Bulletin,* 74(1), 66-80.

Dugas, D. W. (2000). A Program for Oral Testing of EFL Students in Korea. *KOTESOL Proceedings PAC2*, 75-87.

Dugas, D. W. (2001a). Revised Protocols for a Series of Rapid Oral Tests. *KOTESOL Proceedings 2000*, 239-241.

Dugas, D. W. (2001b). *Rapid Testing of Low-Intermediate Spoken English: A Workshop About Teaching EFL in a Way Which Allows Rapid Oral Testing*. (Available from author, 96-3 Yong-un Dong, Dong Gu, Daejeon 300-716 Korea)

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge, England: Cambridge University Press.

Nunan, D. (2001). *English as a global language*. Paper presented at PAC3 (3rd Pan Asia Conference), Kitakyushu, Japan.

Richards, J. C. & Rodgers, T. S. (1986). *Approaches and Methods in Language Teaching*. Cambridge, England: Cambridge University Press.

Siegel, S. J. & Castellan, Jr., N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. (International edition). Singapore: McGraw.

Wiersma, W.(2000). *Research Methods in Education - An Introduction* (seventh edition). Massachusetts: Allyn & Bacon.

Zimmerman, D. W. & Williams, R. H.(1998). Reliability of gain scores under realistic assumptions about properties of pre-test and post-test scores. *British Journal of Mathematical and Statistical Psychology,* 51, 343-351.

Zimmerman, D. W. & Williams, R. H.(1982). Gain scores in research can be highly reliable. *J. Educational Measurement*, 19(2), 149-154.

Appendix 1

Figure 1: Scoring form and two kinds of test results

A: Labelled example of a single unit of the Rapid Scoring Form

Student #1								Student #2						
	AAA												AAA	
Sums	Wt	1st	2nd	3rd	4th	5th	Topics from list#	1st	2nd	3rd	4th	5th	Wt	Sums
	3x						Correct start with Q or R						x3	
	2x						Answer: complete sentence						x2	
	1x						Answer: complex sentence						x1	
	1x						Affirming interaction						x1	
	2x						Self-correction						x2	
BBB							points earned							BBB
	-1x						Usage errors (see protocol)						x-1	
	-1x						Delay (each 5 seconds)						x-1	
	-1x						Prompting (any form						x-1	
CCC							penalties							CCC
DDD							points minus penalties							DDD
+	15 =	EEE	< Score		WO = word order VC = vocabulary PN = pronunciation WA = wrong answer					Score >		EEE	= 15	+

B: Scoring patterns on a Type 2 Exam (extended answer; n.m. of 60)*

Student #1: Student A								Student #2: Student B						
Sums	Wt	1st	2nd	3rd	4th	5th	Topics from list#	1st	2nd	3rd	4th	5th	Wt	Sums
15	3 x 5	I	I	I	I	I	Correct start with Q or R	I	I	I		I	4 x 3	12
30	2 x 15	III	III	III	III	III	Answer: complete sentence	I	I	II	I		5 x 2	10
8	1 x 8	I	III	II	II		Answer: complex sentence	I	I				2 x 1	2
	1 x						Affirming interaction						x 1	
	2 x						Self-correction				I		1 x 2	2
53							points earned							26
-1	-1 x 1				I		Usage errors (see protocol)	IIIII	II	III	I		11 x -1	-11
	-1 x						Delay (each 5 seconds)	IIII			IIIII	IIIII	14 x -1	-14
	-1 x						Prompting (any form						x -1	
-1							penalties				PN: Permanent			-25
52							points minus penalties							1
+	15 =	67	< Score		WO = word order VC = vocabulary PN = pronunciation WA = wrong answer						Score >	16	= 15	+

C: Scoring patterns on a Type 3 Exam (interactive chain; n.m. of 90)

Student #1: Student C								Student #2: Student D						
Sums	Wt	1st	2nd	3rd	4th	5th	Topics from list#	1st	2nd	3rd	4th	5th	Wt	Sums
45	3 x 15	III	III	III	III	III	Correct start with Q or R	III	III	III	III	I	13 x 3	39
30	2 x 15	III	III	III	III	III	Answer: complete sentence	III	III	III	II	III	14 x 2	28
12	1 x 12	II	II	II	III	III	Answer: complex sentence	I	II	I			4 x 1	4
	1 x						Affirming interaction						x 1	
	2 x						Self-correction						1 x 2	
87							points earned							BBB
-5	-1 x 5		II	II		I	Usage errors (see protocol)	II	IIII	IIIII	IIIII	I	17 x -1	-17
	-1 x						Delay (each 5 seconds)	II	IIIII	III	II	IIII	15 x -1	-15
	-1 x						Prompting (any form		I				1 x -1	1
-5			PN: describe				penalties			WO				-33
82							points minus penalties							40
+	15 =	97	< Score		WO = word order VC = vocabulary PN = pronunciation WA = wrong answer						Score >	55	= 15	+

*n.m. = nominal maximum: highest score possible with no errors and no special credits (production rows 3-5).