



筑紫女学園大学リポジット

Diagnosing Communicative Ability in the Classroom : A holistic approach

メタデータ	言語: eng 出版者: 公開日: 2014-02-17 キーワード (Ja): キーワード (En): 作成者: STEWART, Jan, STEWART, Jan メールアドレス: 所属:
URL	https://chikushi-u.repo.nii.ac.jp/records/252

Diagnosing Communicative Ability in the Classroom

— A holistic approach

Jan STEWART

1. Introduction

The Communicative Diagnostic Tests (CDT 01–05) are the conceptual progeny of the English Diagnostic Test (see Stewart 1999; 2000; 2001). Whereas the English Diagnostic Test remained unchanged throughout its three-year life span, these second-generation communicative tests have changed every year since their inception. Although they retain certain aspects of their progenitor, namely, sections which test structure and reading comprehension, the Communicative Diagnostic Tests offer innovation in that they test purely communicative skills that focus on meaning, not form. Most significantly, in the case of CDT-05, a test of speaking skills was added.

In this paper I will present the results of the CDT-05, comparing students' average scores in structure, reading, listening, speaking, and communication skills. These results will be analyzed according to performances by 1st-, 2nd-, 3rd-, 4th- and 5th-year students.¹ Items proving to be most difficult will then be analyzed. The Diagnostic Test itself will be examined in light of item facility, item discrimination, and reliability, in order to determine its strengths and weaknesses, and to decide whether it has accurately described the sample of students in question.

2. Method

The Communicative Diagnostic Test (CDT-05) consists of five parts: Structure, Communication, Reading, Listening and Speaking. Part 1 has 15 multiple-choice structure questions (1 correct answer and 3 distractors). Part 2 has three sets of communicative matching questions. Part 3 has a reading passage and 5 multiple-choice questions (two points each). Part 4 has 5 listening questions (1 correct answer and 2 distractors). Part 5 consists of a speaking test, scored at 5 points. The Test therefore has a total of 50 points, which are doubled to arrive at a total score of 100%.

The CDT-05 was administered on the first day of class, Spring Semester, 2005. Out of 93 respondents, 30 (32.3%) scored lower than 60% (compared to 18.1% of students who had “failed” the EDT). The average score on the CDT-05 was 65.57%, with scores ranging from 28% to 97%.

Each item was analyzed according to the number of students who answered correctly. This is known as *Item Facility*. Twelve items were answered incorrectly by more than 40% of the sample.

A statistical measurement known as *Item Discrimination* was used to compare students who did well on the entire test with those who did poorly (Brown, 1996). First the students were divided into three groups – the top 27%, the bottom 27%, and the middle 46% (in this study 27% of 93 respondents came out to 27 students in the upper group and 25 students in the lower group). Then the percentage of correct answers of the bottom 27% was subtracted from that of the top 27%. The remainder was expressed as a decimal figure. These figures were compared to Ebel’s (1979) Item Discrimination scale to determine the strengths of individual test items.

Finally, the reliability of the test was evaluated. A split-half reliability measurement, which compared students’ answers on the odd-numbered items with their answers on the even-numbered items, was applied. An initial reliability estimate was made using the *Spearman-Brown Prophecy formula*, based on the correlation between the two halves of the test (Brown, 1996). Two further measures of internal consistency, based on the standard deviation, the mean of the test scores, the Item Facility and the Item Difficulty, known as *K-R 20* and *K-R 21*, were used to give additional estimates of the reliability of the test (Kuder and Richardson, 1937). Finally, these results were verified using Brennan’s (1984) phi-coefficient measure for test consistency.

3. Result

In Part 1 (Structure), four items were answered correctly by fewer than 60% of the sample. In Part 2 (Communication) also, four items showed poor results. In Part 3 (Reading), all five items were remarkably poor. In Part 4 (Listening), 2 items scored rather poorly. Table 1 gives the general results of the test, in terms of average scores:

Table 1. Mean Scores (%): Total, and Section by Section

Category	No.	Total	Struct	Comm	Read	Lstng	Spkg
1st year	(48)	63.17	64.03	69.44	45.00	72.08	69.17
2nd year	(17)	65.53	60.78	65.88	63.52	72.94	74.12
3rd year	(2)	51.00	60.00	63.33	10.00	60.00	60.00
4th year	(21)	70.29	66.67	73.65	54.29	84.76	88.57
5th year	(5)	74.80	72.00	84.00	52.00	84.00	92.00
TOTAL	(93)	65.57	64.40	70.50	50.10	75.41	75.40

Structure

Four structure questions were answered correctly by fewer than 60% of the students. These questions are detailed below, along with their analyses, shown in Tables 2–5.

d . That's a charge for room service

e . I'm going to do it tomorrow.

Table 7. Communication Question 10

	a	b	c	d	e	n.a.
1st year (48)	12.50	14.58	12.50	43.75	12.50	4.17
2nd year (17)	17.65	17.65	17.65	35.29	11.76	0
3rd year (2)	0	0	50.00	50.00	0	0
4th year (21)	9.52	4.76	14.29	61.91	4.76	4.76
5th year (5)	0	0	0	100	0	0
TOTAL (93)	11.83	11.83	13.98	49.46	9.68	3.22

(Percentage of students who answered each option)

Two additional questions, from *Passport Plus*, read:

12. And what's the message?

a . I can't make it to the meeting.

15. What did you think of the meeting?.

b . It went very well.

c . There's one every hour.

d . I know where there's a sale.

e . About an hour.

Table 8. Communication Question 12.

	a	b	c	d	e	n.a.
1st year (48)	47.92	41.67	4.17	2.08	2.08	2.08
2nd year (17)	52.95	35.29	5.88	5.88	0	0
3rd year (2)	50.00	50.00	0	0	0	0
4th year (21)	47.62	42.86	0	0	0	9.52
5th year (5)	80.00	0	20.0	0	0	0
TOTAL (93)	50.54	38.71	4.30	2.15	1.07	3.23

(Percentage of students who answered each option)

Table 9. Communication Question 15

	a	b	c	d	e	n.a.
1st year (48)	45.83	45.83	2.08	4.18	0	2.08
2nd year (17)	47.06	47.06	5.88	0	0	0
3rd year (2)	50.00	50.00	0	0	0	0
4th year (21)	38.10	47.62	0	0	4.76	9.52
5th year (5)	20.00	80.00	0	0	0	0
TOTAL (93)	43.01	48.39	2.15	2.15	1.07	3.23

(Percentage of students who answered each option)

Reading

The Reading section consists of a 1-page passage (430 words) about a popular baseball player. The text includes information about his offensive and defensive performances, awards received, records broken, his career history, and some personal details. The questions concerned reading for information (#1), finding the main idea (#2), inferencing (#3), determining cause and effect (#4), and drawing conclusions (#5). Questions 1 and 2 were answered by about 64% of the students. Questions 3, 4 and 5 were answered correctly by 39%, 47%, and 35%, respectively. The low scores in this section indicate that critical thinking skills are a particular weakness among the sample tested (see Stewart, 2003).

Table 10. Reading (%): Question by Question

	Q #1	Q #2	Q #3	Q #4	Q #5	Total
1st year (48)	54.17	56.25	37.50	43.75	33.33	45.00
2nd year (17)	76.47	88.24	29.41	64.71	58.82	63.52
3rd year (2)	50.00	0	0	0	0	10.00
4th year (21)	71.43	71.43	47.62	47.62	28.57	54.29
5th year (5)	80.00	60.00	60.00	40.00	20.00	52.00
TOTAL (93)	63.44	64.52	38.71	47.31	35.48	50.10

(Percentage of students who answered each question correctly)

Listening

The Listening section was based on a script, adapted from *Systematic Listening for the TOEIC Test* (Hokuseido Press, 2005). It was about a lady who went into an ice cream shop and ordered a vanilla cone. While waiting, she noticed that a famous actor was also there, but she tried to act nonchalant in front of him. Outside the shop, she realized that she had forgotten her ice cream cone. She went back, but the actor told her she had put the cone into her purse along with her change. The questions follow:

1. What did the lady order in the ice cream shop?
2. Whom did the lady see there?
3. What did the lady do inside the shop?
4. What did she do outside the shop?
5. What happened in the end?

Table 11. Listening: Question by Question, and Total (%)

	Q #1	Q #2	Q #3	Q #4	Q #5	Total
1st year (48)	52.08	97.92	66.67	56.25	89.58	72.08
2nd year (17)	47.06	100	70.59	70.59	76.47	72.94
3rd year (2)	0	100	100	50.00	50.00	60.00
4th year (21)	80.95	95.24	80.95	80.95	85.71	84.76
5th year (5)	80.00	100	100	80.00	60.00	84.00
TOTAL (93)	58.06	97.85	73.12	65.59	83.87	75.41

(Percentage of students who answered correctly)

Speaking

The Speaking section of the test consisted of an oral interview, in which students were called to the front of the room in pairs (or by threes) and interviewed by the examiner.

The oral interview took the form of a question-and-answer session. First the examiner asked each student questions. Then the students asked questions of each other, and answered them. Content was based on a set of eight illustrations: 1) a brief narrative, 2) an office scene, 3) a number of people telling what they would bring to a picnic, 4) a family meeting their homestay guest at the airport, 5) another office scene, 6) a teacher standing in front of her class, 7) a map of a city, and 8) a man and a woman in various locations, including a restaurant, a library, a bank, and at home in the kitchen.

The illustrations lent themselves easily to testing communicative functions such as describing people, asking and telling the time, asking and giving information about routines and habits, what someone is doing at the moment, simple narratives, reporting what people say, future plans or intentions, signs and notices, giving directions, information about places, comparisons, expressing cause and result, making/granting/refusing simple requests, and expressing ability in the past or present.³

Scoring of the oral interview was done in accordance with the criteria set by the Cambridge Examination Board for the *Preliminary English Test* (P.E.T.). The criteria included Fluency, Accuracy, Pronunciation, and the degree to which the students accomplished the Task at hand. These criteria were awarded a total of 25 points each, for a total of 100 points (%) on the Speaking Section (although this counted as only 10% of the entire score of the CDT-05). The results are shown in Table 12.

Table 12. Speaking: Section by Section (25 points each), and Total (%)

		Fluency (/25)	Accuracy (/25)	Pronunc. (/25)	Task (/25)	Total (100%)
1st year	(48)	16.67	16.92	17.69	17.71	68.98
2nd year	(17)	17.24	18.41	18.94	19.88	74.47
3rd year	(2)	13.00	15.00	16.50	18.00	62.50
4th year	(21)	21.29	21.38	21.52	21.52	85.71
5th year	(5)	21.00	21.60	21.80	21.00	85.40
TOTAL	(93)	17.97	18.41	18.98	19.15	74.51

4. Discussion

a. Item Facility

Item Facility analysis is important because it tells us, at a glance, how many students answered each question correctly (Brown, 1996, 64-66). In CDT-05, Item Facility ranges from about 28% of students on some items to over 97% on other items. Those questions that were answered by more than 60% of the respondents are considered to have been adequately learned by the sample in general. Questions that were answered by lower percentages, however, indicate that remedial instruction is necessary.

There is always the possibility that individual test items render low Item Facility scores for reasons *other than* low student proficiency levels. Test items that contain internal flaws may mislead students. One way to determine whether a test question is “good” or “bad” is to conduct an *Item Discrimination* analysis.

b. Item Discrimination

Item Discrimination compares students who did well on the entire test with those who did poorly (Brown, 1996, 66-69). First we divide the students into three groups: the top 27%, the bottom 27%, and the middle 46%. Then we subtract the number of correct answers of the bottom 27% from those of the top 27%. The remainder gives us the Item Discrimination score, expressed as a decimal figure. Usually, an item which falls within a range of .30 – .70 is considered a “good” test item. Ebel (1979, 267) proposed the following guidelines for evaluating test items:

.40 and above	Very good
.30 to .39	Good, but may need improvement
.20 to .29	Needs improvement
≤ .19	Reject

Items with low Item Frequency and high Item Discrimination may be considered effective indicators of poor student performance. Examples on the CDT-05 include Structure questions #1 (I.F. = 50.53, I.D. = 0.46), #6 (I.F. = 34.41, I.D. = 0.43); Communication questions #3 (I.F. = 54.84, I.D. = 0.56), #10 (I.F. = 49.46, I.D. = 0.66), #12 (I.F. = 50.54, I.D. = 0.46), #15 (I.F. = 48.39, I.D. = 0.54); Reading questions #4 (I.F. = 47.31, I.D. = 0.54), #5 (I.F. = 35.48, I.D. = 0.48); Listening question #1 (I.F. = 58.06, I.D. = 0.57).

c. Reliability

The *reliability* of a test tells us to what extent the results obtained on one occasion will be consistent with the results obtained if the test is given on another occasion (Brown, 1996, 185). For several reasons, however, giving the test on a different occasion presents insurmountable difficulties. The most obvious is that this particular test was given on the first day of class. Students' expectations are high, and the element of surprise cannot be reproduced at a later date. Even if students are to be given exactly the same test later, there is always the possibility that by that time they will have *learned* something, either as a result of attending lectures, or possibly as a result of having taken the test the first time. Moreover, it is difficult to ensure consistent attendance, not to mention the possibility of illness or other personal circumstances that may tend to affect performance on a second test.

Therefore it is necessary to *simulate* the second test. In order to do this, one single test can be divided into two halves, one half consisting of the odd-numbered questions, the other half consisting of the even-numbered questions (Brown, 1996, 194-6). Using the *Spearman-Brown Prophecy formula*, the CDT-05 yielded a reliability estimate of 0.80, with a Standard Error of Measurement of 2.57. The Cronbach alpha score, based on the standard deviation of each half, rendered a reliability estimate of 0.81, with a Standard Error of Measurement of 2.50.

However, the split-half method itself may not be the most reliable instrument for evaluating reliability, inasmuch as a random re-ordering of test items would not consistently produce the same odd-even combination. Theoretically, if all the high-facility test items were aligned in the odd-numbered positions, and all the low-facility items in the even-numbered positions (or vice versa), a quite different set of figures would emerge for the reliability estimates obtained using this method.

Whereas the Spearman-Brown Prophecy formula and the Cronbach alpha use the split-half method, Kuder and Richardson (1937) devised a method which divides the test into two *different* halves: Item Facility (percent correct) vs. Item Difficulty ($1 - I.F.$). Calculations using these halves, combined with the standard deviation, produce more reliable results than the split-half (odd-even) method (Brown, 1996, 197-203).

Kuder-Richardson formula 21 tends to *underestimate* the reliability of a test. Using *K-R 21*, the CDT-05 had a reliability estimate of 0.75, with a 2.87 Standard Error of Measurement. *K-R 20*, a much more accurate measure of reliability (based on the Item Variance), gave the CTD-05 a reliability estimate of 0.78, with a 2.69 Standard Error of Measurement.

The above calculations apply to criterion-referenced interpretations. While the CDT-05 is a criterion-referenced test, it must be pointed out that there is a difference between tests in which items are sampled from specific course objectives, and those in which items are sampled from a general but well-defined domain of behaviors. The results on a domain-referenced test can describe a student's status relative to that domain of behaviors, rather than subtests of specific course objectives. Analysis of the consistency of domain-referenced tests, however, is also appropriate for analyzing objectives-referenced tests (Brown, 1996, 218-19).

Brennan (1984, 315–316), reported a general-purpose estimate of the dependability of a domain-referenced test based on the proportion scores (the percentage of correct answers on the actual number of test items, not weighted). He demonstrated that for any given test, $K-R\ 21 < \phi < K-R\ 20$. The Communicative Diagnostic Test, therefore, should have a test *consistency* of between .75 and .78. This was borne out in calculations, which gave the CDT-05 a *phi* coefficient of 0.756616.

5. Conclusion

In general, the Communicative Diagnostic Test 2005 proved to be more difficult than the English Diagnostic Test 1998-1999 (having average scores of 65.6% and 72.2%, respectively). This imbalance was offset somewhat by the higher scores in the Speaking and Listening sections of the CDT-05, but it was exacerbated by much lower scores in the Reading section. Scores in the newly added Communication section were about five percentage points higher than the whole-test average.

Five out of fifteen Structure items on the CDT-05 had Item Discrimination values over 0.40; nine out of fifteen items on the Communication section were also rated "very good." Three items in the Reading section were rated "very good," while the other two items were rated "good, but may need improvement." Two items in the Listening section scored over 0.40, but the other three items scored very poorly in terms of Item Discrimination.

One of the most significant innovations of the CDT-05 was the addition of an oral interview, which strengthens the focus on "communication" in testing.⁴ Average scores on the Oral Interview were nearly 10 points higher than the whole-test total average score, yet (interestingly) nearly equal to the average scores on the Listening section.

The CDT-05 is a potentially reliable instrument, yet students' performances on a test will vary for a number of reasons. The components of language competence that may affect meaningful

variance on tests include Grammatical Competence, such as vocabulary, morphology, syntax, phonology; Textual Competence, such as cohesion and rhetorical organization; Illocutionary Competence, such as ideational, manipulative, heuristic, and imaginative functions; Sociological Competence, such as sensitivity to dialect, register, naturalness, and the ability to interpret cultural references and figures of speech (Bachman, 1990). The reliability (or lack of reliability) of the Communicative Diagnostic Test may result from a combination of these and other factors.

A diagnostic test can be useful in two ways. Teachers can use it to determine whether students have been properly placed, in programs where placement occurs, or they can use it to determine students' strengths and weaknesses once instruction begins. Diagnostic tests eliminate redundancy, allowing teachers to focus on areas that need the most attention.

As such the diagnostic test forms part of an overall testing program, which progresses in four stages. The first stage consists of the initial screening procedures (entrance exams). The second stage entails placement of students. The third stage involves diagnosis. The fourth stage concerns achievement. In all stages of testing, teachers or program planners may gather information that helps them improve the curriculum.

A "systems approach" to language curriculum planning includes needs analysis, setting goals and objectives, materials development, teaching methods, testing, and program evaluation. Needs include not only the needs of the students; there are many other factors involved. Teachers, administrators, employers, institutions, societies, even nations may have a stake in what transpires in the classroom. Program goals are statements of what must be accomplished to satisfy students' needs. Objectives, on the other hand, are statements of the exact skills that students must possess in order to achieve these goals. Materials may be developed or selected according to decisions made concerning the above factors. Testing, including a diagnostic test, affords us the opportunity to analyze the appropriateness of the course objectives for students' abilities. When all relevant information has been systematically collected, we are able to evaluate the effectiveness of the language program, and promote the improvement of the curriculum. These six factors, needs analysis, goals and objectives, materials, methods, testing, and evaluation, interact with one another to form a dynamic, ongoing process of curriculum development. In programs that emphasize *communication*, a Communicative Diagnostic Test plays an important role in that process.

Notes

- 1 The “years” were determined by Student Identification Numbers, which reflect the date students entered the university. Some, due to their participation in study abroad programs (or other reasons) found themselves in their fifth year of undergraduate study at the time of testing.
- 2 Distractor (a) is ruled out, being the only correct answer for item # 3: “Mr. Parker is in a meeting.”
- 3 A complete list of functions appears in the Examiner’s Manual for the Cambridge *Preliminary English Test* (P.E.T.).
- 4 Underhill (1987) writes, “When a learner says something that is relevant and true (for himself at least), to someone else who is interested and has not heard it before (from that speaker, at least), then that act of speech is communicative” (p.8). However each part of this statement may be contested. Interest and relevance may be important in learning and retention, but they have no major impact on communication. Too, is a lie any less “communicative” than the truth? Of course the truth is more desirable, but both transmit meaning from one person to another. Finally, repetition of a statement does not necessarily negate its communicative qualities.

References

- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Brennan, R. (1984). Estimating the dependability of the scores. In R. Berk (Ed.), *A guide to criterion-referenced test construction*. Baltimore: Johns Hopkins University Press.
- Brown, J.D. (1996). *Testing in language programs*. Upper Saddle River, New Jersey: Prentice Hall Regents.
- Ebel, R. (1979). *Essentials of educational measurement* (3rd ed.). Englewood Cliffs, New Jersey: Prentice-Hall.
- Kuder, G., and M. Richardson (1937). The theory of estimation of test reliability. *Psychometrika*, 2, 151–160.
- Stewart, J. (1999). Diagnosis of students’ English needs in three universities, 1998–1999. Paper presented at the JACET Kyushu-Okinawa Chapter 15th Annual Conference at Miyazaki, November 27, 1999.
- _____ (2000). Item facility, item discrimination, and reliability of the English Diagnostic Test. JACET Kyushu-Okinawa Chapter, *Annual Review of English Teaching and Learning*, No.5, 65-83.
- _____ (2001). Increasing instructional efficiency from day one – The English Diagnostic Test. Chikushi Jogakuen University *Bulletin*, No.13, 205-226.
- _____ (2003). Using critical and creative thinking skills to transcend the grammar barrier – Convergent questioning. *Bulletin of the International Cultural Research Institute*, Chikushi Jogakuen University & Junior College, No.15, 13-25.
- Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques*. Cambridge: Cambridge University Press.