# The Use of Data Analysis in Preparation for the TOEIC

## John A. Dolan

# I . Action Research Introduction

This paper sets out to show how the use of data obtained from in-class practice tests can be analyzed to benefit instruction for a TOEIC preparation class. It is the result of a real need on the part of the instructor to find a way to sift through the many prescriptions for TOEIC success to find the bare minimum number of principles that might be applicable to the particular group he is working with, and for the restricted time period allowed. It is not an extensive survey, either in the number involved in the control group, or in the total number of tests conducted and examined. I believe, however, that correlation from the test data will, over time, reveal Central Tendencies (see Hirota) which will in turn lead to more effective pedagogical principles. The main focus, then, is the demonstration of how certain test data might be analyzed and applied, in concrete terms, to meet the perceived needs of the learners involved. Such a focus would likely bring it under the heading of 'action research' (see Boyle), an increasingly fruitful area of classroom teachers' efforts to increase their effectiveness through monitoring actual classroom procedures.

# II . Class Parameters and TOEIC Format

The original aim of the course in question was to prepare second to fourth year students for the TOEIC, the Test of English for International Communication. There were no specific prerequisites for admission to the course, simply a genuine interest in the Toeic and the determination to work through the many quizzes and several full practice tests as outlined in the syllabus given out before registration. Although a fourth of the twenty students have attended only sporadically, four sophomores, six juniors, and five seniors, including one student auditing the course, have attended regularly and participated in all classroom activities. None of the students have taken the Toeic previously, and it was one of the objectives of the course, as stated in the syllabus, to familiarize them with the Toeic format through the model tests. The results were then to be analyzed for particular weak points in the individual student as well as the class as a whole, and appropriate remedial instruction provided.

The syllabus was designed to focus on each area of the test in the same order that these areas actually appear in the Toeic, with four parts comprising the first, or aural, section, and three parts comprising the second, or reading section of the test. Therefore, the first semester was to be devoted to listening skills and the second to reading. To give the students an overall view of the first section however, all four

parts of that section were tested throughout the semester, sometimes at one sitting and sometimes not. Using a popular primer for the Toiec, the students took an abbreviated test on the whole aural section during the second class period of the term (see Barron's, pgs. 82-91). The four parts of the aural section take the form of 1) identifying  correct descriptions of pictures  2) identifying correct responses to questions  3) answering questions based on short conversations  4) answering questions based on short monologues. The results of this abbreviated test appear in the tables below.

## Ⅲ. Test Results for Abbreviated Aural Section

a) Correct answer percentage per student analysis

---

Table 1      Test Results for Abbreviated Aural Section ※

**Correct answer percentage per student ※※**

Part 1 (10 questions; 13 students)
90  90  70  70  80  70  80  70  80  80  80  80  80

[mean=78]

Part 2 (15 questions)
67  73  80  80  73  67  80  34  60  93  53  60  80

[mean=69]

Part 3 (15 questions)
34  27  34  53  34  27  47  27  27  40  34  40  40

[mean=36]

Part 4 (10 questions)
50  30  30  80  40  40  50  40  50  60  20  20  50

[mean=43]

※ percentages rounded to the nearest whole number
※※ the order of the students remains the same throughout the table

---

What these scores show first of all is a grouping of the mean scores of the first two and last two parts of the test at opposite ends of the distribution. Part one is well within an acceptable range for students exposed to the Toeic for the very first time, while both one and two indicate a facility on the part of the students that is in contrast with the results of Parts 3 and 4.. However, while the mean scores of these

first two parts would seem to indicate a common degree of ease in comparison with Parts 3 and 4, examination of the individual scores in Part 2 shows that this is not necessarily the case. There is a skewed distribution from the mean in Part 2, making for a range of 59 points, or practically double that of the range in Part 1. Although such a distribution is extreme, it is not unusual in studies of this size (the range in Part 4 is even larger). In fact, Hirota claims that a subject group of at least 30 is usually necessary to obtain something closer to normal distribution (Hirota, in Griffee & Nunan, eds., pgs 123-135). As mentioned above, the nature of the problems between Parts 1 and 2 differs not only in that the first part incorporates visual prompts, but that the correct response in the second part depends on the processing of logical relations that are often dependent on function and transition words. Examination of individual problems, which will be taken up later in this paper, will also show certain certain disparities even within the same question formats, due to the students' facility with abstract questions, or with markers of logical relations.

The results of Parts 3 and 4 exhibit mean scores that are closer together than Parts 1 and 2, which makes it natural to consider them together, especially as the scores both indicate a level of increased difficulty for the students. Again, as in the comparison between Parts 1 and 2, there is a major difference in point distribution: Part 3 shows a distribution of 26, while Part 4 shows one of 60. Even if the exceptional scores of student number four (the fourth from the left) are excluded, the ratio of difference remains significant.

Overall, the most salient feature is, of course, the drop in scores on the last two sections of the test. The reasons for this may include fatigue and unfamiliarity with the test procedures, but inexperience with processing extended aural content, unaided by visual or conversational context, in both dialogue and monologue form, seems the most natural explanation and is no doubt what the aim of the questions was intended to test. Whether questions so designed truly test language use capabilities in a learner is another matter however; improvement in test taking skills alone, rather than increasing facility in handling logical relations markers, might also alter scores in these areas, as will be noted in the examination of subsequent results (see below).

One last avenue of analysis is the consistency, or lack thereof, in scores across the different parts of the test by the same students. Mention was made earlier of the high scores of student four, who stood at the top of the distribution in Parts 3 and 4. This

same student, however, was at the bottom of the distribution in Part 1. Likewise, the two students with the highest scores in Part 1 scored below the mean in two out of the remaining three parts, with one of the students being at the very bottom of the distribution in Part 3. In reverse, the two students with the very lowest scores in Part 4 were above the mean and at the median in Part 1. In averaging the scores over all four parts of the test, student four did indeed rank the highest, at 71, while student ten, who scored highest in Part 2 ranked second overall, at 68, and student seven, who did not reach the top of the distribution in any of the parts but scored above the mean in all, ranked third, at 6. The student showing the lowest average for the four combined scores, student eight, was at the low end of the distribution in all but one part, though was joined at these low ends by other students two out of three times. Such an array of performances would seem to indicate drastically different aptitudes, or cognitive styles (see Richards), or might simply be superficial tendencies that iron out with repeated exposure to different kinds of question formats. More immediately useful knowledge for addressing the students' particular weak spots should come from analyzing the nature of the individual questions, and from asking why they were  most often answered correctly or incorrectly.  The numbers in Table 2 represent the percentage of students who answered the questions correctly, and also represent the place of the questions themselves, going from left to right.

b) Correct answer percentage per question

1.  Part 1 Analysis

Table 2 ·    Test Results for Abbreviated Aural Section ※

**Correct answer percentage per problem:**

Part 1 (questions on pictures)
46  85  77  92  100  85  54  77  61  92

Part 2 (identifying correct responses)
92  69  92  61  85  69  77  77  31  61  69  23  85  54  85

Part 3 (questions on short conversations)
 23   8  38  38  85  46  15  38  38  38  15  23  31  31  62

Part 4 (questions on short monologues)
92  23  54  23  38  23  69  54  46  23

※ percentages rounded to the nearest whole number

In looking at these figures, the one generalization that can be made for all four parts is that there are extreme divergences among ratios for correct answers within those parts. It is most profitable here to examine those questions individually, to determine the students' strengths and weaknesses and see if correlations exist with the data in Table 1.

First of all, it may be instructive to compare question #1, which was answered correctly by less than fifty percent of the students, with question # 5, which was answered correctly by all the students. One might also look at questions #4 and #10 for similarities with question #5. As with all the questions in Part 1, the student hears four statements, one of which most accurately describes what he sees in the photograph that accompanies the question. The photograph for question #1 (Barrows', pg. 82) is of two men standing next to each other, as if in line, beside a long table on which they appear to be placing dishes or removing food from the dishes and bowls that are already on the table. They both appear to be employees attending to the setting of a buffet table: one man is wearing a dark jacket and one man a white jacket; they are both wearing bow ties. While the students are studying this scene they hear the following four statements:

(A) There are plates on the tray.
(B) The men have identical jackets.
(C) Both waiters are carrying dishes.
(D) They are waiting for a bus.

The correct response was (A), which was made by six out of the thirteen students; four students chose (C), while three chose (B). To characterize the three answers chosen, (A), though correct, is a description of what is in fact a rather innocuous aspect of the picture as a whole; (B) contains the easily understood word jacket, and the jackets the men are wearing do stand out in the picture, with one of them even bearing a form of rank insignia on the epaulets; (C), the most common mistake, begins with"Both waiters" , which the only people in the picture undoubtedly are, and so perhaps led students to decide on it as the most plausible answer. Their confidence in (C) was perhaps further increased by the content of the predicate, as

one man is indeed carrying a tray loaded with dishes, while the other is extending his hand in the direction of a metal bowl which he might or might not have in his grasp. To sum up the analysis of this question,  the fact that the highest single answer chosen was the correct one and the least plausible one was chosen by none, seems to validate it as a question testing aural comprehension. Analysis of the mistakes however is less clear-cut:  were they made for lexical reasons, such as not knowing the word 'identical', or from extraneous ones, such as misinterpreting the visual cues? Words the students were most likely familiar with, such as 'jackets', 'both', and 'waiters', may even have served as distractors. 'There can be no final answer from the data alone, though analysis may serve to point out certain intermediate vocabulary items that need to be addressed. Other insights might be had by considering why students made no mistakes at all, as in question #5.

The photograph that accompanies question #5 is of a lone man in an elevator. He is wearing a jacket with an emblem on the shoulder, as if he is an employee, and is pushing a button for a floor. The possible responses the students hear are as follows:

(A) The elevator operator pushes a button.
(B) Movies are made in Hollywood.
(C) He's pushing a button on his shirt.
(D) Stocks are going up.

The simplicity of the photo's content would seem to be the primary reason for the 100% correct response rate for this question. The fact that everyone in  Japan would be familiar with an elevator and uses a phonetic approximation in the Japanese language, and that the term 'elevator', with its distinctive phonemes appears in only one of the possible responses, makes its selection a relatively obvious one.  The information it provides on why the students did not choose the wrong responses seems to point to familiarity with a core lexical item, in this case, elevator.  The results for question #10 supports the hypothesis that core words, when representing a feature prominent in the photograph, and when they appear in only one of the choices, serves as the main criterion for selection. This question had a 92% correct response rate; the choice of responses is as follows:

(A) The attendant offers a newspaper
(B) The passenger is sleeping.
(C) The tourists are taking photos.
(D) There is not a cloud in the sky.

This strategy, however, can easily be shown not to be a viable one. Although successful in questions #5 and #10, it proved just the opposite in the case of question #1: there the prominent features 'jackets' and 'waiters' , which were indeed selected over the innocuous, and correct, 'plates on the tray'. Nevertheless, this does not invalidate the hypothesis that such a principle for selection is actually being adhered to by students. With the support of such data, the students can actually be shown the effectiveness of their strategies and be persuaded to alter them as necessary.

## 2. Part 2 Analysis

In looking for the Central Tendency (see Hirota) in the questions of Part 2, for brevity's sake it may be best to concentrate on the high and low extremes, questions #11#13#19#22 (see Table 2, above). In Part 2, students must complete a conversation by choosing an appropriate response to the question they hear; neither question nor responses appear in written form on the test. The questions and responses for all four questions appears below (Barrows', pgs. 567-568):

#11

Where do you live?

(A) About five o'clock.

(B) On Third Avenue.

(C) In ten minutes.

#13

Who are you waiting for?

(A) I've lost a lot of weight.

(B) I don't know the way.

(C) My wife; she is meeting me here.

#19

What should I do now?

(A) Why don't you make some coffee?

(B) The wood is on the fire.

(C) You were the first one here.

#22

How large is your company?

(A) It's about 10 inches.

(B) It's as big as a penny.

(C) We have offices in ten countries.

As shown in Table 2, these questions had a correct response ratio of 92%, 92%, 31%, and 23%. The correct answer for #11 is (B), and is, as the Explanatory Answers in Barrows' points out (pg. 482), a logical response to the question about location. The distractors were meant to confuse with like sounds in (A) and (C) and a logic infraction in (C). For question #13 the correct answer is (C). Again, the distraction strategy is to confuse with similar sounds, though again, as the 92% rate shows, the specific question-word was picked up on easily by the students, who then made the only logical connection.

Although questions #19 and #22 also begin with the commonplace question words 'What' and 'How', the results were dramatically different from #11 and #13. What is the difference between these interrogatory words and the others? For question #19, the Explanatory Answers states that 'wood' and 'should' were meant to be confusing, and that a tense difference was made to rule out (C). The abstract nature of the response that the 'What' word demands in this case must surely have contributed to the students' confusion as well. As seen above, students, at least at this level, tend to select responses with a tendency to the logically direct and concrete. An examination of the mistakes made here might also be instructive: an equal number of students chose (C) as chose the correct answer (A), while an even larger number chose the other option, (B), suggesting that extra work on on auxiliaries, especially when working with abstract responses, might be in order. For question #22 the tendency is again to the direct and interrogative-specific: 'How large?' seems to demand a specific size, in units or the form of a size adjective. An equal number of students chose the two wrong responses according to this principle, combined they were more than three times the total who chose the correct answer (C).

3. Part 3 Analysis

The questions on short conversations in Part 3 proved the most difficult for students, as is borne out by averaging the percentage of times each question in a Part was answered correctly (Table 3).

| Table 3 | Average of Percentages Individual Questions Answered Correctly |
|---|---|
| **Correct answer percentage per Part:** | |
| Part 1 (questions on pictures) | 69% |
| Part 2 (identifying correct responses) | 69% |
| Part 3 (questions on short conversations) | 35% |
| Part 4 (questions on short monologues) | 44% |
| ※ percentages rounded to the nearest whole number | |

Interesting that the figures are identical for the two easiest Parts , though this is not inevitable statistically, as percentages from a full length practice test given later in the semester will attest (see below).

As we see in Table 2, only three of the fifteen questions were answered correctly by more than 40% of the students. Why the poorest showing on Part 3? Inference from the data is made below, but increasing exposure to the particular format and practice in focused listening, would, I thought, raise scores in this area, on the later test. This in fact happened, with the percentage of correct responses going up 13 percent (see Table 4). Nevertheless, inspection of the two lowest scoring questions, #27 and #32, and the highest, #30, does add useful information. The conversation for #27 is as follows:

Man:     What seems to be the matter with your forearm?
Woman:   I was getting off the bus and I fell down the stairs.
Man:     It looks swollen. Let me take an X ray.

After hearing this conversation, the students read in their test booklets the following:

Who are the speakers?
(A) A doctor and patient.
(B) A carpenter and foreman.
(C) A mother and son.

(D)  A bus driver and a rider.

As seen in Table 1, only 8%, or one out of thirteen students answered this question correctly.  The other next most difficult question, #32, is as follows:

The students heard:

Man:       Last year 30% of our budget was spent on advertising.

Woman:  Was that all in TV advertising?

Man:       No, half of it was in print.

The students read:

What percentage was spent on print advertising?

(A) 5%

(B) 10%

(C) 15%

(D) 30%

Only 15% percents of the students answered this correctly, or two out of the thirteen. Why were the students so unsuccessful?  The medical and advertising words make a lack of vocabulary seem the obvious reason, but there are also similarities with other mistakes which might lead to a pattern. In question #27 there are multiple speaker-subjects being inquired after, a feature shared by only one other question in the same Part, question #36, which was answered incorrectly second most often, tied with #32, at 15%.  In contrast, the question which proved the least difficult was one with a single subject and easily visualized, #30.

The students hear:

Man:       Did I have any calls?

Woman:  Yes--about twenty.  I left all the messages on your desk.

Man:       Thanks.  Would you get me a purchase order form, please?

The students read:

Where does this conversation take place?

(A)  On the street.

(B)  In an office.

(C)  On a bus.

(D)  At home.

Eighty-five percent of the students answered this question correctly, again in an inverse ratio to the degree of abstraction.

4. Part 4 Analysis

Part 4 showed the second lowest average of correct responses per question, at 44%. Here the students heard monologues, often announcements and advertising, before reading the questions and possible responses. More than one question was asked about each monologue. Four questions tied for the lowest percentage of correct response, at 23%, with certain characteristics in common. One example will suffice to illustrate the format; the monologue and questions for #41 and #42 are as follows:

Monologue: This train makes all stops between Grand Central Station and 125th Street. The express train is across the station. Please step in and watch the closing doors.

#41.Where was this announcement made?
(A) On a train.
(B) At an intersection.
(C) On an escalator.
(D) In a grain elevator.

#42.What kind of service is provided?
(A) Express--no stops.
(B) Local--all stops.
(C) Limited--some stops.
(D) Out-of-service.

Question #42 was typical of the four questions most often missed in this Part, in that it was the second of the two questions asked about the same monologue. The ratio of correct answers in the first question in comparison with the second was nearly 2 to 1 in four out of the five monologues. It was hoped that focused listening would correct this trend in subsequent tests, but the eight percent drop as recorded in Table 4 shows that exposure alone was not sufficient. Other recurring features we find that seem to correlate with incorrect responses are 'What' questions with abstract predicates, especially those referring to action, a trend also seen in earlier parts. Correct answers, on the other hand, fell again into the pattern of the easily visualized and concrete, as in the question of location in #43, which had the highest success rate in all of Part 4 at 92%, and #47, which had the second highest rate at 69%.

# Ⅳ. Comparison of Abbreviated and Full Test Results

The data analyzed so far was obtained from one test given at the beginning of the term. After seeing what areas seemed to pose the most difficulties for the students, instruction was given on the separate Parts and then a full practice test was given for each Part on a different week. The full test consisted of twice the number of questions found in the abbreviated test. Comparison of the data between the two tests confirmed the existence of some trends, but gave results contrary to expectations about others.

Table 4    Average of Percentages Individual Questions Answered Correctly

**Correct answer percentage per Part:**

|          | **Abbreviated** | **Full** |
|----------|-----------------|----------|
| Part 1   | 69%             | 87%      |
| Part 2   | 69%             | 59%      |
| Part 3   | 35%             | 48%      |
| Part 4   | 44%             | 36%      |

※ percentages rounded to the nearest whole number

The most striking increase was in the visually oriented Part 1.  Considering the trends in the data themselves, this was not unexpected, except for the extent of the rise.  More surprising was the 13% improvement in Part 3, where students scored the lowest on the abbreviated exam. In-class practice on focused reading seems to have brought results here, though the same could not be said for Part 4, which fell 8 percentage points.  More detailed analysis is necessary here and in regard to Part 2, to understand these falls in scores. Finally, a slight rise was seen in the average of the four parts taken together as a composite score, with the students averaging 54% on the first test and 57% on the second.

# V. Conclusion

As stated in section one of this paper, the focus was to analyze test data in order to identify trends in the abilities being tested. I believe some success was had here, especially in regard to question type. The analysis is ongoing however, which, if it makes for an incomplete survey here, offers the promise that a more comprehensive and useful analysis may be possible later using this as a foundation. Ideally, results from the test to be given at the end of term would be correlated with the results from the two exams discussed here, to prove or disprove the suppositions made as well as to evaluate classroom techniques. One other component of the TOEIC that wasn't examined here is the tests for reading comprehension. This same group of students has gone through an Extensive Reading program (see Bamford) throughout the year as a supplement to their classroom work. Along with an analysis of the final exam of aural section, I would in future like to examine the correlative results between the ER program and the scores on the reading comprehension section.

# References

Bamford, J., & Day, R. R.
*Extensive Reading in the Second Language Classroom.*
Cambridge:  Cambridge University Press, 1997.

John Boyle
Report on a Lecture by Dr. Graham Crookes:  'Action Research'
*The Language Teacher,* Volume 21, Number 12(December, 1997)

Griffee, Dale T. & Nunan, David, editors
*Classroom Teachers and Classroom Research*
The Jalt Applied Materials Series.  Japan Association for Language Teaching, 1997.

John Honeyfield
*Simplification .*
TESOL Quarterly, Vol.11. No. 4
December 1977.

Lin Lougheed
*How to Prepare for the Toeic Test*
Barron's, 1995.

David Nunan
*Language Teaching Methodology*
Prentice Hall, 1998.

Jack C. Richards, John Platt, Heidi Platt
*Longman Dictionary of Language Teaching & Applied Linguistics.*
Longman, second edition  1992.