Sophia University Junior College Division Faculty Journal
創立 40 周年記念　第 35 号, 2014, 15-23

TOEIC Scores: How Many Points Are Enough to Show Progress?

Melvin Andrade

For several years in the mid-2000s, Sophia Junior College had a TOEIC "score up policy" that required students to gain 50 points within one academic year. Students not gaining 50 points were required to participate in a 3-week, 15-hour English Support Program. The aim of the policy was to motivate students to study on their own to improve the listening and reading skills that TOEIC measures. The expected outcomes were that the number of students achieving at least a 50-point gain would increase, the number of students taking the English Support Program would decrease, and the average TOEIC scores after one and two years of study would increase. However, in 2010 it became clear that the actual outcomes were the opposite of the expected outcomes: Gains in TOEIC scores were decreasing; more students needed to take the English Support Program not fewer; and the average TOEIC score after one and two years of study was decreasing. Another observation at that time was that the content of the English Support Program was not necessarily related to the skills tested by TOEIC. Consequently, there was no convincing evidence that the English Support Program as a whole was improving the students TOEIC skills or motivating them to do well on TOEIC the following year, although there were a few exceptions. Although 50 points up was thought to be reasonable and attainable improvement for all students within one year, experience showed that was not the case.

In comparison, Osaka Jogakuin Women's Junior College in their online article "Regarding the 2005 Accreditation Report" (p. 5), reported an average gain of 30 points from the end of the first year (average score = 415.2) to the end of the second year (average score = 445.7) in 2004. They considered 30 points, to be an "excellent" learning outcome for their college (Appendix 1). However, as will be explained below, statistical analysis shows that to be confident of a gain in proficiency for most test takers, there must be at least an 80-point increase (assuming equal gains in listening and reading). The aims of this paper is to help clear up some common misunderstanding about TOEIC scores and suggest a way of using TOEIC data to better understand our students' actual progress in English. Specifically, this paper will address the following questions: (1) How can we more accurately report student gains on TOEIC as the measure of the difference

between the April (pretest) and December (posttest) TOEIC scores? (2) How we "clean up" the TOEIC score data by appropriately dealing with outliers? (3) Is improvement in one's TOEIC score, really improvement in one's English proficiency?

Interpreting TOEIC Scores

In a pretest-posttest situation, according to the Educational Testing Service (ETS), "the errors of measurement associated with two administrations are called the Standard Error of Difference (SEdiff). The SEdiff for each of the TOEIC Listening and Reading sections is about 35 scaled score points" (*TOEIC User Guide*, 2007, p. 10). This means that a student who scored 40 points or more on EACH section has probably improved in both skills. "Probably" means a 68 percent confidence level. To reach a 97 percent confidence level, we have to double the points (80 points EACH). At present, ETS does not provide an SEdiff score for the total TOEIC score, so we have to interpret Listening and Reading scores separately. Thus, considering the information above, we can divide students into four groups:

Group 1 (High Achievers):	Gained 40 or more points on BOTH Listening and Reading (40 points + 40 points).
Group 2 (Listening Achievers):	Gained 40 or more points on Listening but NOT Reading (40 points + 35 points or fewer).
Group 3 (Reading Achievers):	Gained 40 or more points on Reading but NOT Listening (40 points + 35 points or fewer).
Group 4 (Low or Non-Achievers):	Gained fewer than 40 points on BOTH Listening and Reading (equal to or less than 35 + 35 points).

By comparing the number of students in each group, we can obtain a more accurate picture of the progress of each cohort of students who enter our college. I recommend that we report our data this way.

In contrast, if we want an estimate of a student's "true" score in April (pretest score), we have to use the Standard Error of Measurement (SEM), which is about 25 points *each* for Listening and Reading (*TOEIC User Guide*, 2007, p. 10). For example, the "true" score of a student with a Listening score of 200 and a Reading score of 200 (400 total) could range from Listening 175-225 and Reading 175-225 (350-450 total). Thus, if that student in December gains 40 points *each* in Listening and Reading and her "true

April score" is 400, we would expect her to obtain 480 points. However, if her "true April score" was 350, a score of 430 would indicate improvement (350 + 40 + 40); or, on the other hand, if her "true April score" was 450, then 530 would indicate improvement (450 + 40 + 40). These ranges, however, would vary because her "true December score" would range as well. Consider the following example.

In April, the student's score is 500 points, but in December it drops to 420. The April "true score" could be as low as 450 (500-25 for Listening and another -25 from Reading), and the December "true score" could be as high as 470 (420 +25 for Listening and another +25 for Reading). The result would be 470 – 450 = 20, which is actually a 20-point gain! What this means is that when considering a student's progress or lack of progress, we have to think in very broad terms. Unfortunately, when it comes to cut-off scores for entrance examinations, a student's "true score" may not be factored in when the admission or rejection decision is made. For more details on SEdiff and SEM, see Appendix 2.

Dealing with Outliers

Below are three explanations of "outliers" and why we need to considered them carefully:

> An outlier is defined as an observation that "appears" to be inconsistent with other observations in the data set. An outlier has a low probability that it originates from the same statistical distribution as the other observations in the data set. On the other hand, an extreme value is an observation that might have a low probability of occurrence but cannot be statistically shown to originate from a different distribution than the rest of the data. (Walfish, 2006) (See Appendix 3.)

> Outliers should be investigated carefully. Often they contain valuable information about the process under investigation or the data gathering and recording process. Before considering the possible elimination of these points from the data, one should try to understand why they appeared and whether it is likely similar values will continue to appear. Of course, outliers are often bad data points. (NIST/SEMATECH, 2012)

> An outlier is an observation that lies an abnormal distance from other values in

> a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations. (NIST/SEMATECH, 2012)

As a rule of thumb in statistics, scores that are three or more standard deviations (SD) above or below the average can be considered outliers (Osborne & Overbay, 2004). That is very a strict rule, but commonsense should apply as well. See Appendix 4 for details.

All things considered, what would be a reasonable course of action in dealing with outliers in our program? First, extremely low scores (perhaps from students who fell asleep during the test) and extremely high scores (probably students who are native English speakers) that qualify as "outliers" should not be included in the aggregate average score for the April and December TOEIC tests given to all students. Instead, we should use footnotes to identify these scores and explain that they were not included when calculating the average for the group. Second, some students are absent and do not take the TOEIC IP. Instead, it is sometimes the case that they take a "practice TOEIC" of some kind later or submit an Official TOEIC score. If so, those scores should be reported separately from the TOEIC IP data and not included in the TOEIC IP data analysis. The two different test scores have not been statistically equated to account for differences in difficulty. If we mix them, we muddle our data. Finally, it should be mentioned that the college gives an annual award to the student who makes the most progress in TOEIC. There is a possibility, however, that this student is an outlier who slept through her April TOEIC test or underperformed for another reason. There is no way to confirm that the winner of this award truly excelled.

Improvement in TOEIC scores vs. improvement in English proficiency

Do TOEIC test score gains correlate with increased communicative competence? Cunningham (2002), for example, in a study of first-year university students, found that there was no correlation between TOEIC test scores and communicative abilities. Similarly, there was no correlation between TOEIC test score gains and improved communicative in a direct assessment of their listening, reading, and writing abilities. In addition, the findings suggested that TOEIC test-preparation did not improve accurate use of structure. As the TOEIC test and TOEIC preparation focus on listening skills, a restricted number of reading genres (e.g., business letters, announcements,

emails), and test-taking strategies, it is not surprising that there may be little effect on expressive skills. Accordingly, recognizing the importance of TOEIC scores for employment and university transfer, as well as the reality that TOEIC score gains and English proficiency are not necessary linked, Sophia University Junior College Division in 2013 took the positive step of establishing a required TOEIC preparation courses for first-year students.

Conclusions

When guiding individual learners, we need to keep in mind that TOEIC is not a diagnostic test. It does not aim to reliably identify the array of strengths and weaknesses in a learner's English proficiency. Provided that we understand the statistical variability explained above, it is useful as a rough measure of overall ability (Prolingua). We need to keep in mind that because of the Standard Error of Measurement (SEM), a student's scores can fluctuate greatly up or down, especially when the test is taken successively after short intervals of study. A few points up or down (plus or minus 35) is most probably not an indication that a learner's English has become better or worse. If this normal fluctuation is not understood, the results can be demotivating for many learners.

In an effective English language program there should be a clear distinction between *achievement* linked to courses on the one hand and *general language proficiency* on the other. In addition, a distinction should be made between goals established for the *program* as a whole (e.g., average gain) and goals established for *individual students* depending on their placement level or personal goals. Furthermore, when setting achievement and proficiency goals, the structure and contents of the curriculum must be taken into consideration to ensure that the goals are realistic and attainable. Although our initial approach to TOEIC did not go as well as expected, we have learned from that experience and undertaken measures that preliminary results indicate will yield positive results.

References

Cunningham, Cynthia R. (2002). The TOEIC Test and Communicative Competence: Do Test Score Gains Correlate With Increased Competence? Unpublished master's dissertation, School of Humanities of the University of Birmingham, United

Kingdom. Retrieved 22 October 2006, from http://www.cels.bham.ac.uk/resources/essays/Cunndiss.pdf

"Interpreting Scores" and "Repeat Test Takers." (2007). In *TOEIC User Guide: Listening and Reading*. Educational Testing Service (page 19). Available: http://www.ets.org/Media/Tests/Test_of_English_for_International_Communication/TOEIC_User_Gd.pdf

NIST/SEMATECH. (2012, April 1). "What are outliers in the data?" *e-Handbook of Statistical Methods*. United States Commerce Department. Available: http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

Osaka Jogakuin Tanki Daigaku.(2005). "*Gaibu hyouka 'jiko tenken-hyouka houkokusho' wo yomu*." Retrieved Oct. 24 2006, from http://www.wilmina.ac.jp/2yrs/special/valuation.html

Osborne, Jason W. & Amy Overbay. (2004). "The power of outliers (and why researchers should always check for them)." *Practical Assessment, Research & Evaluation,* 9(6). Retrieved May 13, 2012 from http://pareonline.net/getvn.asp?v=9&n=6

Prolingua Executive Language Services. "TOEIC Information." Retrieved Dec. 3, 2013, from http://www.prolingua.co.jp/toeic_e.html

TOEIC User Guide. (2007). Princeton, NJ: Educational Testing Service.

Walfish, Steven. (2006, Nov. 2). "A review of statistical outlier methods." *Pharmaceutical Technology*. Available: http://statisticaloutsourcingservices.com/Outlier2.pdf

Appendix 1

Osaka Jogakuin Women's Junior College: "Regarding the 2005 Accreditation Report"
大阪女学院短期大学：『外部評価「自己点検・評価報告書」を読む』（2005 年度） P 5
http://www.wilmina.ac.jp/2yrs/special/valuation.html

「学習成果 learning outcome という観点からみると、本学の教育を受ける前と受けた後の英語運用能力を比較しどれだけの進歩があったかということで、英語教育の成果をみることができる。TOEIC-IP のスコアの変化は、定量的尺度としては標準化された尺度といえるだろう。資料 4 をみると、全学生の平均値が 1 年修了時の 415．2 から 2 年修了時の 445．7 へと、平均で 30 点の上昇は優れた成果である。しかし、気になる点もある。最高点は 900 から 845 へ、最低点は 195 から 125 へと、それぞれ大きく下降している。クラス別の平均点の比較をみると、クラスごとの変化にバラツキがあるが、本文ではこのあたりの説明は特に

なされない。資料 13 をみると 2 学年の間の分布の変化はわかるが、その内容は学生個人の変化にバラツキがあるのは容易に予想できるが、その分析や説明されることが望ましい。」
[*Gist:* An average gain of 30 points from the end of the first year (average score = 415.2) to the end of the second year (average score = 445.7) was observed in 2004 and is considered to be an "excellent" learning outcome for this college. However, some student scores actually decreased, and there was wide variation in the amount of individual improvement that needs explanation.]

Appendix 2

From the *TOEIC User Guide* (p.10):

Interpreting Scores. TOEIC test scores are determined by the number of questions answered correctly. There is no penalty for wrong answers. The number of correct responses on each section, Listening and Reading, is converted to a number on a scale of 5 to 495. The statistical procedure used to convert scores to a common scale for each section seeks to ensure that TOEIC Listening and Reading scores obtained on different administration dates mean the same thing in terms of the level of English proficiency indicated.

If you were to take several versions of the test within a short period of time, you would obtain a number of scores that center around an average value known as your "true" score. Two-thirds of the time, your listening score would be within 25 points of your true score on the listening section, and your reading score would be within 25 points of your true score on the reading section.

Repeat Test Takers. Test takers who take another version of the TOEIC test may obtain slightly different scores from those they received the first time. A question like this may arise, "How much of a difference must there be between two Listening scores or between two Reading scores before I can say that there is a real difference in my level of proficiency?" This question involves two independent tests given at two different times. The error of measurement associated with the score obtained from one administration is called the Standard Error of Measurement (SEM). The errors of measurement associated with two administrations are called the Standard Error of Difference (SEdiff). The SEdiff for each of the TOEIC Listening and Reading sections is about 35 scaled score points.

Another question that may arise, "If a person began training with a Listening score of 300 and, following training, received a score of 340 on a different test form, has

that test taker really improved in Listening or is this increase just a statistical fluke?" To determine whether this is a true increase in the TOEIC score, the test taker would construct a band of ± 1 SEdiff, or ± 35 points, around the obtained scores. In this case, the test taker has truly improved because the posttraining score fell outside the SEdiff (i.e., 265-335). Using this band, we can say with 68 percent confidence that the test taker's proficiency level has truly increased in the time between the two tests.

Appendix 3

"How to deal with outliers" from Walfish (2006, p. 1):

"Once an observation is identified—by means of graphical or visual inspection—as a potential outlier, root cause analysis should begin to determine whether an assignable cause can be found for the spurious result. If no root cause can be determined, and a retest can be justified, the potential outlier should be recorded for future evaluation as more data become available. Often, values that seem to be outliers are the right tail of a skewed distribution. When reporting results, it is prudent to report conclusions with and without the suspected outlier in the analysis. Removing data points on the basis of statistical analysis without an assignable cause is not sufficient to throw data away."

Appendix 4

Excerpts from Osborne & Overbay (2004):

"Identification of Outliers: There is as much controversy over what constitutes an outlier as whether to remove them or not. Simple rules of thumb (e.g., data points three or more standard deviations from the mean) are good starting points. Some researchers prefer visual inspection of the data. Others (e.g., Lornez, 1987) argue that outlier detection is merely a special case of the examination of data for influential data points."

. . .

"Although some authors argue that removal of extreme scores produces undesirable outcomes, they are in the minority, especially when the outliers are illegitimate. When the data points are suspected of being legitimate, some authors (e.g., Orr, Sackett, & DuBois, 1991) argue that data are more likely to be representative of the population as a whole if outliers are not removed."

. . .

"Conceptually, there are strong arguments for removal or alteration of outliers. The analyses reported in this paper also empirically demonstrate the benefits of outlier removal. Both correlations and t-tests tended to show significant changes in statistics as a function of removal of outliers, and in the overwhelming majority of analyses accuracy of estimates were enhanced. In most cases errors of inference were significantly reduced, a prime argument for screening and removal of outliers."