*Compendium Study*

# Validating *TOEIC Bridge*™ Scores Against Teacher Ratings for Vocational Students in China

*Donald E. Powers, Regina Mercadante, and Fred Yan*

*September 2013*

ETS TOEIC

*Know English. Know Success.*

The *TOEIC Bridge*™ test is a 100-item multiple-choice test designed to measure the ability to listen to and read everyday English (http://www.ets.org/toeic_bridge). Targeting beginning and intermediate learners of English, the test includes questions/items that are significantly easier than those found on the *TOEIC*® test. The *TOEIC Bridge* test has been found to be most appropriate for learners at the (basic) A1 and A2 levels and the (intermediate) B1 level of the Common European Framework (Tannenbaum & Wylie, 2008). It has been deemed appropriate for students in South America by virtue of (a) its factorial structure and (b) its association with other relevant tests, student self-assessments, and teacher ratings of students (Sinharay et al., 2009). Moreover, in a large-scale study in South America involving an extensive self-assessment inventory and more than 4,000 students, *TOEIC Bridge* scores correlated moderately with test takers' self-assessments of their ability to perform a variety of listening and reading tasks in English (Powers et al., 2008).

Although positive evidence of the validity of *TOEIC Bridge* scores is available from the studies cited above, the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999) dictate that multiple sources of validity evidence are always desirable. This is true especially when a test is used in a context that differs from the one in which validity evidence was gathered previously.

The objective of this study was to examine another source of validity evidence for *TOEIC Bridge* scores. Specifically, we sought to compare students' *TOEIC Bridge* scores with teachers' assessments of students' English-language skills, particularly their ability to perform a variety of everyday language tasks. The context was vocational schools in China.

## Method

In late November 2011, more than 2,000 vocational school students took the *TOEIC Bridge* test in 10 cities in China. This sample was then stratified according to total score on the test, and approximately 60 students were randomly selected from each of nine 10-point *TOEIC Bridge* score intervals from scores of 100 to 180. (*TOEIC Bridge* total scores range from 20 to 180.) An additional 60 students were selected from TOEIC scores ranging from 20 to 90. This additional (10th), larger score interval was specified because there were too few students in any single 10-point interval below 90 for a meaningful analysis.

For each of the selected students, a teacher was asked to complete a brief language task inventory. This "can-do" inventory consisted of everyday English-language reading and listening tasks that were adapted from an earlier, longer inventory devised for a validity study of the TOEIC test. Because the *TOEIC Bridge* test was designed to measure *emerging* English-language competencies, the inventory used here included tasks that are significantly easier than those comprising the TOEIC test inventory.

The abbreviated 14-item inventory (six listening tasks and eight reading tasks) was completed by teachers at several vocational schools in the winter/spring of 2011–2012. Teachers were asked to complete the inventory for each of several students in their classes, indicating for each student the extent to which that student could perform each of the tasks in English. Teachers were allowed to omit a task if they felt unable to judge a student's ability to perform a task. The questions were translated into Mandarin Chinese to facilitate understanding. Response choices were on a five-point scale as follows:

1. Cannot do at all,
2. Can do with great difficulty,
3. Can do with some difficulty,
4. Can do with little difficulty, and
5. Can do easily.

Analyses of data entailed computing correlations (for reading, listening, and totals) between *TOEIC Bridge* scores and teacher ratings, both for individual tasks and for composites based on all reading and all listening tasks. Because there was no opportunity to train teachers to apply the same standard when providing ratings, we have chosen to present results for individual teachers when the number of students rated was adequate (more than five students). We have also pooled data across all teachers and have reported results in the aggregate, when appropriate. First, however, we inspected the data in order to identify possible statistical outliers that might unduly affect overall results.

## Results

A total of 87 teachers provided ratings for one or more of the 613 students who were selected for study. By chance, teachers rated variable numbers of students. At one extreme, one teacher rated 117 different students (the next greatest numbers were 58 and 40). At the other extreme, 28 other teachers each rated only one student. A total of 26 teachers each provided ratings for more than five students.

Because they lacked any formal training to make their ratings or any benchmarks to guide them, teachers may have applied very different standards when rating their students. Therefore, we conducted a preliminary analysis of each teacher's ratings in order to ascertain the extent to which standards may have varied across raters. Specifically, for reading, listening, and total *TOEIC Bridge* scores, we examined scatter plots of (a) mean *TOEIC Bridge* score for each teacher's students against (b) mean composite rating (sum over all task ratings) for reading, listening, and total ratings for each teacher. Figures 1–3 show these plots for the 26 teachers who provided more than five student ratings. Each plot reveals a moderately strong relationship between mean rating and mean *TOEIC Bridge* score of students rated ($r = .58$ for both reading and listening and $r = .60$ for total). However, it appears that one teacher (whose data point is designated by a larger square in each figure) was clearly an outlier when compared with the other 25 teachers, especially with regard to reading

and total ratings. This teacher's students had lower mean *TOEIC Bridge* scores on both reading and listening than did the students of any other teacher. Despite their low scores, however, this teacher's students received ratings that were relatively high on average. This teacher was also atypical of other teachers in providing ratings for 117 students—far more than any other teacher. This discrepancy leaves open the possibility that this teacher may have applied a significantly different rating process than did teachers who rated many fewer students. Finally, we also examined the scatter plots (see Figures 4–6) for this outlying teacher (whom we call Teacher A), and we find that the patterns of relationships are quite irregular. Thus, although we report results individually for this teacher, we have also chosen to exclude these ratings when reporting results aggregated over all teachers.

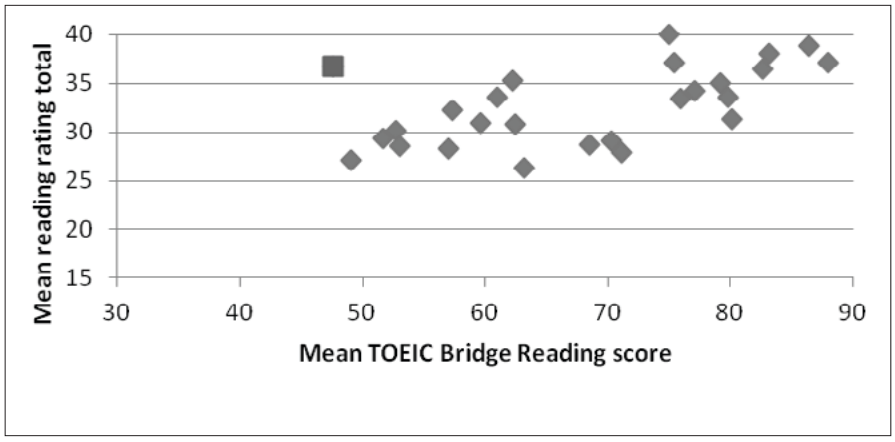


***Figure 1.*** **Mean teacher reading total rating vs. mean *TOEIC Bridge* Reading score.**

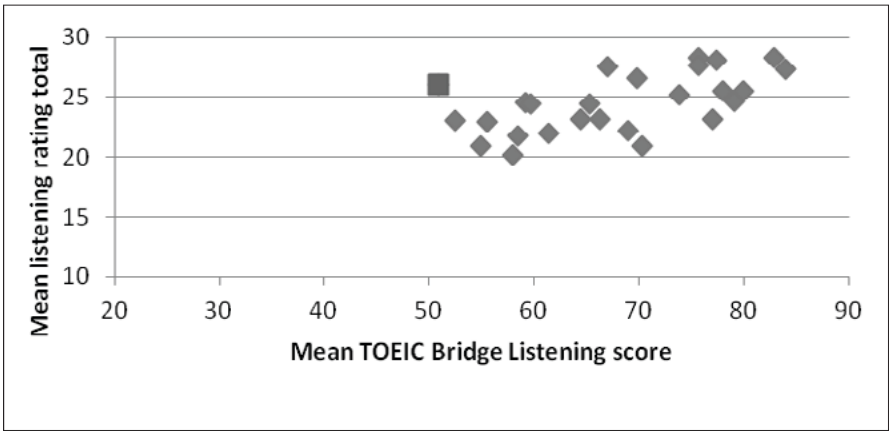


***Figure 2.*** **Mean teacher listening total rating vs. mean *TOEIC Bridge* Listening score.**

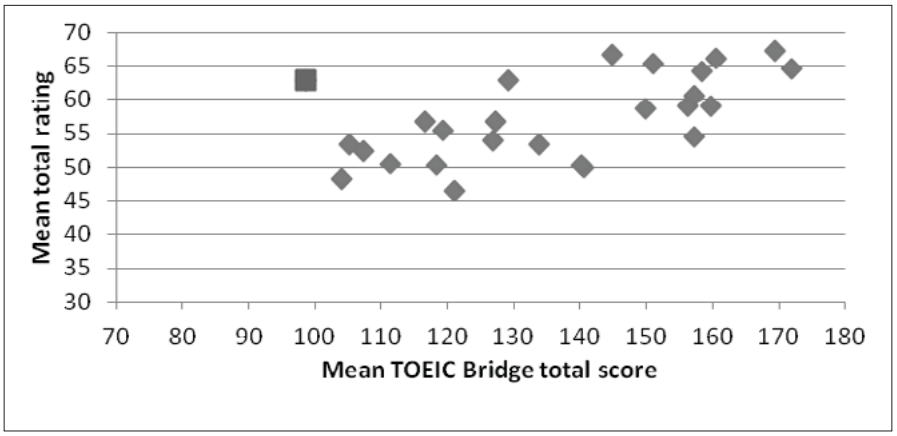


***Figure 3*. Mean teacher total rating vs. mean *TOEIC Bridge* total score.**

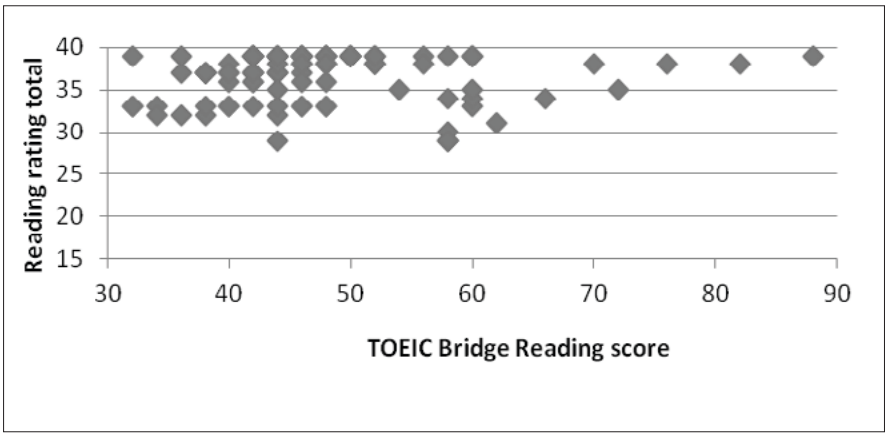


***Figure 4*. Reading rating total vs. *TOEIC Bridge* Reading score for Teacher A.**

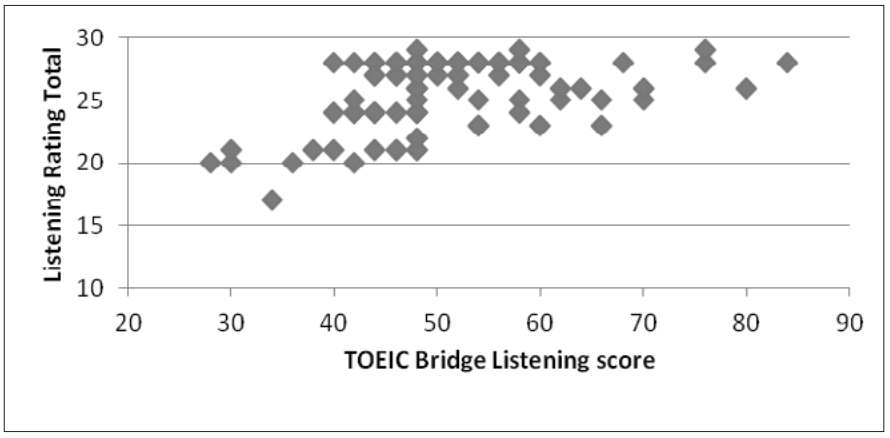


***Figure 5*. Listening rating total vs. *TOEIC Bridge* Listening score for Teacher A.**

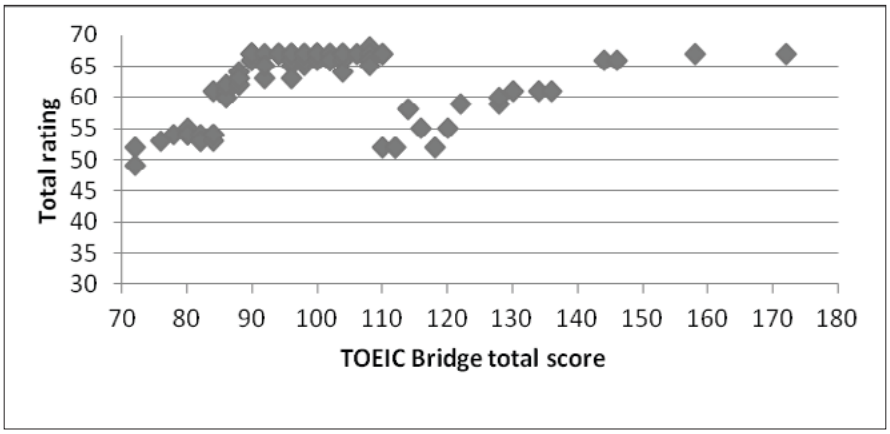


***Figure 6*. Total rating vs. *TOEIC Bridge* total score for Teacher A.**

The ratings made by the teachers were estimated to be reasonably reliable, with Cronbach alpha reliability estimates of .91 for the eight reading tasks and .81 for the six listening tasks for the total sample of teachers. Table 1 provides a high-level summary of findings. The mean *TOEIC Bridge* scores were relatively high, indicating that, in general, test takers were relatively able in terms of English-language proficiency. Over all reading tasks and over all listening tasks, mean teacher ratings were also relatively high, indicating that teachers also perceived these students to be relatively able, on average.

## Table 1

***Correlations Among Can-Do Teacher Assessments and TOEIC Bridge Scores***

| No. | Measure | M (SD) | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| 1 | *TOEIC Bridge* Listening | 69.0 (11.5) | 1.00 | | | |
| 2 | *TOEIC Bridge* Reading | 69.7 (14.6) | .84* | 1.00 | | |
| 3 | Can-do listening | 24.1 (3.2) | .58* | .59* | 1.00 | |
| 4 | Can-do reading | 32.2 (5.0) | .57* | .58* | .84* | 1.00 |

*Note.* $N = 496$.

*$p < .01$

As can be seen from Table 1, *TOEIC Bridge* Reading and Listening scores were highly related ($r = .84$) in our sample. Teacher ratings (composites) of reading skills and listening skills were also strongly related ($r = .84$), suggesting that the two rating scales generally reflected the same trait. Of greater interest, the correlations among the composite ratings and *TOEIC Bridge* scores ($r = .57$ to $.59$) were relatively strong. Again, however, these correlations suggested no differentiation between reading and listening tasks/scores. The correlations computed here between ratings and test scores can, according to conventional standards for effect sizes in the social sciences, be characterized as large (Cohen, 1988).

Table 2 provides the means (and standard deviations) of teacher ratings for each of the 14 individual can-do tasks. (In these analyses, we have not included the ratings provided by the one outlying teacher that was discussed above.) Teacher ratings for the tasks corresponded relatively well with our own expectations of the difficulty of the tasks. For example, for the reading task, "Understand the viewpoints expressed in articles and reports about contemporary issues or problems" was rated as the most difficult task for these students, while "Recognize memorized words and phrases (for example, 'Exit,' 'Entrance,' and 'Stop')" was rated as easiest. Thus, we have some confidence in the validity of these ratings.

Table 2 also contains the correlation of teacher ratings with *TOEIC Bridge* score for each individual language task. As can be seen, for reading tasks, the correlation of *TOEIC Bridge* Reading scores with teachers' ratings ranged from .29 to .52. The correlation with a composite index based on all eight reading tasks was .58. For individual listening tasks, the correlations ranged from .13 to .51 and .58 for a composite index based on all six listening tasks.

## Table 2

***Means (and Standard Deviations) for Teacher Ratings and Correlations with TOEIC Bridge Scores***

| Can-do statement | Mean | SD | Correlation with *TOEIC Bridge* |
|---|---|---|---|
| | Reading | | |
| Read and understand a popular novel. | 3.44 | 0.9 | .48 |
| Understand the viewpoints expressed in articles and reports about contemporary issues or problems. | 3.24 | 1.0 | .38 |
| Read and understand a travel brochure. | 3.63 | 1.0 | .46 |
| Understand short, simple texts (e.g., personal letters). | 4.19 | 0.8 | .52 |
| Understand the main point of simple messages and short, clear announcements. | 4.26 | 0.7 | .47 |
| Read and understand a simple postcard from a friend. | 4.31 | 0.7 | .48 |
| Understand any (a) reading and (b) written instructions (in English) required as part of his/her schoolwork. | 4.31 | 0.7 | .51 |
| Recognize memorized words and phrases (for example, "Exit," "Entrance," and "Stop"). | 4.78 | 0.5 | .29 |
| Total reading | 4.02 | 0.7 | .58 |
| | Listening | | |
| Understand extended speech and lectures, and follow complex arguments on familiar topics. | 3.48 | 0.8 | .49 |
| Understand an extended debate on a relatively complex topic. | 3.16 | 0.9 | .32 |
| Understand a person in social situations talking about his/her background, family, or interests. | 3.95 | 0.9 | .50 |
| Understand someone who is speaking slowly and deliberately about his or her hobbies and interests. | 4.43 | 0.7 | .43 |
| Understand discussions and oral instructions (in English) required as part of his/her schoolwork. | 4.25 | 0.7 | .51 |
| Understand simple questions in social situations such as "How are you?" and "Where do you live?" | 4.81 | 0.5 | .13 |
| Total listening | 4.01 | 0.6 | .58 |
| Total | 4.01 | 0.6 | .62 |

*Note.* All correlations are significant at the .01 level. $N = 496$ students.

As noted above, because teachers may have differed with respect to the leniency (or stringency) of the standards they used for rating students, pooling over all teachers may have depressed the correlations between *TOEIC Bridge* scores and teacher ratings. Therefore, we also examined this relationship for each of the 26 teachers who provided ratings for more than five students. For individual teachers, the correlation of *TOEIC Bridge* scores with the sum of ratings over all tasks ranged from -.03 to 1.00 for reading (median = .67) and from .00 to .98 for listening (median = .72). The substantial variability among correlations across individual teachers is illustrated in Figures 7–12, which show scatter plots for two teachers, one (Figures 7–9) whose ratings had virtually no relationship to *TOEIC Bridge* scores and another (Figures 10–12) whose ratings were very highly correlated with *TOEIC Bridge* scores. These correlations for individual teachers are therefore somewhat higher on average than those based on an aggregation over all students. These individual teacher results should not, however, be overemphasized because, with few exceptions, they are based on quite small samples. (We note that, even for the outlying teacher, whose ratings are included in these results, the correlation of composite listening ratings with *TOEIC Bridge* Listening scores was significant [$r = .41$]).

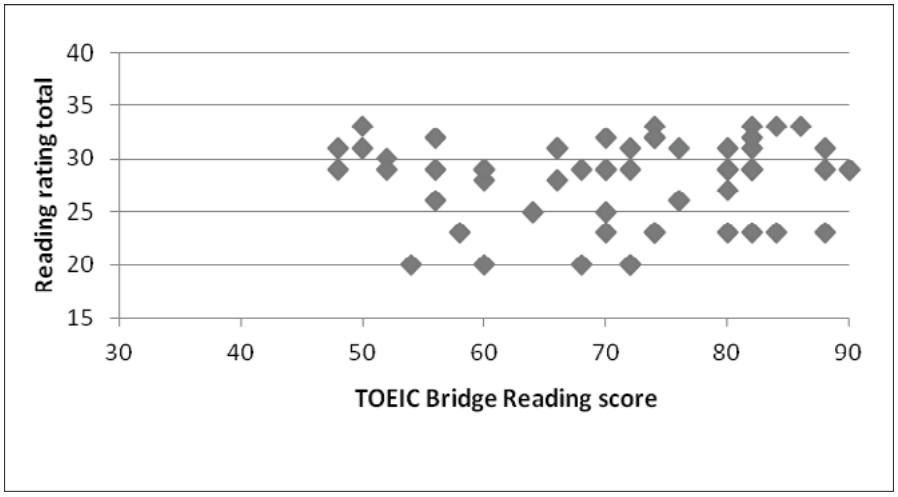


***Figure 7*. Reading rating total vs. *TOEIC Bridge* Reading score for Teacher B.**

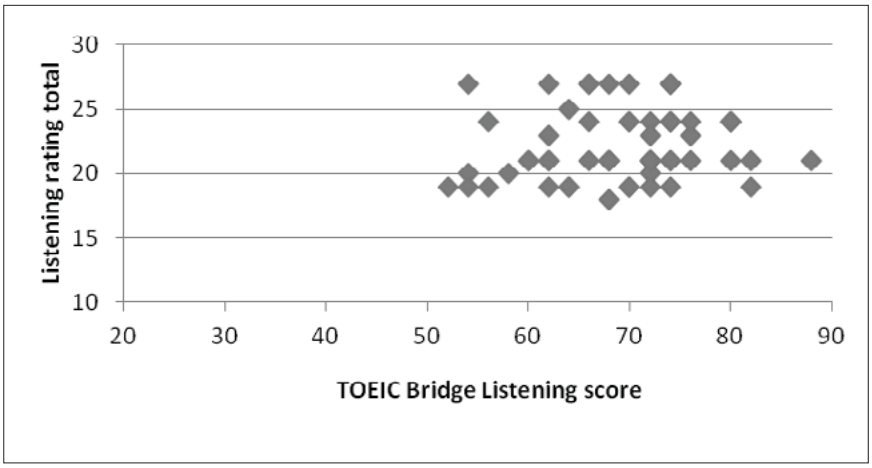


***Figure 8.*** **Listening rating total vs. *TOEIC Bridge* Listening score for Teacher B.**

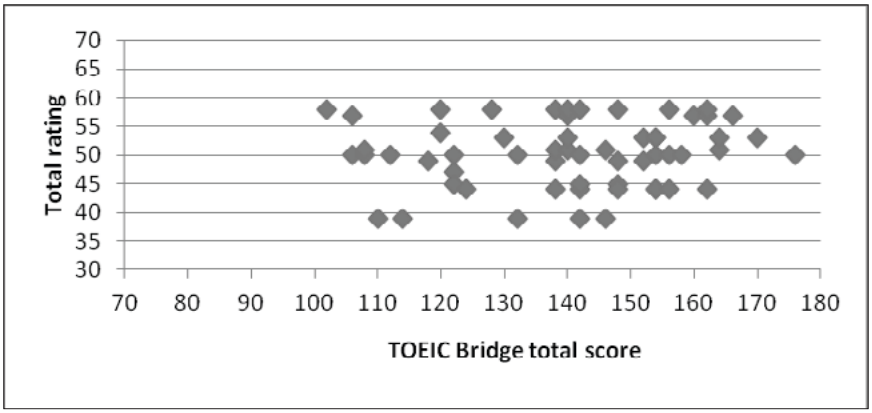


***Figure 9.*** **Total rating vs. *TOEIC Bridge* total score for Teacher B.**

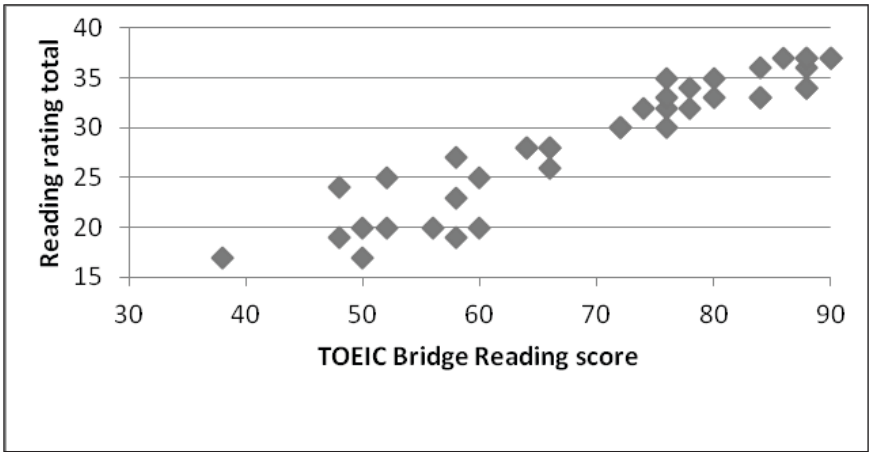


***Figure 10.*** **Reading rating total vs. *TOEIC Bridge* Reading score for Teacher C.**

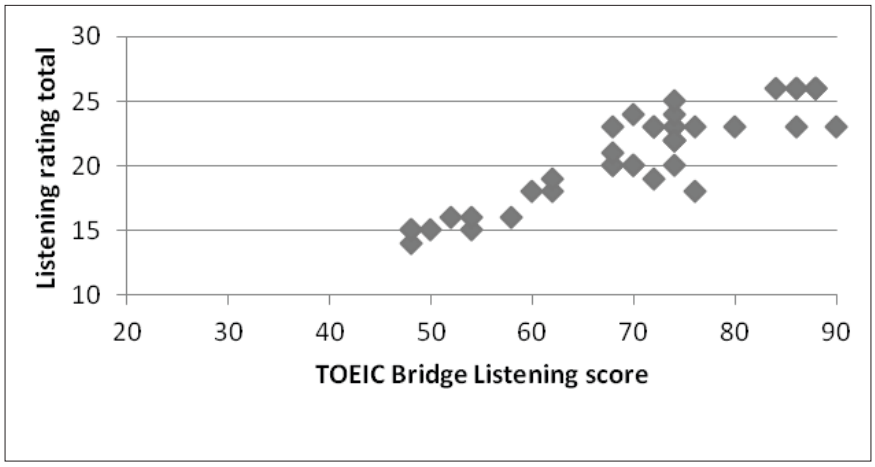


***Figure 11*. Listening rating total vs. *TOEIC Bridge* Listening score for Teacher C.**

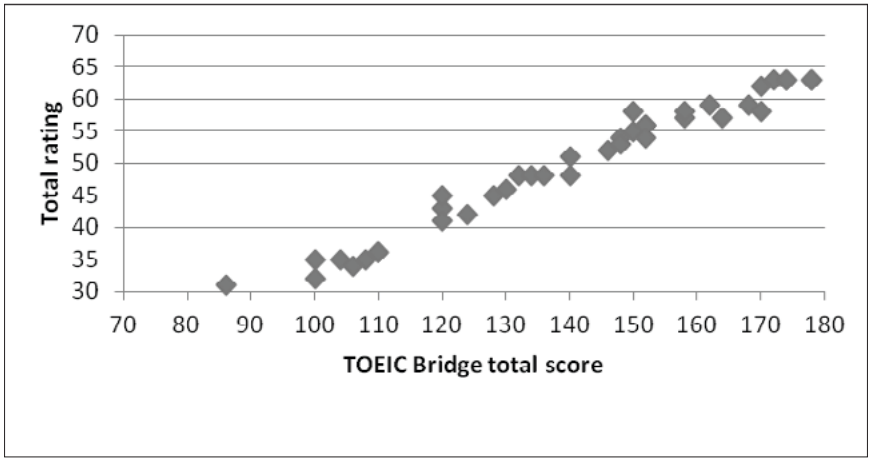


***Figure 12*. Total rating vs. *TOEIC Bridge* total score for Teacher C.**

Finally, Table 3 displays the study data in a somewhat different way. Shown are the percentages of students at each of several *TOEIC Bridge* score levels (for both Reading and Listening) who were judged by teachers to be able to perform each language task either easily or with little difficulty. As can be seen, almost without exception, the percentages increase with each higher *TOEIC Bridge* score level. For some tasks (e.g., "Read and understand a popular novel"), the increases are sharp, in this case from 8% at the lowest test score level to 75% at the highest. For other tasks (e.g., "Recognize memorized words and phrases [for example, 'Exit,' 'Entrance,' and 'Stop']") the increases are less, largely because some tasks were judged to be relatively easy for all students even at the lowest *TOEIC Bridge* score levels.

**Table 3**

***Percentages of TOEIC Bridge Test Takers, by Score Level, Whose Teachers Indicated That They Could Perform Various English-Language Reading and Listening Tasks Either Easily or With Little Difficulty***

| | TOEIC Bridge score | | | | |
|---|---|---|---|---|---|
| Tasks | 10–50 | 51–60 | 61–70 | 71–80 | 81–90 |
| | Reading | | | | |
| Read and understand a popular novel. | 8 | 25 | 35 | 55 | 75 |
| Understand the viewpoints expressed in articles and reports about contemporary issues or problems. | 9 | 18 | 34 | 55 | 72 |
| Read and understand a travel brochure. | 14 | 47 | 61 | 71 | 89 |
| Understand short, simple texts (e.g., personal letters). | 66 | 70 | 85 | 90 | 97 |
| Understand the main point of simple messages and short, clear announcements. | 72 | 76 | 88 | 88 | 94 |
| Read and understand a simple postcard from a friend. | 74 | 77 | 91 | 89 | 97 |
| Understand any (a) reading and (b) written instructions (in English) required as part of his/her schoolwork. | 73 | 77 | 91 | 94 | 99 |
| Recognize memorized words and phrases (for example, "Exit," "Entrance," and "Stop"). | 96 | 98 | 97 | 97 | 100 |
| *N* for score interval. | 74 | 83 | 93 | 104 | 142 |
| | Listening | | | | |
| Understand extended speech and lectures, and follow complex arguments on familiar topics. | 3 | 29 | 44 | 55 | 89 |
| Understand an extended debate on a complex topic. | 3 | 19 | 27 | 41 | 75 |
| Understand a person in social situations talking about his/her background, family, or interests. | 24 | 52 | 73 | 82 | 94 |
| Understand someone who is speaking slowly and deliberately about his or her hobbies and interests. | 74 | 78 | 93 | 97 | 98 |
| Understand discussions and oral instructions (in English) required as part of his/her schoolwork. | 72 | 73 | 87 | 97 | 99 |
| Understand simple questions in social situations such as "How are you?" and "Where do you live?" | 97 | 94 | 95 | 99 | 100 |
| *N* for score interval. | 34 | 89 | 133 | 145 | 95 |

*Note.* $N = 496$ students. Responses were on a five-point scale, from 5 = *can do easily*, 4 = *can do with little difficulty* to 1 = *cannot do at all*.

# Discussion

The results of the study presented here suggest that *TOEIC Bridge* scores relate moderately well to teacher assessments of students' English-language skills. In terms of conventional standards for characterizing effect sizes in the social sciences, the correlations observed here (in the high .50s for composite ratings) can best be described as *large* (Cohen, 1988). The data support this conclusion despite certain study limitations that may have hindered finding even stronger relationships. In particular, the study sample proved to be relatively capable in terms of *TOEIC Bridge* scores, and in general the sample also was judged to be relatively proficient with regard to the language tasks for which teachers provided ratings. This restriction in the range of ability found in our sample may

have depressed the correlations that were computed. In addition, there was a significant period of time between the time at which students were tested and the time at which they were rated by their teachers. Because students' proficiency may have changed significantly during the gap, the relationship between test scores and ratings may have diminished. Finally, there was no attempt to train teachers with respect to how to make their ratings. Nor was any information available to us about either teachers' own proficiency with English or the degree to which they were familiar with the students they rated. Thus, if (a) the study sample had been more heterogeneous with regard to the students' tested English-language ability, (b) the language tasks had spanned a wider range of difficulty, (c) teachers had been trained to apply a common standard when making their ratings, and (d) both test scores and ratings had been obtained at the same time, the relationships would, predictably, have been even stronger than those noted here. Relationships might also have been stronger for some teachers, depending on their own English proficiency and their familiarity with students.

Nonetheless, within the limits of the study, the results provide reasonably compelling evidence of the validity of the *TOEIC Bridge* test as an indicator of English-language proficiency. These results are both statistically significant, and more importantly, practically significant with regard to making decisions about students' emerging proficiency in English.

## References


American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Powers, D. E., Bravo, G. M., Sinharay, S., Valdivia, L. E., Simpson, A. G., & Weng, V. Z. (2008). *Relating scores on the TOEIC Bridge to student perceptions of proficiency in English* (Research Memorandum No. RM-08-02). Princeton, NJ: Educational Testing Service.

Sinharay, S., Powers, D. E., Feng, Y., Saldivia, L., Guinta, A., Simpson, A., & Weng, V. (2009). Appropriateness of the *TOEIC Bridge* test for students in three countries of South America. *Language Testing*, *26*, 589–619.

Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English-language test scores onto the Common European Framework of Reference: An application of standard-setting methodology* (TOEFL iBT® Research Report No. 06). Princeton, NJ: Educational Testing Service.