Research Memorandum

# Repeater Analyses for TOEFL® iBT

Yanling Zhang

February 2008

ETS RM-08-05

# Repeater Analyses for TOEFL® iBT

Yanling Zhang
ETS, Princeton, NJ

February 2008

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html



**Abstract**

The operational administrations of TOEFL® iBT began in September 2005 and gradually expanded to the rest of the world in 2006. Some TOEFL iBT test-takers chose to take the test more than once and they were categorized as repeaters. Knowledge about the performance of these repeaters as a group provides useful information in evaluation of the consistency and validity of the test scores resulting from test forms given at different administrations. This study evaluated the repeater performance on the TOEFL iBT tests in 2007. The findings indicated that small mean score changes were observed between performance on the two tests and the correlations between the two tests were moderate to high.

Key words: TOEFL iBT, repeaters, score change, validity

**Acknowledgments**

The author would like to acknowledge the statistical support given by Jill B. Carey. Special thanks also go to Daniel Eignor, Mary Enright, Anna Kubiak, Bradley Moulder, and Lin Wang for their careful review and helpful comments on the paper.

## Introduction

After more than a decade of research and development, the Internet-based Test of English as a Foreign Language™ (TOEFL® iBT) was first launched in the United States on September 24, 2005. Since then, it has gradually rolled out worldwide and has replaced the TOEFL computer-based test (TOEFL CBT). Like with any other large-scale assessment, it is not uncommon throughout the history of TOEFL that some candidates for a variety of reasons choose to take the test more than once. It is also not uncommon that these candidates who repeat the test (called repeaters in this report) may get different scores when they repeat the test. This is simply because no test scores can be free of measurement error even if the time interval between the two tests is too short for any learning to occur. Under normal circumstances, however, scores of repeaters are expected to vary to a small extent if a test is valid and reliable and the tests are repeated within a short period of time (i.e., no more than a month or so) and no intensive training occurs during this period of time. Large observed score differences should be investigated as it is important to examine such large variations in the scores of repeaters to evaluate test score validity. In addition, it is helpful to learn about the characteristics of repeaters. Hence, this repeater analysis study was expected to answer the following questions:

1. Who were the repeaters?
2. How did the repeaters as a group perform on their first and second tests?
3. To what extent did repeaters' scores change between the two tests?
4. What was the relationship between the scores of the two tests?

## Data Source

TOEFL iBT is composed of four sections: reading, listening, speaking, and writing. The scaled scores of each section range from 0-30. A total score is formed by a summary of the scaled scores of each section, thus ranging from 0-120.

The repeaters selected for this study were the candidates who took one TOEFL iBT test in one month and took the other in the next month. The rationale of selecting repeaters this way was that learning would be unlikely to occur within a month in the usual course of language acquisition. Thus any score changes between two test results could be attributed to such factors as measurement error and the quality of the test scores.

The data for this study included section and total scaled scores from the January 2007 to August 2007 operational administrations. Since TOEFL iBT did not completely roll out worldwide until late 2006, it is believed that data from the selected 2007 administrations would better represent the test performance of a more stable testing population than the data from earlier administrations in 2005 and 2006. It should be pointed out the repeater sample analyzed here was a self-selected sample as not everyone is going to repeat the test in reality. However, the results could be deemed generalizable to the group of people who would repeat the test under similar circumstances such as within one month.

Outlier observations (i.e., those examinees who had extremely large score gains or losses) were excluded from this study. The criteria to remove outliers were based on the same rules that are used to identify test takers with large score differences (LSDs) in the operational setting (Lewis, 2007). The criteria are ± 22 for reading, ± 21 for listening, ± 13 for speaking, ± 20 writing, and ± 45 for total scores. Out of approximately 12,300 repeaters in the data set, 22 (0.18%) cases were outliers who were removed from further analyses.

In the following discussion, the first test a repeater took is named Test 1 and the repeated test is referred to as Test 2. Approximately 250,000 candidates took TOEFL iBT from January to August 2007. About 10% of the candidates repeated the test at least once. About 12,000 candidates repeated the test once within 30 days, and they formed the sample of interest for this study. A small percentage of candidates took the test three times or more.

## Results

Tables 1–3 provide information about some characteristics of the repeaters in response to the first question of interest in this study. Table 1 shows the repeater counts sorted by both test center and native language. It is obvious that the majority of repeaters were Korean and Japanese candidates who took the test in the United States and Canada or in their own native countries. Chinese candidates comprised the next largest group retaking the test in China, Taiwan, the United States, and Canada. When looking at native languages only in Table 2, one can see that the same subgroups of examinees constituted the higher percentages in the repeater sample. Table 3 shows that most repeaters were also from the test centers in Korea, the United States, Japan, and Canada. To make sure the repeater data are representative of the regular administration data in the same time period, the largest language groups were also checked. Similar proportions of largest language groups were found in the regular administration data.

**Table 1**

***Repeaters by Test Centers and Native Language***

| Test center | Native language | Frequency | Percent |
|---|---|---|---|
| Korea, Republic of | Korean | 3,690 | 29.79 |
| Japan | Japanese | 1,845 | 14.90 |
| United States | Korean | 1,043 | 8.42 |
| China | Chinese | 511 | 4.13 |
| Canada | Korean | 400 | 3.23 |
| Canada | Chinese | 357 | 2.88 |
| United States | Japanese | 331 | 2.67 |
| United States | Arabic | 322 | 2.60 |
| Taiwan | Chinese | 286 | 2.31 |
| United States | Chinese | 237 | 1.91 |
| Turkey | Turkish | 199 | 1.61 |
| Saudi Arabia | Arabic | 126 | 1.02 |
| Italy | Italian | 99 | 0.80 |
| Indonesia | Indonesian | 99 | 0.80 |

*Note.* Total $N = 12{,}385$, showing here $N \geq 100$.

**Table 2**

***Repeaters by Native Language***

| Native language | Frequency | Percent |
|---|---|---|
| Korean | 5,523 | 44.59 |
| Japanese | 2,381 | 19.22 |
| Chinese | 1,476 | 11.92 |
| Arabic | 753 | 6.08 |
| Turkish | 263 | 2.12 |
| Spanish | 213 | 1.72 |
| Telugu | 145 | 1.17 |
| Thai | 130 | 1.05 |
| Farsi | 122 | 0.99 |
| Italian | 118 | 0.95 |
| French | 116 | 0.94 |
| Portuguese | 106 | 0.86 |
| Indonesian | 105 | 0.85 |

*Note.* Total $N = 12{,}385$, showing here $N \geq 100$.

**Table 3**

***Repeaters by Test Centers***

| Test center | Frequency | Percent |
|---|---|---|
| Korea, Republic of | 3,697 | 29.85 |
| United States | 2,580 | 20.83 |
| Japan | 1,955 | 15.79 |
| Canada | 1,030 | 8.32 |
| China | 559 | 4.51 |
| Taiwan | 321 | 2.59 |
| Turkey | 204 | 1.65 |
| India | 166 | 1.34 |
| Saudi Arabia | 133 | 1.07 |
| Indonesia | 126 | 1.02 |
| Italy | 110 | 0.89 |
| France | 107 | 0.86 |

*Note.* Total $N$ = 12,385, showing here $N$ >= 100.

The second to fourth questions of interest to this study were addressed next. Tables 4–8 summarize the Test 1 and Test 2 mean scaled scores and the mean score changes (Test 2 score minus Test 1 score) for each section and the total test scaled score. Since the standard deviations of the four sections vary to some extent, it is more appropriate to compare the magnitude of the mean score changes in terms of an effect size (ES), which is a standardized and scale-free measure of the relative size of the mean differences across the testings. ES is defined as the difference in Test 2 score mean ($X_2$) and Test 1 score mean ($X_1$), divided by pooled standard deviation of Test 1's standard deviation ($S_1$) and Test 2's standard deviation ($S_2$), $\frac{X_2 - X_1}{\sqrt{\frac{S_1^2 + S_2^2}{2}}}$. It is equivalent to a Z-score (standardized score) of a standard normal distribution.

For reference, the mean scores and standard error of measurement (SEM) of all operational forms that were administered within the time period are also provided in the bottom two rows of these tables. The correlations between Test 1 and Test 2 scores are shown in the last column of each table.

**Table 4**

***Summary Statistics of Repeater Performance on Test 1 and Test 2 for Reading***

| | Mean | SD | Minimum | Maximum | Correlation |
|---|---|---|---|---|---|
| Test 1 | 17.06 | 8.31 | 0 | 30 | 0.78 |
| Test 2 | 18.43 | 8.31 | 0 | 30 | |
| Mean score change | 1.37 | 5.51 | -20 | 21 | |
| ES of mean score change | 0.17 | | | | |
| Operational test score[a] | 18.47 | 8.54 | -- | -- | |
| Operational test SEM[a] | 3.33 | -- | -- | -- | |

*Note.* $N$ = 12,364.

[a] Operational test score and operational SEM (standard error of measurement) refer to the mean scaled scores and the SEM of the forms administered within the same period. The sample sizes of the operational forms range from 3,500 to 15,000. ES = effect size.

**Table 5**

***Summary Statistics of Repeater Performance on Test 1 and Test 2 for Listening***

| | Mean | SD | Minimum | Maximum | Correlation |
|---|---|---|---|---|---|
| Test 1 | 17.95 | 7.97 | 0 | 30 | 0.77 |
| Test 2 | 18.87 | 7.85 | 0 | 30 | |
| Mean score change | 0.92 | 5.36 | -20 | 20 | -- |
| ES of mean score change | 0.12 | | | | |
| Operational test score[a] | 19.48 | 8.05 | -- | -- | -- |
| Operational test SEM[a] | 3.15 | -- | -- | -- | -- |

*Note.* $N$ = 12,363.

[a] Operational test score and operational SEM (standard error of measurement) refer to the mean scaled scores and the SEM of the forms administered within the same period. The sample sizes of the operational forms range from 3,500 to 15,000. ES = effect size.

**Table 6**

***Summary Statistics of Repeater Performance on Test 1 and Test 2 for Speaking***

| | Mean | SD | Minimum | Maximum | Correlation |
|---|---|---|---|---|---|
| Test 1 | 17.38 | 4.48 | 0 | 30 | 0.84 |
| Test 2 | 17.99 | 4.35 | 0 | 30 | |
| Mean score change | 0.62 | 2.48 | -12 | 12 | -- |
| ES of mean score change | 0.14 | | | | |
| Operational test score | 19.07 | 4.84 | -- | -- | -- |
| Operational test SEM | 1.62 | -- | -- | -- | -- |

*Note.* $N$ = 12,364.

[a] Operational test score and operational SEM (standard error of measurement) refer to the mean scaled scores and the SEM of the forms administered within the same period. The sample sizes of the operational forms range from 3,500 to 15,000. ES = effect size.

**Table 7**

***Summary Statistics of Repeater Performance on Test 1 and Test 2 for Writing***

| | Mean | SD | Minimum | Maximum | Correlation |
|---|---|---|---|---|---|
| Test 1 | 18.91 | 5.06 | 0 | 30 | 0.77 |
| Test 2 | 19.74 | 4.96 | 0 | 30 | |
| Mean score change | 0.83 | 3.39 | -18 | 17 | -- |
| ES of mean score change | 0.17 | | | | |
| Operational test score | 20.10 | 5.46 | -- | -- | -- |
| Operational test SEM | 2.81 | -- | -- | -- | -- |

*Note.* $N$ = 12,364.

[a] Operational test score and operational SEM (standard error of measurement) refer to the mean scaled scores and the SEM of the forms administered within the same period. The sample sizes of the operational forms range from 3,500 to 15,000. ES = effect size.

**Table 8**

***Summary Statistics of Repeater Performance on Test 1 and Test 2 for Total Score***

| | Mean | SD | Minimum | Maximum | Correlation |
|---|---|---|---|---|---|
| Test 1 | 71.27 | 22.03 | 7 | 119 | |
| Test 2 | 75.01 | 21.88 | 7 | 120 | 0.91 |
| Mean score change | 3.74 | 9.50 | -42 | 44 | -- |
| ES of mean score change | 0.17 | | | | |
| Operational test score | 77.13 | 23.67 | -- | -- | -- |
| Operational test SEM | 5.62 | -- | -- | -- | -- |

*Note.* $N$ = 12,370.

[a] Operational test score and operational SEM (standard error of measurement) refer to the mean scaled scores and the SEM of the forms administered within the same period. The sample sizes of the operational forms range from 3,500 to 15,000. ES = effect size.

Overall, the repeaters' performance on their first test was lower than that of the total operational group. On Test 2, the repeaters' performance on all sections improved slightly and their Test 2 mean scores were closer to but were still lower than the total operational group means. The mean scores of Test 1 and Test 2 showed small differences. For the sections, reading had highest score change (1.37), whereas speaking had the lowest score change (0.62). The ESs of the mean score changes of reading and writing (both 0.17) are slightly higher than those of listening and speaking. Even though speaking appeared to have the smallest mean score change of the four sections, listening actually has the smallest ES (0.12), followed by speaking (0.14). Cohen's rule of thumb (1988) labels effect sizes of 0.20, 0.50, and 0.80 to be indicative of small, medium, and large effects, respectively. Following Cohen's rule, the ESs for all four sections are less than 0.20, suggesting that the score changes of the four sections and the total scores were fairly small.

The correlations between Test 1 and Test 2 appear to be moderate to high: very high for the total scores and moderately high for the four section scores. This suggests that the rank ordering of the repeaters using their scores varied to a small to medium small extent over the test taken first and test taken second.

Figures 1–5 depict the score distributions of the score changes between Test 1 and Test 2. All the change score distributions are centered on zero and are symmetrical around the center. They are approximately bell-shaped.

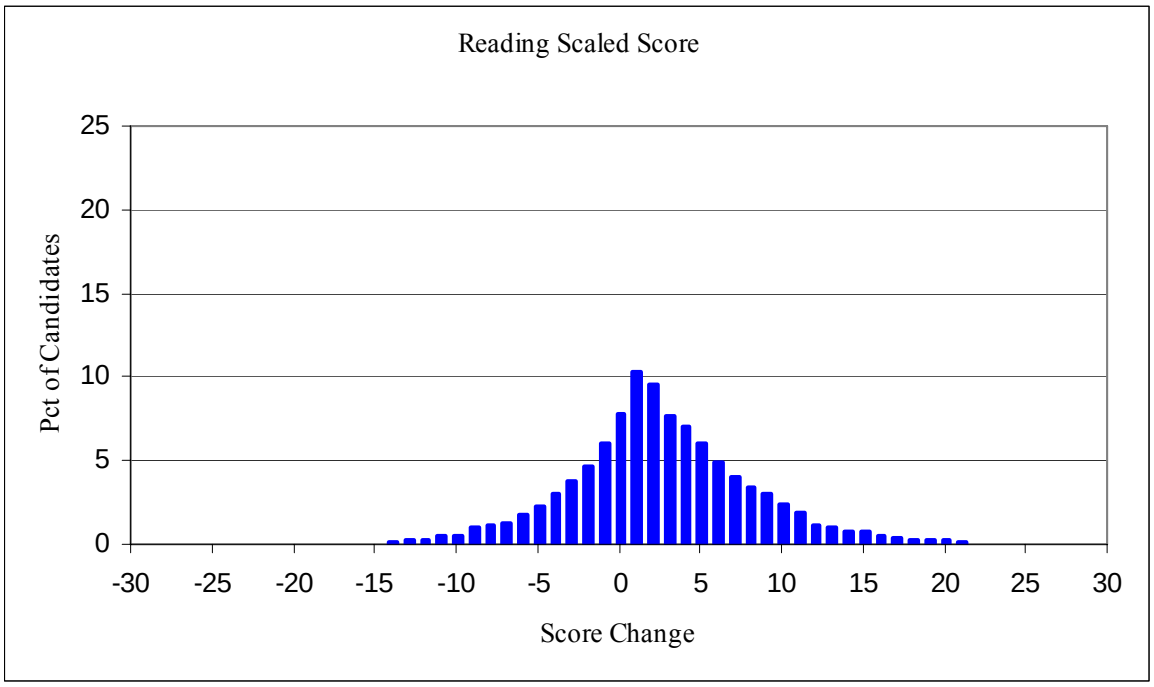


***Figure 1.*** **Frequency distributions of reading score changes.**

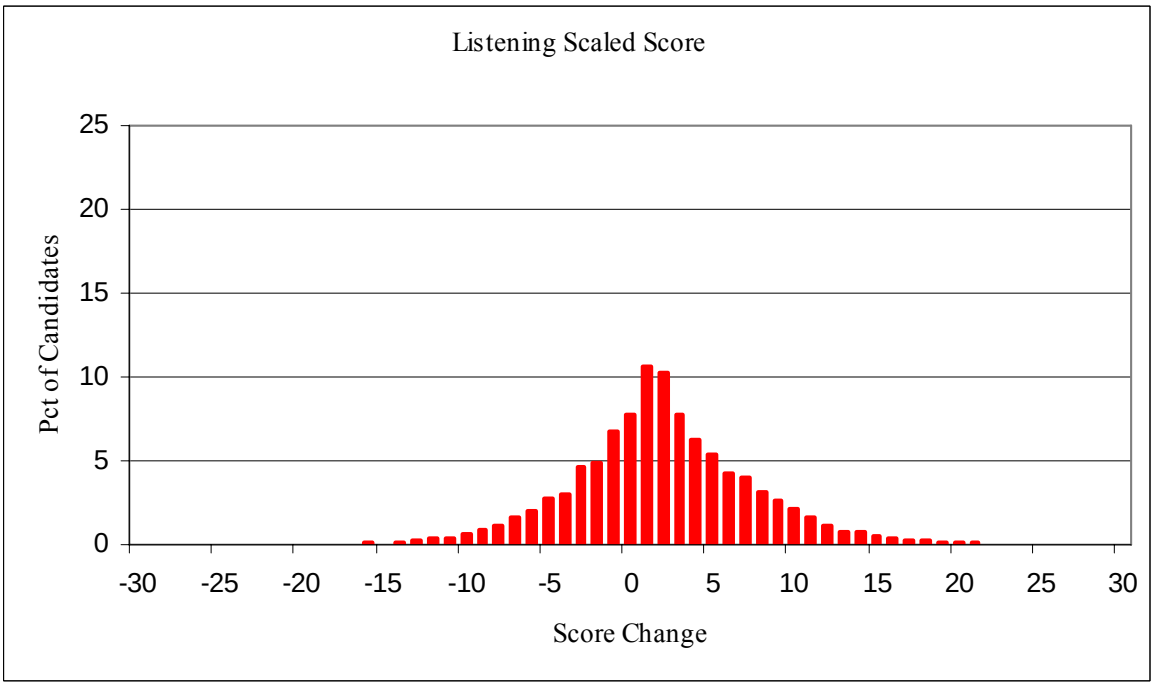


***Figure 2.*** **Frequency distributions of listening score changes.**

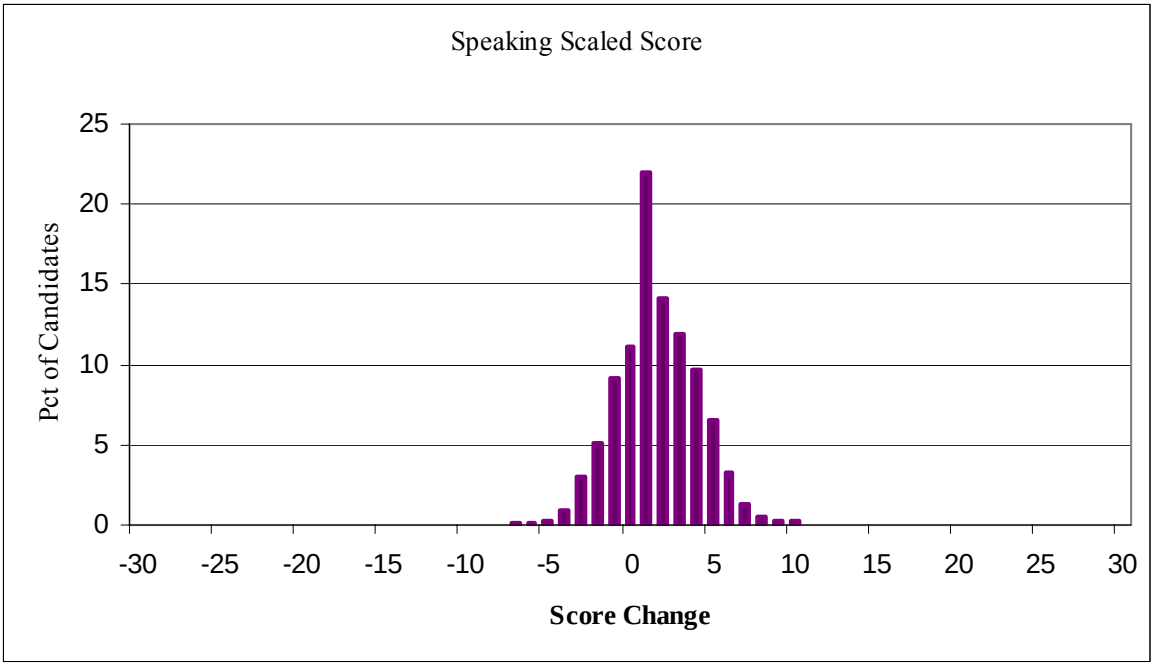


***Figure 3.*** **Frequency distributions of speaking score changes.**

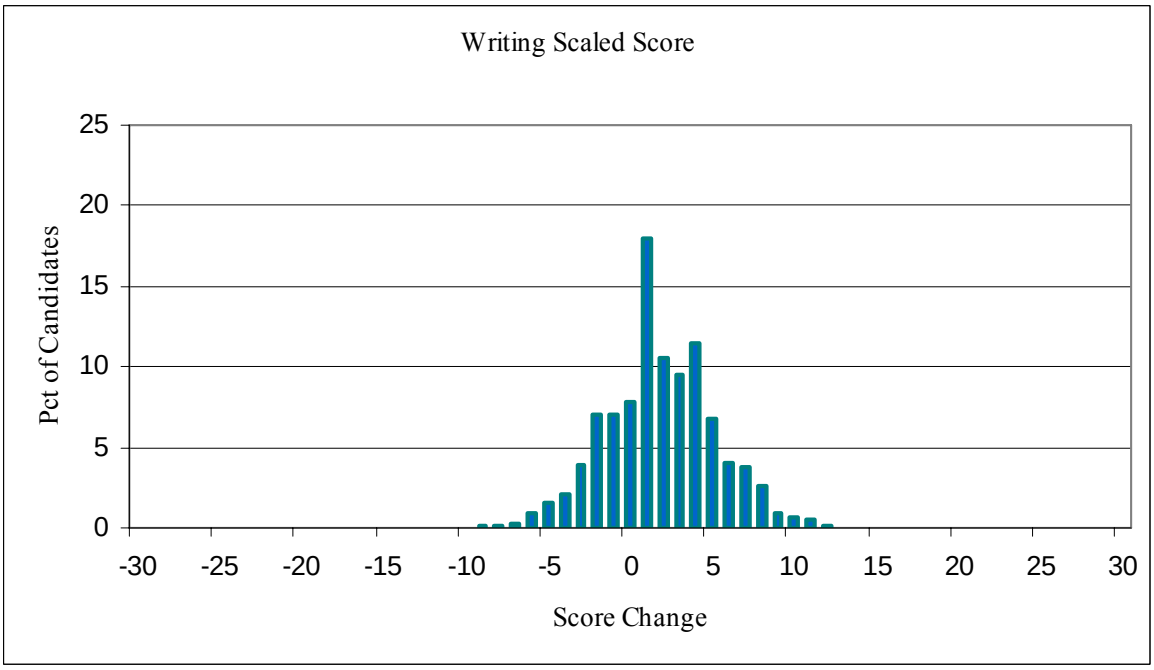


***Figure 4.*** **Frequency distributions of writing score changes.**

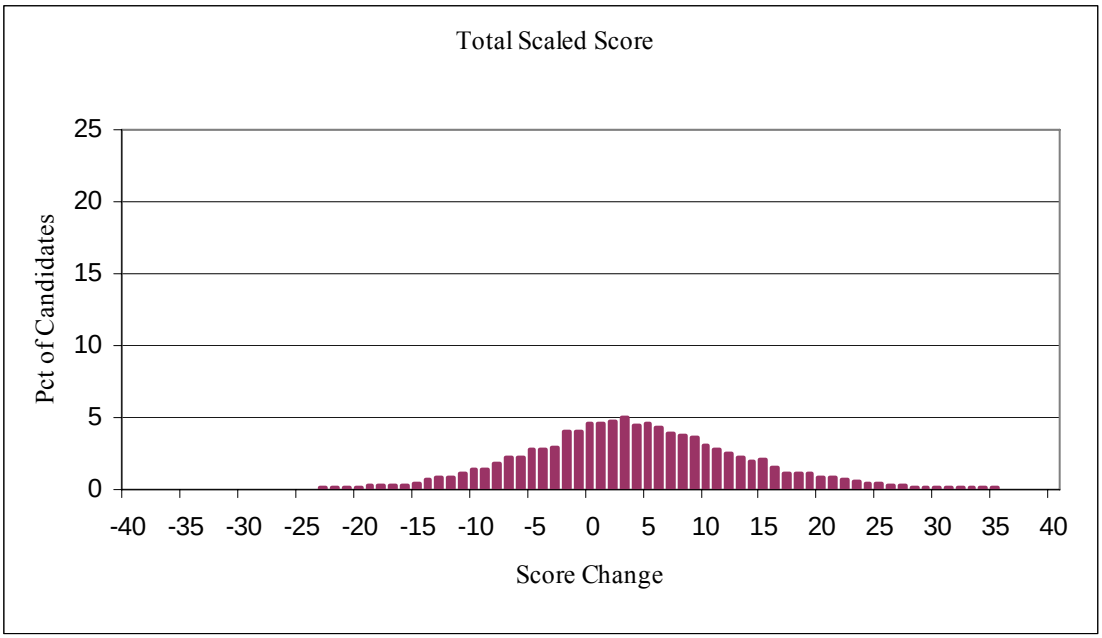


***Figure 5.*** **Frequency distributions of total scaled score changes.**

## Conclusions

In this special study, the test performance of repeaters who took a second test within 30 days of having taken a first test in the period from January to August 2007 was examined and evaluated. Small changes were observed in the test scores between the repeaters' first tests and their second tests. In addition, the effect sizes of the mean score changes of the four sections and the total score were found to be small, reinforcing the fact that the mean score changes are negligible. High to moderate correlations between the two test scores indicated a high degree of consistency in repeaters' rank orders of their scores. In the context of the data used in the study, the correlations are reflective of the test-retest reliability of alternate forms except that the data were not collected from a controlled design.

The distributions of score changes between the two tests resemble a symmetric bell-shaped distribution, suggesting that the majority of repeaters' scores changed slightly since they are clustered in the middle where the differences are from zero to a few points in each direction. There can be a number of factors accounting for the large score changes but it is hard to pinpoint these factors given the nature of the data that were available for this study. By examining some specific cases of large score changes in the data, it was noticed that some test-takers were only interested in a specific section in the first test and a different section in the second test, resulting

in large score differences on such sections between the two tests. These repeaters' performances were definitely not typical of regular TOEFL iBT test-takers.

The information on repeaters' performance may help candidates make informed decisions about the need for repeating the test. The findings from this study suggest that TOEFL iBT scores appear to remain stable across test forms within the studied period of time. However, these findings constitute only one piece of empirical evidence about the stability and validity of the TOEFL iBT test scores. In the future, a specially designed research study and a more extensive analysis will be necessary to validate these findings.

## References

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

Lewis, C. (2007). *Setting large score difference (LSD) thresholds for TOEFL.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.