Research Memorandum
ETS RM–23-02

TOEFL iBT® MyBest® Scores

TOEFL® Psychometrics Team

May 2023

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

TOEFL iBT® MyBest® Scores

TOEFL® Psychometrics Team
ETS, Princeton, New Jersey, United States

May 2023

Corresponding author: T. Li, E-mail: tli002@ets.org

Suggested citation: TOEFL® Psychometrics Team. (2023). *TOEFL iBT® MyBest® scores* (Research Memorandum No. RM-23-02). ETS.

Find other ETS-published reports by searching the
ETS ReSEARCHER database.

To obtain a copy of an ETS research report, please visit
https://www.ets.org/contact/additional/research.html

Action Editors: Jonathan Schmidgall

Reviewers: Larry Davis and Ikkyu Choi

Abstract

Since 2019, TOEFL iBT® score reports have included MyBest® scores, which are a test taker's highest section scores from all test dates within the past 2 years and a total score that is the sum of MyBest scores for each of the four sections of the TOEFL iBT test. Using data from test takers in 2021 who had taken the TOEFL iBT more than once in a 2-year span, this study investigated how the MyBest scores of these test takers, or repeaters, differed both from their most recent scores and from their test date best total scores, which are their best scores in a single test administration. In addition, the characteristics of the 2021 repeater sample were compared to the characteristics of a 2016 sample. Findings indicate that around 30% of TOEFL iBT test takers in 2021 were repeaters, and the differences between first and most recent total scores increased as the number of testing occasions increased. Approximately 22% of these repeaters obtained their MyBest total scores after they achieved their test date best total scores. At the total score level, the average difference between MyBest and most recent total scores was 5.69 scale score points. Between a repeater's MyBest score and the test date best total score, the average difference was 2.56 points. These findings indicate that that most repeaters can anticipate small to moderate differences between their MyBest scores and test scores from a single test administration date. Overall, the repeater sample composition and score patterns were very similar between 2021 and 2016 cohorts.

Keywords: superscores, TOEFL iBT® test, score reports, repeaters, MyBest® scores

In August 2019, TOEFL iBT® score reports debuted a new set of scores called MyBest® scores, in addition to the scale scores achieved on the test takers' current test date (ETS, 2019). MyBest scores offer repeaters, test takers who take a test more than once, a way to show their best overall performance on the TOEFL iBT by presenting their highest section scores from all test dates within the past 2 years and a total score that is the sum of MyBest section scores.

The use of MyBest scores, also referred to as superscores, reflects an increasingly popular trend in university admissions of reviewing the highest section scores of applicants from multiple testing occasions. This practice provides opportunities for test takers to leverage their best performance on each section and compensate for suboptimal performance caused by factors unrelated to the test. It also enhances the meaning of score interpretations by providing score users with test takers' highest level of ability (Bachman & Palmer, 2010).

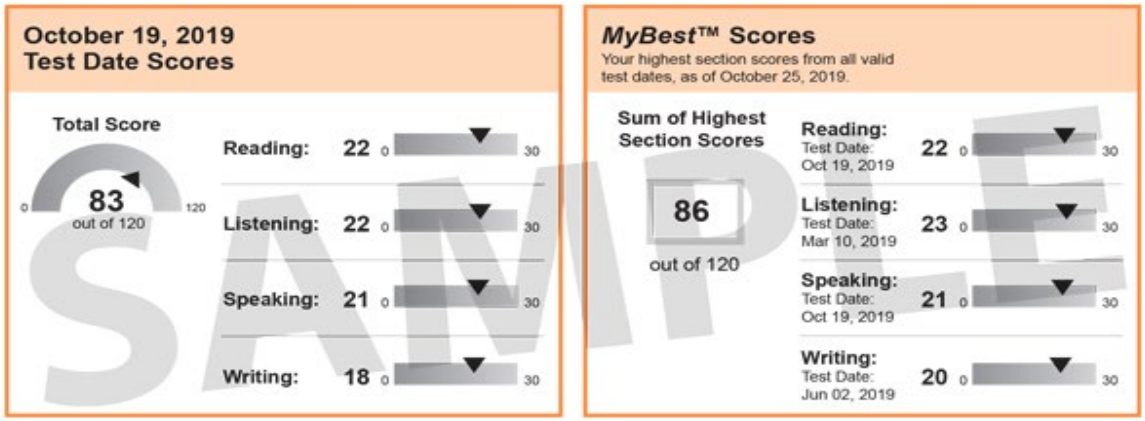
Table 1 shows an example of TOEFL iBT MyBest scores for a test taker who has taken the test three times. A sample score report is also shown in Figure 1.

Table 1. An Example of TOEFL iBT MyBest Scores for a Repeat Test Taker

Section	Test date 03/10/2019	Test date 06/02/2019	Test date 10/19/2019	MyBest scores
Reading	16	18	**22** [a]	22
Listening	**23** [a]	20	22	23
Speaking	20	19	**21** [a]	21
Writing	19	**20** [a]	18	20
Total	78	77	83	86

[a] Indicates MyBest score (in bold).

Figure 1. Sample TOEFL iBT Score Report



Studies have shown that superscores provide an indication of test taker proficiency that is largely consistent with other commonly used scoring methods, such as the use of most recent test scores or the use of average scores across all testing occasions. Results of these research studies indicated that observed score increases between superscores and scores from a single test administration reflect the positive effects of retesting (Cruce & Mattern, 2020; Mattern & Radunzel, 2019; Mattern et al., 2018; Roszkowski & Spreat, 2016).

In 2018, prior to the launch of TOEFL iBT MyBest scores, we conducted a study based on the 2016 test takers who took TOEFL iBT more than once within a 2-year timeframe (i.e., 29.25% of the test takers in 2016 were repeaters). That study included a set of analyses evaluating sample composition and test-taking patterns of the repeaters, as part of an effort to inform the feasibility of providing MyBest scores for the TOEFL iBT program. Based on the 2016 sample, we found:

- Most repeaters had small to moderate differences between their MyBest and most recent test scores. On average, the difference between MyBest and most recent section scores was greater for Reading and Listening than for Speaking and Writing. The average differences were within 1 standard error of measurement (*SEM*) for all four sections.
- At the total score level, the average difference between MyBest and most recent total score was 4.09 points, and between MyBest and test date best score was 2.44 points. The differences were also within 1 *SEM* for the total score.
- Note, test date best total score refers to the total score from a single test-administration date when the maximum total score was achieved. In the example as shown in Table 1, the test date best total score was 83 and was obtained on October 19, 2019; the MyBest total score was 86 and was achieved on the same date.

The purpose of the present study was to reevaluate the previous findings using data from TOEFL iBT repeaters in 2020–2021, who took the test following the launch of MyBest scores. Specifically, our goal was to understand how repeaters' MyBest scores differed from their most recent scores and from their test date best total scores. In addition, the repeater sample characteristics of 2021 were compared with the characteristics of the 2016 sample to

investigate if MyBest scores have led to any changes of test takers' testing behaviors based on their score patterns, such as the number of retests in relation to score changes.

Method

Data

TOEFL iBT test scores are certified by ETS to be valid for up to 2 years (see Powers & Lall, 2013). Thus, this study used TOEFL iBT data that reflected this 2-year timeframe. Specifically, all of the valid 2021 TOEFL iBT test takers' records and their previous records, determined by their most recent 2021 test date and extending back 2 years, were included in the analyses. Among them, 69.28% of the test takers tested only once within the 2-year timeframe, and the remaining 30.72% test takers took TOEFL iBT more than once. The data from these repeaters were the focus of the current analyses.

Analyses

Two sets of analyses were conducted. The first set focused on describing the characteristics of repeaters in 2021, including their test-taking patterns and demographic information. Also included was the corresponding information of their counterparts in 2016. The second set of analyses focused on examining MyBest scores of the 2021 cohort, especially how these scores differed from test takers' most recent scores and test date best total scores.

Results

Characteristics of Repeaters

Comparison of Average Total Scores by Number of Testing Occasions

Table 2 shows the average total score at the first testing occasion compared to that at the most recent testing occasion for the 2021 repeater sample grouped by the number of testing occasions within the 2-year timeframe. As shown, more than half of the repeaters (52%) took TOEFL iBT only twice, followed by 20% of the test takers who tested three times, and 10%, four times. In comparison with the average score of nonrepeaters (mean 88.11, standard deviation 22.01), repeaters' average total scores were lower at both the first and most recent testing occasions. At the first testing occasion, the average total scores ranged from 70.39 for those who tested 10 or more times to 76.58 for those who tested only twice, with the scores

increasing as the number of testing occasions decreased. With respect to the most recent scores, smaller differences were seen across repeater groups. They ranged from 84.98 for those who tested 10 or more times to 86.13 for those who tested twice.

Table 2. Comparison of First and Most Recent Total Scores by Number of Testing Occasions Based on the 2021 Repeater Sample

Number of testing occasions	Percentage	Mean and *SD* of first total score	Mean and *SD* of most recent total score
2	52	76.58 (24.21)	86.13 (20.75)
3	20	74.96 (22.76)	85.23 (20.49)
4	10	73.79 (21.77)	85.32 (20.09)
5	6	72.71 (21.28)	85.58 (19.34)
6	4	71.98 (20.78)	85.02 (20.05)
7	2	71.10 (20.78)	85.33 (19.99)
8	2	71.09 (19.90)	85.57 (19.28)
9	1	70.42 (19.91)	85.81 (19.91)
10^{+}	3	70.39 (20.03)	84.98 (20.59)

Note. The average score of nonrepeaters was 88.11 (22.01). The numbers in parentheses are the standard deviations (*SD*s).

Figure 2. Test-Taking Patterns of Repeaters in 2021 Versus 2016

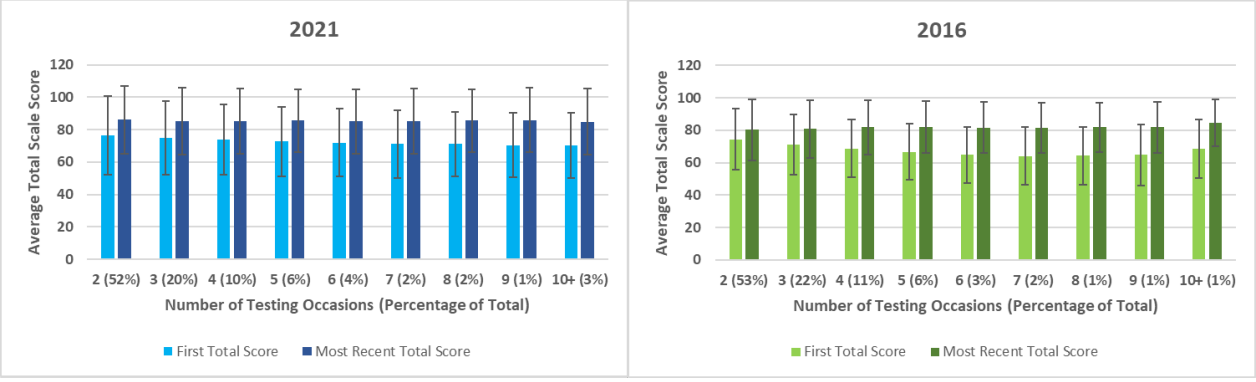
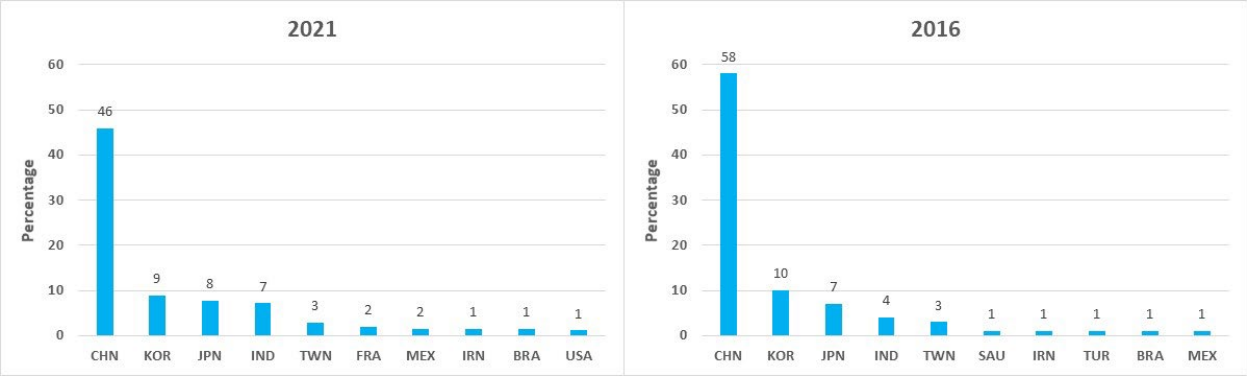


Figure 2 displays two graphs visualizing the composition of repeater cohorts and the differences between first and most recent total scores grouped by the number of testing occasions, for 2021 and 2016. Overall, the repeater sample composition and score patterns were very similar between 2021 and 2016, with the differences between first and most recent total scores increasing as the number of testing occasions increased.

Percentage of Repeaters by Top 10 Native Countries

Figure 3 shows a comparison of the percentages of repeaters for top 10 native countries between 2021 and 2016. From the graph, it can be seen that in 2021 for those who took TOEFL iBT more than once, the majority were Chinese (46%), followed by Korean (9%), and Japanese (8%). The overall pattern was similar between 2021 and 2016, but the percentage of Chinese repeaters slightly dropped from 58% in 2016 to 46% in 2021, which may be due to a test-taking population change during the COVID-19 pandemic.

Figure 3. Percentage of Repeaters for Top 10 Native Countries



Note. BRA = Brazil, CHN = China, FRA = France, IND = India, IRA = Iran, JPN= Japan, KOR = Korea, MEX = Mexico, SAU = Saudi Arabia, TUR = Turkey, TWN = Taiwan, USA = United States of America.

MyBest Score—A Closer Look

This section provides a closer look at MyBest scores of the 2021 repeater sample as pertains to the timing of achieving MyBest scores and the pattern of score differences from multiple testing occasions.

Timing of MyBest Scores

To investigate the timing of when MyBest scores were achieved, test-taker patterns were examined with a comparison of when MyBest total score was obtained relative to test date best total score. Note that because a test taker may achieve MyBest section scores at different testing occasions, the latest date among four MyBest section score dates was counted as the date when MyBest total score was obtained.

As shown in Table 3, 75.95% of 2021 repeaters' MyBest total scores was achieved on the same test date as the test date best scores, and 22.13% achieved their MyBest total scores after their test date best scores. This suggests that 22.13% of the test takers were able to improve their performance on at least one section when they repeated the test after achieving their test date best total scores.

Table 3. Comparison of When MyBest Total Score Was Obtained Relative to Test Date Best Score

When	Percentage
MyBest before test date best	1.92
MyBest on test date best	75.95
MyBest after test date best	22.13

Differences Between MyBest and Most Recent Section Scores

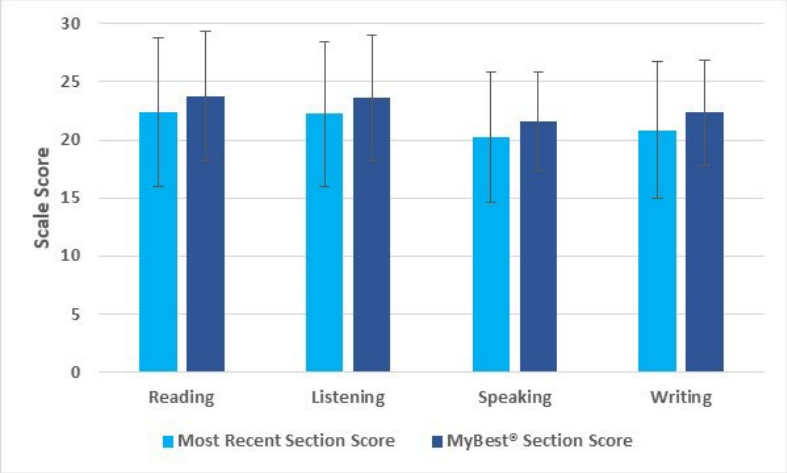
Descriptive statistics for MyBest and most recent section scores and their differences are summarized in Table 4 and Figure 4. Specifically, for the Reading and Listening sections, the average differences were 1.39 and 1.43 points, with standard deviations of 2.67 and 2.80, respectively. For the Speaking and Writing sections, the average differences were 1.36 and 1.51 points, with standard deviations of 3.62 and 3.84, respectively. The average differences were within 1 *SEM* for all four sections.

Table 4. Descriptive Statistics of MyBest and Most Recent Section Scores, and Their Differences

Section	Score	Mean	*SD*	Min	25th	Median	75th	95th	Max
Reading	Most recent	22.39	6.37	0	19	24	28	30	30
SEM = 2.34	MyBest	23.79	5.55	0	20	25	28	30	30
	MyBest: most recent	1.39	2.67	0	0	0	2	6	30
Listening	Most recent	22.25	6.25	0	19	24	27	30	30
SEM = 2.38	MyBest	23.67	5.39	0	20	25	28	30	30
	MyBest: most recent	1.43	2.80	0	0	0	2	6	30
Speaking	Most recent	20.36	5.60	0	18	21	23	27	30
SEM = 1.57	MyBest	21.61	4.24	0	19	22	24	28	30
	MyBest: most recent	1.36	3.62	0	0	0	1	5	30
Writing	Most recent	20.83	5.91	0	19	22	25	28	30
SEM = 2.14	MyBest	22.34	4.51	0	20	23	25	28	30
	MyBest: most recent	1.51	3.84	0	0	0	2	6	30

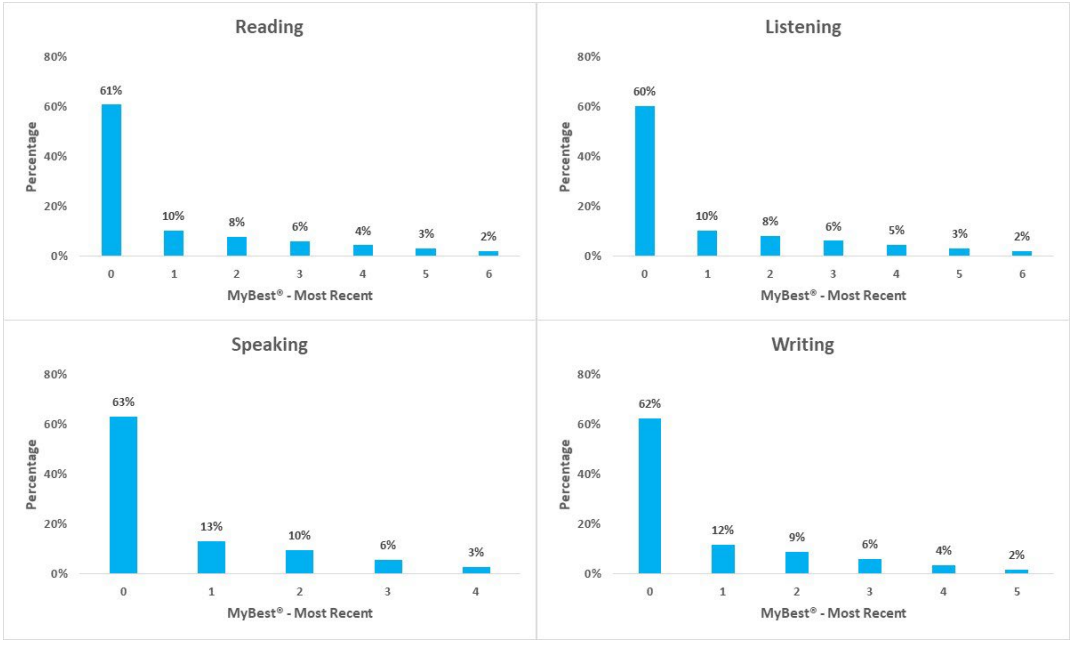
Note. Max = maximum, min= minimum, *SD* = standard deviation, *SEM* = standard error of measurement, 25th, 75th, and 95th = quartiles.

Figure 4. The Average MyBest and Most Recent Section Score and Their Differences



Distributions of the differences are plotted in Figure 5. Results suggest that the distributions of the differences between MyBest and most recent section scores were similar for the Reading and Listening sections. The distributions for the Speaking and Writing sections were similar, but with larger standard deviations than those for Reading and Listening. As shown in Figure 5, 60% to 63% of the repeaters had their MyBest section scores at the most recent testing occasion.

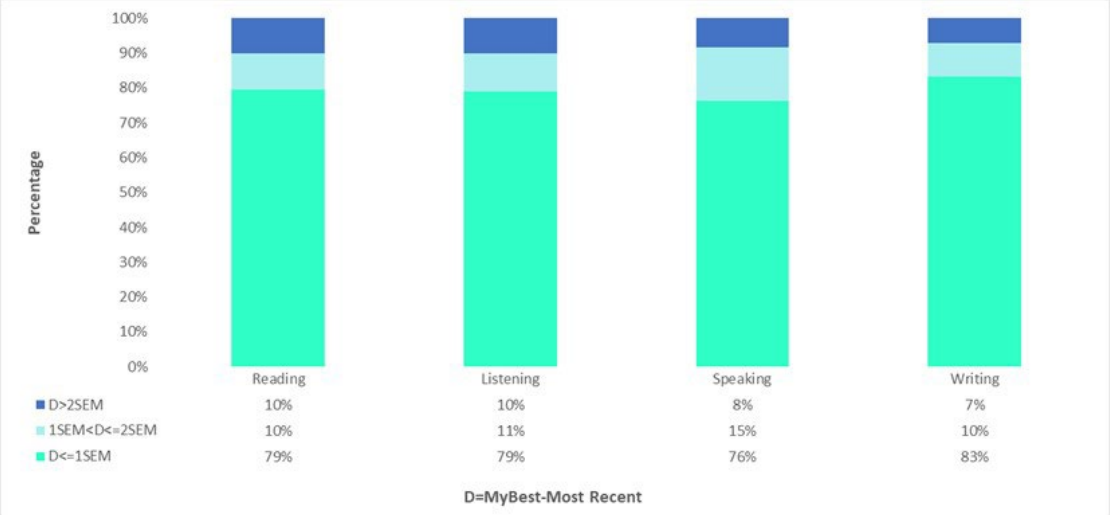
Figure 5. Distributions of the Differences Between MyBest and Most Recent Section Scores



Note. The largest value on the x-axis is the 95th percentile of the difference.

For each section, differences between MyBest and most recent scores were compared with their respective SEMs, as plotted in Figure 6. Results suggest that across all four sections, the majority of the test takers had score differences within 1 *SEM*. The Writing section had the greatest percentage of score differences within 1 *SEM*, while the Reading, Listening, and Speaking sections had similar percentages of score differences within 1 *SEM*.

Figure 6. Percentages of the Differences Between MyBest and Most Recent Sections Scores in Comparison to the Standard Error of Measurement



Differences Between MyBest, Most Recent, and Test Date Best Total Scores

In this section, the focus is on three types of total scores—MyBest, most recent, and test date best. Descriptive statistics for the three types of total scores and the differences between MyBest and the other two types of scores are summarized in Table 5. These results show that MyBest score was on average 5.69 points higher than most recent score, with a standard deviation of 9.86 points. Also, the MyBest total score was on average 2.56 points higher than the test date best score, with a standard deviation of 2.53. The average difference between MyBest and most recent scores was greater than 1 *SEM* (4.26) for total scores, and the average difference between MyBest and test date best scores was within 1 *SEM* for total scores.

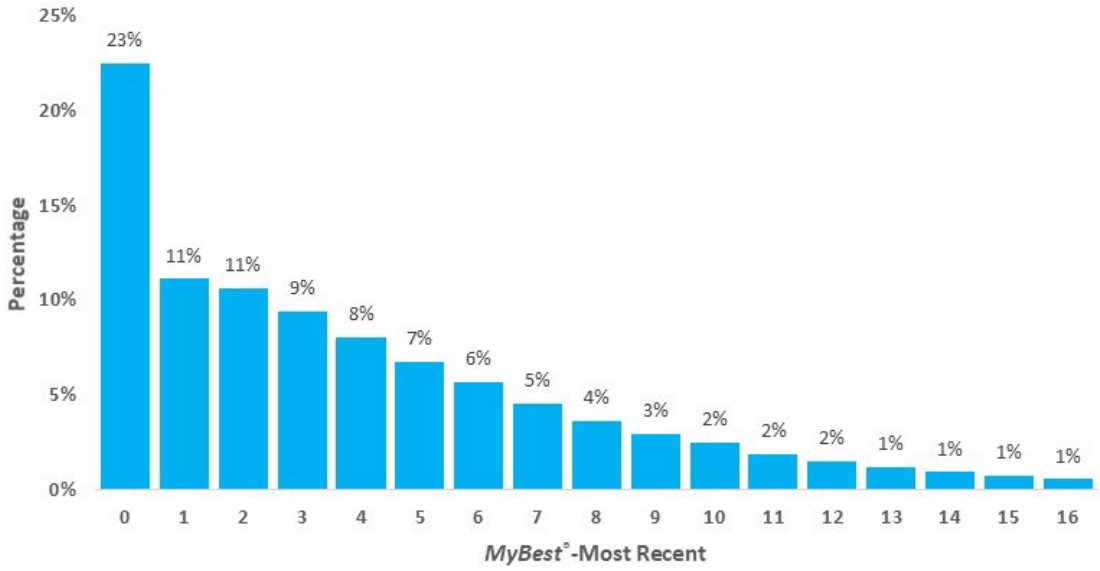
Table 5. Descriptive Statistics of Most Recent, Test Date Best, and MyBest Total Scores and Score Differences

Scores	*M*	*SD*	Min	25th	Med	75th	95th	Max
Most recent	85.73	20.47	0	76	90	101	110	120
Test date best	88.85	17.41	0	79	92	102	111	120
MyBest	91.41	17.00	0	82	95	104	113	120
MyBest: most recent	5.69	9.86	0	1	3	7	17	114
MyBest: test date best	2.56	2.53	0	0	2	4	7	38

Note. M = mean; max = maximum; med = medium; min = minimum; *SD* = standard deviation; 25th, 75th, and 95th = quartiles.

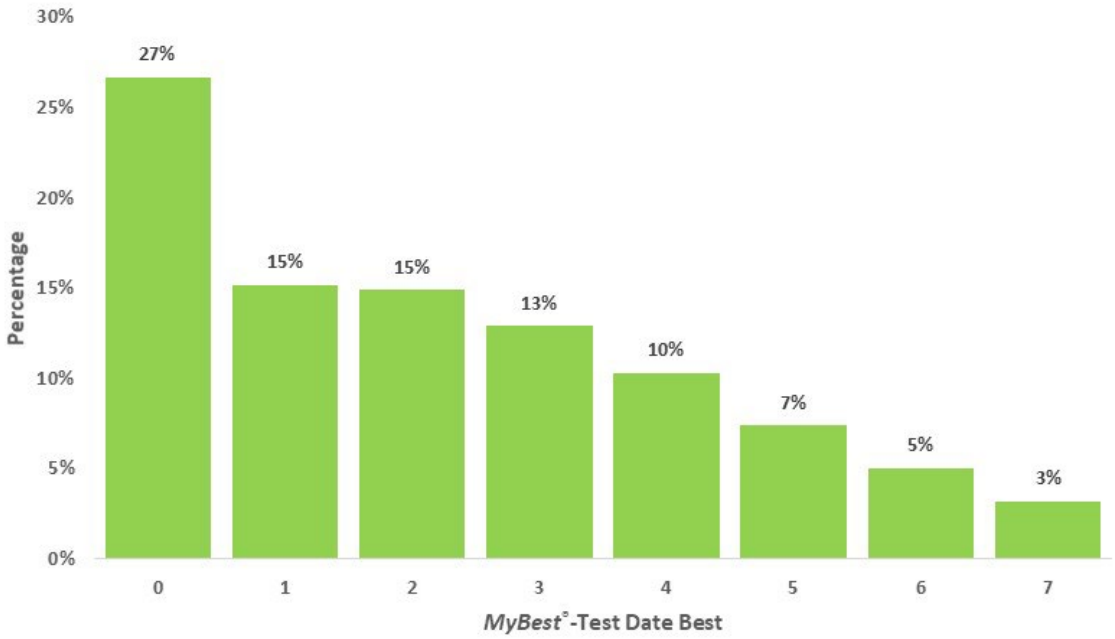
Figure 7 displays the distribution of the differences between MyBest and most recent total scores. Approximately 23% of the repeaters had a score difference of 0 between their MyBest and most recent total scores. Figure 8 shows the distribution of the differences between MyBest and test date best total scores. Approximately 27% of the repeaters had a score difference of 0 between their MyBest and test date best total scores. About 95% of the repeaters had a score difference within 7 scale score points.

Figure 7. Distribution of the Differences Between MyBest and Most Recent Total Scores



Note. The largest value on the x-axis is the 95th percentile of the difference.

Figure 8. Distribution of the Differences Between MyBest and Test Date Best Total Scores



Note. The largest value on the x-axis is the 95th percentile of the difference.

Summary

The current research memorandum examined the characteristics of TOEFL iBT repeaters and investigated if MyBest scores have led to any changes of test takers' retesting behaviors based on their score patterns. In summary, the analysis results obtained from all valid 2021 TOEFL iBT test takers' records and their previous records within the 2-year time window showed the following:

- About 30% of the TOEFL iBT test takers in 2021 were repeaters, and on average they had lower performance than nonrepeaters. The majority of the repeaters took the test four times or less (82%), and the differences between first and most recent total scores increased as the number of testing occasions increased. This trend was within our expectation and comparable between the 2021 and 2016 repeater cohorts.
- Approximately 22% of the repeaters obtained their MyBest total scores after they achieved their test date best total scores, indicating that they were able to improve their performance on at least one of the four sections. The average differences between MyBest and most recent section scores were 1.39, 1.43, 1.36, and 1.51 for the Reading,

Listening, Speaking, and Writing sections, respectively. They were all within 1 *SEM* of the corresponding section scale scores. In 2016, the average differences were 1.18, 1.29, 0.79, and 0.82 for the four sections, respectively.

- At the total score level, the average difference between MyBest and most recent total scores was 5.69 scale score points, which was greater than 1 *SEM* for the total scale score, and between MyBest and test date best total scores was 2.56 points and within 1 *SEM* for the total scale score. Overall, 62% of the repeaters in 2021 had a total scale score difference between MyBest and most recent scores within 1 *SEM*. In 2016, the average difference between MyBest and most recent total score was 4.09 points, and between MyBest and test date best total score was 2.44 points.

MyBest scores are useful indicators of test-taker performance as they are drawn from all valid test scores within a 2-year period. Based on the findings from the current analyses, it is believed that most repeat test takers can anticipate small to moderate differences between their MyBest scores and test scores from a single test administration date. These small differences may be important to test takers if they are retesting to meet a specific score requirement for admission purposes. On the other hand, as shown above, most score changes were within 1 *SEM*, so MyBest scores can help mitigate the impact of the small amount of noise that is inevitably present in any type of measurement and help test takers demonstrate their best performance.

Lastly, it is worth noting that due to the COVID-19 pandemic, TOEFL iBT test-taker composition changed in the years of 2020 and 2021. We recommend that the change in test-taker composition be taken into consideration when interpreting the results in this study.

References

Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford University Press.

Cruce, T., & Mattern, K. (2020). *The impact of superscoring on the distribution of ACT scores* [Issue brief]. ACT.

ETS. (2019). *MyBest Scores: A rationale for using TOEFL iBT superscores.*

Mattern, K., & Radunzel, J. (2019). *Does superscoring increase subgroup differences?* [Technical brief]. ACT.

Mattern, K., Radunzel, J., Bertling, *M*., & Ho, A. D. (2018). How should colleges treat multiple admissions test scores? *Educational Measurement: Issues and Practice, 37*(3), 11–23. https://doi.org/10.1111/emip.12199

Powers, D. E., & Lall, V. (2013). *Supporting an expiration policy for English language proficiency test scores* (Research Memorandum No. RM-13-09). ETS.

Roszkowski, *M*., & Spreat, S. (2016). Retaking the SAT may boost scores but this doesn't hurt validity. *Journal of the National College Testing Association, 2*(1), 1–16.