ETS TOEFL

Quality Beyond Measure.

TOEFL iBT® Research Report
TOEFL iBT–29
ETS RR–17-31

The Effects of Different Levels of Performance Feedback on *TOEFL iBT*® Reading Practice Test Performance

Yasuyo Sawaki

December 2017

The *TOEFL*® test was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the *Graduate Record Examinations*® (*GRE*®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education. The test is now wholly owned and operated by ETS.

ETS administers the TOEFL program under the general direction of a policy board that was established by, and is affiliated with, the sponsoring organizations. Members of the TOEFL Board (previously the Policy Council) represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, 2-year colleges, and nonprofit educational exchange agencies.

❖ ❖ ❖

Since its inception in 1963, the TOEFL has evolved from a paper-based test to a computer-based test and, in 2005, to an Internet-based test, the *TOEFL iBT*® test. One constant throughout this evolution has been a continuing program of research related to the TOEFL test. From 1977 to 2005, nearly 100 research reports on the early versions of TOEFL were published. In 1997, a monograph series that laid the groundwork for the development of TOEFL iBT was launched. With the release of TOEFL iBT, a TOEFL iBT report series has been introduced.

Currently this research is carried out in consultation with the TOEFL Committee of Examiners (COE). Its members include representatives of the TOEFL Board and distinguished English as a second language specialists from academia. The committee advises the TOEFL program about research needs and, through the research subcommittee, solicits, reviews, and approves proposals for funding and reports for publication. Members of the TOEFL COE serve 4-year terms at the invitation of the Board; the chair of the committee serves on the Board.

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

ETS is an Equal Opportunity/Affirmative Action Employer.

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 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.

TOEFL iBT Research Report Series and ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

The Effects of Different Levels of Performance Feedback on *TOEFL iBT*® Reading Practice Test Performance

Yasuyo Sawaki

Waseda University, Tokyo, Japan

The purpose of the present study is to examine whether performance on the *TOEFL iBT*® Reading practice test is affected by 3 different levels of feedback provided to learners upon completion of reading exercises: (a) correctness of learner response (the knowledge of correct results [KCR] feedback), (b) KCR feedback and rationales for correct/incorrect answers, and (c) KCR feedback and the rationales along with on-demand video lectures. Japanese learners of English completed 18 reading exercises and reviewed the results by using the type of feedback provided to the group to which each learner had been assigned. They also completed 2 TOEFL iBT Reading test forms: 1 as a pretest and 1 as a posttest. A multilevel modeling analysis of data from 193 participants who completed the study showed that despite a gain of over 3 scaled score points on the TOEFL iBT Reading test from the pretest to the posttest, the level of feedback was not found to affect the test score. A learner background variable, previous knowledge about the TOEFL iBT test, was identified as a predictor positively related to the test score. Moreover, a series of ANOVAs conducted on learner survey data suggested that the level of feedback affected the degree to which study participants perceived the usefulness of the study materials in understanding different item types that appear on the TOEFL iBT Reading section and in developing strategies for understanding the gist of the text within a short time.

Keywords *TOEFL iBT*® test; reading; test preparation; feedback; multilevel modeling

doi:10.1002/ets2.12159

The *TOEFL*® test is an international assessment of academic English often used for making decisions to admit nonnative speakers of English to academic programs. The latest version of the test, *TOEFL iBT*®, was launched in 2005 with major changes to the test design, including the introduction of a speaking section, tasks in the speaking and writing sections that integrate modalities, and new item types in the reading and listening sections. The rationale behind these design changes is for the new test to better reflect what international students actually do with the language in academia and thus promote a positive test impact on stakeholders (C. A. Taylor & Angelis, 2008).

Due to the high-stakes nature of decisions made based on TOEFL iBT test scores, many TOEFL test takers engage in various activities to prepare for the test, ranging from studying off-the-shelf test preparation materials by themselves to attending intensive TOEFL test preparation courses. However, currently researchers know little about the extent to which such test preparation activities lead to score gain on the TOEFL iBT test or to the improvement of general academic language ability (Hamp-Lyons, 1998; Spratt, 2005; Wall, 2000). Furthermore, although such investigations should work in tandem with the consideration of the nature of the materials used for test preparation by teachers and learners, previous studies that examined the effects of specific features of test preparation material on test score improvement are scarce. In particular, test preparation activities typically entail learners working on language exercises that are designed to be similar to those on the test and receiving performance feedback on the practice items. However, currently little is known about exactly what kinds of feedback may be effective in improving test scores. As an attempt to address this gap in the literature, this study examines relative effectiveness of different levels of feedback on learners' language test performance with a specific focus on preparing for the TOEFL iBT Reading section.

Effectiveness of Test Preparation on Score Improvement and Test Preparation Materials

Previous empirical studies, particularly those in educational measurement, shed light on the effects of test preparation programs on cognitive test performance. For instance, Powers's (1993) summary of four previously published meta-analyses

Corresponding author: Y. Sawaki, E-mail: ysawaki@waseda.jp

of preparation courses for the *SAT*® and later studies showed that empirical results do not necessarily support a common belief that test preparation can lead to dramatic score improvement. According to Powers, the expected average score gains attributable to coaching were only 3 points on the SAT Verbal section and 17 points on the SAT Mathematics section (both on the scale of 200 – 800 points), suggesting that the effects of coaching on SAT scores may be modest at best.

By contrast, only a few published studies addressing the effects of language test preparation on test score gain are currently available. One is Read and Hayes's (2003) 4-week observation study of IELTS preparation courses at two different language schools in New Zealand with different instructional foci (one short-term course focusing on test preparation activities and one longer course taking a more general approach to fostering overall language ability). A statistical comparison of the scores on two retired IELTS test forms administered as the pretest and the posttest to the students in the two programs resulted in no statistically significant increase of the IELTS band score after the instruction, except that the listening section score for the group of students that received more extensive listening practice showed a significant improvement on the posttest. However, these results were inconclusive due to the small sample size used for the statistical analyses.

In a study conducted to examine the effects of introducing a new test on learners' language ability, Andrews, Fullilove, and Wong (2002) cautioned that a score gain may not necessarily indicate meaningful improvement of language ability. Andrews et al. examined the impact of a newly introduced high-stakes public English-speaking test in Hong Kong on secondary students' oral language ability. A speaking test that simulated the new oral test was administered to three cohorts of Secondary 7 students: the 1993 cohort (the last group not taking the new oral test) and the 1994 and 1995 cohorts (the first two groups taking the new test). A comparison of the mean scores for matched samples drawn from the three cohorts showed a noticeable difference between the 1993 and 1995 cohorts, indicating a delayed potential positive effect of the new test on students' speaking ability. A subsequent qualitative analysis of the students' speech samples indicated that the score improvement might be attributable to students' improved time management and the use of certain formulaic expressions and strategies for organizing the discourse that appeared in commercial textbooks that were available around that time. Andrews et al. found, however, that the students' use of such expressions was not always appropriate to the context, suggesting that their learning was rather superficial.

Although more studies on the effects of language test preparation on score gain are certainly needed, such investigations should be conducted in tandem with the consideration of various factors that may affect learning. One area that is particularly worthy of a focused investigation is the nature of instructional materials and how they are used by teachers and learners for test preparation. Chapelle (2008) and Wang, Eignor, and Enright (2008) explicated a key role that instructional materials play in building an argument for the TOEFL iBT test. For Chapelle, the claim that the TOEFL iBT score is useful for making various decisions about candidates, which needs to be examined to build an argument for the test, is based on the assumption that the test gives a positive impact on English instruction. This "requires evidence that appropriate instructional materials are available and that they are being used to improve English language teaching" (Chapelle, 2008, p. 346). This point is important because, as described in Bailey's (1996) model of washback, instructional materials can directly affect what and how teachers teach, which in turn can affect students' process of learning and the product of their learning. Furthermore, Andrews et al.'s (2002) results cited previously suggest how the content of instructional materials can directly affect students' responses to test tasks as well.

Previous studies give some insight into the content and use of instructional materials for language test preparation. A notable point about the use of such materials reported in multiple studies is teachers' heavy reliance on commercially available course books (Alderson & Hamp-Lyons, 1996; Read & Hayes, 2003) for reasons such as meeting student expectations, taking advantage of information about methods for instruction, and lack of confidence and time to develop materials on their own (Alderson & Hamp-Lyons, 1996; Cheng, 1997; Wall & Horak, 2011). Despite teachers' reliance on commercial textbooks, however, an issue of particular importance about the content of test preparation materials raised in the previous literature is the lack of information that helps learners and their teachers identify areas that require further practice. Hamp-Lyons (1998) reported that only one of the five textbooks for the TOEFL paper-based test reviewed in her study consistently included rationales for correct and incorrect answers, suggesting that students would not receive any diagnostic information about their performance beyond the correctness of their answers in the majority of textbooks. Coupled with the unavailability of clear explanations about underlying constructs assessed on the TOEFL test, Hamp-Lyons concluded that the textbooks that she reviewed offered little guidance on instructional planning for teachers. Similarly, Wall and Horak's (2011) recent comparison of textbooks for TOEFL computer-based tests and TOEFL iBT textbooks used by teachers before and after the launch of the TOEFL iBT test in Central and Eastern Europe found that only half of the TOEFL textbooks they reviewed included rationales for correct and incorrect answers.

To summarize, empirical evidence currently available about the effectiveness of second language (L2) test preparation on score improvement is limited. Moreover, despite teachers' tendency to rely on commercial test preparation materials reported in previous studies, using such materials may not promote short-term improvement of test scores or long-term development of language ability in general due to the lack of helpful diagnostic information. In particular, it is worth noting that little empirical evidence is available as to how specific features of the test preparation material content affect test score improvement. Although the argument for more diagnostic information to better guide teachers and learners discussed in the previous studies is reasonable, currently we know little about whether such diagnostic information affects test performance at all, and if so, exactly what type of feedback makes test preparation more effective in improving test scores.

Providing Feedback on Second Language (L2) Reading Comprehension Performance

In L2 acquisition research, feedback provided to the learner in various forms is considered to play an important role in the learner's L2 development. According to Lyster and Ranta's (1997) framework of corrective feedback in oral interaction in the L2 classroom, for instance, when a learner is having a conversation with an interlocutor, feedback from the interlocutor for negotiation of meaning (e.g., a clarification request such as the interlocutor's request to repeat a word) would lead to a learner uptake (the learner's noticing the presence of a pronunciation problem), which in turn may result in a repair, or the learner modifying his or her output (e.g., a more targetlike pronunciation of the word that the interlocutor has requested to be repeated). A large number of studies have been published on the effects of feedback on the development of productive aspects of language ability (e.g., see Biber, Nekrasova, & Horn, 2011; Lyster & Sato, 2010; Russell & Spada, 2006, for recent research syntheses on the effects of feedback on writing and speaking performance). However, so far, only a few studies have examined how feedback provided to the learner affects L2 reading comprehension. Two studies, which examined the effects of feedback given to learners on their reading comprehension performance during instruction on their subsequent reading test or exercise performance, are particularly relevant to the present purpose.

The first is Kozulin and Garb's (2001) study on the effects of feedback provided in the form of teacher mediation in one-on-one tutoring on Israeli students' standardized English reading comprehension test performance. Kozulin and Garb's study was conceptualized within the framework of dynamic assessment, a type of classroom assessment based on sociocultural theory, in particular, on Vygotsky's zone of proximal development (e.g., Lantolf & Poehner, 2008; Leung, 2007; Poehner, 2007). Kozulin and Garb developed teacher intervention materials based on a detailed task analysis of the test items. After student participants completed a pretest form, each student met with a teacher individually for reviewing the test using the teacher materials. This intervention resulted in a large positive effect on the students' performance on an alternate test form administered as a posttest, although this result should be interpreted with caution due to the lack of a control group.

The second study is Murphy's (2007) experimental study on the effects of L2 proficiency level, manner of study, and type of feedback on L2 learners' performance on multiple-choice reading comprehension items provided in a computer-assisted language learning environment. Although not conducted in the context of test preparation, this study is relevant because it compared the effects of two different types of feedback on learners' reading comprehension exercise performance. First-year English majors at a university in Japan were randomly assigned to four different study conditions where they completed a first reading comprehension exercise either individually or in pairs while receiving one of two types of feedback: correct answers to individual items or elaborative feedback that offered hints until they answered each item correctly. ANOVA results of 225 students' performance data showed that the higher proficiency group (determined based on the English course placement level) performed better than did the lower proficiency group on a second reading comprehension exercise and that those who completed the first exercise in pairs with the elaborative feedback performed significantly better than did the others on the same exercise. Coupled with the results of a qualitative analysis of the interaction of six pairs that showed a high quality of the interaction under the elaborative feedback condition, Murphy concluded that a combined use of pair work and elaborative feedback might promote L2 reading comprehension.

These studies by Kozulin and Garb (2001) and Murphy (2007) are highly informative in conceptualizing feedback conditions for the present study. First, the feedback conditions employed by Murphy are based on a traditional taxonomy of feedback on the learner's performance on multiple-choice items employed in cognitive and educational measurement

(Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Fuchs & Fuchs, 1986; Kluger & DeNisi, 1996; Nyquist, 2003). In one such approach, the amount of information provided to the learner is varied to devise study conditions with different levels of feedback. In such studies, a no feedback condition, where the learner receives no feedback on his or her responses to individual items, might serve as the control condition. Other feedback conditions might present different levels of information about a given item in a cumulative manner: (a) the knowledge of results (KR) condition, where the correctness of the learner's responses is reported; (b) the knowledge of correct results (KCR) condition, where the information offered in the KR condition is supplemented with correct answers to individual items; and (c) the rationales condition, where more elaborative feedback, such as explanations as to why a given option is correct or incorrect (rationales), is provided along with the information available in the KR and KCR conditions. Furthermore, the one-on-one tutoring condition employed by Kozulin and Garb was conceptualized as a form of feedback offering an extensive amount of information as instruction in previous feedback studies of this type (e.g., Nyquist, 2003).

Findings on the relative effectiveness of the different types of performance feedback in the extant literature are rather mixed partly due to complex interactions among the nature of feedback (e.g., content, method, timing) as well as the context in which feedback is provided to learners (Hattie & Timperley, 2007). In general, previous results suggest that providing some feedback is better than providing no feedback, but results are inconsistent as to whether more elaborative feedback necessarily leads to better learner performance (e.g., Bangert-Drowns et al., 1991; Kulhavy, White, Topp, Chan, & Adams, 1985; Nyquist, 2003; Shute, Hansen, & Almond, 2007). However, the dearth of empirical results on L2 language comprehension performance makes it difficult to gauge the extent to which the results obtained in other fields may apply to L2 reading test preparation. Thus, systematically examining the relative effectiveness of different types of feedback could inform how to design L2 reading test preparation materials that better promote student learning.

Research Questions and Their Rationales

As an attempt to explore the degree to which different types of feedback affect learner reading comprehension performance, the present study compared three feedback conditions that varied in terms of the extensiveness of the information provided to the learner. Participants were randomly assigned to feedback conditions, completed a series of reading exercises, and then received the level of feedback for the group to which they had been assigned. Then, the effects of the three different levels of feedback on learner performance were examined by comparing the difference on practice TOEFL iBT Reading test scores administered as the pretest and the posttest across the groups.

The three feedback conditions compared in the study were the KCR condition, the rationales condition described previously, and an on-demand lecture condition. The on-demand lecture condition offered a type of instruction, which was conceptualized as the most extensive feedback and thus supposedly the most effective feedback condition among the three. As described in more detail in the "Materials" section that follows, this condition offered an on-demand video lecture focusing on reading and test-taking strategy instruction and a review of general academic vocabulary for each reading exercise along with the information available in the KCR and rationales conditions.

The three feedback conditions were selected for use in this study in consideration of a fairness issue and authenticity of the performance feedback, namely, the correspondence between the nature of feedback participants receive in this study and the characteristics of L2 reading test preparation activities with which learners typically engage in real life. First, employing the no feedback condition or the KR condition mentioned in the review previously was deemed to raise fairness issues among participants in this study because many of them had actually been planning to take the TOEFL iBT test in the near future. Although the feedback provided to all groups was made available to all participants after study completion, there was a possibility that waiting until then did not leave enough time for them to review the materials before taking an operational TOEFL iBT test. Second, comparing the KCR condition and the rationales condition allowed an exploration of the important issue about the lack of diagnostic information in many test preparation textbooks noted in the literature review, which is the effects of the availability of diagnostic information on learner performance. On the one hand, the KCR condition resembles a situation where a learner uses a course textbook that offers only correct answers to exercise items. On the other hand, the rationales condition resembles a situation where a learner uses a course textbook that includes not only correct answers but also explanations as to why specific answers are correct or incorrect. Third, the on-demand video condition reflects a situation where a learner attends a test preparation course. While some teacher–learner interaction is expected in a face-to-face test preparation course, it is quite possible that such a course is offered in a monologic lecture style, particularly in a course in which a large number of learners are enrolled. In addition, with the advancement of long-distance learning technologies, watching on-demand lectures through a learning management system (LMS) is becoming a common activity for learners of English today. Among these three feedback conditions, the KCR condition was treated as the control condition because (a) it offered the least amount of feedback among the three conditions and (b) it reflected a condition resembling the level of feedback offered in a typical course textbook without diagnostic information.

In addition, as part of a larger study, data for another feedback condition—one-on-one tutoring—were collected from a small group of participants ($N = 13$). However, results of this feedback condition will not be discussed in this report because directly comparing this tutoring condition against the KCR, rationales, and on-demand lecture feedback conditions was not feasible due to the tutoring group's small sample size and the assignment of participants to this condition by self-selection.

In addition to the level of performance feedback provided to participants during reading practice, other learner variables, such as their previous knowledge and experience with the TOEFL test and involvement in other TOEFL iBT Reading test preparation activities during study participation, were hypothesized to affect their practice TOEFL iBT Reading test scores. Examining the effects of such learner variables was important, particularly because a vast amount of information for test preparation is available to learners, and study participants were expected to come with different degrees of exposure to such information and previous test-taking experience. Accordingly, two specific research questions were addressed in this study:

1. Do learners' practice TOEFL iBT Reading test scores improve as the extensiveness of the information provided as feedback increases from the knowledge of correct answer (KCR) condition to the rationales condition and the on-demand lecture condition? Do any other learner variables predict scores independently or by interacting with the feedback level as well?
2. Does learners' perception of the usefulness of study materials differ depending on the extensiveness of the information provided to them as feedback?

Methods

Participants

Participants in this study were Japanese learners of English recruited from all over Japan. English learners preparing to take the TOEFL test or those interested in it were eligible. However, only those who had no previous experience using the *Official Guide to the TOEFL iBT Test, Third Edition* (Educational Testing Service [ETS], 2009; henceforth, the guide) were recruited. This criterion was deemed necessary because part of the reading materials employed in this study involved those that appeared in the guide. The guide provides rationales for a portion of the reading materials. Although no information was collected as to whether learners who had reported using it actually remembered the content, previous exposure to that information was considered to undermine the study design because the amount of performance feedback provided to participants had to be controlled carefully. Another reason to adopt this criterion was that having learners work on materials to which they had been previously exposed could negatively affect their motivation to focus on the materials for this study.

Recruitment announcements were made through an ETS mailing list of TOEFL test takers and at TOEFL seminars held in different locations in Japan. Recruitment flyers were placed for examinee pickup at selected TOEFL iBT test centers in different cities. The announcements were made in various English classes at eight universities as well. As a result of these recruitment activities, a total of 336 participants initiated the study, while only 199 of them (59.2%) completed it. Some attrition of participants was anticipated because this study required participants to self-study materials online for weeks. However, a combination of logistical and technical issues as well as problems for other participants to complete the entire study in a timely manner contributed to the attrition of a fairly large number of participants. The mean scaled scores for the TOEFL iBT Reading section conducted as the pretest were 13.91 ($SD = 9.00$; $n = 198$)[1] for the 199 participants who completed the study and 11.02 ($SD = 8.48$; $n = 126$)[2] for the 137 participants who did not. An independent-sample t-test indicated that this mean difference was statistically significant ($t = 2.89$; $df = 322$, $p < .05$), suggesting that, on average, reading ability as measured by the TOEFL iBT Reading section for those who completed the study was significantly higher than for those who did not. Detailed background variables were not available for those participants who dropped the study because a participant survey was conducted at the end of the study. However, participants who completed the study and those who did not were compared on age and occupation, the information regarding which was available on the project application form. Results showed that these two participant groups were similar to each other regarding these two variables, where about half were undergraduate students and another 30% were employed workers with the largest group being participants in their 20s (see Appendix A for further details.)

Meanwhile, all participants who completed the study ($N = 199$) responded to the participant survey, and, thus, their background is briefly summarized. Among them, 92 participants (46.2%) were males and 104 (52.3%) were females, with three participants' responses missing. The median age was 22 with a range of 16–64. Roughly half (101 participants; 50.8%) were in their 20s, while those in their teens (57 participants; 28.6%) and in their 30s (28 participants; 14.1%) were the next largest groups. In terms of occupation, the two largest groups were undergraduate students (93 participants; 46.7%) and employed workers (67 participants; 33.7%), followed by high school students (18 participants, 9.0%) and graduate students (14 participants; 7.0%) with data for seven participants missing. A total of 106 participants (53.3%) reported that they had no experience of living in English-speaking environments abroad, whereas 89 participants (44.7%) reported having had such experience. The total length of stay abroad was less than 1 year for 56 participants, from 1 year to less than 3 years for 18 participants, and 3 years or longer for 12 participants, with three participants' responses missing.

In terms of the participants' knowledge about the TOEFL test, 91 participants (45.7%) reported that they had had no or little knowledge about the TOEFL iBT test before participating in this study, whereas 100 participants (50.3%) responded that they knew the TOEFL iBT test well or very well. As for previous experience with the TOEFL test, 150 participants (75.4%) reported having taken at least one version of the test previously, with 144 of them having taken the TOEFL iBT test. Meanwhile, 46 participants (23.1%) had no previous experience of taking any version of the test. Furthermore, 87 participants (43.7%) reported having prepared for taking the TOEFL iBT test previously, whereas 110 participants (55.3%) had never done so. The most frequently reported methods to prepare for the TOEFL iBT Reading section were studying commercial TOEFL test preparation materials (71 participants), referring to information available through the Internet (29 participants), and attending TOEFL test preparation courses (27 participants). Participants used these methods alone or in combination. In total, 150 participants (75.4%) reported having plans to take the TOEFL test in the future for studying abroad (112 participants), applying for universities and graduate schools in Japan (13 participants), licensure (13 participants), job hunting (6 participants), or for other purposes (7 participants).

Materials

The study employed the TOEFL iBT Reading materials provided by ETS and various types of feedback materials developed for the purpose of this study. First, the TOEFL iBT Reading materials made available to this study by ETS were two forms of the TOEFL iBT Reading section released for research purposes and 15 TOEFL iBT Reading sets that had appeared in operational TOEFL iBT administrations previously, including those that were part of the guide (ETS, 2009). Each reading set comprised a reading text accompanied by 12–14 multiple-choice reading comprehension items. The reading exercises fully reflected the design of the actual TOEFL iBT Reading section. That is, the test items were designed to assess academic reading ability to fulfill three specific purposes described by ETS (2006): (a) understanding basic information stated in the text (basic comprehension), (b) drawing inferences (inferencing), and (c) reading to summarize main ideas and important details (reading to learn). These three aspects of academic reading ability were assessed by means of 10 different item types: five for basic comprehension (vocabulary, factual information, negative fact, sentence simplification, and pronoun reference); three for inferencing (rhetorical purpose, inferencing, and insert text); and two for reading to learn (prose summary and schematic table).

Second, various types of feedback materials were developed for a total of 18 TOEFL iBT Reading sets: the three sets included in the pretest and the 15 sets used only for the intervention (how these sets were administered will be discussed more in the "Procedure" section). Four university faculty members served as materials developers. They were all female, native speakers of Japanese with extensive English as a Foreign Language (EFL) teaching experience in Japan and advanced degrees in applied linguistics and related fields from institutions overseas. Among the 18 reading sets, two were used by the present author to prepare sample materials for discussion with the team, and each of the four members developed the two types of feedback materials, rationales for correct/incorrect answers and on-demand video lectures, on four randomly assigned reading sets. The team met in half-day monthly meetings during the 4 months of the materials development period. In initial meetings, this author provided an orientation to the general purposes and goals of the present study and steps for developing the feedback materials. In later meetings, the members had extensive discussions on the sample materials and draft feedback materials that they had developed. The content and format of the feedback materials were finalized based on the discussion results.

Rationales for Correct/Incorrect Answers

For each item in the 18 reading sets, the materials developers wrote explanations as to why an answer to a given item was correct or incorrect. First, each member developed rationales in Japanese by following general guidelines to prepare rationales for the 10 different TOEFL iBT Reading item types developed by this author. The draft rationales were reviewed by two independent reviewers. The Japanese version was reviewed by a Japanese team member, and the English translation of the draft rationales was reviewed by a native speaker of English. A small portion of the English translation of the draft rationales was reviewed by an ETS assessment development specialist as well. Based on the review results, each team member finalized the rationales she was in charge of.

On-Demand Video Lectures

Two types of videos were prepared. The first was a 20-minute video lecture for test familiarization that all participants were required to watch at the beginning of their study participation. This author developed and delivered the lecture in Japanese, which was recorded digitally for video streaming. This video provided a brief introduction to the TOEFL iBT Reading section purposes, structure, target constructs, and item types and a brief section providing general suggestions to prepare for the test. Few previous feedback studies reported how well learners understood constructs assessed in a test for which they were preparing. However, this study adopted a position that familiarizing participants with these issues is essential for facilitating test preparation. In other words, promoting learners' better understanding of how the target constructs of the TOEFL iBT Reading section and the test items relate to important reading skills required for academic work was hypothesized to help motivate students to prepare for the test.

Second, for each of the 18 reading sets, 30-minute on-demand lectures, each delivered in Japanese by this author or one of the four materials developers, was developed for use as part of the on-demand video feedback condition. The on-demand video lectures were designed with two specific goals in mind: to provide training on reading and test-taking strategies to optimize the test-taking process and to promote the learner's awareness of the need for building general academic vocabulary. The design of these videos was informed by two sources. The first was Kozulin and Garb's (2001) study reviewed previously, which argued that reading comprehension ability is a cognitive function that is generally considered difficult to improve in the short term and that short-term reading comprehension instruction should focus primarily on areas that are amenable to instruction, such as cognitive strategy use. Previous research has suggested that L2 learners can be trained on their cognitive and metacognitive strategy use, as shown by an effect size of .54 (Hedges' *g*) reported in a recent meta-analysis of L2 reading strategy training studies (A. Taylor, Stevens, & Asher, 2006).

Findings of the author's previous observation studies of Japanese English learners' process of completing TOEFL iBT Reading items were taken into consideration as well. In these studies, behaviors, such as spending too much time on the initial reading of the text with attention to every single detail and reading only parts of the text directly related to the items, were frequently observed. Both behaviors resulted in a fragmented understanding of the global text structure, which did not allow those learners to quickly locate specific information being asked for or to identify main ideas and important details. Such reading behaviors may partly be explained by the historical emphasis on *yakudoku*, the careful reading for word-by-word translation of English text into Japanese, in EFL education in Japan (e.g., Henrichsen, 1989; Hino, 1988). Thus, it was deemed important to include reading exercises that help learners foster reading skills and strategies that are more applicable to selective reading of academic text for main ideas and important details within a limited time, something they would encounter in the TOEFL test as well as in academia. In the meantime, cognitive skills and processes required for successfully performing on TOEFL iBT Reading items were explicated in recent studies of strategy use and cognitive diagnosis (Cohen & Upton, 2006; Jang, 2009; Sawaki, Kim, & Gentile, 2009). Thus, it was hypothesized that instructing students on how to read the text effectively as well as on the purpose of different items and general steps involved in responding to individual items contributes to the development of a cognitive model of task performance (Leighton & Gierl, 2007) that allows learners to arrive at a correct answer effectively.

The rationale for including a vocabulary exercise in the instructional materials in this study is based on previous studies that suggested the importance of vocabulary size in effective reading comprehension. In particular, Hu and Nation (2000) and Schmitt, Jiang, and Grabe (2011) concluded that the reader should know as much as 98% of the vocabulary items that appear in a text in order to comprehend it satisfactorily. The author's preliminary studies mentioned previously also identified the lack of general academic vocabulary that often appears in academic texts across different disciplines as a source of comprehension problems for lower level Japanese students.

The videos were designed as review materials for viewing after the study participant had completed all items in a given set and reviewed rationales for correct/incorrect answers to them. The present author and the four materials developers delivered lectures on those sets for which they were in charge of developing the lecture content. All on-demand video lectures were recorded digitally for video streaming. The on-demand lecture for each reading set comprised three sections: review text, review questions, and review vocabulary. PowerPoint slides and a handout were prepared for each set.

The first part of the lecture, review text, was devoted to a timed reading exercise that aimed to promote the learner's strategy to quickly grasp the gist of the text while keeping the paragraph structure and discourse markers that often appear in academic texts in mind. In Cohen and Upton's (2006) terms, this exercise offers practice in using reading strategies, such as considering prior knowledge of the topic by using the title, reading the whole text rapidly, and looking for markers of meaning in the passage. This part of the video started with a brief explanation of the topic and key words of the text by the instructor. Then, the learner was instructed to read an excerpt of the text within 3 minutes and refer to a handout. The handout contained parts of the text where essential information for understanding the gist of the text is often found: the title, the entire first paragraph, the first and last sentences of subsequent paragraphs, and the entire last paragraph. Additional information, such as transitional sentences including discourse markers, was also included, where appropriate. The text presented on the handout was approximately 300–350 words long, requiring the learner to read the excerpt at least at the speed of approximately 100–110 words per minute. After the 3 minutes had passed, the instructor explained the main points of the entire text by showing a schematic outline of the text and by describing how text coherence was developed through the use of discourse markers as well as repetitions and paraphrases of keywords.

The purpose of the second section of the video, review questions, was to promote the learner's awareness of the nature and purpose of each TOEFL iBT Reading item type and of the general steps involved in responding to different types of items. In this part, the instructor demonstrated steps she would take to arrive at the correct answer to selected items representing different item types, along with slides noting the purpose of each item type and what to pay attention to in the task completion process. In addition, for items comprising multiple steps, the instructor demonstrated how they could be completed in parts. For instance, a prose summary item requires identification of three options that together best complete a summary of the text. In the first step, the instructor confirmed the meaning of the lead sentence for the summary provided. Then, she demonstrated confirming each option for its accuracy in representing the text content and examining each option identified as accurate in the previous step to see whether it represented a main idea of the text. This demonstration accompanied a table on the slide that summarized the results. This approach was based on results of previous one-on-one tutoring studies in various subject areas where effective tutors often broke large tasks into smaller sequential parts that are manageable to learners (McArthur, Stasz, & Zmuidzinas, 1990) and demonstrated how a big task can be completed in steps.

The last part of the lecture, review vocabulary, focused on general academic words. Coxhead's (2000) Academic Word List (AWL) was used as the primary criterion for selecting the items to be covered in this exercise. Among the 570 word families included in AWL, 287 appeared in at least one of the 18 reading sets used in this study. For each reading set, 10–12 words listed on AWL that appeared in the text were covered. A brief summary on the use of discourse markers that appeared in the text was included as part of this section as well. Discipline-specific vocabulary (e.g., *iceberg*, *overglazing*, *enzyme*) was not covered in this exercise because many of these words were explained while discussing the text and items in the earlier parts of the lecture. In this exercise, the English words and their Japanese translations were presented in blocks on the slide. Then, the learner was instructed to match each English word with the most appropriate Japanese translation of the word. This step was followed by the instructor's presentation of the correct answers and a brief explanation about the vocabulary items. This exercise was rather short with limited cognitive demands. Thus, it was hypothesized that exposing the learner to an exercise like this one would promote his or her awareness of general academic words, albeit not necessarily being effective in building vocabulary within the short lecture time.

Table 1 Research Design

	Treatment condition		
Phase	A	B	C
1	Introductory video; pretest (Form 1)	Introductory video; pretest (Form 1)	Introductory video; pretest (Form 1)
2	Practice with KCR feedback	Practice with KCR feedback and rationales	Practice with KCR feedback, rationales, and on-demand video
3	Posttest (Form 2); participant survey	Posttest (Form 2); participant survey	Posttest (Form 2); participant survey

Note. KCR = knowledge of correct response.

Procedure

All study materials were delivered through an Internet project Web site based on an LMS called Quon Juku. As can be seen in Table 1, the study procedure had a pretest-treatment-posttest design with the three experimental conditions described previously. Group A (the KCR group for short) was the control group that received the scoring results (correctness of their responses) with correct answers, or the KCR feedback as described previously. The other two were experimental groups that received additional feedback. Group B (the rationales group) received the KCR feedback with rationales for correct/incorrect answers. Group C (the on-demand video group) watched the on-demand lecture after reviewing the KCR feedback and the rationales for correct/incorrect answers.

Each participant accessed the secure project Web site from convenient locations to complete the study materials. Participants were instructed to complete all study materials within 2 months. In Phase 1, all participants first viewed the 20-minute introductory video lecture for test familiarization described previously and then completed a TOEFL iBT Reading form administered as the pretest with the standard time limit of 60 minutes. Once completion of the Phase 1 materials was confirmed on the project database, the participant was instructed to move on to Phase 2 within half a month from completion of the pretest.

In Phase 2, participants completed the 18 TOEFL iBT Reading sets. As noted already, the first three sets were those that appeared on the pretest. By working on these sets for the second time at the beginning of Phase 2, each participant had an opportunity to retry the items and carefully review parts of the texts or items with which they had had trouble. This design was adopted to take advantage of test preparation activities that were generally assumed to be familiar to study participants in the Japanese context (i.e., reviewing test results to self-assess areas that require further study). Such strategies are likely to be used by learners of English in Japan. As noted by Kuramoto and Koizumi (2016), test items that appear on major language tests in Japan (e.g., major English language proficiency tests such as EIKEN as well as high school/university entrance examinations) are often released to the public immediately after administration for self-scoring responses (by those who have actually taken the exams) and for test familiarization and practice (by prospective test takers and their instructors). The remaining 15 sets were practice sets that study participants had not been exposed to previously.

All groups completed the sets in the same order while receiving the three different levels of feedback concerning their performance on individual reading items depending on the group to which they were assigned. The participants were instructed to complete roughly six reading sets per week, so that all the Phase 2 materials could be completed within a month. Moreover, in order to control the total length of time the participant was exposed to the study materials, all groups were instructed to spend no more than 1 hour studying each reading set. Within the 1 hour assigned to each set, the participant responded to the items in the reading set within the first 20 minutes and then submitted his or her answers online to receive scoring results immediately. The remaining time in the hour was spent to review the feedback provided to the designated group. If any time was left after reviewing the feedback, the participant was instructed to review the reading set content further by using whatever method he or she would normally use. In the participant survey, a majority reported looking up unknown words in the dictionary (138 participants), re-reading the text (120 participants), and/or going over the items (105 participants) for at least six of the 18 reading sets. In contrast, relatively few students reported referring to any other reference materials (14 participants). Other activities reported by a few students were reading the text aloud and making their own vocabulary lists for future use.

Upon confirmation of completing Phase 2 on the project database, the participants were instructed to complete Phase 3 within half a month from when they had completed Phase 2. Participants took a second TOEFL iBT Reading form with a 60-minute time limit as the posttest and then completed the participant survey. The survey included items concerning participant background variables, how the participant spent time while working on the study materials individually, and the perceived usefulness of the study materials for TOEFL test preparation and the development of more general English reading skills.

A detailed analysis of participants' responses to the items on time spent on the study materials and duration of their study participation recorded on the project database showed, however, that the actual time spent on Phase 2 exercises and the duration of study participation varied greatly across participants. For this reason, a decision was made to take these two timing variables into account in subsequent analyses, as described in more detail in the following section.

Analysis

SPSS Version 18.0.0 was employed for all analyses conducted for this study. The raw scores for the pretest and the posttest were converted to the TOEFL iBT Reading scaled scores (0–30) first. The Cronbach's α values were .81 for both occasions, suggesting a satisfactory level of internal consistency reliability for the present sample. Various types of preliminary analyses conducted for the pretest and posttest scores and for participants' responses to the individual survey items and timing data associated with the completion of various study materials as well as main analyses conducted to address the research questions are described in detail in the following sections.

Internal Validity of the Data

As noted previously in the "Procedure" section, a close inspection of the timing data for the 199 participants who completed the study revealed a few important issues concerning the internal validity of the data obtained. First, two participants did not complete the study materials in the order specified. Thus, it was decided to exclude these cases from subsequent analyses. Second, while participants were instructed to complete all study materials within 2 months, only 112 (56.9%) of the remaining 197 participants who completed the entire study were able to meet the time requirement. The other participants took longer, as long as up to 139 days, to complete the study. Because excluding these cases will result in a loss of a large proportion of the data, and also because the large variability in study participation duration might have affected study results, a decision was made to include days in the study as a predictor variable in subsequent analyses.

Finally, while the direction given to the participant at the beginning of each reading set was to spend 1 hour for each set, the average time actually spent on a reading set varied across participants. Because accurate timing data for completing each reading set were not available from the LMS employed in this study, this information was obtained in the participant survey by having each participant rate the average time spent per reading set based on an 8-point rating scale: 1 = *15 minutes or less*; 2 = *16–25 minutes*; 3 = *26–35 minutes*; 4 = *36–45 minutes*; 5 = *46–55 minutes*; 6 = *56–65 minutes*; 7 = *66–75 minutes*; and 8 = *76 minutes or more*. Participants' responses to this question varied greatly while the distribution of the responses across the scale and the mean time spent on each set were similar across the three groups (see Appendix B). A majority of participants spent roughly 25–65 minutes. This would allow the participant reasonable time to complete the reading set, check the correctness of his or her responses, and review the feedback materials. Participants who spent a relatively short time on each set might have completed the reading items fairly quickly, reviewed the performance feedback only selectively, or both. However, additional data that would have allowed a further exploration of this issue were not collected. Meanwhile, a small number of participants reported spending an extremely short (15 minutes or less; 6 participants) or long (76 minutes or more; 3 participants) time per set. While this greatly deviated from the suggested time, a decision was made to retain these cases in further analyses because they were not flagged as outliers given the fairly large variability of this variable.

Effects of the Attrition of Participants on a Random Assignment

Participants were assigned randomly to Groups A, B, and C when they signed up for the study. Accordingly, the attrition of a fairly large number of the participants can be a concern because it may affect the comparability of the groups established by the random assignment. Results of an ANOVA with the pretest score as the dependent variable and group as the independent variable showed that the pretest score means were not statistically significantly different across the groups ($F = 1.89$, $df_{between} = 2$, $df_{within} = 193$, $p > .05$). Moreover, frequency distributions were examined for various background variables reported in the participant survey. Results showed no statistically significant differences across the experimental

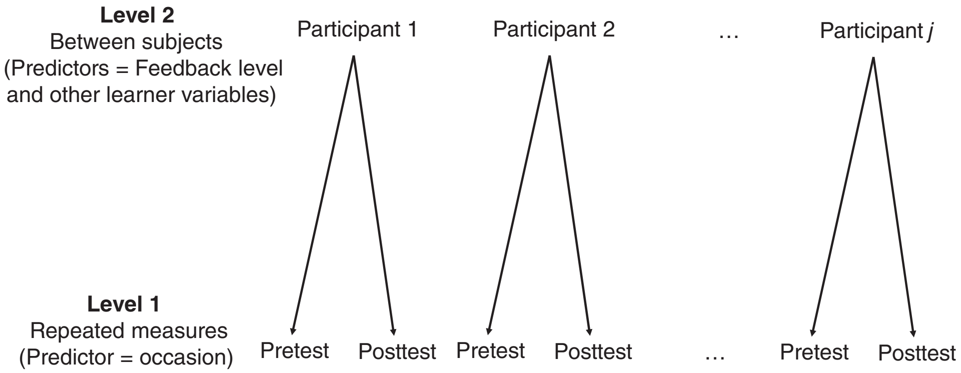

Figure 1 Graphical representation of the two-level multilevel model.

groups in terms of any of the following background variables: participants' age, gender, occupation, previous experience living in English-speaking environments abroad, previous knowledge about the TOEFL iBT test, previous experience of taking the TOEFL test, future plans for taking the TOEFL test, and involvement in other test preparation activities during study participation. To sum up, although the participant attrition rate was high, the three groups were comparable at least in terms of the pretest score and the background variables noted here.

Analyses to Address Research Question 1

Multilevel modeling was conducted to address Research Question 1. A common analytic approach employed in a pretest-posttest design for comparing means across multiple groups, as in this study, is an ANCOVA with the posttest score as the dependent variable, the group as the between-subject independent variable, and the pretest score as the covariate. An advantage of using multilevel modeling over ANCOVA is that multilevel modeling does not impose the homogeneity of regression slopes assumption, which is often difficult to satisfy with real data (Tabachnick & Fidell, 2007, p. 781). Satisfying this assumption was not an issue with the present data because the regression slopes were equal across groups ($p = .055$; $df = 2$). However, notable advantages of using multilevel modeling were an explicit way in which nested structures could be modeled along with predictor variables collected at different levels and the availability of overall model fit indices based on the maximum likelihood estimation that allowed systematic statistical comparisons among alternative models. Another advantage of this approach was the fact that multilevel modeling allowed analyzing the degree to which additional variables (the two timing variables of concern described previously and some learner background variables discussed later) could serve as reasonable predictors of the dependent variable (practice TOEFL iBT Reading test score) in addition to the experimental condition without aggregating the data at the group level. The initially hypothesized multilevel model was a two-level regression model with repeated measures as shown in Figure 1.

As shown in Figure 1, at Level 1, the scaled TOEFL iBT Reading test score (0–30) was specified as the dependent variable with occasion (pretest vs. posttest) as a repeated-measures unit. Occasion was fixed and was coded dichotomously (0 = pretest; 1 = posttest). At Level 2, participants were modeled as a random effect as they can be considered as a sample of a large number of the Japanese TOEFL iBT population. These participants were nested within groups representing the different levels of feedback, which was modeled as a categorical variable with three levels (1 = Group A; 2 = Group B; 3 = Group C). Because the definitions of the feedback conditions were unique to this study, it was not intended to generalize the results of this study to other feedback conditions. Thus, group was modeled as a fixed predictor at Level 2 rather than as a random effect at the third level. Also considered for inclusion as Level 2 predictors were the two timing variables (discussed previously) and the three learner background variables (discussed later) that were deemed to be substantively related to practice the TOEFL iBT Reading test score:

- Average time spent per reading set in Phase 2: the 8-point scale described previously.
- Days in the study: the number of days with the pretest completion date as the first day and the posttest completion date as the last day.
- Previous knowledge about the TOEFL iBT test: a 4-point scale (1 = *hardly knew*; 2 = *knew a little*; 3 = *knew well*; 4 = *knew very well*).

- Previous experience of taking the TOEFL test: coded dichotomously (0 = *never took it before*; 1 = *took it before*).
- Involvement in other TOEFL iBT Reading test preparation activities during study participation: the sum of ratings on five types of activities, each on a 6-point scale (0 = *0 hour*; 1 = *1–5 hours*; 2 = *6–10 hours*; 3 = *11–15 hours*; 4 = *16–20 hours*; 5 = *21 hours or more*)
    - self-study using commercially available reference books
    - online search for information about the TOEFL iBT test via Web sites and blogs
    - self-study by using TOEFL iBT correspondence course materials
    - attending TOEFL iBT preparation courses at an English language school or preparatory school
    - other

In addition to the five variables, length of stay in English-speaking environments abroad was another Level 2 variable of interest. However, it was decided to exclude this variable from further analyses due to the presence of a large number of missing scores on this variable.

Prior to the multilevel modeling analysis, a series of preliminary analyses were conducted on all variables considered for inclusion in the multilevel modeling: pretest score, posttest score, previous knowledge about the TOEFL iBT test, previous experience of taking the TOEFL test, other TOEFL iBT Reading test preparation activities engaged with during study participation, the average time spent across reading sets in Phase 2, and days in the study ($N = 197$). An inspection of missing scores on these variables showed that survey data, including the variables to be modeled in the multilevel modeling, were mostly or completely missing for two cases. Thus, these cases were excluded listwise from subsequent analyses. Meanwhile, the following data were missing for the remaining 195 cases: one score each for pretest score and average time spent across reading sets in Phase 2, six scores for previous knowledge about the TOEFL iBT test, and seven scores for other TOEFL iBT Reading test preparation activities engaged with during study participation. These cases were retained by imputing each missing score for the mean of the relevant variable. In addition, some cases had extremely low pretest and posttest scaled scores that were zero or near zero. Although there is a possibility that these extremely low scores suggest the associated participants' lack of motivation, these cases were retained in subsequent analyses for two reasons. First, a scaled score of zero was not equivalent to the raw score of zero, suggesting that even participants who had answered some items correctly earned those extremely low scaled scores. Second, it was deemed possible to observe such low performance on the TOEFL iBT Reading section in the present study sample, which included participants whose reading comprehension level had not yet reached the reading ability level of the TOEFL test-taker population.

Next, univariate and multivariate outliers and score distributions were inspected on the resulting dataset. First, an inspection of the Mahalanobis distance with all variables considered for inclusion as Level 2 predictors identified two multivariate outliers. These two cases were also identified as univariate outliers on other TOEFL iBT Reading test preparation activities engaged, during study participation, with extremely long hours of other test preparation activities. For this reason, these two cases were excluded from subsequent analyses ($N = 193$). After removing these two cases, the univariate normality of the score distributions was inspected for all variables except the dichotomous variable, previous experience of taking the TOEFL test. The histograms and the skewness and kurtosis values for these variables indicated that all variables were roughly normal. The only exception was a large positive kurtosis value (+4.0) for other TOEFL iBT Reading test preparation activities engaged with during study participation. Second, bivariate score distributions among the variables were inspected based on scatter plots for all pairs of variables. No noticeable trends deviating from linearity were identified.

The relationships among the pretest and posttest scores and the five predictor variables listed previously were explored further in order to identify a small set of predictor variables for inclusion in the multilevel modeling analysis. This analysis was necessary so as not to make the models too complex. First, bivariate correlation coefficients among all variables of interest were obtained to examine their interrelationships. Second, a hierarchical stepwise regression analysis was conducted with the posttest score as the dependent variable and the pretest score, group, and all the five timing and learner background variables as the predictor variables. Based on the results of these two analyses, variables that were significantly related to the TOEFL iBT Reading scaled score (the dependent variable in the multilevel modeling) and contributed to explaining its variance were selected.

The multilevel modeling was conducted in the bottom-up fashion suggested by Hox (2010). In this approach, the simplest model was tested first, followed by testing the relative goodness-of-fit of five alternative models of increased complexity:

- Model 1 (intercept-only model): This model was tested for calculating an intraclass correlation coefficient, which is the ratio of between-participant variance to the total variance in the two-level model proposed previously. A sizable intraclass correlation coefficient suggests the feasibility of continuing with multilevel model testing by including participants as the random effect at Level 2.
- Model 2 (Level 1 predictor model): This model was obtained by adding occasion to the model as the Level 1 predictor. Comparing this model against Model 1 allowed testing the significance of the fixed component of the slope for occasion.
- Model 3 (Level 1 predictor and group effect model): This model was obtained by adding to Model 2 a Level 2 predictor of primary concern: group (feedback level). Comparing this model against Model 2 made it possible to test the significance of the fixed component of the slope for group.
- Model 4 (Level 1 predictor and all Level 2 predictors model): This model was built by adding all the other Level 2 predictor variables to Model 3, along with Level 2 interactions of interest. Comparing this model against Model 3 allowed testing the significance of the fixed component of the slopes associated with the newly added Level 2 predictors.
- Model 5 (parsimonious model): This model was obtained by removing Level 2 predictors that were not statistically significant in Model 4. This model was the parsimonious model representing the relationship among the dependent variable and the fixed components of the Level 1 and Level 2 predictors. Following Hox (2010), variants of this model with random components of the slopes for the predictor variables were also tested at this stage.
- Model 6 (final model): This model was obtained by adding the cross-level interaction (occasion by group) to the most viable version of Model 5. Comparing these two models allowed testing the significance of the cross-level interaction.

For testing these six models, full maximum likelihood was employed as the model estimation method. Four overall model fit indices available in SPSS were obtained for each model: −2 log likelihood, which is equivalent to deviance (Hox, 2010, p. 47); Akaike's Information Criterion (AIC); Bozdogan's Criterion (CAIC); and Schwarz's Bayesian Criterion (BIC). To compare the relative goodness-of-fit of each model pair, a deviance test (chi-square difference test) was conducted by taking the difference between −2 log likelihood estimates for the relevant models. The values of AIC, CAIC, and BIC were not used for statistical model comparisons because no comparisons of nonnested model pairs were performed.

Analyses to Address Research Question 2

The final sample of 193 cases used for the multilevel modeling was also employed for the series of univariate ANOVAs to address Research Question 2. These analyses were conducted on the part of the participant survey containing eight items designed to elicit participants' perception of the usefulness of the study materials. All items in this section of the survey were based on the same stem, "What is your opinion about the materials for this study concerning the point below? Please choose one of the following that best applies to you." This stem was followed by the eight statements presented in Table 8, and the participant rated each on a 5-point rating scale (1 = *strongly disagree*; 2 = *disagree*; 3 = *neutral*; 4 = *agree*; 5 = *strongly agree*).

The descriptive statistics for the eight variables showed that a small number of cases included missing scores for some of the eight survey items analyzed, with the number of available data points ranging from 188 to 193 across the items. Missing scores were not imputed, and univariate ANOVAs were conducted on all available data points on the individual variables. Moreover, bivariate Spearman rank-order correlation coefficients among the eight variables were significant for 26 out of 28 coefficients, but they were mostly low to moderate, in the .20s to the .40s. The generally low correlations among the variables suggested the appropriateness of conducting separate ANOVAs rather than a MANOVA.

For the ANOVA runs, each of these items was used as the dependent variable, and the feedback condition was the independent variable. A Bonferroni correction was used to control the overall alpha level across multiple runs. Because using the conventional critical value of .05 as the overall alpha level for a Bonferroni correction here makes the statistical test too stringent, the overall alpha was set at $p = .10$, where the critical level for each run was thus set at $p = .10/8 = .0125$.

Table 2 Descriptive Statistics

Variable	Statistic	Group A	Group B	Group C	All
Continuous variables					
Pretest	*N*	62	63	67	192
	Mean	15.03	12.25	14.31	13.87
	SD	8.78	9.06	8.78	8.90
	Min	0	0	0	0
	Max	29	28	28	29
Posttest	*N*	63	63	67	193
	Mean	16.71	16.65	17.82	17.08
	SD	7.80	8.91	8.19	8.29
	Min	0	0	1	0
	Max	29	29	29	29
Previous knowledge about TOEFL iBT	*N*	60	63	64	187
	Mean	2.58	2.51	2.37	2.49
	SD	.96	.93	1.03	.98
	Min	1	1	1	1
	Max	4	4	4	4
Average time per set in Phase 2	*N*	62	63	67	192
	Mean	4.29	4.54	4.37	4.40
	SD	2.70	2.49	2.54	2.57
	Min	1	1	1	1
	Max	8	8	7	8
Involvement in other TOEFL iBT preparation activities	*N*	60	63	63	186
	Mean	1.78	2.20	2.14	2.08
	SD	2.69	3.06	2.95	2.90
	Min	0	0	0	0
	Max	14	15	12	15
Days in the study	*N*	63	63	67	193
	Mean	65.87	57.63	65.37	63.01
	SD	30.03	28.33	28.53	29.06
	Min	24	20	28	20
	Max	135	131	139	139
Categorical variable					
Previous experience of taking TOEFL	No	9	17	20	46
	Yes	54	46	47	147
	Total	63	63	67	193

Note. Min = minimum; Max = maximum.

For post-hoc analyses Tukey's honestly significant difference (HSD) tests were conducted to identify significantly different group mean difference pairs.

Results

Results on Research Question 1

Descriptive Statistics

The descriptive statistics for the final sample ($N = 193$), without imputation, are presented in Table 2. Overall, the group mean differences were small, reflecting the random assignment of participants to the three groups. The only noticeable group mean difference was seen in the pretest score, ranging from 12.25 ($SD = 9.06$) for Group B to 15.03 ($SD = 8.78$) for Group A. As noted previously, however, this group mean difference was not statistically significant. On the posttest, the group means were closer to each other, ranging from 16.65 ($SD = 8.91$) for Group B to 17.82 ($SD = 8.19$) for Group C. A comparison of the grand means between the pretest and the posttest shows that, overall, participants' practice TOEFL iBT Reading test performance improved by 3.2 points. This pretest-posttest score difference is explored further in the multilevel modeling.

Given the small differences among group means on the other variables, as noted in the "Methods" section, only the grand mean of the remaining variables will be discussed. With regard to previous knowledge about the TOEFL iBT test,

Table 3 Correlation Coefficients Among the Dependent Variable Plus Feedback Level, Timing, and Learner Background Variables

Variable	Pretest	Posttest	Know	Ave time	Other	Days	Exp
Pretest	1.00						
	(192)						
Posttest	.65**	1.00					
	(192)	(193)					
TOEFL iBT knowledge	.41**	.41**	1.00				
	(186)	(187)	(187)				
Ave time/set in Phase 2	−.15*	−.01	.01	1.00			
	(191)	(192)	(186)	(192)			
Other test prep activities	.05	−.04	.21**	.03	1.00		
	(185)	(186)	(181)	(185)	(186)		
Days in study	.09	.09	.13	−.05	.20	1.00	
	(192)	(193)	(187)	(192)	(186)	(193)	
TOEFL taking experience	.34**	.35**	.52**	.09	.21**	.30**	1.00
	(192)	(193)	(187)	(192)	(186)	(193)	(193)

Note. Ave = average; Exp = experience.
*$p < .05$; **$p < .01$.

overall, the level of participants' previous knowledge was moderate with a grand mean of 2.49 (*SD* = .98). The grand mean of 4.40 (*SD* = 1.57) for the average time spent per set in Phase 2 is between the two rating categories, 36–45 minutes and 46–55 minutes. As for involvement in other TOEFL iBT Reading test preparation activities during study participation, the grand mean of 2.08 (*SD* = 2.90) represents engagement in other activities for approximately 2–10 hours. While a large number of participants reported not being engaged in other activities at all, the presence of a small number of participants who spent considerably longer on other activities made the distribution positively skewed. The length of study participation duration (days in the study) was, on average, 63.01 days (*SD* = 29.06), although there was a fairly large variation as noted in the "Methods" section. Finally, 147 participants, or 76.2% of the cases included in the final sample, represented previous TOEFL test takers.

Relationships Among the Variables

Bivariate Pearson correlation coefficients among the variables are presented in Table 3, in which a few notable patterns were observed.[3] First, the correlation between the pretest and posttest scores was moderate (.65), suggesting some rank-ordering change among participants across the two occasions. Second, two variables, knowledge about the TOEFL iBT test and previous experience of taking the TOEFL test, were significantly correlated with the pretest and posttest scores with the correlation coefficients with the pretest and posttest scores in the .40s for the former and in the .30s for the latter. Third, these two variables were moderately correlated with each other (.52), suggesting that those who had previously taken the TOEFL test tended to report knowing more about the TOEFL iBT test. Finally, these two predictors were related to some other variables. For example, both were positively and weakly related to engagement in other TOEFL iBT Reading test preparation activities (.21 for both). As for the timing variables, the average time spent per set in Phase 2 was negatively and weakly related to the pretest score (−.15), suggesting that those who scored higher tended to spend less time on each set. In addition, those who had taken the TOEFL test before tended to spend longer to complete the study (.30), which may be related to their engagement in other TOEFL iBT Reading test preparation activities.

The relationships of the pretest and posttest scores and these learner background and timing variables were further explored in the hierarchical stepwise regression analysis conducted with the posttest score as the dependent variable. In Block 1, the pretest score and group were entered as the predictor variables, and all the other variables were entered as predictors in Block 2. As can be seen in Table 4, the adjusted *R*-square for the model with Block 1 variables was .45. When the other predictors were added to the model in Block 2, the adjusted *R*-square for the model increased to .47 with previous knowledge about the TOEFL iBT test as the only variable entering the equation as a significant predictor. This means that, when the pretest score and group had already been taken into account, previous knowledge about the TOEFL iBT test explained 2% of the total score variance on the dependent variable, while all the other timing and learner background variables did not.

Table 4 Hierarchical Stepwise Regression Analysis Result With the Posttest Score as the Dependent Variable

Model	Variable	β	*t*	*p*	Adjusted R^2
1	Constant		4.85	.00	.45
	Pretest	.68	12.19	.00	
	Group	.06	1.03	.31	
2	Constant		2.31	.02	.47
	Pretest	.61	10.21	.00	
	Group	.07	1.33	.19	
	Knowledge[a]	.16	2.63	.01	

[a] Previous knowledge about TOEFL iBT.

Table 5 Evaluation of Overall Model Fit

Statistic	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
df	3	4	6	10	7	9
Deviance	2683.87	2648.90	2647.25	2591.52	2603.12	2598.48
AIC	2689.87	2656.90	2659.25	2611.52	2617.12	2616.48
CAIC	2704.73	2676.72	2688.99	2661.07	2651.81	2661.08
BIC	2701.73	2672.72	2682.99	2651.07	2644.81	2652.08

Note. AIC = Akaike's Information Criterion; CAIC = Bozdogan's Criterion; BIC = Schwarz's Bayesian Criterion.

The results of the correlation analysis and the hierarchical regression analysis suggest that previous knowledge about the TOEFL iBT test, previous experience of taking the TOEFL test, and average time spent per set in Phase 2 were significantly related to the pretest and/or posttest scores. Given that previous knowledge about the TOEFL iBT test was the only significant predictor of the posttest score after taking account of the pretest score and group, it was possible to retain only this variable as a predictor in the multilevel modeling among the five timing and learner background variables considered. However, it was decided to use all the three variables that were significantly related to the pretest and/or posttest scores in the multilevel modeling by taking a conservative approach.

Multilevel Modeling

Tables 5 and 6 present model fit statistics and unstandardized parameter estimates along with their standard errors for all models tested for the multilevel modeling analysis, respectively. To aid the interpretability of the results, two continuous variables entered as Level 2 predictors (previous knowledge about the TOEFL iBT test and average time spent per set in Phase 2) were centered by subtracting their grand means from individual scores (grand mean centering; Hox, 2010, p. 61). Moreover, model comparisons based on deviance tests are presented in Table 7.

Parameter estimates for Model 1, the intercept-only model, were used to calculate the intraclass correlation coefficient, which was obtained as the ratio of the between-participant variance to the total variance (44.89/[44.89 + 31.06] = .591). The obtained value suggests that as much as 59.1% of the total score variance on the dependent variable was accounted for by variance across participants, suggesting that modeling participants as the random effect in Level 2 is tenable. In Model 2 the occasion variable (pretest vs. posttest) was added to Model 1 as a fixed Level 1 predictor. The significantly better fit of Model 2 over Model 1, as shown in Table 7 (deviance = 34.97; $df = 1$), suggested that the fixed component of the slope for occasion contributed significantly to the model. In Model 3, the fixed effect for group was added to Model 2 as a Level 2 predictor. However, the nonsignificant deviance test result (deviance = 1.65; $df = 2$) for the comparison of this model against Model 2 indicated that adding the fixed component of the slope for the group did not significantly improve the model fit, suggesting the lack of group mean differences on the scaled TOEFL iBT Reading test score. Although continuing the model testing while dropping this nonsignificant parameter was a possibility, it was retained in subsequent models because its effect on participants' practice TOEFL iBT Reading test performance was of primary concern in this study. In Model 4, the remaining two Level 2 predictors were added to Model 3 in order to see whether any other predictor could significantly contribute to explaining the between-participant differences on the dependent variable. The fixed components of the slopes for the two variables, previous knowledge about the TOEFL iBT test and previous experience of

Table 6 Multilevel Model Testing Result Summary (Unstandardized Model Parameter Estimates)

Effect	Parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Fixed effects	*Level 1*						
	Intercept	15.47 (.56)	13.87 (.62)	14.46 (.98)	12.94 (1.61)	14.87 (.88)	14.72 (.95)
	Occasion		3.21 (.52)	3.21 (.52)	3.21 (.52)	3.21 (.52)	3.51 (.87)
	Level 2						
	Group A			−.20 (1.36)	−1.40 (1.19)	−.93 (1.22)	−.04 (1.37)
	Group B			−1.61 (1.36)	−1.91 (1.18)	−2.08 (1.21)	−2.53 (1.36)
	Group C			0[a]	0[a]	0[a]	0[a]
	Experience				3.06 (1.62)		
	Knowledge				3.98 (1.29)	3.67 (.52)	3.67 (.52)
	Average time/set in Phase 2				−.54 (.31)		
	Knowledge × experience				−1.51 (1.45)		
	Cross−level interaction						
	Group A × occasion						−1.81 (1.25)
	Group B × occasion						.89 (1.25)
	Group C × occasion						0[a]
Random effects	Between-person variance	44.89 (6.35)	47.46 (6.29)	46.95 (6.24)	31.92 (4.75)	34.70 (5.03)	35.01 (5.02)
	Within-person variance	31.06 (3.16)	25.91 (2.64)	25.91 (2.64)	25.91 (2.63)	25.91 (2.64)	25.30 (2.58)

Note. [a]Fixed at zero due to redundancy; underlined model parameter estimates were significant ($p < .05$); figures in parentheses are standard errors.

Table 7 Model Comparisons Based on Deviance Tests

Models compared	Deviance test (*df*)	Significance
Model 1 vs. Model 2	34.97 (1)	$p < .05$
Model 2 vs. Model 3	1.65 (2)	$p > .05$
Model 3 vs. Model 4	55.73 (4)	$p < .05$
Model 4 vs. Model 5	11.60 (3)	$p < .05$
Model 5 vs. Model 6	4.64 (2)	$p > .05$

taking the TOEFL test were added. Also introduced to this model was the interaction between previous knowledge about the TOEFL iBT test and previous experience of taking the TOEFL test due to the moderate correlation observed between these two variables in the correlation analysis (see Table 3), suggesting that whether the participant had previously taken the TOEFL test affected the degree of perceived knowledge about the TOEFL iBT test. Model 4 resulted in a significantly better fit over Model 3 (deviance = 55.73; *df* = 4). However, the only significant predictor was the fixed component of the slope for previous knowledge about the TOEFL iBT test (3.98, $p < .001$; see Table 6). For this reason, Model 5, which included only the two variables of primary interest (occasion and group) and the significant learner background variable (previous knowledge about the TOEFL iBT test), was obtained as a more parsimonious model, although the fit of Model 5 was significantly worse than that of Model 4 (deviance = 11.60; *df* = 3).

At this stage, variants of Model 5 were also tested by adding random slopes for the predictors remaining in the model. As suggested by Hox (2010), the significance of the random component of the slope was entered to the model for only one variable each time. However, none of the Model 5 variants including the random slopes converged, suggesting model estimation problems. For this reason, Model 6 was built by adding the cross-level interaction (the occasion by group interaction) on Model 5 presented in Table 6 to obtain the final model. The nonsignificant deviance test (deviance = 4.62; *df* = 2) suggested that the introduction of the cross-level interaction did not significantly improve the overall model fit.

The model parameter estimates for the final model (Model 6) are presented in the last column of Table 6. The estimated fixed component of the intercept (grand mean) for the scaled TOEFL iBT Reading test score (the dependent variable) was 14.72, while the significant random component of the intercept suggests that individual participants' intercepts deviated significantly from the fixed component of the intercept. The standard deviation of the random intercept (the square root of the between-person variance, or 5.92) suggests that, on average, 95% of the intercepts for individual participants fell between the TOEFL iBT Reading scaled score of 2.90 – 26.56. Meanwhile, neither the fixed group-by-occasion interaction

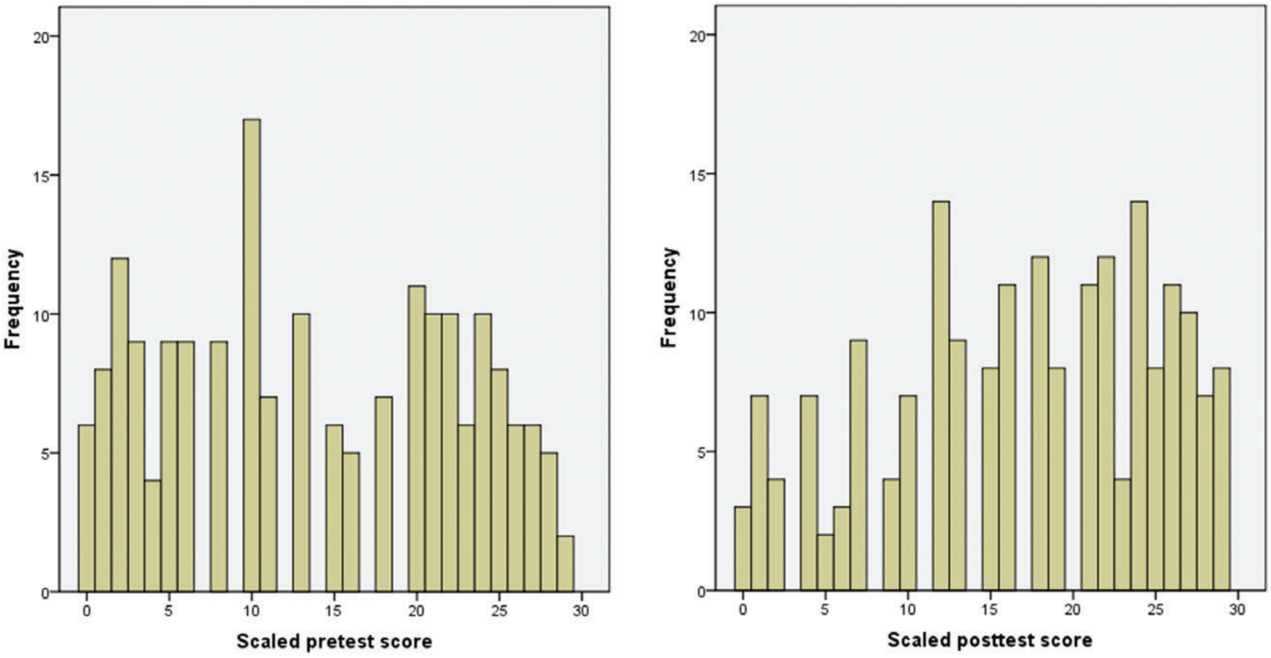

Figure 2 Pretest and posttest score distributions (TOEFL iBT Reading scaled score).

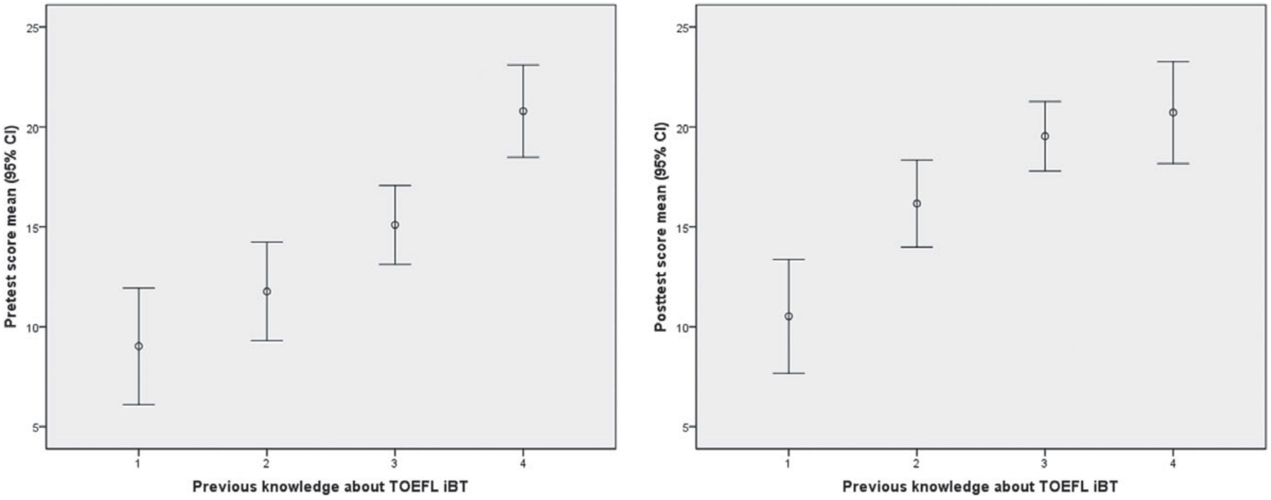

Figure 3 Pretest and posttest score means (TOEFL iBT Reading scaled score) by the level of previous knowledge about the TOEFL iBT test.

effect nor the fixed group main effect significantly predicted the between-participant variance. The only two significant predictors were occasion and previous knowledge about the TOEFL iBT test. The parameter estimate of 3.51 for occasion indicates that, on average, the practice TOEFL iBT Reading test score was 3.51 scaled score points higher on the posttest than on the pretest. Moreover, the parameter estimate of 3.67 for previous knowledge about the TOEFL iBT test suggests that, on average, the scaled TOEFL iBT Reading test score increased by 3.67 points as the rating increased by 1 point on this survey item.

The effect of the two significant Level 2 predictors of score variation among participants is illustrated in Figures 2 and 3. Figure 2 shows the difference in the practice TOEFL iBT Reading test score distribution between the pretest and the posttest. The pretest score distribution (the left panel) shows a peak oriented toward the lower end of the scale. In contrast, the peak of the posttest score distribution (the right panel) shifted upward, suggesting the increase of higher scores in the middle range, despite the wide overall score distribution across the scale on both occasions.

Table 8 Descriptive Statistics on Perceived Usefulness of Study Materials

			Group				
	Survey item	Statistic	A	B	C	All	*p*
1	I was able to understand abilities being assessed in the TOEFL iBT Reading section and its structure.	Mean	3.99	4.16	4.30	4.15	.063
		SD	.84	.77	.65	.76	
		N	62	63	67	192	
2	I became able to pace myself effectively.	Mean	3.37	3.51	3.33	3.40	.593
		SD	.96	1.05	1.07	1.02	
		N	63	63	66	192	
3	I was able to understand different types of items that appear in the TOEFL iBT Reading section.	Mean	4.11	4.32	4.57	4.34	.000
		SD	.76	.59	.53	.66	
		N	63	62	65	190	
4	I became able to understand the gist of the text within a short time.	Mean	3.06	3.11	3.63	3.27	.002
		SD	.93	1.07	.94	1.01	
		N	63	62	65	190	
5	I became able to understand how to approach different types of items.	Mean	3.71	3.83	4.05	3.86	.047
		SD	.71	.85	.74	.78	
		N	63	63	65	191	
6	The materials were helpful for me to build vocabulary required for taking the TOEFL iBT Reading section.	Mean	3.60	3.58	3.42	3.53	.569
		SD	1.00	1.03	1.04	1.02	
		N	62	62	64	188	
7	I was able to learn skills that lead to improvement of my TOEFL iBT Reading section score.	Mean	3.46	3.70	3.79	3.65	.120
		SD	.91	.91	.95	.93	
		N	63	63	66	192	
8	I was able to learn skills that are helpful for me to develop academic reading ability in the long run.	Mean	3.48	3.56	3.81	3.62	.077
		SD	.88	.93	.78	.87	
		N	63	63	67	193	

Figure 3 includes plots of the mean practice TOEFL iBT Reading test score by the level of previous knowledge about the TOEFL iBT test for the pretest and the posttest. The trend seen on the pretest data (the left panel) is a linear, positive relationship between the two variables, suggesting that those who had more knowledge about the TOEFL iBT test tended to score higher on the pretest. The monotonous pattern of improvement of practice TOEFL iBT Reading test performance is clear on the posttest (the right panel) as well, but one noticeable trend observed is the slightly larger mean TOEFL iBT Reading score increase for those who responded that they had previously known about the TOEFL iBT test a little or knew about it well (the two middle response categories) than for those who responded that they had previously had no knowledge about the TOEFL iBT test or knew it very well (the two extreme response categories).

Results on Research Question 2

Participants' mean ratings on the perceived usefulness of the study materials on eight different survey items are presented in Table 8, along with probability values for the *F* tests from the ANOVA runs. Key findings can be summarized in terms of 3 points. First, as for the results for all examinees, the mean ratings were above 3.0, suggesting that the participants' perception of the usefulness of the study materials was in the direction of neutral to positive on all items. The mean ratings were high, across groups, on two of the eight items in particular: "I was able to understand the abilities being assessed in the TOEFL iBT Reading section and its structure (Item 1)" and "I was able to understand the different types of items that appear in the TOEFL iBT Reading section (Item 3)." The grand means for both items exceeded 4.0, suggesting the participants' agreement with these items. Second, except for two items ("I became able to pace myself effectively" [Item 2] and "The materials were helpful for me to build vocabulary required for taking the TOEFL iBT Reading section" [Item 6]), the pattern of the mean score increase from Group A to Group C was in the expected direction, corresponding to the increase in the extensiveness of the performance feedback across the groups.

Results of the one-way ANOVAs with the ratings on the given item as the dependent variable and experimental condition as the independent variable showed that the group mean differences were significant for two items with medium effect sizes: Item 3 ("I was able to understand the different types of items that appear in the TOEFL iBT Reading section," $F = 8.32$, $df = 2$, $p < .0125$, partial $\eta^2 = .08$) and Item 4 ("I became able to understand the gist of the text within a short time," $F = 6.34$, $df = 2$, $p < .0125$, partial $\eta^2 = .07$). Post-hoc Tukey's HSD tests on the mean ratings were conducted for Items 3 and 4. First, on Item 3, the mean for Group C was significantly higher than that of the control group (Group A), suggesting that the on-demand video lecture condition was perceived as more useful for understanding different types of items than the control condition. Second, on Item 4, the mean for Group C was significantly higher than that of Group B, which was, in turn, significantly higher than that of the control group. This result suggests that the perceived usefulness of the study material for understanding the gist of the text within a limited time increased with the level of feedback, whereas the on-demand lecture condition was perceived as the most useful, followed by the rationales for correct answers condition, then by the control condition.

In addition to Items 3 and 4, relatively notable increases in the group means corresponding to the increase in the intensity of feedback were observed for understanding how to approach different types of items (Item 5). However, these mean differences did not reach statistical significance when a Bonferroni correction was applied to the *F* test.

Discussion and Conclusion

The present study examined the effects of three different levels of feedback, along with a set of learner variables, on the TOEFL iBT Reading practice test score of Japanese learners of English. Their perception of the usefulness of the study materials was compared across the performance feedback conditions as well. All participants in this study completed the same reading practice in the same format and content as the TOEFL iBT Reading section, while receiving different levels of performance feedback after completing the exercises. The three levels of feedback examined were KCR (the control condition), rationales for correct/incorrect answers, and on-demand video lectures. Results are discussed in relation to each research question in the following sections.

Research Question 1

> Do learners' practice TOEFL iBT Reading test scores improve as the extensiveness of the information provided as feedback increases from the KCR condition to the rationales condition and the on-demand lecture condition? Do any other learner variables predict learners' scores independently or by interacting with the feedback level as well?

The final model adopted in the multilevel modeling analysis conducted in this study showed, first, that the estimated overall mean of the TOEFL iBT Reading score for the present sample was 14.72, although the mean across the pretest and the posttest varied significantly among participants. The overall score increase from the pretest to the posttest was estimated to be 3.51 scaled score points, a statistically significant improvement. Second, the level of feedback participants received during reading practice did not make any difference to the TOEFL iBT Reading test score, while previous knowledge about the TOEFL iBT test was identified as a significant predictor that positively influenced the TOEFL iBT Reading test score. Three specific issues are worthy of further discussion in relation to these findings.

Among the key results, first, the overall score improvement of over 3 points between the pretest and the posttest regardless of the feedback level deserves further explanation. In terms of the reading skill levels reported in the TOEFL iBT Examinee Score Report, where the scaled score is divided into low (0–14), intermediate (15–21), and high (22–30) skill levels, the pretest grand mean was in the low level, whereas the posttest grand mean was in the intermediate level.[4] Although this 3-point improvement on the 30-point scale is notable, it may not be dramatic. In this sense, this result seems consistent with the modest effect of SAT preparation found in Powers's (1993) meta-analysis cited in the introduction.

However, this overall score gain should not be interpreted as a positive effect of the learners' engagement in this study because the present study did not include a control group that did not receive any treatment, namely, any reading exercises whatsoever. Thus, the present study design did not allow a statistical test of the effects of learners engaging in the reading exercises in this study per se on their TOEFL iBT Reading test performance. A previous meta-analysis on test preparation for cognitive assessment by Kulik, Kulik, and Bangert (1984) suggested that the mere exposure to practice tests that are parallel to the actual test can lead to score improvement. Although the reading exercises participants completed in this study were not administered in the standard format, it may be the case that their continued exposure to the test-like exercises might have contributed to this result. Thus, further research is required to explore to what factors—maturity, exposure to reading exercises that are highly similar in format to the actual test, or others—the overall score improvement observed in the posttest can be attributed.

Second, at first glance, the nonsignificant effect of the level of feedback on TOEFL iBT Reading test performance seems to contradict findings of previous feedback studies. This finding is not necessarily the case, however. A key difference between the previous studies and the present study lies in how the control group was defined. On the one hand, previous feedback studies designated a group that received no feedback (not even scoring results, namely, correctness of learner responses) as the control group. Thus, when the performance of the control group was compared with that of the other groups receiving more feedback, statistical differences were detected, resulting in a situation where some feedback was better than no feedback. On the other hand, because quite a few study participants were planning to take the TOEFL test in the near future and not providing any feedback in a specific group would have caused fairness concerns, this study designated the group receiving information about the correctness of learner responses along with answer keys (the KCR condition) as the control group. This made it more difficult to detect group differences in this study. Moreover, previous findings about the relative effectiveness of different types of feedback (i.e., KCR and more elaborate feedback) are inconsistent. Hence, taking these issues into consideration, the results of this study may not be surprising.

Meanwhile, the observed significant effect of learners' previous knowledge about the TOEFL iBT test on the TOEFL iBT Reading test score is worth noting. This variable was also found to be significantly related to other variables such as previous experience of taking the TOEFL test and involvement in other TOEFL iBT Reading test preparation activities. Collectively, these results suggest two things. First, those who were involved in other test preparation activities and had taken the TOEFL test tended to report knowing more about the TOEFL iBT test. Second, accumulating such knowledge about the test may potentially be more important than the level of performance feedback learners receive during reading practice when effects on their TOEFL iBT Reading score gain are concerned.

Research Question 2

> Do learners' perceptions of the usefulness of study materials differ depending on the extensiveness of the information provided to them as feedback?

This research question was addressed by conducting ANOVAs to examine the effects of the level of performance feedback provided to participants after reading practice on their responses to survey items on their perception of the usefulness of the study materials. These analyses identified some features of the materials employed in this study that study participants perceived as useful. First, participants' endorsement of the usefulness of the study materials for understanding abilities being assessed in the TOEFL iBT Reading section (Item 1 on the survey) and different types of items that appear on the test (Item 3) was generally high across groups. This seems to suggest that, regardless of the feedback participants received across the different feedback conditions, their exposure to the on-demand video lecture for test familiarization before the pretest as well as the experience of working on a total of 21 TOEFL iBT Reading sets throughout the study (i.e., 18 sets included in the pretest and the reading exercises and 3 sets in the posttest) might have provided enough information for them to feel that they had sufficient understanding of what was being tested. Furthermore, a significant group difference was found among the generally high group means on the perceived usefulness of the study materials for understanding different item types (Item 3). The significantly higher mean rating for the on-demand lecture group than for the control group indicates that receiving the video-based instruction provided with scoring results (correctness of learner responses) and rationales for correct/incorrect answers enhanced learners' perceived understanding of the item types over and above what they could gain from their exposure to the test familiarization video and the series of reading sets with minimal feedback (correctness of their responses). This may be because the lecturer of each on-demand video grouped items in a set by type and discussed the purpose of selected item types, followed by a demonstration of effective ways to approach some sample items in the review items section of the video lecture.

In contrast, participants' mean ratings were relatively lower for other survey items. In particular, those for being able to keep pace effectively (Item 2) and understanding the gist of the text (Item 4) were moderate, indicating that participants did not feel that the study materials helped on these aspects of their performance. It is worth noting, however, that the on-demand video group's mean rating on the perceived improvement in understanding the gist of the text within a short time was higher than that of the rationales group, which was, in turn, higher than that of the control group. This result seems to suggest that it may be helpful to receive more extensive feedback on this relatively difficult area. In other words, focused instruction providing a timed reading practice and an explicit explanation of the text organization, like that used in this study, may be effective in order for learners to feel that they have learned how to grasp the gist of the text within a limited time; even if that is not possible, providing written rationales for correct/incorrect answers could be of some help.

Finally, the generally lower ratings on Item 6 (building vocabulary required for taking the test) deserve some explanation. As described previously, the vocabulary exercises were based primarily on general academic vocabulary. While learning the basic vocabulary often used in academic text is essential for lower level students to develop their reading skills, the exercises seemed to have been too easy particularly for those participants with relatively high reading levels in the present study. This result has implications concerning the importance of developing instructional materials customized to the needs of students at different English ability levels.

To summarize the key results concerning the two research questions, it can be concluded, in general, that 18 hours of test-like reading exercises resulted in a score improvement of 3.51 scaled score points on the practice TOEFL iBT Reading test. Although previous knowledge about the TOEFL iBT test was found to be a significant predictor of the test score, no difference in the score improvement was observed across the three different feedback conditions examined in this study. Meanwhile, receiving the more extensive feedback led to learners' greater perceived understanding of different item types included in the test and how to grasp the gist of the text within a short time. However, this latter point cannot be overinterpreted given that participants' perceived usefulness of the study materials concerning these points did not lead to actual score gain on the test.

However, it is still premature to draw a conclusion that the level of performance feedback provided to learners during reading practice does not make any difference to their reading test performance for a few reasons. First, as rightly pointed out by some study participants in their survey responses, it seems reasonable to think that 18 hours of involvement in reading comprehension exercises in the different treatment conditions of the present study might simply not be long enough to induce visible change in learners' reading comprehension test performance in the first place. Second, there were three important limitations regarding the study design. The first limitation was the exclusion of participants who had previous exposure to the Official Guide (ETS, 2009), which might have affected the degree to which the study sample represented the TOEFL test-taker population in Japan. The second limitation was the attrition of a large number of participants. It turned out that the participants who completed the study were a more able group than those who did not based on the pretest score. It is also possible, for instance, that those who completed the study might have been more motivated to prepare for the TOEFL iBT Reading test than those who did not, and this difference in motivational intensity might have affected the study results in subtle ways, as pointed out by a reviewer of a previous version of this report. Fortunately, the homogeneity of the three feedback groups based on the participants' key background variables for the final data analysis sample suggested that the random assignment of participants to the different feedback conditions was not negatively affected by the attrition of participants. However, it should be kept in mind that the feedback groups might have differed systematically on other variables not examined here. The third limitation of the study design was the lack of sufficient control of two important timing variables (time spent on each reading exercise and study participation duration), which may have washed out the feedback level effect on learner test performance. In particular, more information should have been collected as to exactly how and to what extent participants paid attention to different parts of the performance feedback provided because being exposed to feedback does not necessarily mean that a learner pays attention to and processes every single piece of information. Accordingly, it is necessary for this study to be replicated in the future with a larger sample and for a longer duration with due attention paid to all the issues raised here about the study design. Such a study may reveal that the relatively more extensive feedback, which the study participants perceived to be helpful in improving certain aspects of their reading performance, may actually start to have positive effects on reading test performance.

The present results suggest some directions for how to design preparation materials for the TOEFL iBT Reading section in the future. First, test preparation materials should include sufficient information and guidance about the target constructs to be assessed. This finding supports the design of currently available published TOEFL iBT test preparation materials such as the Official Guide (ETS, 2009), which contains this type of information. Second, the degree to which more extensive forms of feedback, such as rationales for correct/incorrect answers and giving instruction on reading as done in the on-demand lecture condition in this study, are effective requires careful evaluation. Developing these materials is resource intensive, but the degree to which such elaborate feedback may lead to score improvement may be limited as far as the results of this study are concerned. On the one hand, including these types of more elaborate feedback may help boost learners' familiarity with the test content and boost their confidence on some features of their reading performance, such as how to quickly grasp the gist of the text under time pressure. On the other hand, however, such a positive effect on their perception may not necessarily lead to actual learning as evidenced by test score gain.

Another issue that future research should address is the meaningfulness of highly test-focused reading exercises like those employed in this study. The fairly narrow range of language exercises that may not necessarily be conducive to the development of general language ability of such materials has been pointed out by previous authors (e.g., Hamp-Lyons, 1998; Hilke & Wadden, 1997; Wall & Horak, 2011). This issue may also explain study participants' only moderate endorsement of statements on the survey that the study materials helped them learn skills that would lead to improvement of their TOEFL iBT Reading section score and that they were able to learn skills that would help them develop academic reading ability in the long run (Items 7 and 8; see Table 8).

Consistent with the line of argument provided by Hamp-Lyons (1998), different positions were taken by educators on the ethicality and justifiability of test preparation exercises using test-like exercises (e.g., Lai & Waltman, 2008; Mehrens & Kaminski, 1989; Popham, 1991). In particular, it is generally considered in this literature that it is problematic if test-like exercises become part of regular instruction. The previous discussions that mainly took place concerning achievement testing in grade schools in the United States may not be directly applicable to language test preparation that often takes place outside regular language curricula at school. However, discussion on this issue is surprisingly sparse in language assessment. While supplemental use of this type of reading exercise in conjunction with regular language instruction might be reasonable, it is imperative to deepen our understanding as to how best to utilize test-like reading exercises for meaningful test preparation that can contribute not only to short-term improvement of test scores but also to long-term development of general academic reading ability.

Acknowledgments

The author is thankful to the TOEFL Committee of Examiners (COE) at Educational Testing Service (ETS) for providing funding to this study. The author's special thanks go as well to many people without whom this study could never have been completed: Jeanne Malloy at ETS Assessment Development for her help with the design and development of the study materials; Motoki Adachi, Tomoko Agematsu, Yoko Asari, Kaoru Imai, Yutaka Ishii, Koichi Sasaki, Mitsuya Tanaka, and Takayuki Yagi at Waseda University as well as staff at CIEE Japan and Prometric Japan for their administrative support during various phases of the project; Tomoko Ishii, Mika Ishizuka, Dennis Koyama, Megumi Kobayashi, Kahoko Matsumoto, and Eric Setoguchi for their contribution to the materials development work; and all the Japanese learners of English who participated in this study.

Notes

1 The scaled score for one participant was missing due to database error.

2 The scaled scores for 11 participants were missing due to database error.

3 Given the nonnormal distribution of other TOEFL iBT Reading test preparation activities engaged with during study participation, Pearson correlation coefficients based on a square-root transformation of this variable and Spearman rank-order correlations based on the original variable were also examined. However, Table 3 reports Pearson correlation coefficients based on the original variable because the patterns of the relationships among the variables were similar across the methods.

4 The three reading skill levels reported in the TOEFL iBT Examinee Score Report were established by dividing the TOEFL iBT population into three equal percentiles. Thus, the low level corresponds to the lowest 33% of the examinees, the intermediate level to the middle 33%, and the high level to the highest 33%. For further details, see Garcia Gomez, Noah, Schedl, Wright, and Yolkut (2007).

References

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing, 13*(3), 280–297.

Andrews, S., Fullilove, J., & Wong, Y. (2002). Targeting washback—A case-study. *System, 30,* 207–223.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*(3), 257–279.

Bangert-Drowns, R., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213–238.

Biber, D., Nekrasova, T, & Horn, B. (2011). *The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis* (Research Report No. TOEFLiBT-14). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02241.x

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–352). New York, NY: Routledge.

Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education, 11*(1), 38–54.

Cohen, A. D., & Upton, T. A. (2006). 'I want to go back to the text': Response strategies on the reading subtest of the new TOEFL. *Language Testing, 24*(2), 209–250.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34,* 213–238.

Educational Testing Service. (2006). *Propell™ workshop for TOEFL iBT*. Princeton, NJ: Author.

Educational Testing Service. (2009). *The official guide to the TOEFL test* (3rd ed.). New York, NY: McGraw-Hill.

Fuchs, L., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53,* 617–641.

Garcia Gomez, P., Noah, A., Schedl, M., Wright, C., & Yolkut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT Reading test. *Language Testing, 24*(3), 417–444.

Hamp-Lyons, L. (1998). Ethical test preparation practice: The case of the TOEFL. *TESOL Quarterly, 32*(2), 329–337.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Henrichsen, L. E. (1989). *Diffusion of innovations in English language teaching: The ELEC effort in Japan, 1956–1968.* Westport, CT: Greenwood Press.

Hilke, R., & Wadden, P. (1997). The TOEFL and its imitators: Analyzing the TOEFL and evaluating TOEFL-prep texts. *RELC Journal, 28,* 28–52.

Hino, N. (1988). Yakudoku: Japanese dominant tradition in foreign language learning. *JALT Journal, 10*(1), 45–53.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Hu, M. H., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430.

Jang, E. E. (2009). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly, 6*(3), 210–238.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284.

Kozulin, A., & Garb, E. (2001, August). *Dynamic assessment of EFL text comprehension of at-risk students.* Paper presented at the 9th conference of the European Association for Research on Learning and Instruction, Fribourg, Switzerland.

Kulhavy, R. W., White, M. R., Topp, B. W., Chan, A. L., & Adams, A. J. (1985). Feedback complexity and corrective efficiency. *Contemporary Educational Psychology, 10,* 285–291.

Kulik, J. A., Kulik, C.-L. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21*(2), 435–447.

Kuramoto, N. & Koizumi, R. (2016). Current issues in large-scale educational assessment in Japan: Focus on national assessment of academic ability and university entrance examinations. *Assessment in Education: Principles, Policy & Practice.* https://doi.org/10.1080/0969594X.2016.1225667

Lai, E. R., & Waltman, K. (2008). Test preparation: Examining teacher perceptions and practices. *Educational Measurement: Issues and Practice, 27*(2), 28–45.

Lantolf, J. P., & Poehner, M. E. (2008). Dynamic assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed.)*:* Vol. 7. Language testing and assessment (pp. 273–284). New York, NY: Springer.

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26*(2), 3–16.

Leung, C. (2007). Dynamic assessment: Assessment *for* and *as* testing? *Language Assessment Quarterly, 4,* 257–278.

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition, 20,* 37–66.

Lyster, R., & Sato, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition, 32,* 265–302.

McArthur, D., Stasz, C., & Zmuidzinas, M. (1990). Tutoring techniques in algebra. *Cognition and Instruction, 7,* 197–224.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and Practice 8*(1), 14–22.

Murphy, P. (2007). Reading comprehension exercises online: The effects of feedback, proficiency and interaction. *Language, Learning and Technology, 11*(3), 107–129.

Nyquist, J. B. (2003). *The benefits of reconstructing feedback as a larger system of formative assessment: A meta-analysis* (Unpublished master's thesis). Vanderbilt University, Nashville, TN.

Poehner, M. E. (2007). Beyond the test: L2 dynamic assessment and the transcendence of mediated learning. *The Modern Language Journal, 91,* 323–340.

Popham, W. J. (1991). Appropriateness of teachers' test-preparation practices. *Educational Measurement: Issues and Practice, 10*(4), 12–15.

Powers, D. (1993). Coaching for the SAT: A summary of the summaries and an update. *Educational Measurement: Issues and Practice, 12,* 24–30.

Read, J., & Hayes, B. (2003). The impact of IELTS on preparation for academic study in New Zealand. *IELTS International English Language Testing System Research Reports, 4,* 153–206.

Russell, J., & Spada, N. (2006). The effectiveness of corrective feedback for the acquisition of L2 grammar. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 133–164). Amsterdam, The Netherlands: John Benjamin.

Sawaki, Y., Kim, H.-J., & Gentile, C. (2009). Q-matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly, 6*(3), 190–209.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26–43.

Shute, V. J., Hansen, E. G., & Almond, R. G. (2007). *An assessment for learning system called ACED: Designed for learning effectiveness and accessibility* (Research Report No. RR 07-26). Princeton, NJ: Educational Testing Service.

Spratt, M. (2005). Washback and the classroom: The implications for teaching and learning of studies on washback from exams. *Language Teaching Research, 9*(1), 5–29.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). New York, NY: Pearson.

Taylor, A., Stevens, J. R., & Asher, J. W. (2006). The effects of explicit reading strategy training on L2 reading comprehension: A meta-analysis. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 213–244). Amsterdam, The Netherlands: John Benjamin.

Taylor, C. A., & Angelis, P. (2008). The evolution of the TOEFL. In C. A. Chapelle, M. K. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 27–54). New York, NY: Routledge.

Wall, D. (2000). The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled? *System, 28,* 499–509.

Wall, D., & Horak, T. (2011). *The impact of changes in the TOEFL examination in the TOEFL exam on teaching in a sample of countries in Europe—Phase 3: The role of the coursebook; Phase 4: Describing change* (Research Report No. TOEFLiBT-17). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02277.x

Wang, L., Eignor, D., & Enright, M. K. (2008). A final analysis. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 259—318). New York, NY: Routledge.

Appendix A

Participants Who Completed or Dropped the Study by Occupation and Age

Variable	Completed (*N*%)	Dropped (*N*%)
Occupation		
High school student	18 (9.0)	7 (5.1)
Undergraduate student	93 (46.7)	78 (56.9)
Graduate student	14 (7.0)	9 (6.6)
Employed workers	67 (33.7)	37 (27.0)
Other or missing	7 (3.5)	5 (4.4)
Age		
16–19	57 (28.6)	29 (21.2)
20–29	101 (50.8)	62 (45.3)
30–39	28 (14.1)	20 (14.6)
40–49	5 (2.5)	1 (0.7)
50–59	3 (1.5)	0 (0.0)
60 and above	1 (0.5)	0 (0.0)
Missing	4 (2.0)	25 (18.2)
Total (*N*)	199	137

Appendix B

Frequency Data on the Average Time Spent per Set in Phase 2

Average time	Group A	Group B	Group C	All
15 min or less	2	2	2	6
16—25 min	9	5	6	20
26—35 min	9	8	11	28
36—45 min	15	15	18	48
46—55 min	11	18	12	41
56—65 min	10	11	14	35
66—75 min	4	4	5	13
76 min or more	2	1	0	3
Missing	1	0	2	3
Total	63	64	70	197

Suggested citation:

Sawaki, Y. (2017). *The effects of different levels of performance feedback on TOEFL iBT® Reading practice test performance* (TOEFL iBT Research Report No. 29). Princeton, NJ: Educational Testing Service. https://dx.doi.org/10.1002/ets2.12159

Action Editor: Don Powers

Reviewers: This report was reviewed by the Research Subcommittee of the TOEFL Committee of Examiners