Routledge
Taylor & Francis Group

Language Assessment Quarterly

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hlaq20

Maintaining access to a large-scale test of academic language proficiency during the pandemic: The launch of TOEFL iBT Home Edition

Spiros Papageorgiou & Venessa F. Manna

To cite this article: Spiros Papageorgiou & Venessa F. Manna (2021) Maintaining access to a large-scale test of academic language proficiency during the pandemic: The launch of TOEFL iBT Home Edition, Language Assessment Quarterly, 18:1, 36-41, DOI: 10.1080/15434303.2020.1864376

To link to this article: https://doi.org/10.1080/15434303.2020.1864376

LANGUAGE ASSESSMENT QUARTERLY
2021, VOL. 18, NO. 1, 36–41
https://doi.org/10.1080/15434303.2020.1864376

Routledge
Taylor & Francis Group

Maintaining access to a large-scale test of academic language proficiency during the pandemic: The launch of TOEFL iBT Home Edition

Spiros Papageorgiou and Venessa F. Manna

Educational Testing Service, Princeton, New Jersey, USA

ABSTRACT

The TOEFL iBT test was introduced in 2005 to better reflect the language demands of real-life academic tasks than did previous versions of the test. The task-based design of the test was intended to support the interpretation of its scores as a trustworthy measure of international students' ability to use English in an academic environment. Until recently, the TOEFL iBT test was exclusively administered online at secure test centers around the world. In response to the disruption caused by the pandemic in early 2020, the TOEFL iBT Home Edition was launched, which is identical to the version administered in test centers, but taken at home through the examinee's computer, in a proctored environment. We present key features of content delivery and the security of the test. We then discuss preliminary findings related to the comparability of scores across modes of delivery, and important implications for the design and score interpretation of at-home language proficiency tests.

The TOEFL iBT test

The TOEFL iBT test was introduced in 2005 by Educational Testing Service (ETS) to better reflect the language demands of real-life academic tasks than did previous versions of the test. The purpose of the TOEFL iBT test is to evaluate the English proficiency of people whose first language is not English, and its scores are used in more than 150 countries primarily as a measure of the ability of international students to use English in an academic environment. The test content was based on extensive research of the tasks students typically perform in a university context (see Chapelle, Enright, & Jamieson, 2008). Core to the task-based design philosophy of the test were a number of key design features, including integrated speaking and writing tasks that engage multiple skills to simulate language use in academic settings, and test materials that reflect the reading, listening, speaking, and writing demands of real-world academic environments. The TOEFL iBT test contains four sections covering all four language skills mentioned above, and it takes approximately 3 hours to complete.

The domain definition inference is the starting point in the TOEFL validity argument and it is based on the warrant that observations of performances on the test reveal the knowledge, skills and abilities in situations that are representative of the target language use

CONTACT Spiros Papageorgiou spapageorgiou@ets.org Center for Language Education and Assessment Research (CLEAR), Research and Development Division, Educational Testing Service, 660 Rosedale Road, Princeton NJ, 08541, USA.

domain Chapelle (2008). To support the above warrant, test tasks are designed to simulate the types of tasks typically encountered in such institutions, based on domain analysis by applied linguistics experts. In addition, the design of the test aimed to generate positive washback by encouraging good practices in learning and teaching academic English. Research findings of positive washback in the classroom included the use of tasks that integrate the four language skills and emphasis on the development of speaking skills, which reflected the key design features of the TOEFL iBT test (Wall & Horák, 2011).

Since the launch of the TOEFL iBT test, an extensive research program has resulted so far in over 200 publications by ETS and external researchers (a reference list of all publications to date related to TOEFL research can be found at www.ets.org/toefl/research). This research program has accumulated strong empirical evidence supporting the inferences in the TOEFL validity argument laid out in Chapelle (2008).

Launch of the TOEFL iBT Home Edition

The TOEFL iBT test is computer-delivered and had been primarily administered online at secure test centers around the world. The pandemic started disrupting administrations at test centers in February 2020. In response to this disruption, ETS launched the *TOEFL iBT Special Home Edition* in March 2020 (www.ets.org/s/cv/toefl/at-home). The test became a permanent offering of the TOEFL program in November 2020, and the name was modified to *TOEFL iBT Home Edition*. In terms of content, order of presentation of test tasks and questions, and scoring, the Home Edition test is identical to the TOEFL iBT test administered in test centers and draws on the same pool of test items, but it is taken at home, and examinees use their own computer. Similar to the test version offered at test centers, the TOEFL iBT Home Edition offers accommodations for examinees who require them, such as extended time and extra breaks.

Security considerations for the TOEFL iBT Home Edition

While stakeholders such as institutions and examinees welcomed the addition of an at-home solution during the pandemic, it was critical for ETS to ensure that test scores can be trusted in the same way as scores from the test version administered at test centers. Test security measures in place for the test-center version are also followed for the at-home version, based on the 3-pronged approach of prevention, detection, and communication (www.ets.org/toefl/score-users/about/security/prevention). For example, test content is delivered using secure transmission protocols, and test forms are assigned through centrally controlled algorithms, taking into account the location of the examinees and their time zone. Scoring is also controlled centrally for both versions of the test to further support security. For example, responses to the speaking and writing tasks are evaluated by proprietary artificial intelligence (AI) scoring engines as well as certified raters, whose scores are recorded and constantly monitored for rating quality by scoring leaders through a proprietary online platform. The use of the online platform helps ensure that raters will not know the examinees whose responses are being evaluated. Scores are also reviewed and analyzed statistically to identify suspicious patterns of test responses.

In addition to the above security measures already in place for the version of the test administered in test centers, additional arrangements were made specific to the TOEFL iBT

Home Edition. Prior to test administration, examinees are required to download a secure web browser on the computer they will take the test, run a system check, and fix any issues before the test date. On the test date, test security is safeguarded throughout the session by online human proctors and artificial intelligence measures offered by Proctor U, a third party remote-proctoring company. Prior to starting the test, examinees are required to show a photo ID to their proctor and a workspace that meets several requirements (for details see https://www.ets.org/s/cv/toefl/at-home/test-day). The proctor reviews the exam rules and requests access to the computer screen for monitoring purposes. Examinees are also asked to use either a hand-held mirror or a cell phone to show the proctor the computer screen. The proctor then instructs the examinee to launch the secure browser and provides the ID and password to access the test. Throughout the test, the proctor monitors the computer screen and watches the examinee via the computer camera. Examinees may only leave their seats for a 10-minute break after they complete the reading and listening sections and are required to return on-time to continue with the remaining test sections. The proctor can communicate with the examinee and examinees can also contact the proctor during the test. In addition to synchronous video-based human proctoring of examinees, there are technological innovations for monitoring activity and settings on the examinee's computer and alerts are sent to proctors about unusual behavior or room conditions (for example, outside noises, communicating with someone other than the proctor, looking away from the screen, and moving away from the screen).

Implications for test validation of at-home testing

Mode comparability for test-center and at-home versions

The TOEFL iBT Home Edition is identical in content and scoring to the TOEFL iBT test administered in test centers. Therefore, the two versions can be considered as two different modes of delivery of the same test. Mode effects could be present because of differences in the testing environment (home v. test center), the computer equipment, and steps in setting up the test for delivery through a computer. However, to be useful for decision-making, scores across the two modes of delivery should be interchangeable, and if an examinee were able to take the test in both modes, then scores should be equivalent after accounting for measurement error. From a practical perspective, examinees' level of language proficiency might change between test administrations, even when the time interval is short, because of factors such as additional language instruction or familiarity with the test format; therefore, directly comparing test scores for the same test-takers across the two modes of delivery can be challenging. One additional challenge is that the mode effect can be confounded with differences in language proficiency between the two groups of examinees (the group that took the test at home and the group that took the test at a test center).

To address the above challenges, one way to explore the comparability of scores across modes of delivery could be to examine the distribution and reliability of their test scores. Preliminary data show that the TOEFL iBT Home Edition and the TOEFL iBT test administered in test centers have comparable reliability and standard error of measurement (for details on how reliability and SEM are calculated for the TOEFL iBT test see Educational Testing Service, 2020, pp. 6–7). Although preliminary findings offer some support for mode comparability, further analysis is planned to investigate test-taker

performance by subgroups, such as the country where the test was taken, first language, and reason for taking the test (for example, to attend undergraduate or graduate degree courses). Ongoing monitoring of test administrations and analysis of test data can also help address any future mode comparability issues by employing measures such as statistically adjusting for population differences.

Addressing fairness and privacy concerns

The TOEFL iBT test is available on multiple dates each month through a large network of test centers in more than 180 countries around the world. Examinees can generally expect to obtain a test date in a test center that meets their needs, sometimes with a little advance planning. Despite increases in test center availability in recent years, a language test that can be taken at home can be more convenient and has the potential to be financially less demanding for those students who need to travel some distance to the closest test center. At the same time, a high-quality at-home test comes with some important requirements. Examinees need to be able to take the test in a quiet room with an uncluttered desk space. Technological requirements include an up-to-date computer and a reliable, high-speed internet connection. Not all examinees may have the financial resources to meet such requirements; thus, unfairness could be introduced to the test-taking process in ways that are not common to standardized tests administered in test centers. In addition, remote proctoring of an at-home test and monitoring of an examinee's computer can raise concerns about privacy that are less likely to be the case with tests administered in test centers.

Language tests that are only delivered at home cannot accommodate examinees whose space or personal computer do not meet specific requirements; in other words, an at-home test might not necessarily be accessible to all examinees. However, in the case of the TOEFL iBT Home Edition, examinees who cannot meet space or technology requirements, or who have privacy concerns, can instead take the TOEFL iBT test at a test center. Irrespective of whether a test center is available to examinees, future work for the TOEFL iBT Home Edition will need to include analyses of data investigating issues of fairness and privacy. The goal of such analyses should be to investigate if testing at home has differential effects on the performance of different groups of examinees based on factors such as test location, quality of high-speed internet, and use of an up-to-date computer.

Balancing between the need to minimize content exposure and to adequately support the domain inference

As stated earlier, the TOEFL iBT test is delivered through secure transmission protocols, and test forms are assigned through centrally controlled algorithms, taking into account the location of the examinees and their time zone. These measures help minimize content exposure, along with creating a large pool of test tasks and offering frequent test administrations. Providers of other at-home language tests have resorted to short, discrete-point test items, arguably to minimize content exposure (for review of at-home tests, see Isbell & Kremmel, 2020). Discrete-point item types found in other at-home tests include identifying real words as opposed to pseudo-words from a list, typing sentences that have been dictated, or read aloud a sentence shown on the screen. These item types are attractive to some testing companies because their design is simple, and it is straightforward to display them

on a computer screen. Discrete-point items can be produced at low cost, in many cases automatically generated by computer programs, thus allowing the creation of large item pools.

While producing large numbers of discrete-point items could help minimize content exposure, such emphasis on discrete-point design comes with serious shortcomings for evaluating academic language proficiency. First, the lack of a task-based approach in the design of tests used to inform higher education admissions can lead to the absence of a design basis to support the interpretation of the test scores as trustworthy indicators of academic language ability (Norris, 2018). In other words, when the test design is based on short, discrete-point items that only test rudimentary English, construct underrepresentation is inevitable, as core language abilities related to performing real-life academic tasks cannot be evaluated. In the end, support for the domain inference for tests relying on discrete-point items will be weak, and the test scores generated based on performance on such items cannot be trusted for making decisions about academic language proficiency. Second, negative washback is likely to occur, because a discrete-point test design is typically associated with ineffective language learning practices, such as memorizing lists of words out of context. Given the lack of correspondence between discrete-point test design and real-life academic language use, examinees might perform well on a test consisting of discrete-point items but are unlikely to have developed the language skills and abilities they need in order to be able to cope with the demands of English language instruction on campus.

Some providers of at-home tests seem to argue that their discrete-point design taps relevant language skills and abilities because their test scores correlate moderately with the scores of established tests of academic language proficiency, such as the TOEFL iBT test. However, as Isbell and Kremmel (2020) note, it is well known that correlations between scores of two tests are not sufficient to establish interchangeability of scores. Consequently, score correlations cannot fully support the usefulness of a language proficiency test for important decisions such as admissions of international students to higher education institutions. Instead, a comprehensive validity argument needs to be developed and supported empirically, covering all aspects of test design and validation of tests scores (Chapelle, 2008).

When it comes to decisions about academic language proficiency, the shortcomings of tests that primarily rely on discrete-point design outweigh any benefits of producing large pools of items to minimize content exposure, because, simply put, these tests cannot measure important language skills and abilities in the context of academic English. In the case of the at-home version of the TOEFL iBT test, it was critical to maintain the task-based philosophy that underpins it. Delivery of the TOEFL iBT test through the computer allows for the inclusion of important dimensions of the academic context such as visual elements supporting reading passages and audio stimuli, response formats beyond multiple-choice ones to allow for the production of extended oral and written responses, and integration of language skills when responding to speaking and writing tasks. These design features, which are central to the task-based approach of the TOEFL iBT test, help support the domain inference with test tasks that resemble those in the target language use domain. Because the TOEFL iBT test was already an internet-based, computer-delivered assessment, it was possible to take swift and effective action to launch an at-home version with the extensive test security measures described earlier, without any compromise in the test design or the support for the domain inference. The security measures for TOEFL iBT Home Edition appear to be sufficiently effective, but as

with security measures for the version of the test administered at test centers, they will be monitored on an ongoing basis as more data become available for analysis.

Conclusion

The task-based design of the TOEFL IBT test was intended to support the interpretation of its scores as a trustworthy measure of international students' ability to use English in an academic environment. The internet-based, computer-delivered format of the test allowed for the launch of the TOEFL iBT Home Edition within weeks from the initial disruption the global pandemic caused to test centers. The TOEFL iBT Home Edition aims to offer an at-home solution with extensive security measures while avoiding the design limitations of at-home tests that rely on discrete-point items. Mode comparability between the two versions of the test and fairness are topics that will continue to be explored as more examinees select the TOEFL iBT Home Edition over the version of the test administered in test centers.

Acknowledgments

We thank the associate editor and the reviewers for their comments. We also thank our colleagues John Norris, Dan McCaffrey, and Brent Bridgeman for their careful review of an earlier version of the manuscript. The authors are responsible for any errors in this publication. Any opinions expressed in this article are those of the authors and not necessarily of Educational Testing Service.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Spiros Papageorgiou http://orcid.org/0000-0002-7940-3472

References

Chapelle, C. A. (2008). The TOEFL validity argument. In C. Chapelle, M. Enright, & J. Jamieson (Eds.), *Building a validity argument for the test of English as a Foreign language* (pp. 319–352). London, UK: Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the test of English as a Foreign language*. New York, NY: Routledge.

Educational Testing Service. (2020). *Reliability and comparability of TOEFL iBT® scores*. Retrieved from www.ets.org/toefl/research/insight-series

Isbell, D. R., & Kremmel, B. (2020). Test review: Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing*, *37*(4), 600–619. doi:10.1177/0265532220943483

Norris, J. M. (2018). Task-based language assessment: Aligning designs with intended uses and consequences. *JLTA Journal*, *21*, 3–20. doi:10.20622/jltajournal.21.0_3

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