Program Evaluation and Triangulation: What Three Data Sets Show about How Participants Respond to the Course *Teaching Academic English with the TOEFL iBT® Test*

Gerriet Janssen*, Renka Ohta, Jeremy Lee, Michael Suhan

ETS, Princeton, USA

Abstract

Drawing from a larger CIPP program evaluation (*Context, Input, Processes, Products*) comprised of 11 analyses of the ETS online teacher training course titled *Teaching Academic English with the TOEFL iBT® Test* (TAE), this paper triangulates three qualitatively different data sets and seeks to understand how participants responded to the TAE course. The research team considered self-reported data (closed- and open-answer survey questions), engagement data, and learning data. The self-reported data were strongly positive and illustrated the value participants reported having about the course content, especially ETS-created materials. Some gaps were indicated: calls were made for more materials—especially additional resources participants could use when teaching—and for increased interactions during the course workshop. In partial juxtaposition, the engagement data indicated that participants were engaging the course less than anticipated. The learning data simultaneously indicated that some assessments were quite easy for participants while elsewhere participants did not have the desired uptake of some topics. Viewing the three data sets together, the research team arrived at a more nuanced conclusion: despite participant satisfaction, development processes should continue, to ensure the highest quality teacher training program possible, in terms of its content, resources, and interactions.

Keywords: *Triangulation, Program Evaluation, TESOL, Teacher Training, EAP, English for Academic Purposes, Task-Based Language Teaching, Self-Reported Data, Engagement Data, Learning Data*

Introduction

TAE

ETS offers teacher development opportunities for English language instructors worldwide in accordance with its mission to support and advance English language teaching and learning.

As part of that effort, *Teaching Academic English with the TOEFL iBT Test* (TAE) was developed to provide a short, online teacher training course that presents in-service teachers with the best practices for the instruction of English for Academic Purposes (EAP). Developed by ETS researchers John Norris and John Davis—and based heavily within the thinking developed in the volume by Norris, Davis, and Timpe-Laughlin (2017) titled *Second language educational experiences for adult learners*—TAE focuses on the pedagogical principles and practices that are instrumental when teaching English for academic purposes in higher education (i.e., universities). As there is a strong, purposeful link between the academic language tasks that English language learners complete in university classrooms and the test tasks developed on the TOEFL iBT test (see Jamieson et al., 2008), TAE helps teachers understand how to implement a task-based language teaching (TBLT) methodology in their classrooms, developing the types of academic tasks English assessed on the TOEFL iBT test.

The TAE course considered in this article was the first publicly available course version. Delivered between September 2019 and December 2022 in an asynchronous, self-paced format oriented towards advanced users of English, cohorts of participants from China completed the 11 online modules in the course, following the same 10-week study plan. Module activities consisted principally of readings, discussion board questions, and quizzes. Each of the 11 course modules required approximately three hours to complete; Appendix A lists the titles of the 11 course modules.

After completing the online self-study, participants participated in a three-day workshop, which since 2020 was delivered as an online collaboration between ETS researchers and local Chinese professors specialized in EAP teaching. During the workshop, these presenters highlighted the key elements of the TAE coursework for participants and demonstrated how to incorporate these core concepts in actual lesson sequences. In each presentation, participants interacted with the workshop facilitators and other participants during small and whole group discussions, Q & A sessions, and through the chat box on the meeting platform (see Appendix B for an overview of the TAE workshop). Over the course of the workshop, assigned groups of five to eight people collaborated and developed an EAP lesson, implementing workshop strategies. Participants who completed the online study and the workshop received a certificate of completion for the TAE course.

This version of TAE had an optional third segment: a final project that participants completed after the workshop. In this project, participants independently developed a detailed EAP lesson plan for their own classroom, which received individual feedback twice from an ETS facilitator. Those who received a passing grade on the final project were awarded an additional distinction on their already-earned TAE certificate.

As part of the initial organization of the TAE course, the course developers built structures to collect data, ensuring that the TAE course would benefit from ongoing, systematic review. Brown described this type of ongoing program evaluation as *systematic curriculum development*, which "makes possible the assessment of the quality of the curriculum once it is put in place as well as the maintenance of that curriculum on an ongoing basis" (1995, p. 24), safeguarding against what he called a program becoming "inflexible." Broadly speaking, the overarching program evaluation gathered data to ensure that TAE was well-adapted to its context: the Chinese context of English language teaching and learning. This specific paper reports on the following evaluation question: How did participants respond to the TAE course,

as seen through the following three data sets: (1) self-reported data; (2) engagement data; (3) learning products? As this is the first large-scale program evaluation of the TAE course, we think that it can signal future actions to be taken by course developers. Additionally, we think that the triangulation of these three qualitatively different data sets has the potential to depict various facets of the course participants' experiences.

Review of the Literature

Educational evaluation includes foci as diverse as testing, measurement, accreditation, and program evaluation (Kellaghan et al., 2003). Within these areas, the following statement applies: "the root term in evaluation is *value*…. Essentially, evaluators assess the services of an institution, program, or person against a pertinent set of societal, institutional, program, and professional/technical values" (Stufflebeam, 2003, p. 33). *Program evaluation*, then, is a specific area of educational evaluation that uses "an applied inquiry process for collecting and synthesizing evidence that culminates in conclusions about the state of affairs, value, merit ... of a program, product, person, policy, proposal, or plan" (Fournier, 2005, p. 139). Stufflebeam and Fournier's definitions each include four axes: the entity, people, setting, or process under study; the aspects being considered; the processes used to collect and analyze data; and finally the values that are used to make interpretations about the data. Leskes and Wright order these axes into a four-step cycle (2005). First, goals, questions, and values should be posed; next, evidence should be gathered; then, findings should be analyzed; afterwards, the findings should be used to make improvements, "closing the loop" (Sylwester, 2017, p. 23), so that finally—and critically—this cycle can begin again with new and reframed goals, questions, and values. In this study, we look at participants valued the TAE course, as seen through self-reported data, platform engagement data, and learning data. Interpretations and suggestions from the analysis of these different data sets are intended to inform future iterations of the TAE course.

Situated within different worldviews (e.g., rational-positivist, post-positivist, constructivist, transformative; for an overview see House, 2003) are an ever-increasing number of different evaluation frameworks. These respond to the diverse entities, people, settings, or processes being evaluated—as well as the different values that can be applied when make interpretations (see Patton, 2012 for a comprehensive list of frameworks). Within the post-positivist worldview is Stufflebeam's CIPP evaluation, which emphasizes the importance of a program's *context*, *inputs*, *processes*, *products* (and thus the acronym CIPP). CIPP is "one of the oldest and most thoroughly tested approaches to evaluation" (House, 2003, p. 10), and CIPP evaluations generally have "improvement/ formative and accountability/summative roles" (Stufflebeam, 2003, p. 32). CIPP is the type of program evaluation that informs this study, and this program evaluation has a formative focus. CIPP program evaluations have been frequently used with medical education programs (cf. Lee et al., 2019; Lippe & Carter, 2018; Mirzazadeh et al., 2016; for a review, see Toosi et al., 2021) and also in the field of education, such as the evaluation of an English curriculum in Turkey (Karataş & Fer, 2009), a program evaluation and redesign of an online MA program (Sancar Tokmak et al., 2013), and a program evaluation of a service-learning program development, implementation, and evaluation (Zhang et al., 2011).

Many scholars have described the benefits of purposefully using mixed methods research frameworks when conducting an evaluation (see Greene, 2005, pp. 255–256). Elsewhere,

Mackey and Gass more specifically list 16 benefits of using mixed-methods frameworks (2016, pp. 278–279), and from these benefits we employ mixed-methods in order to have greater validity, to offset methodological weaknesses in one area with strengths in others, to build a more complete representation of the phenomenon under study, and to have greater credibility. As a recommended practice of any mixed-methods project, we consider findings independently, but also across data sets (Creswell & Creswell, 2018, p. 300); we do this by presenting a summative scorecard at the paper's end. This scorecard presents the *strengths*, *gaps*, and *concerns* the findings imply, with *gaps* indicating elements that could be added to the course and *concerns* referring to issues that should be addressed. Similar scorecards were presented internally as part of an in-house final report. The overarching CIPP evaluation guiding this project was developed within a convergent mixed-methods framework (Creswell & Creswell, 2018).

Triangulation is a particularly important strategy when building program evaluations, and it can be used to build the credibility, transferability, and dependability of qualitative data (Brown, 2001; Mackey & Gass, 2016). Though Brown presents seven categories of triangulation (2001, pp. 228–229), we specifically implement *theoretical triangulation* (i.e., multiple conceptual stances) and substantiate the opinions and beliefs presented in the self-reported data (open and closed survey questions), with engagement data (learning platform timestamps) and learning data (group projects). These different data sets require *methodological triangulation* (different procedures). We employ *investigator triangulation* in our analysis of all qualitative data, ensuring that multiple people arrive at the same qualitative finding. This "maximize[s] the possibility of obtaining credible findings" (Brown, 2001, p. 228).

Most CIPP evaluations employ at least some triangulation. Indeed, the CIPP evaluation conducted by Karataş & Fer (2009) employed several iterations of surveys with teachers and students to make refinements to an English course, a strategy that was similarly carried out by Sankar Tokmak et al. (2013) in their evaluation of an online MA program. Much more complex is the CIPP evaluation conducted by Zhang et al. (2011). These authors included more than 20 different data collections and processes in a multi-stage framework, with the *process* segment of their CIPP evaluation seven data collections, including pre-service teacher interviews (self-reported data), and classroom observations (ethnographic data, not included in this study), and document analysis of student work (learning data). The larger CIPP evaluation we conducted emulated Zhang et al., similarly gathering multiple different sets of data. The 11 data collections we carried out are listed in Appendix C.

Methodology

Participants

Between 2019–2022, data was collected from 548 participants from 10 public cohorts (open-enrollment from across China) and two private cohorts (teachers from a single Chinese institution). As our data collections were completed at slightly different times, the specific number of participants have been noted for each data set.

Biodata collected in the end-of-course survey (n = 540) characterize TAE participants as mostly female (81.5%) and between the ages of 26–40 (80.0%). While the online format allows participation from all over China, 66.3% were from *Tier 1* (Beijing, Shanghai, Guangzhou,

Shenzhen) or *New Tier 1* (e.g., Xi'an, Chengdu, Wuhan) level cities. These cities are described as having the highest levels of income, consumer sophistication, talent, and business opportunities in China (Yicai Global, 2017). Participants typically brought an important educational foundation to the course: 69.1% had completed an MA degree; 4.8% had completed a PhD program; the remaining 26.1% had completed a BA degree. Additionally, 38.5% reported having a certification in teaching English, (e.g., the ACT teaching certificate; CELTA certificate). In terms of teaching experience, 58.0% were what we thought of as being *experienced*, with seven or more years of teaching. Participants taught in very different teaching programs: while 42.2% taught at private language institutes, 27.6% taught in high-schools, and 28.1% taught at universities. The most frequent reasons for course enrolment related to what we coded as *improve teaching skills* (36.7%), *better understanding the TOEFL iBT test* (27.2%), and *personal growth* (11.6%).

Materials
Self-Reported Data
After completing the course, course participants responded to a compulsory exit survey (for sample survey questions, see Appendix D). This was comprised of 80 items: 13 biodata questions and 67 questions that asked participants to rate the different course elements (e.g., discussion boards, modules, workshop sessions, quizzes). While most questions were in a Likert-scale format, participants were also provided open-ended items, where they could make suggestions concerning future course versions or describe what they valued most.

Engagement Data
As a proxy measure for engagement, we considered the timestamps recorded by the TAE learning management system. These measured the time each participant had their browser open for each module. As this measure did not "time out" if participants clicked into a different browser, the timestamp can only be used as a rough measure of participants' engagement.

Learning Data
We considered two sets of learning data: assessments and lesson plans. All 11 course modules each included three knowledge check quizzes (2–3 questions each) and an end-of-module assessment (10 questions); all assessments were selected-response. We also considered the group lesson plans created during the workshop; although the optional individual lesson plans were studied for the larger program evaluation, space does not permit their consideration here.

Procedures and Analyses
Self-Reported Data
For the Likert-scale survey questions, descriptive statistics are provided (M, SD, SE); 95% confidence intervals have been placed around relevant items to help make comparisons. Open-ended questions were coded by two raters. For each question, they first trained their coding on the initial 100 lines of data, refining a pre-established code set. After reaching an acceptably high value of agreement (.90), the remaining data were rated independently; all discrepant ratings were resolved through discussion (see Trace et al., 2016). In our findings for each open-

ended question, we present the major trends from the data and their percentages of occurrence in the data set.

Engagement Data

As a proxy measure for the time spent on a module, we used the time each participant had their browser open for the module. This time measurement was classified into a corresponding 30-minute time bucket, for instance 0–30 minutes; 31–60 minutes; 61–90 minutes, and so forth. As an example, if a participant had their browser open for 27.35 minutes on Module 10, their time would be placed into the 0–30 minute time bucket for Module 10. With all time measurements classified into buckets, tallies were then made of the number of participants in each time bucket, for each module. Bar-graphs of these tallies illustrate the distribution of how much time course participants had their browser open for each individual module. Viewing these bar-graphs collectively, trends concerning time use across the entire course are depicted. As an additional illustration of the time participants spent in each module across the course, the modes for each module (i.e., the "most popular" time bucket for each module) are represented cumulatively in a separate graphic.

Learning Data

To understand the trends within the knowledge check quizzes and end-of-module assessment scores, we provide descriptive statistics (M, SD, SE) for each module. Additionally, we compare means across modules, to explore whether there are any trends in learning performance across the course (i.e., increases or decreases). The assessments went through a large reformulation after Cohort 5, based on the participant performance on the assessment items. In this reformulation, we removed or modified the items that received exceptionally low average scores across the five cohorts and added new items as necessary. In addition, we interspersed knowledge check quizzes throughout the module so that course participants could check their understanding of the content after completing a small section, rather than at the end of the module. Because of these differences, data from Cohorts 1–5 are presented separately from later cohorts.

We analyzed workshop group lesson plans from seven early cohorts (Cohorts 2–7 and one private cohort). During initial data collection, these were chosen for study because of their structural simplicity; more recent data collection has focused on individual lesson plans, which are more complex (not included here). The scoring rubric for the group lesson plans focused on five central elements in the project: assessment users, purpose, relevance, lesson objectives, and task authenticity (see Appendix E). Three raters trained and reached an interrater reliability of .90 when using the rubric. The raters all hold master's or doctoral degrees in applied linguistics, and have an average of seven years of EFL/ESL teaching experience and at least four cohorts of experience facilitating the TAE course. These dimensions of each lesson plan were rated using a dichotomous scoring (0/1). The results of coding were converted to percentages, and the average percentages of ratings were compared across different cohorts.

Results

Self-Reported Data

Likert-Scale Survey

Participants were asked to rate their satisfaction with the TAE course generally, its content, its organization, and the online self-study. For these items, the overall averages were nearly the same, ranging between 3.08–3.14 on a 0–4 Likert scale (Figure 1, top data grouping). However,

Figure 1

No Matter Which Cohort, Positive Ratings were Given to the Global TAE Course Elements
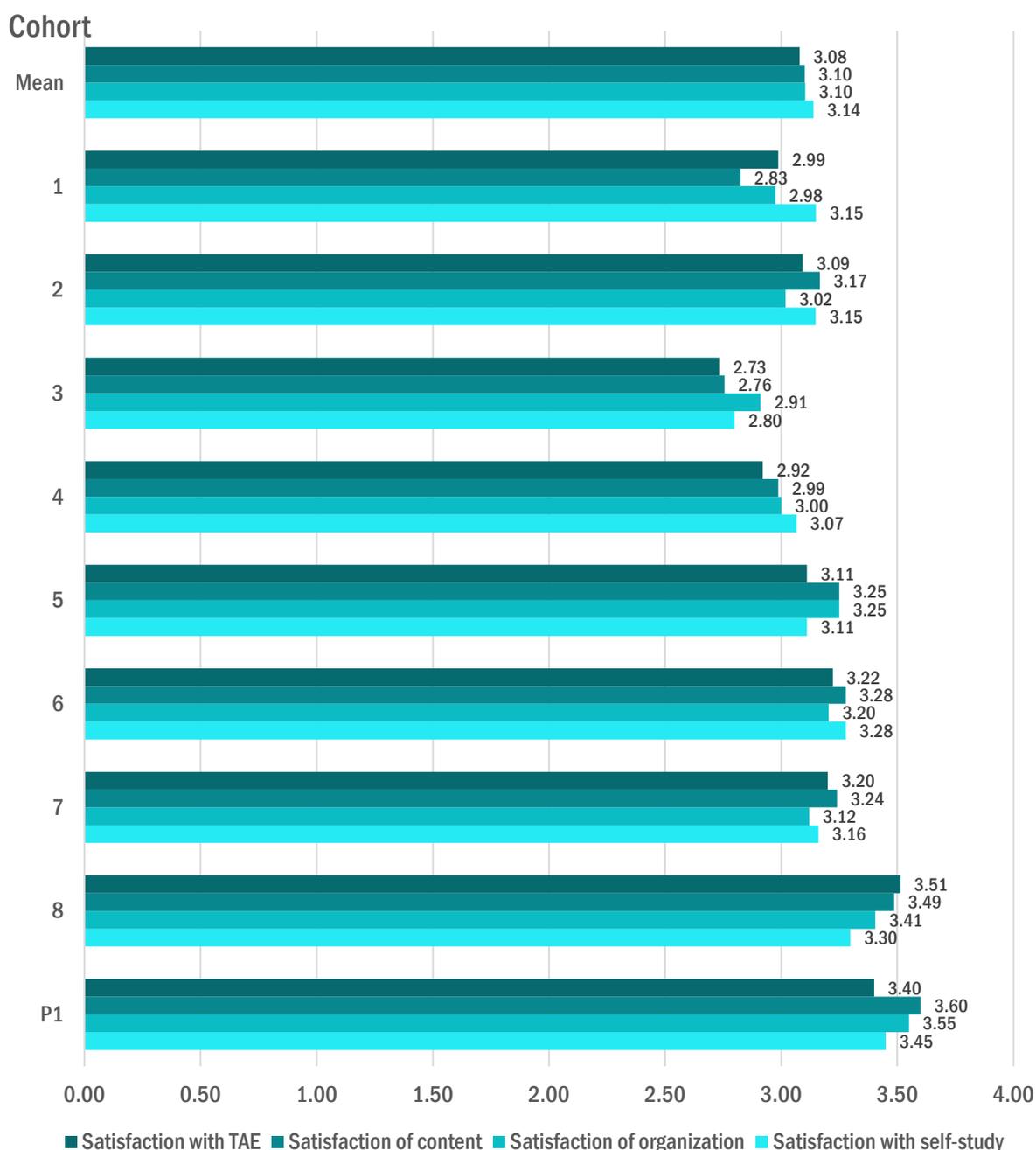
different cohorts presented different levels of satisfaction with these global course elements. For instance, cohorts 3 and 4 did not rate their satisfaction as favorably as other cohorts, with most mean ratings for these cohorts being below 3.00. In contrast, recent cohorts 8 and P1 were much more satisfied with the course, with ratings between 3.41–3.60 (Figure 1, bottom two data groupings). As one possible explanation, the delivery of the TAE program may be improving over time, resulting in improved ratings for the course. As a second possible interpretation, the lower ratings from cohorts 2 and 3 coincide with—and may be due to—the onset of the COVID pandemic and its strict sanitary measures, which potentially lowered ratings.

In terms of the different course modules, all were rated positively by participants, with the minimum rating being 3.12 on a 0–4 Likert-scale (compare the darkest shaded upper bars in each data grouping in Figure 2). It is important to highlight that two course contributions, Module 3 (TBLT) and Module 4 (Developing Academic Learning Objectives) were rated slightly higher than the other modules (3.41; 3.43 respectively). When compared to similar online teacher training courses, these two modules are unique to the TAE course, and this finding indicates that participants notice and value distinct TAE course offerings. While most modules were rated between 3.24 and 3.34, Module 1 was rated the lowest at 3.12. One possible explanation for this lower rating is that Module 1 is an introduction to the course; as such, it does not offer much new information to the students. Future iterations of course development should consider how innovative and distinctive content is valued, and how more performative course elements (i.e., introductions) are valued less. Accordingly, any information found in these performative course elements may be better positioned within course documents, such as the syllabus, leaving the modules to present content. Finally, in terms of the overall usefulness of the self-study course segment (Figure 2, lowest data grouping), this was also rated lower than the individual modules (3.14). While still positive, this relatively lower rating indicates that developers may want to consider the design of the online instructional elements in the course, to maximize the self-study section's positive characteristics.

Ratings of the usefulness of course modules also varied according to the education levels of the participants. While the large majority of course participants had an MA (69.1%) or a BA (26.8%), there were several participants with a PhD (6 with a PhD from a foreign country and 13 with a PhD from China). As can be seen in Figure 2, it is no surprise that the ratings of usefulness by participants with MA degrees—being such a high percentage of the course participants—is quite similar to the average rating. Then, it is encouraging to note that participants who only have achieved a BA degree tended to rate the different course modules more positively than the average rating, as did persons who have achieved a PhD in a foreign country.

The higher ratings by these two participant groups suggest two things. First, we speculate that persons with BA degrees may value more strongly the new teaching concepts and techniques introduced by the TAE course; this makes sense as the course materials may make a larger relative contribution to their overall knowledge of teaching. Persons with only a BA level of education could be considered as an important participant group to focus on during future iterations of the course, and it is worth considering whether the course content should be oriented towards this group in terms of the course's tone and linguistic complexity.

The high ratings of the different course modules by persons with PhD degrees from a foreign country potentially suggests something different: we speculate that the course content resonates with what this group of students has learned abroad. This may be sensible as this course was developed by two scholars situated within North American teaching traditions. In contrast were the ratings by persons with PhD degree from Chinese institutions, which

Figure 2

Each TAE Module was Rated Positively for Usefulness, though these Ratings Varied Somewhat Depending on Participant Education Level (n = 427)
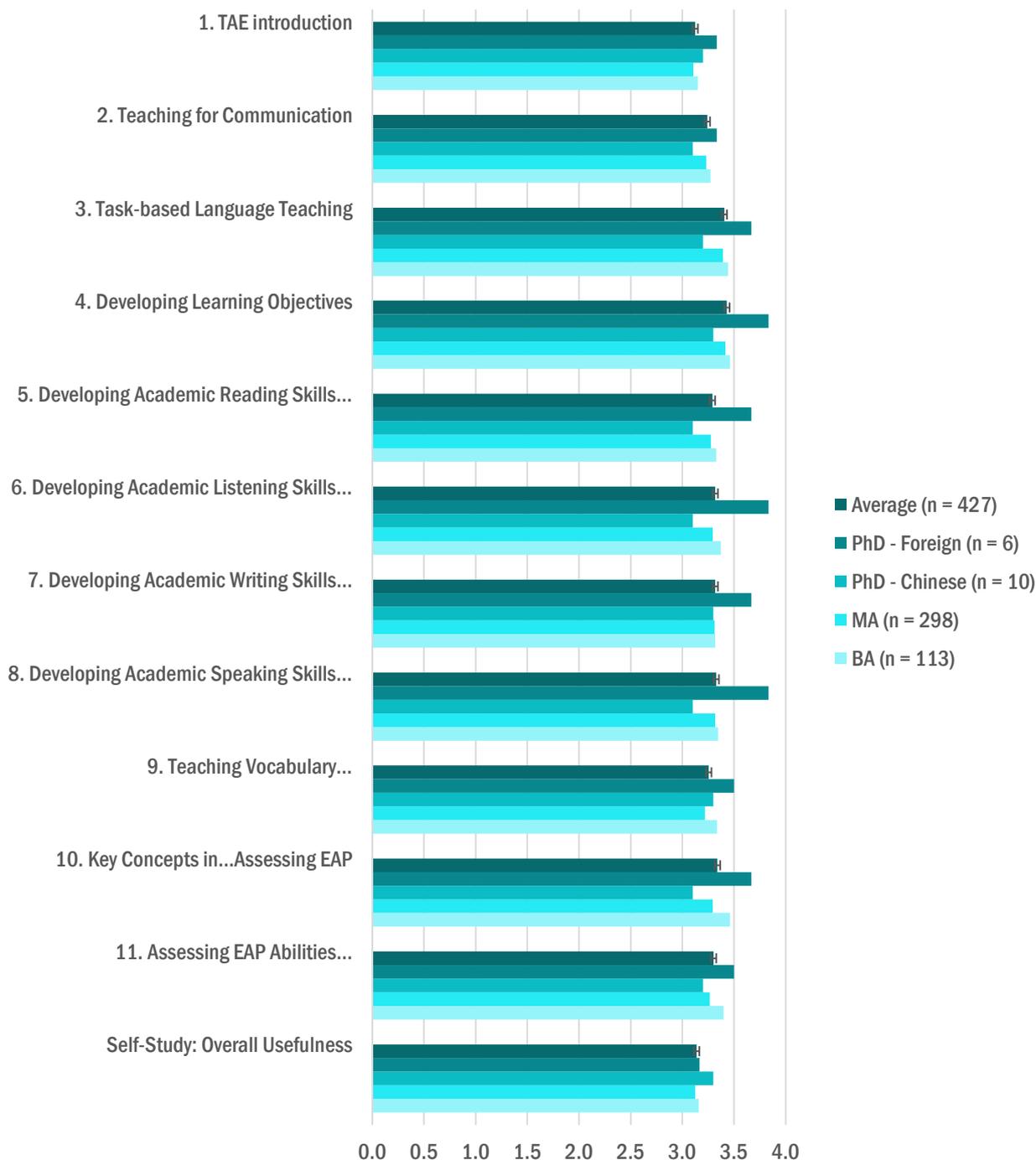
frequently were below the average module ratings. This may indicate that these persons—because of their depth of experience in China—may have had difficulty seeing how the course content could be implemented in their context. Although the sample sizes are too small to make definitive conclusions, it is still noteworthy to notice the difference in ratings between these different participant subgroups ($n = 427$). Reasons for these differences should be confirmed in future studies.

In terms of the usefulness of the workshop sessions, course participants generally rated each of the individual sessions as useful, with all sessions with average ratings of 3.20 or higher (see Figure 3). Participants highly rated the sessions on assessment, language form, and key concepts for academic English, with all ratings for these sessions being between 3.34–3.36. It should be noted that these three sessions are unique content contributions of the TAE course. This suggests again that participants place a special value on this unique content. Other highly rated components of the workshop include the workshop slides and the workshop booklet. This similarly suggests that unique TAE content created by ETS is of particular value. In contrast, the four sessions on each different language skill were rated lower than other workshop sessions. This may be because participants may already have had experience teaching the four language skills, while the sessions on key concepts and teaching language forms provided new information.
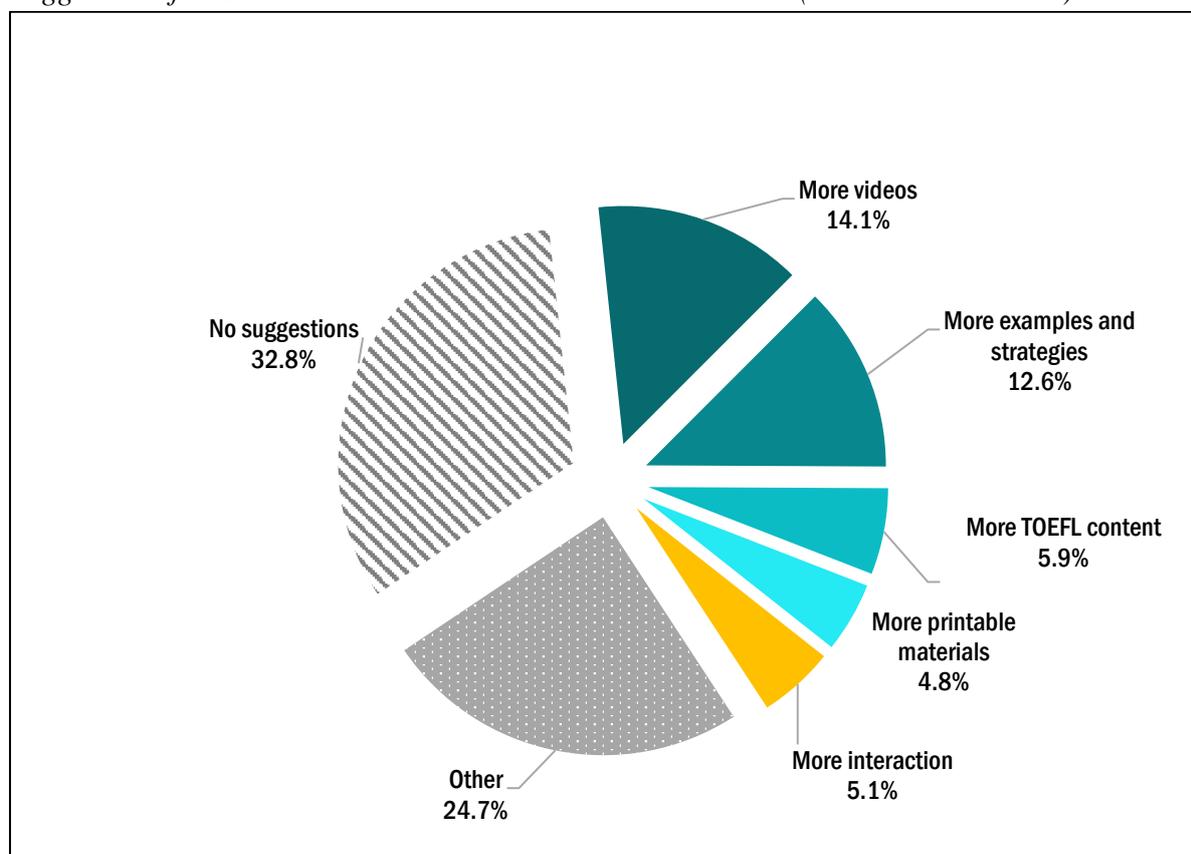
Figure 3
Ratings of Workshop Components are Consistently High. S1, S2, etc., Refer to the Specific Session of the Workshop (n = 427)
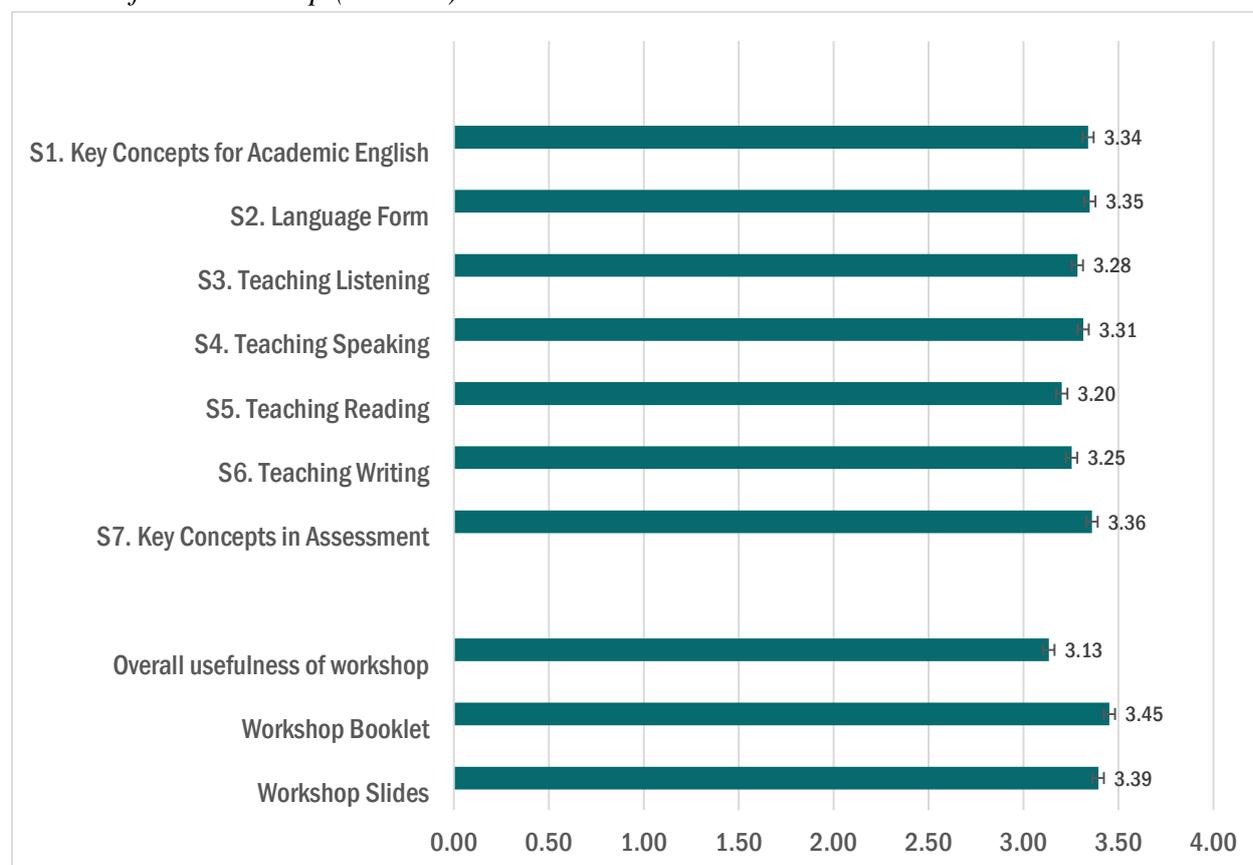
Open-Answer Survey Data

Course participants were asked to provide their suggestions for improving the online self-study modules and the workshops. In terms of the self-study modules (Figure 4), we coded 37.4% of participant suggestions as focusing on the different types of additional course content they would like to see. Though these suggestions varied widely in their scope, they centered on additional videos in the course (14.1%), additional examples and strategies (12.6%), more content about the TOEFL test (5.9%), and more printable materials (4.8%). These comments reinforce how participants value concrete content provided by the TAE course that can directly inform and support their teaching, and they suggest that the development of these sorts of materials will bring additional value to the TAE course. As a second trend, nearly one third of course participants (32.8%) made no comment about how the course could be improved. This may suggest that participants were satisfied with the course content. As a last major trend, 24.7% of participant responses were coded as *Other*. These comments did not converge around any larger theme and included suggestions such as requiring participants to submit an assignment for individualized feedback, modifying the content to cater to the needs of young learners, or allowing access to online modules for a longer period of time. Finally, it should be noted that there was an important isolated trend signaled by a number of participants: a call for more interaction (5.1%). While qualitatively smaller than the other codes we found, this recommendation will become important in this paper's next section.

Figure 4

Suggestions for Online Modules Center on 'More Materials' (Blue-Toned Sectors)

Participants were also asked to provide suggestions about the workshop (Figure 5). As a first tendency, we coded that nearly 40% of course participants had no suggestions to improve the workshops. This lack of commentary seems to confirm the qualitative finding that a large percentage of the course participants were at least relatively satisfied with the workshop. The next large group of findings (24.5%) were suggestions for increased interactions in the workshops. As part of this, 16.0% of participants made suggestions concerning increased interaction with presenters during the workshop, or that the workshop sessions be more interactive. Specific recommendations included having a longer Q&A session, more frequent and longer group discussions and activities, more involvement and feedback from the workshop facilitators, and an improved organization and content for some of the lecture sessions. In a similar vein, 8.5% of participants indicated that they were interested in a live workshop, which also is suggestive of increased interaction and the benefits of interacting with the professors and colleagues. As a second large trend, 18.9% of participants suggested that they would like to have an increase in content. 13.6% of these comments centered on teaching content, such as teaching strategies, tips, example lessons, or teaching demonstrations (13.6%).

Figure 5

Suggestions for Improving the Course Workshop Focus on Increased Content and Interactions (n = 427)
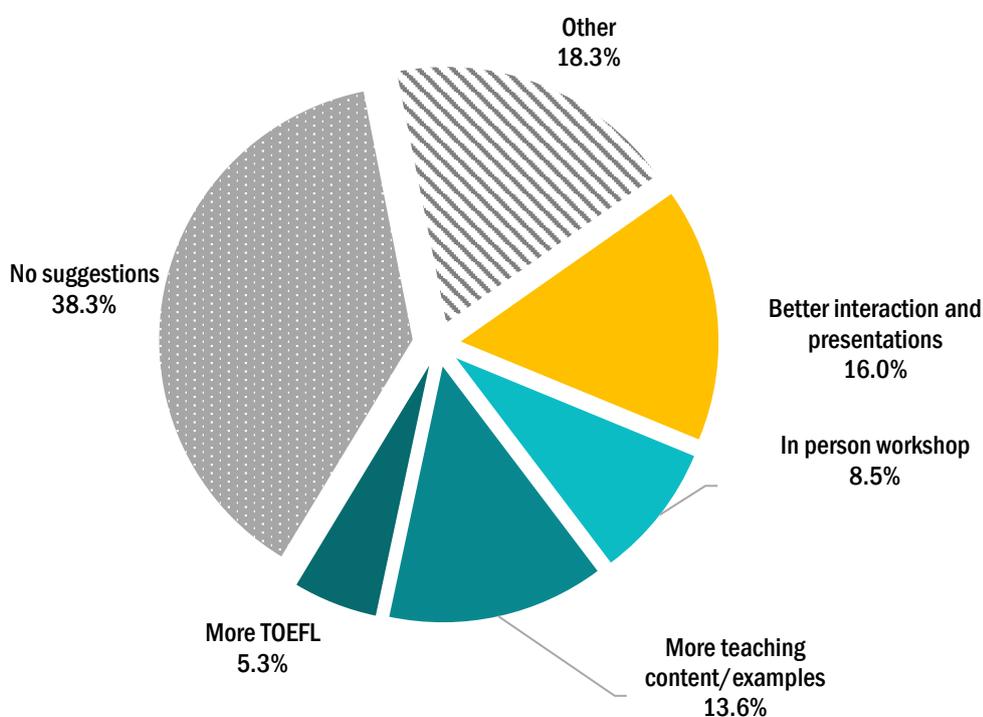
5.3% of participants suggested that they wanted more TOEFL-related content. As a last major finding, 18.3% of the responses were categorized under the *other* category, which includes various comments that did not fall under the major categories. Examples include improving the group arrangement (e.g., grouping team members with the same teaching background), reducing the course fee, and increasing the frequency and quality of communication from the course platform administrative team. In sum, while participants generally seem content with the workshop process, it should be highlighted that course participants continue to value content that is specific to ETS. Also worth noting is a request for increased interaction.

Conclusions Regarding Self-Reported Data
In Table 1, we present a visual conclusion of this section of the program evaluation concerning the participants' self-reported data about the TAE course.

In synthesis, there is a general consensus between the different data sets concerning the areas of the course content that are strong and those that require strengthening. While participants are generally satisfied with the TAE course, its content, and workshop, unique units of study and concrete materials created by ETS are rated especially highly by participants. This unique content is an asset that should continue to be developed. This said, when moving forward the course could expand on the types of interactive elements it uses to engage participants. This was signaled in both the Likert-scale and open answer items; attention to increased interaction is likely to increase participant satisfaction with the course.

Table 1
Scorecard for Self-Reported Data

|	Likert-Scale Questions	Open-Answer Questions
Strengths	• General satisfaction with TAE course • Unique content valued (e.g., TBLT) • Workshop materials highly valued • General satisfaction with workshop	• ETS-specific contributions valued
Gaps	• Additional materials desired • Some calls for more interaction	• More ETS / TOEFL content requested • More interaction in the workshop
Concerns	• Performative course elements (e.g., introduction) were valued less • Four skill language sections were less liked; they should have extra value added.	

Engagement Data
While it is vital to ask participants to provide their ratings and feedback about the different course processes, it is similarly important to compare participant beliefs against their actual behavior during the course (engagement data). This data triangulation complements the self-reported data and provides nuance about how the TAE course is being used.

As a first finding of the time-on-task data, participants spent less time on any given TAE course module than what was expected. Although course designers expected participants to spend between 3–4 hours per module (180–240 minutes), the time spent per module across all modules was less than 180 minutes in 54.8% of cases, within the expected range in 13.7% of cases, and greater than or equal to 240 minutes in 31.5% of cases. A similar tendency is

illustrated in Figure 6, which presents the time-on-task data for Module 10—*Key Concepts in Language Assessment*—one of the least engaged modules.

Figure 6

Participant Time-on-task (Engagement) is Less than Expected
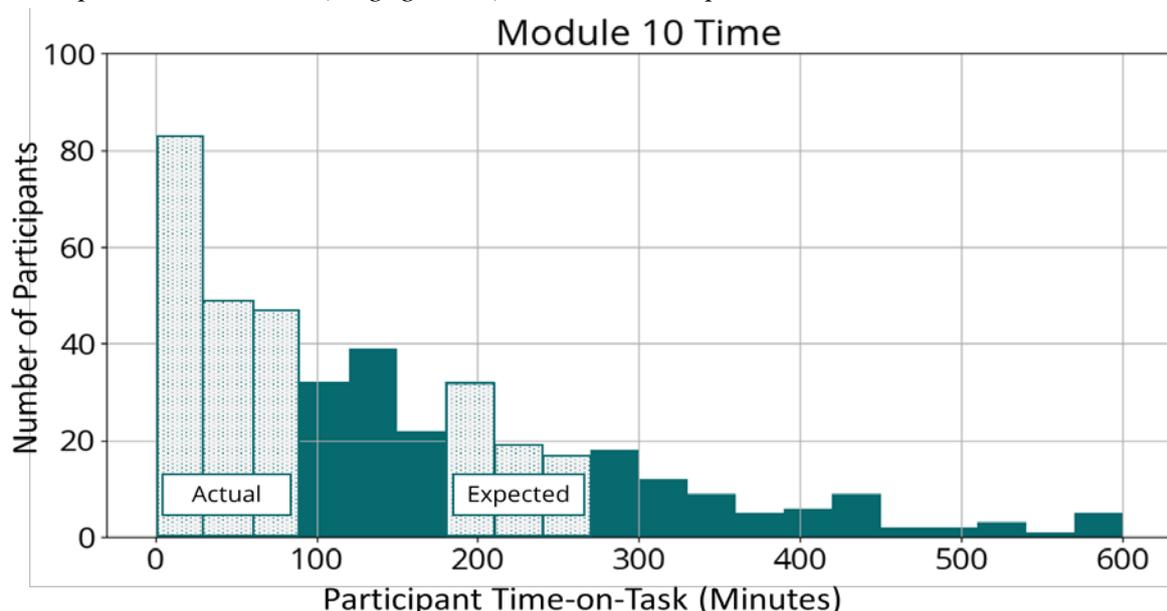
Note. This figure presents 10 cohorts of data concerning Module 10 (n = 435).

There are various possible interpretations of this general trend. First, participants may not require as much time as expected. This may imply that the content coverage is either narrow or shallow in that it does not provoke any new discovery or cognitively challenging content for participants to digest. It may also be the case that the content may consolidate what participants already know, which allows them to move quickly through the module content. This could be remedied by adding additional content, especially the unique type of content provided by ETS that participants described valuing. As a second alternative, participants may not invest as much time as expected because there is no course requirement concerning content mastery, which might in fact require 3–4 hours of dedication. It is also not known the degree to which participants have real extrinsic or intrinsic motivation to master the course material, besides receiving the course certificate.

To these points, course developers should ensure that minimum levels of understanding are met by the course participants, such as by requiring minimum quiz scores, something that was not required in this course version. As a last alternative, the content may not adequately engage the participants. As suggested in the self-reported data, there were some calls for increased interaction, which could positively impact the time invested in the course. To better understand this situation, the causes of this low time-on-task should be explored in future surveys, so that improvements can be made.

Also in terms of engagement, the time students invested in the course lessened as they progressed through the modules. This can be seen in Figure 7, which presents the mode time-on-task bin for each module, this is to say, the "most frequent" 30-minute time bin of time spent per module. While it was most frequent for participants in the first four modules to spend

between 90 and 240 minutes on each, they only spent between 0 and 30 minutes per module during the last seven modules. As stated above, possible reasons for this decline may include little extrinsic motivation. Also plausible is the general tendency to lose interest over time, and indeed many instructors will affirm that students lose motivation and engagement across any

Figure 7

Each Module's Mode 30-minute Time-on-task Bin
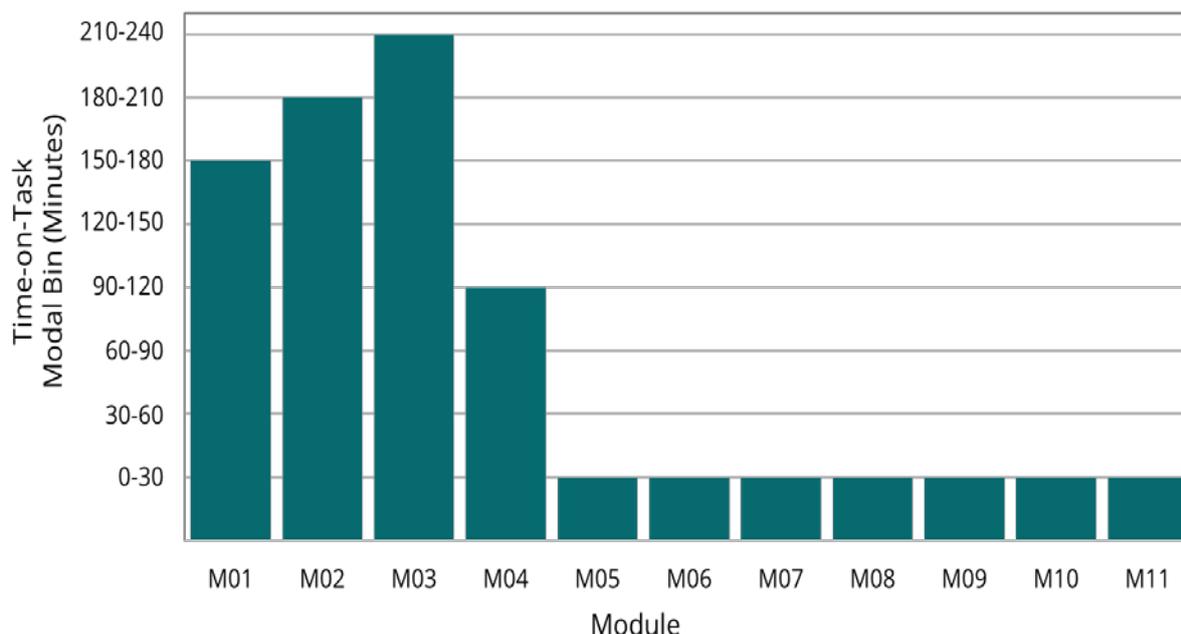
Note. Time bin modes from 10 cohorts indicate that later modules have reduced engagement (n = 435).

study period. Still, this tendency is still a concern to the TAE course, and the reasons for this decline in study time should be explored directly with course participants, in future studies.

Conclusions Regarding Time-on-Task Data

In Table 2, we present a visual conclusion of this section of the program evaluation concerning the TAE course's a time-on-task analysis, which captured how much time participants invested in each of the different modules in the TAE course. Of some concern is a decrease in student time-on-task across the course. Adding new course features or types of content that bolster participant engagement may be one way to maintain their attention in the course. This should be understood in future survey research (questionnaires, interviews, focus groups), to ensure that each module includes materials that engage participants and that will directly inform and support participants' teaching.

Table 2

Scorecard for Engagement Data

|	Time-on-Task Data
Strengths; Gaps Concerns	– • Time-on-task is lower than expected • Time-on-task decreases across course progression

Learning Products

Knowledge Checks and Quizzes

During Cohorts 1–5, each module included a 10-item multiple-choice end-of-module assessment. Starting with Cohort 6, short knowledge check quizzes were added throughout each module, and the end-of-module assessments were reformulated to provide a brief summary statement about key topics with 4–6 fill-in-the-blank exercises using a drop-down question format. Figure 8 presents the mean percent correct for both end-of-module assessment formats. For Cohorts 1–5, while the mean percent correct across modules was 76.92%, the mean percent correct by module varied from 66.92 (Module 9) to 95.84 (Module 1). Except for Module 9, the percent correct for all modules was above 70%. This suggests either that the uptake of important concepts discussed in the reading passages was generally high or that the questions were relatively easy or both.

Figure 8 also presents the mean percent correct for latter cohorts (Cohorts 6–9) in the new end-of-module assessment format. The mean percent correct for all modules was 85.75, which is substantially higher than the previous end-of-module assessment format used for Cohorts 1–5. Similarly, each module's mean percent correct was also higher, with the lowest percentage being 80.37 (Module 11) and the highest percentage being 91.67 (Module 4). Again, this suggests either that course participants were successful in their learning of course materials, that the end-of-module assessment was very easy, or both. It is also possible that the completion of fill-in-the-blanks using a drop-down menu was easier for participants due to the context clues provided in the summary paragraph, while multiple-choice questions may have had fewer of these clues. Beginning with Cohort 6, the fill-in-the-blank end-of-module assessment was complemented with a set of three or four knowledge-check quizzes, interspersed throughout each module. Each set of quizzes included questions either in a multiple-choice or true/false

Figure 8

Later Cohorts (6–9) Answer End-of-module Fill-in-the-blank Questions Better than Earlier Cohorts

Mean Percent Correct by Module
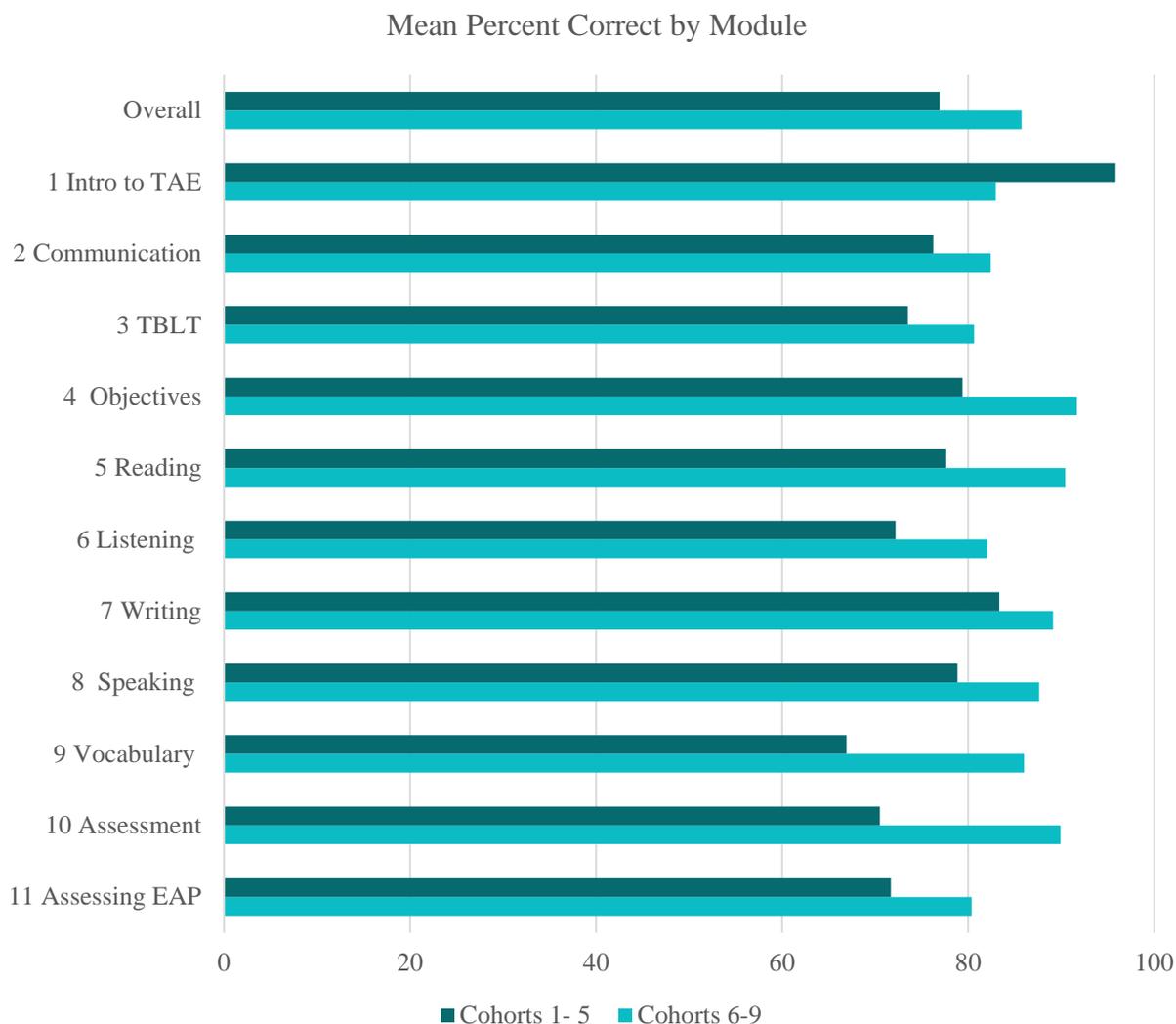
formats. Figure 9 shows the mean percentages of correct answers for the knowledge check quizzes. The mean percent of correct answers for all modules was 78.63, which is similar to the results for the multiple-choice end-of-module assessment for Cohorts 1 through 5. Similarly, means varied by module; Module 10 (assessment) had the lowest mean percentage of 64.33% and Module 4 (writing objectives) had the highest mean of 90.26%. Except for Module 10, all the other modules had mean percentages of 73% and above. In general, these findings suggest that the participants either had a reasonable uptake of course content, that the questions were quite easy for them, or both. Findings suggest that the participants either had a reasonable uptake of course content, that the questions were quite easy for them, or both.

Figure 9

Knowledge Check Quizzes Mean Percent Correct by Module (Cohorts 6–9)
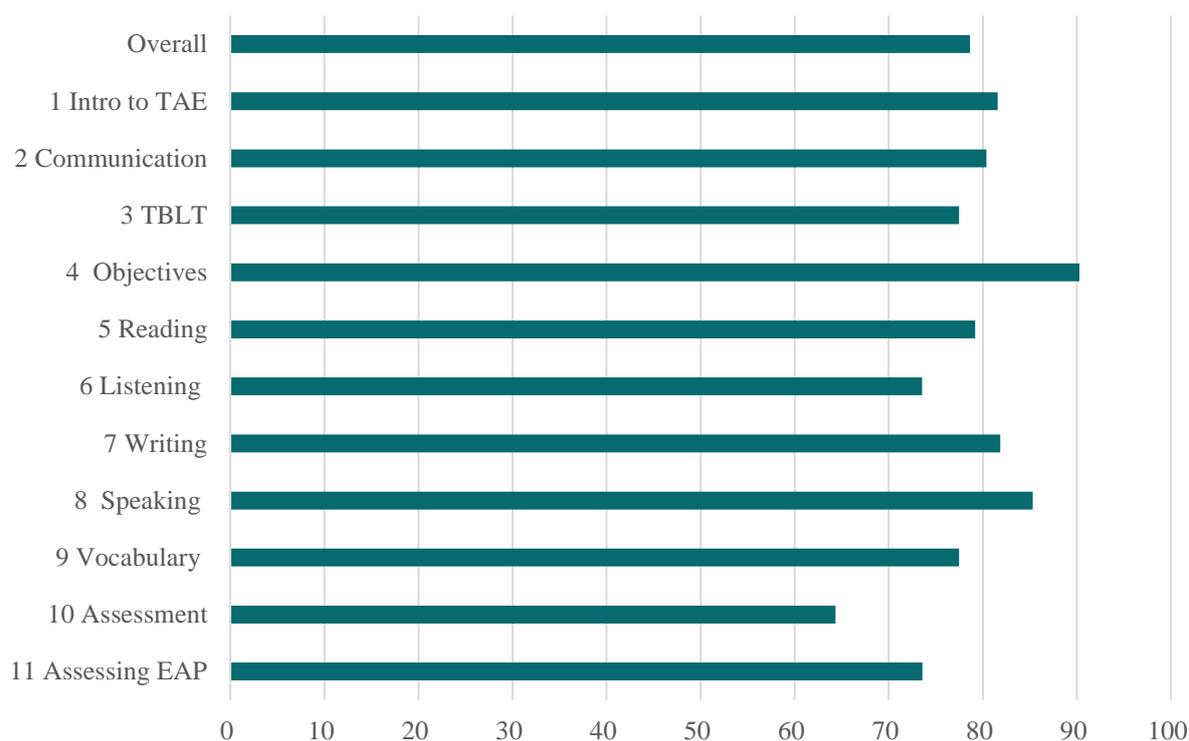
Learning Data–Group Lesson Plans

Complementary data sets were also studied to examine the degree to which participants changed after completing the TAE course. Here, we consider their completion of the group lesson plans, completed by groups of four to nine participants as the last component of the workshop segment of the course. The analyses focused only on the assessment part of the lesson plans, and Table 3 shows the average ratings over all groups in each cohort (Appendix E

Table 3

Percentage of Group Lesson Plans Incorporating Assessment Principles

Cohort	Users	Purpose	Relevance	Objective	Tasks
C2 (n = 8)	75.0%	75.0%	62.5%	87.5%	75.0%
C3 (n = 8)	75.0%	75.0%	62.5%	62.5%	37.5%
C4 (n = 9)	66.7%	33.3%	44.4%	44.4%	77.8%
C5 (n = 6)	16.7%	16.7%	16.7%	100.0%	83.3%
C6 (n = 6)	50.0%	37.5%	50.0%	50.0%	50.0%
C7 (n = 6)	66.7%	66.7%	66.7%	66.7%	100.0%
Yunnan (n = 6)	83.3%	83.3%	33.3%	66.7%	100.0%
Overall	65.3%	57.1%	51.0%	69.4%	75.5%

Note. Group lesson plans were generally successful in incorporating important dimensions of the assessment principles discussed in the course.

presents the rating criteria). Despite some variations in ratings across cohorts, the general trend shows that group lesson plans included many important dimensions of the assessment principles and strategies discussed in the course. Particularly successful dimensions among the five criteria were *tasks* (75.5% overall) and *objectives* (69.4% overall), indicating that

assessment tasks were as close to the real-life target task as possible, and all the learning objectives were assessed.

This indicates that between the online modules and workshop content, participants made important connections between their lesson's target task, its learning objectives, and its assessments. On the other hand, relatively challenging dimensions included *purpose* (57.1% overall) and *relevance* (51.0% overall). Indeed, some group lesson plans did not mention how the assessment results were being used, though this might have been clear to themselves. In terms of *relevance* many assessments did not focus on skills relevant to the target task and learning objectives. In future versions of the course, greater emphasis should be made on providing explicit rationales for assessment design, so that others could understand the rationale behind the assessments and be able to replicate the lesson. Furthermore, additional instruction should be provided to connect assessment constructs to the lesson, and the development of this content through additional course interactions (discussion board responses; additional feedback on the group lesson plans) represents the opportunity to add increased value to the course.

While the average ratings indicate difficulties for certain criteria in a general sense, there are large variations across cohorts for each of the criterion. For example, while five of the six groups in the Yunnan Cohort clearly identified users for their assessments, only one of the six groups in Cohort 4 addressed this construct. This variance may be due to the nature of the group lesson plans and the limited time and slides they had to describe their lessons. It may be the case that some details were delivered orally, instead of written on the slides. Another source of variation may come from the decision to provide additional, targeted feedback on participants' group lesson plans during the workshop beginning in Cohort 7. This may have positively impacted the quality of their final product.

Learning Data Scorecard: A Visual Conclusion
In Table 4 below, we present a visual conclusion of this section of the program evaluation concerning the learning data from the TAE course.

In synthesis, there is a general demonstration across the different data sets concerning the knowledge participants gained in the course. Positively, both the end-of-module assessments and group lesson plans indicated that participants had a general uptake of course content across all modules. This seems to indicate that the course has a solid foundation upon which it can continue to build. While some assessment items (the fill-in-the-blank items) may be too easy for participants—or while some content areas may require additional explication—changes to

Table 4
Scorecard for TAE Course Products

	Knowledge Checks & Quizzes	Group Lesson Plans
Strengths	• High uptake of key course points	• Generally incorporated important assessment principles and best practices
Concerns	• Revised fill-in-the-blank assessment may be too easy due to context clues	• Some content may require further clarification (*assessment purpose, relevance*)

the assessments are likely to add value to the course, especially if they center on providing additional content to participants, or if they ask participants to engage in additional interactions.

Conclusions and Future Directions

In this paper, we looked at three different classes of data—self-reported data, engagement data, and learning data—in order to better understand how participants responded to the TAE course. A synthesis of the main evaluation findings is presented in Table 5.

In terms of course strengths, self-reported data showed that participants valued the course strongly, and they especially valued the highly specific ETS content and material resources. In terms of learning data, the assessments and quizzes indicated that there was a good uptake of most key course topics, and participants were typically able to incorporate key information from the course in their group lesson plans. Despite this positive response, we suggest that program evaluation continue of the TAE course, to "constantly. . . improve each element of [the] curriculum on the basis of what is known about [these] elements, separately as well as collectively" (Brown, 2001, p. 15). Specifically, we hope to continue to collaborate with course administrators to "close the loop" and ensure that these evaluation findings are considered in future course versions. The provision of high quality content, either in terms of making refinements to existing course content, or adding additional topics, such as *using AI technology*, should continue to be a priority.

The data sets—in particular the self-reported data—did, however, signal some gaps in the course. Matching our suggestion above, survey responses illustrated that participants would like to have even more material resources, including specific information about the TOEFL iBT test. There were also some calls for more interaction, both in the online course and the workshop. Paired together, these two findings underscore the importance of the course's content (the *what*), but also of the importance of the interactions required by the course (the *how*). It seems that it would be important to study how the content is delivered, to ensure that it is in pieces that are attractive and easily digestible, avoiding issues of multimedia instruction such as *the wall of words* (Clark & Meyer, 2016). Additionally, interaction could be increased with the personalization of assessment and quiz questions, as could different moments of the workshop.

Finally, the data sets also raised some concerns about some course elements. In terms of course content, the learning data indicated that some assessments may be too easy, while other elements of the course require further explanation. This suggests that course developers should continue to monitor the learning data and adjust the course content according to the tendencies it shows longitudinally, as we recommended above. Furthermore, the engagement data indicated that students are not spending the amount of time predicted on the course tasks. This indicates that course developers should consider not only the content but how it is presented to the participants, to ensure that it matches their needs and learning styles. It is important to highlight that this concern with the course was not evident in the self-reported data. This may be that participants may have unwittingly—or consciously—sought to please the survey writers. This has been found elsewhere, such as in Cao, who documents how

Table 5

TAE Program Evaluation Score-card: Three Data Sets

	Key Findings	S/E/L	Moving Forward
Strengths (to keep)	• General satisfaction • ETS-specific content valued • Concrete material resources valued • Uptake of key course points • Generally incorporated important assessment principles and best practices	• S • S • S • L • L	• Continue to build for general satisfaction by investing in details • Continue to expand the unique content in the course the course (e.g., TBLT, TOEFL, using AI technology) because it is valued by participants! • Expand the resources (e.g., lesson plans, videos) the course provides to continue to add value to the course • Continue to monitor assessment scores • Continue to monitor and respond to gaps found in group lesson plans
Gaps (to add)	• Additional materials desired • Additional ETS/TOEFL content desired • Some general calls for more interaction • More interaction in the workshop • Less value of "four skills" sections	• S • S • S • S • S	• Provide additional worked examples—and potentially videos—that help students organize and build their understanding • Build for more meaningful interactions in the self-study segment of the course, with scenarios that contextualize the content, more personalized discussion board questions • Reframe workshop elements to provide additional interaction • Add unique contributions to typical content sections (i.e., the four language skills + TBLT), so participants feel that they are learning unique content
Concerns (to fix)	• Performative elements valued less • Revised fill-in-the-blank assessment may be too easy due to context clues • Some content may require further clarification (*assessment purpose, relevance*) • Time-on-task lower than expected • Time-on-task decreases across course	• S • L • L • E • E	• Reposition performative elements outside of course modules • Revise assessments to ensure they are challenging. Interactive assessment items may answer concerns about interaction, while providing additional assessment moments. • Add additional content • Explore with students why they have decreasing time-on-task across the course.
In Sum	• Specific, interactive content valued!		• How can the content be more specific and more interactive?

Note. (**S**) = survey data finding; (**E**) = engagement data finding; (**L**) = learning data finding.

"1,559 respondents . . . consistently inflated the interest domain that matched the target" (2016, p. ii). Finding discrepant evidence, in our case the engagement and learning data, indicate that despite the very positive self-reported data (the questionnaires), that course developers should continue to invest in improving the content and course interactions, as indicated above. Second language researchers should continue to provide program evaluations of the TAE course, "the systematic collection and analysis of all relevant information necessary to promote the improvement of the curriculum and analyze its effectiveness within [its] context" (Brown, 1996, p. 277), so that the TAE course can continue to close the loop, always making improvements to better match the course with the needs that the participants have.

ORCID

https://orcid.org/0000-0002-9060-9412

https://orcid.org/0000-0002-6913-0300

https://orcid.org/0000-0003-4854-1116

https://orcid.org/0000-0002-4734-3794

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

Rights and Permissions

Open Access

References

Brown, J. D. (1995). *The elements of language curriculum*. Heinle & Heinle.

Brown, J. D. (1996). *Testing in language programs*. Prentice Hall Regents.

Brown, J. D. (2001). *Using surveys in language programs*. Cambridge University Press.

Cao, M. (2016). *Examining the fakability of forced-choice individual differences measures*. [Unpublished doctoral dissertation]. University of Illinois at Urbana-Champaign.

Clark, R. & Meyer, R. (2016). *e-Learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*. Wiley.

Creswell, J., & Creswell, D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage.

Fournier, D. (2005). *Evaluation*. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 139–140). Sage.

Greene, J. (2005). *Mixed methods.* In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 255–256). Sage.

House, E. (2003). Introduction. In T. Kellaghan, D. Stufflebeam & L. Wingate (Eds.), *International handbook of educational evaluation* (pp. 9–14). Springer.

Jamieson, J. M., Eignor, D., Grabe, W., & Kunnan, A. J. (2008). Frameworks for a New TOEFL. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Karataş, H., & Fer, S. (2009). Evaluation of English curriculum at Yıldız Technical University using CIPP model. *Education & Science, 34*(153), 47–60.

Kelleghan, T., Stufflebeam, D. L., & Wingate, L. A. (2003). Introduction. In T. Kellaghan, D. Stufflebeam & L. Wingate (Eds.), *International handbook of educational evaluation* (pp. 1–6). Springer.

Lee, S. Y., Shin, J.-S., Lee, S.-H. (2019). How to execute context, input, process, and product evaluation model in medical health education. https://doi.org/10.3352/jeehp.2019.16.40rg/10.3352/jeehp.2019.16.40

Leskes, A., & Wright, B. (2005). *The art and science of assessing general education outcomes*. Association of American Colleges and Universities.

Lippe, M., & Carter, P. (2018). Using the CIPP model to assess nursing education program quality and merit. *Teaching and Learning in Nursing, 13*(1), 9–13. https://doi.org/10.1016/j.teln.2017.09.008

Mackey, A., & Gass, S. (2016). *Second language research: Methodology and design*. Routledge.

Mirzazadeh, A., Gandomkar, R., Hejri, S., Hassanzadeh, G., Koochak, H., Golestani, A., … Razavi, S. (2016). Undergraduate medical education programme renewal: A longitudinal context, input, process, and product evaluation study. *Perspectives in Medical Education, 5*, 15–23. https://doi.org/10.1007/s40037-015-0243-3

Norris, J. M., Davis, J. McE., & Timpe-Laughlin, V. (2017). *Second language educational experiences for adult learners*. Routledge.

Patton, M.Q. (2012). *Essentials of utilization-focused evaluation*. Sage.

Sancar Tokmak, H., Baturay, H., & Fadde, P. (2013). Applying the Context, Input, Process, Product evaluation model for evaluation, research and redesign of an online master's program. *The International Review of Research in Open and Distance Learning, 14*(3), 273–293. https://doi.org/10.19173/irrodl.v14i3.1485

Stufflebeam, D. (2003). The CIPP model of evaluation. In T. Kellaghan, D. Stufflebeam, & L. Wingate (Eds.), *International handbook of educational evaluation*. Springer.

Sylwester, B. (2017). *Impact of program assessment in higher education: A case of an applied linguistics program*. [Unpublished doctoral dissertation]. University of Hawai'i, Mānoa.

Toosi, M., Modarres, M., Amini, M., & Geranmayeh, M. (2021). The Context, Input, Process, and Product evaluation model in medical education: A systematic review. *Journal of Education and Health Promotion, 10,* 1–12.

Trace, J., Meier, V., & Janssen, G. (2016). "I can see that": Developing shared rubric category interpretations through score negotiation. *Assessing Writing, 30*(1), 32–43. https://doi.org/10.1016/j.asw.2016.08.001

Yicai Global. (2017, May 31). *The 2017 "New First-Tier" City Ranking was released, with Chengdu, Hangzhou and Wuhan ranking among the top three, and Zhengzhou and Dongguan newly entering the list.* https://www.yicai.com/news/5293378.html

Zhang, G., Zeller, N., Griffith, R., Metcalf, D., Williams, J., Shea, C. & Misulis, K. (2011). Using the context, input, process, and product evaluation model (CIPP) as a comprehensive framework to guide the planning, implementation, and assessment of service-learning programs. *Journal of Higher Education and Outreach Engagement 15*(4), 57–84.

Appendix A

TAE Modules

Module	Title
1	Introduction to Teaching Academic English
2	Teaching for Communication
3	Task-based Language Teaching
4	Developing Academic English Learning Objectives
5	Developing Academic English Reading Skills for the *TOEFL iBT*
6	Developing Academic English Listening Skills for the *TOEFL iBT*
7	Developing Academic English Writing Skills for the *TOEFL iBT*
8	Developing Academic English Speaking Skills for the *TOEFL iBT*
9	Teaching Vocabulary for Academic English Tasks
10	Key Concepts in Language Assessment
11	Assessing EAP Abilities Using *TOEFL iBT®* Test Materials

Appendix B

Overview of TAE Workshops

Day–Time	Workshop Title	Speaker
Friday–AM	• Introduction to the Workshop Sessions	ETS
	• Key Concepts for Teaching EAP skills Assessed on the TOEFL iBT Test	ETS
	• Teaching Language Form for Academic English	ETS
Friday–PM	• Techniques for Developing EAP and TOEFL iBT Test Listening Skills	Local Expert
Saturday–AM	• Techniques for Developing EAP and TOEFL iBT Test Speaking Skills	Local Expert
	• New Trends and Development for the TOEFL iBT Test	Local Coordinator
	• Techniques for Developing EAP and TOEFL iBT Test Reading Skills	Local Expert
Saturday–PM	• Techniques for Developing EAP and TOEFL iBT Test Reading Skills	Local Expert
Sunday–AM	• Key Concepts and Principles of Language Assessment	ETS
	• Supplementary Learning and Teaching materials	Local Coordinator
	• Q&A Session	ETS, Local Experts
	• Feedback session on group lesson plans	ETS
Sunday–PM	• Presentation of Lesson Plans	Student Groups
	• Award to Best Lesson Plan & Closing Ceremony	ETS

Note. All days and times reflect Beijing Time.

Appendix C

Overarching CIPP Data Collections

Context

What do we know about the context in which this program is situated, based on:
- Thematic analyses: project proposal documents
- Likert-scale survey responses, participants (quantitative analyses)
- Thematic analyses: open-answer survey responses, participant
- Thematic analyses: competitive landscape documents

Inputs, Materials

What can we learn about the quality of the input provided by the course materials, based on:
- Likert-scale survey responses, participants (quantitative analyses)
- Thematic analyses: open-answer survey responses, participants
- Thematic analyses: interview data, course administrators
- Content analysis: coursework documents, based on Clark and Mayer's (2016) instructional design rubric.

Processes

What can we learn about the course processes based on:
- Likert-scale survey responses, participants (quantitative analyses)
- Thematic analyses: qualitative survey responses, participants
- Participant time-on-task data

Products

What can we learn about what students learned in the TAE course, based on:
- Learning data: percentage scores, quiz and end-of-module tests
- Thematic analyses: group lesson plans
- Thematic analyses: Final projects

Appendix D
Sample Exit Survey Questions
Background Information (multiple-choice or open-ended questions)
- What is your gender?
- How old are you?
- What is the highest level of education you have completed?
- In what type of educational institution do you teach?
- For how many years have you taught English?
- What experience do you have teaching TOEFL iBT® skills?
- Briefly explain why you enrolled in *Teaching Academic English with the TOEFL iBT® Test*.

Overall Feedback on TAE (Likert-scale questions)
- Overall, how satisfied are you with TAE?
- Overall, how satisfied are you with the organization of TAE?

Feedback on TAE Online Study (Likert-scale questions)
- Overall, how useful was online self-study for your teaching and/or professional development?
- How useful were the online modules for your teaching/professional development?
- How useful were the different parts of the online modules?
- What comments or suggestions do you have for improving the TAE online modules?

Feedback on TAE Online Study (Likert-scale questions)
- How useful were the individual online webinars for your teaching/professional development?

Feedback on TAE Workshop (Likert-scale or open-ended questions)
- Overall, how useful was the TAE workshop for your teaching and/or professional development?
- How useful were the individual workshop sessions for your teaching/professional development?
- How useful were the different workshop activities for your teaching/professional development?
- How useful were the workshop materials for your teaching/professional development?
- What suggestions do you have for improving the TAE workshop in generally or its specific parts?
- Overall, how useful was the TAE course (online study + webinar + workshop) for your teaching and/or professional development?

What did you learn from TAE? (Likert-scale questions)
- As a result of the TAE online course and workshop, how much has your knowledge of teaching academic English and *TOEFL iBT*® skills increased?

Appendix E

Lesson Plan Scoring Rubric

Lesson Plan Coding Scheme for Group Lesson Plans

Dimension	Definition	Score range
Users	Assessments are created for specific users	0-1
Purpose	There is a specific purpose/use for the assessment	0-1
Relevance	Only relevant skills and constructs covered in the lesson are assessed	0-1
Objectives	All objectives are assessed	0-1
Tasks	The tasks in the assessment are as similar as possible to the language tasks performed in the real-world	0-1