Listening. Learning. Leading.®

Research Report
ETS RR–14-36

A Design Framework for the ELTeach Program Assessments

John W. Young

Donald Freeman

Maurice Cogan Hauck

Pablo Garcia Gomez

Spiros Papageorgiou

December 2014

ETS Research Report Series

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

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

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

A Design Framework for the ELTeach Program Assessments

John W. Young,[1] Donald Freeman,[2] Maurice C. Hauck,[1] Pablo Garcia Gomez,[1] & Spiros Papageorgiou[1]

1 Educational Testing Service, Princeton, NJ
2 University of Michigan, Ann Arbor, MI

ELTeach is an online professional development program consisting of two courses, English-for-Teaching and Professional Knowledge for English Language Teaching (ELT). Each course includes a coordinated assessment leading to a score report and certificate for individual teachers. Developed with reference to international and national teaching standards, and drawing on the resources of various national English-language curricula and teaching materials, the program is designed to ensure that teachers have the functional classroom English language and the professional knowledge to support the implementation of the English-language curricula they are expected to teach. In this document, we outline the design and development process followed in creating the two ELTeach assessments—the Test of English-for-Teaching (*TEFT*™) and the Test of Professional Knowledge (*TPK*™). The paper is organized into the following sections: (a) the purposes and intended uses for test scores/certificates, (b) the target populations, (c) the construct definitions, (d) the program frameworks and domain descriptions, (e) the test designs, (f) the scoring systems, and (g) directions for research in support of the program.

Keywords ELTeach; teacher professional development; English language teaching; English-for-Teaching; teacher professional knowledge

doi:10.1002/ets2.12036

In the globalized world of the 21st century, English is a resource that provides increased access to economic and social rewards (e.g., Graddol, 2006, 2010). As many countries around the world expand English language teaching (ELT) and introduce the language to students at lower grade levels, the demand for teachers who are qualified, capable, and confident to teach English is expanding rapidly. For students at the elementary and secondary school levels, access to teachers who have the necessary professional knowledge and functional English language skills to teach English effectively is critical. Enhancing students' motivation contributes to the English proficiency students ultimately can achieve through classroom instruction.

Increasingly, governments and educational authorities need teachers to develop quickly the professional knowledge and functional language skills to teach English. This need creates a scalable demand to provide training for large numbers of teachers, often within the contexts of their ongoing job responsibilities. Thus, the professional development of teachers, and of teacher candidates, is an important step in ensuring that students will have access to high quality English language instruction and the social and economic opportunities such instruction can afford. The challenge, then, is to provide teachers with effective, timely, cost-effective, and scalable professional development that will increase their professional knowledge and language skills to teach English effectively and to successfully reach a growing number of students.

To address this global need, Educational Testing Service (ETS) collaborated with National Geographic Learning (NGL) to develop the ELTeach program, an online professional development program consisting of two courses: the English-for-Teaching course and the Professional Knowledge for ELT course. Each course includes a coordinated assessment leading to a score report and certificate for individual teachers. The coursework and assessment are offered as an integrated program; neither component is independent. Developed with reference to international and national teaching standards, and drawing on the resources of various national English-language curricula and teaching materials, the program is designed to ensure that teachers have the functional classroom English language and the professional knowledge to support the implementation of the English-language curricula they are expected to teach. This alignment bounds the classroom language known as English-for-Teaching (EFT) and defines the domain and the concepts known as professional knowledge for ELT, which are presented in the learning materials and tested in the assessments.

Corresponding author: P. Garcia Gomez, E-mail: pggomez@ets.org

The primary purpose of this document is to describe the design and development of the two ELTeach assessments — the Test of English-for-Teaching (*TEFT*™) and the Test of Professional Knowledge (*TPK*™). In documenting the design and development of these assessments, we seek to ensure that the assessment development process meets the professional standards for the intended uses of the assessments and that the assessments, as part of the integrated program, are designed to be appropriate for their intended uses. In addition, the document can serve as a reference for future investigations of test validity that may be required to support the intended test uses as part of the ELTeach program. This said, however, it must be understood that these assessments were not designed to stand independently of the learning materials. Each test is part of this integrated program of professional learning and is not designed to be used independently of the learning materials. Documentation of the development and piloting of the complete ELTeach program of learning materials and assessments is available in the *ELTeach Global Pilot Report*, *2012* (Freeman, Katz, Le Dréan, Burns, & Hauck, 2013).

In this document, we describe the following aspects of the two assessments: (a) the purposes and intended uses for test scores/certificates, (b) the target populations, (c) the construct definitions, (d) the program frameworks and domain descriptions, (e) the test designs, (f) the scoring systems, and (g) directions for research in support of the program.

Test Purpose and Intended Uses for Test Scores/Certificates

The ELTeach program has been designed and developed for institutional users such as ministries of education (MOEs), local education authorities (LEAs), multilateral funding agencies, universities, and preservice training organizations, among others, to support the professional learning of ELT teachers in both preservice and in-service contexts. The development process has focused particularly on the public-sector teaching force in order to arrive at a scalable solution.

In aligning learning materials and assessments according to common frameworks (see the "Domain Descriptions" section in this paper), the integrated design of the program is intended to serve two functions. For teachers, the program provides training through the two courses, which is linked to independent documentation through the course-coordinated assessment. For administrative authorities, the program offers documentation of candidate learning and performance, which can support institutional users of the program to strengthen planning and evaluation. Thus, the purpose of the ELTeach program is (a) to support teachers in improving their command of the language and skills needed to teach English in English (through the English-for-Teaching course) and/or their professional knowledge (PK; through the Professional Knowledge in ELT course) while (b) providing institutional users with analytic information about teachers' learning and performance.

In terms of goal (a), the assessments in the ELTeach program are designed to provide

- demonstration of what teachers know and can do within the program framework (see the "Domain Descriptions" section in this paper);
- documentation about how teachers' individual levels of language skills and PK in the content compare to the internationally developed descriptors of performance established in the program;
- information about areas within the ELTeach program curriculum on which teachers might focus improvement; and
- a form of documentation of teachers' professional development in ELT that could be included on an individual's resume, curriculum vitae, or professional development portfolio.

In terms of goal (b), the ELTeach program is designed to provide institutional users and educational authorities with information

- to document the overall PK and language skills of groups of teachers in the content of the ELTeach program curriculum;
- to document the degree to which candidates within teacher preparation or training programs have developed the PK and language skills presented in the ELTeach curricula; and
- to help meet policy and accountability goals and benchmarks in preservice and professional development planning in ELT.

The ELTeach program assessments have been designed and developed as integral components of the professional development program. They assess the degree to which the individual test taker has demonstrated PK and/or language skills for teaching — depending on the test — presented in the course learning materials. This alignment between learning materials and assessments is both a conceptual imperative and an operational foundation of the program (see the "Construct

Definitions" section in this paper). The security model, test length, and test reliability are designed to support the test uses described above. Neither test is designed to be appropriate for independent, high-stakes uses such as teacher certification. The ELTeach program recognizes, and indeed anticipates, that over time the uses of these assessments will evolve based on how they are valued locally and the role they come to play in specific educational contexts.

Target Populations

A large population of teachers of English as a foreign/second or additional language exists in the world today, with large numbers of them, perhaps even a majority, working in the public sector as teachers of English in elementary, middle, and high schools. Some, if not most, teach large classes of students, within national curricula, and in preparation for national tests. These teachers may have only a basic command of general English — most are likely at the Common European Framework of Reference (CEFR) A1 or A2 levels (Council of Europe, 2001). They may use the local first language (L1) for a considerable proportion of the class period, either because of the limitations of their own English proficiency (they are more comfortable and less embarrassed speaking in L1) or because they feel their students may not understand them if they use English. These teachers recognize that their immediate command of English is not fully adequate for their professional work, both for classroom teaching of English in English and for potential professional engagement with the global ELT community. Both they and their employers — MOEs and LEAs — want to improve teaching quality. From a workforce development point of view, the challenge is to do this in ways that increase systemic capacity of the current teaching force and can bring direct and immediate results (Butler, 2004).

The ELTeach program is designed to address these needs for both preservice and in-service teachers. The target population for the TEFT and TPK assessments includes teacher candidates who are still in training, novice teachers with limited classroom experience, and also teachers who may have years of classroom experience but would benefit from the increased confidence and effectiveness that can be gained from the ELTeach program.

It is also important to note that the uses of ELTeach do not depend on conventional distinctions drawn between native and nonnative speakers. While a majority of the users of the program will likely be referred to as nonnative speakers of English, ELTeach is appropriate both for teachers who are themselves at various stages of developing English language proficiency and for fully fluent English speakers. The functional English in the English-for-Teaching course has been selected primarily for teachers whose L1 is not English. However, it is also appropriate for those fluent English speakers who need to learn to moderate their language use in order to be effective in classroom contexts. Similarly, the Professional Knowledge for ELT course presents a globally derived curriculum of essential PK that is appropriate as a foundation for any ELT teacher. In fact, in many national and local contexts, the program learning materials and tests may serve as tools to determine and assess professional development needs (see the "Test Purpose and Intended Uses for Test Scores/Certificates" section in this paper).

Construct Definitions

The conceptual design of the entire ELTeach program was a highly collaborative effort among ETS staff; staff from National Geographic Learning (NGL)/Cengage; senior program advisors Donald Freeman, Anne Katz, and Anne Burns; and a panel of 18 experts in English language education from 12 countries around the world. While ETS staff had primary responsibility for designing, developing, and operationalizing the assessment of each construct, the assessment design and development work was done in ongoing consultation with staff from NGL/Cengage and the external advisors.

The core construct definitions for the ELTeach program are derived from an alternative theoretical framework for teacher's knowledge, known as *knowledge-for-teaching*. In the following section, relevant literature underlying knowledge-for-teaching is summarized, followed by the construct definitions for EFT and for PK, respectively. Overviews of the entire design and development process are provided later in Figure 1 (for EFT) and Figure 4 (for PK).

Relevant Literature: Knowledge-for-Teaching

The two core constructs of the ELTeach program, EFT and PK for ELT, both sit within the theoretical framework of knowledge-for-teaching. This section elaborates the antecedents of this construct.

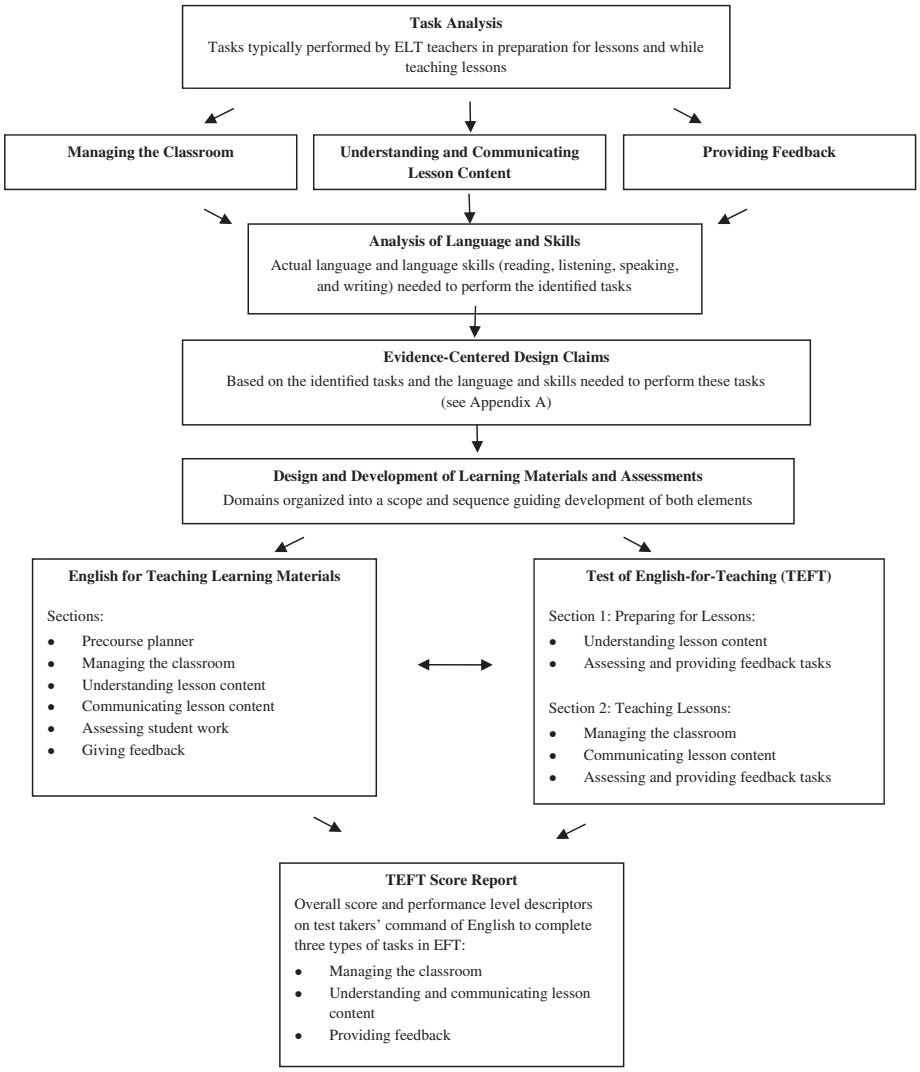


Figure 1 Overview of the English-for-Teaching (EFT) design and development process.

Interest in the types of knowledge teachers may use to inform their teaching dates back to Dewey (1920) and indeed before. However, sustained research into the area dates primarily from the 1980s (see Freeman, 1996), and the construct of content-knowledge-for-teaching is rooted in general educational research in a trajectory of work that dates from the same period. This work sought to conceptualize the knowledge teachers use in teaching as unique and socially situated. The constructs included pedagogical content knowledge (Shulman, 1987), personal knowledge (Elbaz, 1983), personal practice knowledge (Connelly & Clandinin, 1988), instructional knowledge (Cohen, Raudenbush, & Ball, 2003), and the above-mentioned content-knowledge-for-teaching.

In the paper that is generally seen as establishing the present focus on knowledge in teaching, Shulman (1987) argued for a differentiated kind of knowledge that integrated classroom (or pedagogical) knowledge of pedagogy with subject-matter (or content) knowledge through the lens of experience teaching particular subject matter to specific groups of learners (Turner-Bisset, 1999; Wilson, Shulman, & Richert, 1987). His paper outlined seven specific types of knowledge that teachers use in classroom instruction, of which pedagogical content knowledge (PCK) captured the attention of researchers, teacher educators, and policy makers. Since then, PCK has become a central focus of research, with particular attention paid to how PCK functions in teaching and how it is learned by new teachers. Studies have examined teaching of social studies (e.g., Harris & Bain, 2011; Monte-Sano & Budano, 2013), mathematics (e.g., Ball, Thames, & Phelps, 2008), science (e.g., Abell, Rogers, Hanuscin, Lee, & Gagnon, 2009; Magnusson, Krajcik, & Borko, 1999; Van Driel, Verloop, & de Vos, 1998), and English language arts (e.g., Dudley-Marling, Abt-Perkins, Sato, & Selfe, 2006; Grossman, 1990).

Ben-Peretz (2011) conducted a targeted review of research on teacher knowledge and observed that, since Shulman (1987) proposed his categories of teacher knowledge, the notion of teacher knowledge has been written about, expanded upon, and developed significantly to include ideas of personal knowledge and societal issues. For example, Elbaz (1983) argued that teachers hold practical knowledge that is made up of five subtypes of knowledge, incorporating knowledge of instruction but also knowledge of self. Importantly, she argued that knowledge is oriented in different ways and that this orientation influences both its storage and its use. This perspective of teachers' knowledge brings to light the importance of experiences, both before a teacher enters a classroom and experiences as a teacher, and how they influence one's knowledge and subsequent instruction.

However, foreign or second language teaching is notably absent from this work. This absence may be for a number of reasons. Principally, the issue seems to turn on the complex interrelation between how language is learned and used in the world generally and how it becomes distilled as subject matter to be taught, a process that Larsen-Freeman and Freeman (2008) have called creating subject-language. Defining what it is that teachers know—or need to know—in order to teach languages has always been a complicated undertaking. Where language content and teaching methodology have conventionally been seen as independent of one another (Kelly, 1969), in the last decade that notion of separation has been challenged both conceptually and operationally, by consideration of how the two knowledge domains inform, and indeed blend into, each other in the work of teaching (Freeman & Johnson, 1998). This blending, which is now generally termed *content-knowledge-for-teaching*, has become increasingly central in theorizing and implementing teaching in all content areas.

Pioneered in fields such as elementary mathematics (Ball et al., 2008), knowledge-for-teaching mathematics has been documented as distinct in elementary mathematics teachers, as influencing their classroom practices (e.g., Ball, Hill, & Bass, 2005), and as evident in the learning outcomes of their students. In language teaching, the issue of knowledge-for-teaching is complicated by the fact that language is both learned as content knowledge and acquired through individual experience and socialization. In fact, one function of classroom practices in language teaching seems to be to simultaneously establish language as content (through attention to grammar and language forms) while also attempting to replicate language learning through individual experience and socialization (e.g., communicative approaches to teaching). The ELTeach program addresses this paradox of treating language as *subject-language* in its core constructs, EFT and PK for ELT.

Theoretical Background of the English-for-Teaching (EFT) Construct

To improve student-learning outcomes, educational authorities around the world seek ways to improve teachers' command of English. Because the phrase *command of English* has not been clearly defined (viz., the use of the CEFR, which is a framework, as a de facto set of language standards), nor has it been directly tied to the work of classroom teaching, a common strategy has been to focus on improving teachers' general language proficiency. The assumption is that a general capacity in the language will, in turn, lead to improving classroom teaching and student learning. This assertion has not been borne out in research (Elder, 2001), nor has it been proven to be effective in a policy/resource sense.[1]

The ELTeach program takes the approach that to improve teachers' command of classroom language, the focus needs to be on the English needed on a daily basis to carry out classroom instruction and manage predictable interactions. The EFT construct is defined as

> the essential English language skills a teacher needs to be able to prepare and enact the lesson in a standardized (usually national) curriculum in English in a way that is recognizable and understandable to other speakers of the language.

Therefore, EFT can be thought of as a tool, in the sociocultural sense (see Engestrom, 2001), that a teacher uses to teach English as a school subject in English (Larsen-Freeman & Freeman, 2008). The use of English to teach English is significant in a number of ways, perhaps most prominently in that the use of language is to accomplish particular ends in the classroom context, and not in terms of overall or general proficiency. Thus, the construct as defined inherently repositions English as a practical communicative tool rather than simply as an object of study. In characterizing the form of English as what is needed to teach the lesson, it is important to distinguish this from the language a teacher could use to teach that lesson if perhaps he or she had more facility with the language, different goals (e.g., language immersion), and/or students at a higher general proficiency level.

The language learned and practiced within the EFT framework is a subset of the language that a speaker could use to enact the lesson; it is a predictable set of language interactions that allow the teacher to enact the particular lesson in the classroom setting. The challenge, then, is to determine parameters of that language in order to define and select the particular content to be presented in the learning materials and exemplars to be assessed in the test. Sešek (2007) observed that in assessing English as a foreign language (EFL) teachers' command of English language, there is an interaction between content knowledge of English as the subject and knowledge of English as the language of instruction. In the ELTeach program, this process of establishing boundaries for language content is achieved through the learning materials: The test taker meets and can practice the particular language that will be assessed. In this way, the test becomes an assessment of a prescribed set of language classroom uses; it is a test of language proficiency for specific purposes rather than a test of general language proficiency. In this way, it is similar to other tests of professional language use, such as legal cant or medical terminology. The basis for assessment is further described in the "Domain Descriptions" section in this paper.

There is a further important benefit of establishing the boundaries of classroom language in the program; it contributes to what is called *professional confidence*. It is widely recognized and documented that a speaker's self-confidence can influence fluency. In the classroom, teachers use English when they are confident of their control over the specific language they need to conduct classes in the target language (Zakeri & Alavi, 2011). However, there is little evidence that general language proficiency necessarily provides them with that confidence (Butler, 2004; Consolo, 2006; Elder, 2001). Therefore, the ELTeach program is designed to directly address teachers' professional confidence through explicit attention to it in the precourse planner activities included in the learning materials. In these activities, teachers self-rate their confidence to carry out the tasks and routines in the course in English, through the explicit connection of practice activities to classroom settings in which they occur and through the basic premise that the teacher will be tested on what she or he has studied in the learning materials.

The TEFT assessment measures the test taker's control over functional English, as represented in the learning materials, which is needed to teach that country's national curriculum in English at the elementary or secondary school level. The test is based on the assumptions that the test taker

- may or may not use English partially or completely as the medium of instruction, although he or she is familiar with the curricular content;
- is familiar with classroom routines, including basic classroom management and teaching strategies, and can carry out these classroom tasks and routines that are predictable;
- is expected to use a defined (often nationally prescribed) curriculum;
- draws English language support from instructional materials;
- is teaching students who are at the beginning or intermediate levels of general English proficiency; and
- is expected to use English to interact with students in simple and predictable ways.

While the TEFT contains tasks that require the test taker to use the language skills of listening, speaking, reading, and writing, it is not a test of those skills per se, as a test of general proficiency might be said to be. Rather the TEFT tests a group of functional uses for each skill based on classroom practices. These uses are grouped in the EFT construct into three functional areas that serve as crucial organizing principles for the learning materials, the assessment, and the score reporting: managing the classroom; understanding and communicating lesson content; and providing feedback.

In developing the functional categories that anchor the EFT construct, the ETS and NGL teams drew on processes and practices accepted in English for specific purposes, including (a) a definition of specific language use or performance outcomes, (b) an analysis of users' (in this instance, teachers') needs and specific tasks performances, (c) a description of language skills used for these specific purposes, and (d) specific lexical and grammatical nature of the language involved. More information about how these processes and practices informed the development of the TEFT is provided in the sections on domain description and test design.

Theoretical Background of the Professional Knowledge (PK) Construct

The construct of PK in the ELTeach program is a widely accepted one. The framework of domains (see the "Domain Descriptions" section in this paper) was developed through a three-stage process that included (a) reviewing various national, regional, and international ELT teacher standards; (b) comparing this group of categories to a socioprofessional knowledge base in ELT as represented in the NGL/Heinle professional list (the book content was tagged to the chapter

level and then categorized according to the developing framework); and (c) completing an external review by a global panel drawn from universities and MOEs in 10 countries (see the *Global Pilot Report*, *2012* [Freeman et al., 2013] for more details). The process resulted in the identification of a construct for PK that aligns, at a policy level, with national and international ELT teacher standards. The construct also reflects a socioprofessional consensus as represented in the NGL/Heinle professional list and confirmed by an international panel of 10 ELT educators and researchers representing a range of regions of the globe.

The TPK assessment is intended to measure a test taker's PK of ELT as defined in the framework and represented in the learning materials. These materials, which are organized into two modules—foundations of PK and core teaching skills—present concepts specific to ELT in English at a level comprehensible to the target population, which due to the academic nature of the content, is approximately the CEFR A2-B1 reading competency level. The test is based on the assumptions that the test taker

- may or may not be familiar with the professional content;
- if familiar, may have met that content partially or completely in his or her national/first language;
- is familiar with classroom routines, which may be through previous training or simple school socialization; and
- has an A2/B1 level of general English language proficiency in reading, which may be higher than his or her proficiency in speaking.

Test scores indicate how well test takers understand the content presented in the learning modules. As a knowledge test, the scores cannot be used to make any claims about how well a teacher might implement the concepts presented in the course.

Domain Descriptions

The Test of English-for-Teaching (TEFT) Assessment

The language taught in the EFT learning materials is a bounded set of functional words, phrases, and language skills teachers can use to carry out essential classroom activities in English. The language in the course is organized into three areas reflecting the three categories introduced above: managing the classroom, understanding and communicating lesson content, and providing feedback. These categories emerged from the teacher tasks identified in the task analysis and relate well to typical categories in several of the national teacher training and classroom curricula. The language is anchored in the regular, predictable tasks that teachers perform in the course of teaching English rather than in a general English language proficiency framework. The specific language exemplars were gleaned, with direction from the international panel of experts, from national teaching materials, curricula, and information from local classrooms (including assignments, samples of student work, and video recordings of actual classes). The course materials are designed to be self-accessed, thus allowing teachers to focus on their personal areas of need and determine individually what they practice and how they can best make progress. The TEFT assessment has been developed based on the same framework as the course materials; it is designed to measure how test takers perform on typical classroom tasks as represented in the course materials.

The construct of EFT (as a core part of the larger theoretical framework of knowledge-for-teaching) was developed and operationalized through a combination of English for specific purposes (ESP) and evidence-centered design (ECD) frameworks. These two frameworks each offer distinct benefits. The ESP framework focuses on ensuring that the design of learning materials and assessments is centered on learners' needs and that practical outcomes are emphasized. This is, to a large extent, due to ESP's focus on tasks that are typical of the target situation in which language use is contextualized and on detailed analysis of language and skills needed to perform these tasks (Douglas, 2000; Dudley-Evans & St John, 1998; Hutchinson & Waters, 1987).

The ECD approach, on the other hand, provides an overarching framework that requires all components of a training and assessment program to have coherent and logical connections between the intended claims to be made about its users and the evidence that is gathered through learning and test tasks (Mislevy, 1994; Mislevy, Almond, & Lukas, 2003; Mislevy, Steinberg, & Almond, 1999). The ECD framework comprises different stages of systematic evidentiary reasoning for the design of assessments: domain analysis (preliminary synthesis of knowledge or skills involved in real-world situations and theoretical perspectives to explain performance); domain modeling (claims, observable features of performance that

would support the intended claims, and tasks that would involve the type of performance expected); construction of the assessment framework (technical detailing of what is generated in the previous stages, including aspects such as the student model, the evidence model, the task model, the presentation model, and the assembly model); and operational deployment (includes presentation, capturing, and scoring of responses as well as summary scoring). For the purpose of this document, only selected aspects of the ECD stages followed for the TEFT are discussed.

The development of the TEFT followed a multistage process synthesized in Figure 1. First, a thorough list of tasks typically performed by ELT teachers in preparation for lessons and while teaching lessons was created based on input from the global panel of experts as well as detailed analysis of curricula, textbooks, and classroom video recordings from various regions around the world. All of the identified teacher tasks were then organized into the three broad categories around which the construct is built (managing the classroom, understanding and communicating lesson content, and providing feedback). These categories eventually became the core areas for the ECD overall claims as well as the organizational framework for the EFT learning materials, the TEFT assessment, and the TEFT score report. As a next step, each of the ELT teacher tasks identified was linked to specific language, skills, and arrangements of skills.

A set of ECD claims was then created. Claims are statements a program would like to be able to make about test takers' knowledge and skills (or other attributes) on the basis of their performance on a test. The overall claim for the TEFT states that the test taker has the essential English language skills to carry out basic tasks in ELT, supported by English language instructional materials. A candidate who successfully completes the program will typically be able to, in English, complete three types of basic tasks in EFT:

- Managing the classroom (engaging with students in simple, predictable classroom exchanges).
- Understanding and communicating lesson content (understanding content for students and tasks for the teacher as included in instructional materials, and presenting lessons in class based on a defined curriculum and instructional materials).
- Providing feedback (providing basic oral and written feedback to students).

A set of subclaims was also created, organized by the language skills in each of the three functional areas described above. Following ESP frameworks, each subclaim includes the skill/s involved, the purpose for using such skill/s, and the nature of the language involved (see Figure 2).

For example, one of EFT's subclaims reads as follows:

Can comprehend student output for the purposes of:

- Monitoring and assessing comprehension.
- Identifying errors to provide feedback in grammar, vocabulary, punctuation, spelling, etc.

Skill: Reading

Purpose: To monitor and assess comprehension; identify errors and provide feedback

Nature of Language: Text of various types as presented in students' instructional materials; student output

Another subclaim reads as follows:

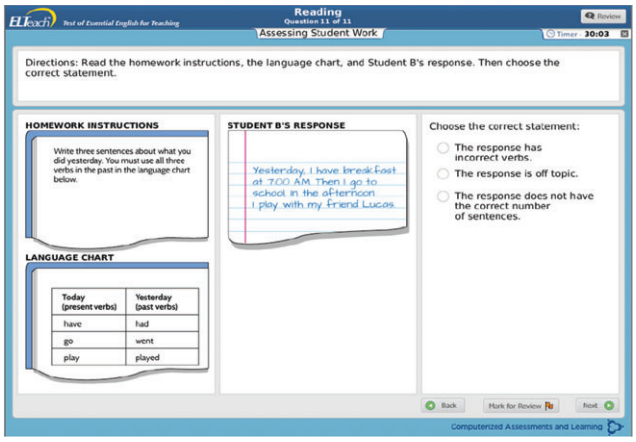


Figure 2 Sample TEFT reading task.

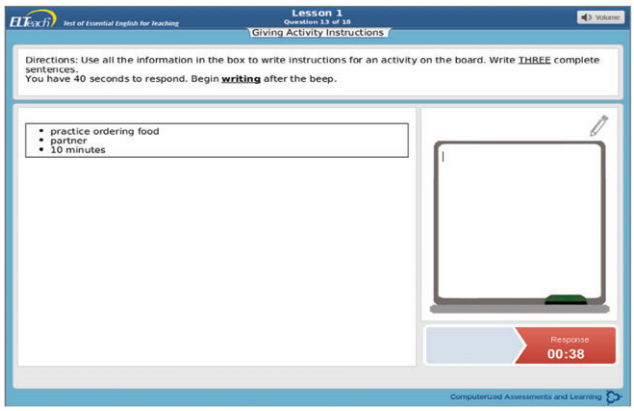


Figure 3 Sample TEFT writing task.

Can write instructions and explanations (on worksheets, assessments, etc.) by using formulaic language for the purpose of preparing and enacting a lesson (see Figure 3).

Skill: Writing

Purpose: To give instructions and provide explanations

Nature of Language: Formulaic/based on phrases typically used by teachers to write instructions and explanations. Phrases for these and other teacher tasks have been collected in an EFT "language bank."

All of the subclaims developed for EFT follow the same pattern (skill, purpose, nature of the language involved). The remaining steps shown in Figure 1, specifically the design and development of the test content itself and the score reports, are described in the sections below on test design and score reporting. Excerpts from the list of the EFT teacher tasks and the associated subclaims is available in Appendix A.

Once the claims and subclaims were established, the learning materials and assessments were designed and developed in tandem and with ongoing cross-consultation between the two technical teams to maximize alignment between the two components of the program.

The final decisions on reporting were made after the program was piloted. Statistical analysis provided support for the type of reporting described in last part of Figure 1 and later in the document. The stages described in Figure 1 are shown in sequential order for explanatory purposes, but the process, as it is the case with any new program design, was not linear. Cycles of evidentiary reasoning took place throughout the design and development effort.

The Test of Professional Knowledge (TPK) Assessment

The content in the Professional Knowledge for ELT course represents a synthesis of knowledge domains drawn initially from a review of international and national teaching standards. The framework for the course was articulated by a global panel of experts from 10 countries and then judged for comprehensiveness against the PK base represented in over 30 titles published by NGL-Heinle. The resulting set of concepts and terminology is organized into two major areas: core teaching practices and foundations of PK. The content is presented at a level of English accessible to most teachers who have acquired English language skills in a largely EFL context.

The framework underlying both the PK learning materials and the TPK course includes seven domains, which are divided into two areas as shown below:

Core Teaching Practices

- Planning: This domain covers the teachers' knowledge of how to plan teaching that meets students' goals and promotes learning and how to modify those plans to keep students engaged and progressing toward learning outcomes.
- Teaching: This domain covers the teachers' knowledge of how to create a classroom environment that helps students learn in explicit ways and promotes respectful interactions among them.

- Assessing: This domain covers the teachers' knowledge of how to gather and interpret information about individual student performance and learning in order to promote ongoing English language development.
- Materials, resources, technology: This domain covers the teachers' knowledge of how to recognize and use a variety of resources to enhance learning.

Foundations of Professional Knowledge (PK)

- Knowledge of students: This domain covers the teachers' knowledge of how to recognize the importance of who students are and how their expectations, backgrounds, and communities can influence and support their English language learning.
- Knowledge of theory and professional commitment: This domain covers the teachers' knowledge of theories and approaches to ELT and learning as well as teachers' ability to recognize the ways in which students acquire new language, both inside and outside classrooms. In addition, this domain covers the teachers' ability to recognize and appreciate the ways in which ELT is shaped by the local teaching and professional environments, as well as the broader professional community.[2]
- Knowledge of English and skills (metalanguage): This domain addresses the understanding to recognize and use appropriate knowledge of English to support students in using English both inside and outside class.

The TPK assessment was developed based on the same framework as the course materials; it measures how well test takers understand the content presented in the course materials. The design and development process followed for both the course and the assessments is synthesized in Figure 4.

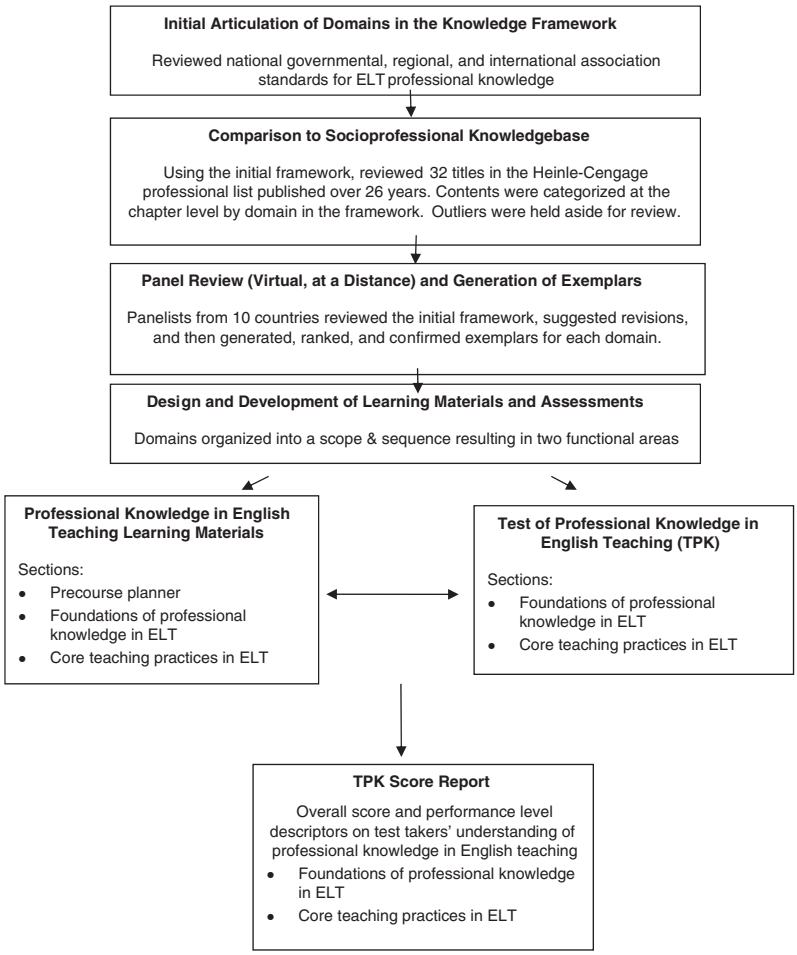


Figure 4 Overview of the professional knowledge (PK) design and development process.

Test Designs

Test of English-for-Teaching (TEFT)

Once the EFT domain was defined, ECD processes also helped determine the types of test items and test structure that would best capture appropriate evidence about the test claims. The EFT claims call for test items that focus on tasks that teachers need to do while preparing for lessons and while teaching lessons and on the language and language skills needed to perform such tasks. Single-selection and multiple-selection multiple-choice items were used when appropriate, but the test also features novel, technology-enhanced item types and sequences of items. Together, these assessment tasks represent the ways in which teachers use English in preparation for lessons (i.e., reading materials in textbooks and student work, listening to textbook recordings, correcting student work, and writing feedback) and while teaching lessons (going over whole lessons, managing the classroom, presenting lesson content, and providing feedback tasks).

Because the focus of TEFT is on measuring teachers' command of English to perform essential ELT teaching tasks, the items in the test differ from those typically used to assess general English language proficiency. For example, teachers are presented with the materials they would typically have at hand in the real world. They are required to handle multiple pieces of information including, among others, lesson goals, activity instructions, a variety of reading texts and scripts, language charts, textbook examples, explanations, and samples of student written and oral work. One result of this is that while the language teachers encounter as input (from curriculum materials, from students) and are asked to produce as output is generally quite simple, TEFT test takers are required to process several ideas and then express them in a manner that faithfully represents the complexity of teaching in English. Every effort was made in the design process to make the tasks on the TEFT look as much like the actual classroom work of language teaching as possible. Thus, the authenticity of these tasks enhances the validity of the test.

The TEFT assessment is organized into two sections: Section A, Preparing for Lessons, includes three parts containing reading, writing, and listening tasks, respectively; Section B, Teaching Lessons, contains four parts, each of which represents the chronology of a single lesson. Section B contains a variety of integrated speaking, writing, and listening tasks. Test takers have 150 min to complete the test. Figure 5 shows the test structure for TEFT.

Test of Professional Knowledge (TPK)

The TPK assessment was developed based on the PK for ELT learning materials and maintains the language accessibility of the course content. The test measures how well test takers understand the pedagogical content presented in the course materials. It does not attempt to assess how well teachers are able to implement this knowledge in actual classroom teaching situations (i.e., it is not intended as a performance measure).

Unlike the TEFT, which required a novel construct definition and innovative task types, PK's knowledge base domain and its content areas could be assessed using multiple-choice items assembled in a traditional test structure.

The TPK assessment contains 80 selected-response test items. The test includes single-selection multiple-choice items with four written options, single-selection multiple-choice items with three graphic options, and multiple-selection multiple-choice items for which the two correct answers must be identified. The test includes questions on the two PK

Section A:	**Part 1: Reading**	**Part 2: Writing**	**Part 3: Listening**
Preparing for Lessons (80 items)	Test takers read and answer questions about textbook materials (lesson goals, activity instructions, reading texts) and student work.	Test takers write instructions (for worksheets, quizzes, and tests) and correct student work.	Test takers listen to textbook recordings (instructions, conversations, talks) and answer questions about them.

Section B:	**Part 1: Lesson 1**	**Part 2: Lesson 2**	**Part 3: Lesson 3**	**Part 4: Lesson 4**
Teaching Lessons (60 items)	Each lesson has three subparts: Beginning the Class, Teaching the Lesson, and Ending the Class. In each lesson, test takers complete a variety of integrated speaking, writing, and listening tasks. They include tasks related to managing the classroom, presenting and explaining lesson content, and providing feedback.			

Figure 5 Test of English-for-Teaching (TEFT) test structure.

content areas: foundations of PK in ELT and core teaching practices in ELT. The TPK is presented in a single test section, and test takers have up to 90 min to complete the assessment.

Scoring Systems

Scaling

Scaling is the process of assigning numbers to the performance of representative samples of target examinees to help test users in interpreting test scores or results. The process of scaling produces a score scale, and the scores produced through scaling are referred to as scaled scores used to reflect examinee performance. Given that ELTeach scale anchoring studies were conducted using the pilot Form 1 data (1,307 test takers), it was decided that the first pilot form should serve as the base form for both the TEFT and the TPK.

The Test of English-for-Teaching (TEFT) Assessment

For the TEFT, there are 61 unique scale score points ranging from 400 to 700, reported in units of 5 points. The highest scaled score is set to the raw score[3] that is equivalent to the 98th percentile of the score distribution, which corresponds to a raw score of 183. The lowest scaled score is set equal to the raw score of 103. Because pilot test scores were highly skewed in a positive direction, this raw score represents the 4th percentile. From a content point of view, this positive skewing is acceptable; it reflects the fact that students who engaged with the course materials were generally able to achieve more than half of the available raw scores,[4] which is consistent with the overall intentions of the program. However, this skewing did create challenges for establishing an appropriate score scale. In order to alleviate the skewing of the scaled scores, every two raw scores are associated with the same reported scaled score between raw scores of 103 and 142. A one-to-one mapping of raw to scale score is utilized from 143 to 183.

In addition to the equated overall total scores, TEFT also features nonequated scaled scores reflecting performance on items that assess each of the four language skills scores (listening, speaking, reading, and writing). These scores resulted from an equipercentile scaling of the total scaled scores to the language skill scores. These scores are reported on a 40 to 70 scale with reporting units of 2 points. As a result, there are 16 unique score points for each language skill.

The Test of Professional Knowledge (TPK) Assessment

A linear scaling method is used for the TPK. There are 41 unique scaled score points ranging from 100 to 300, reported in units of 5 points. The highest scaled score is set to the raw score[5] that is equivalent to the 98th percentile of the score distribution, which corresponds to a raw score of 60. The lowest scale score is set equal to the raw score of 20, which is slightly higher than chance and represents the 3rd percentile.

Item Calibration, Scale Linking, and True Score Equating

Item response theory (IRT) is a statistical theory for analyzing test takers' performance on a set of test questions (items; see, e.g., Hambleton, Swaminathan, & Rogers, 1991). It is characterized by a collection of mathematical models in which a set of item parameters (discrimination, and/or difficulty, and/or guessing) is used to describe the relationship between a test taker's ability level and the probability of a test taker choosing a given response category. One of the most important features of IRT is the property of parameter invariance, which means that ability estimates based on different sets of items that measure the same trait are comparable by means of a linear transformation. Similarly, item parameter estimates are equivalent by means of a linear transformation when different groups of examinees are used.

Multiple operational test forms are being developed for both the TEFT and the TPK. The total raw scores from one test form, however, are not directly comparable to the ones from another form. Once a scale is established, subsequent test forms are placed on the same scale by equating scores on the new form to scores on a previously administrated form that has already been placed on the scale. Equating is a statistical procedure to adjust the form difficulty so that the test scores from different test forms have the same meaning and can be used interchangeably.

The ELTeach program performs equating within the IRT framework. A new test form is equated to the base form using the nonequivalent groups anchor test (NEAT) design. Under the NEAT design, one set of items is administered to one

group of test takers, another set of items is administered to a second group of test takers, and a third set of common items is administered to both groups. The procedures used for equating involve three steps: item calibration, scale linking, and true score equating. All analyses are performed using ETS's proprietary GENASYS II system.

Item Calibration

For both the TEFT and the TPK, the two-parameter logistic (2PL) model (Birnbaum, 1968) is used for all multiple-choice items, and the generalized partial credit model (GPCM) is utilized for constructed-response items. The 2PL model is characterized by two item parameters: item difficulty and item discrimination. In the 2PL model, the probability of a correct response ($x_i = 1$) for person j on item i, given ability θ_j, is denoted

$$P_{ij}\left(x_i = 1 \left| \theta_j \right.\right) = \left[\frac{\exp\left(a_i\left(\theta_j - b_i\right)\right)}{1 + \exp\left(a_i\left(\theta_j - b_i\right)\right)} \right],$$

where b_i is the item difficulty parameter and a_i is the item discrimination parameter for item i, which indicates the effectiveness of an item in separating examinees into different ability levels. Items with higher values for the discrimination parameter are more useful in categorizing examinees by ability levels than are ones with lower values for the discrimination parameter. The 2PL model, instead of the three-parameter logistic (3PL) model, was chosen because the 3PL model tends to encounter problems in estimating the pseudoguessing parameter. The computations with the 3PL model are the hardest when the guessing probabilities are not clearly differentiated from zero even if the 3PL model holds (Haberman, 2006; Hambleton et al., 1991). The less reliable c-parameter estimates would in turn affect the stability of a- and b-parameter estimates.

The GPCM was introduced by Muraki (1992) and is an extension of the 2PL model to the polytomous case. With the GPCM, there are $(m_i + 1)$ category scores (ranging from 0 to m_i) for item i and the probability of an examinee j with ability θ_j receiving a score x on item i is denoted as

$$P_{ix}\left(\theta_j\right) = \frac{\exp\left[\sum_{k=0}^{x} a_i\left(\theta_j - b_{ik}\right)\right]}{\sum_{h=0}^{m_i} \exp\left[\sum_{k=0}^{h} a_i\left(\theta_j - b_{ik}\right)\right]},$$

where a_i is the item discrimination, which is "the degree to which the categorical responses vary among items as θ level changes" (see Muraki, 1992). The variable b_{ik} is the step difficulty between one category of m_i and the next category.

Scale Linking

After the calibration, item parameter estimates for a new test form are linked to the previously obtained base scale using the common set of items between forms. Specifically, the test characteristic curve method (Stocking & Lord, 1983) is used to link item parameters on the new form to the scale of the base form. With the property of parameter invariance of IRT, the Stocking-Lord method seeks to estimate linear transformation coefficients that minimize the squared differences in test characteristic curves across all common items. The transformation process (i.e., scale linking) may require more than one run if any equating items show large differences in item characteristic curves between the new and the base form and are removed from the equating set. After the final run, the resulting linear transformation coefficients are applied to place all items on the base scale for both the TEFT and the TPK.

True Score Equating

Once the new item parameter estimates are placed on the common scale of the base form, ability estimates that are based on item parameter estimates are comparable. The IRT true score equating is then implemented to equate the new form to the scale of the base form. The equating relates true scores on a new test form with true scores on the base form using the

IRT definition of true score. The procedure is conducted in three stages: specifying a true score on the new form for which the corresponding true score on the base form is desired; determining the ability level which corresponds to the given true score on the new form; and determining the true score on the base form, which corresponds to the ability level. Although no theoretical justification exists for applying the true score relationships to observed scores, this is often conducted in practice (Kolen & Brennan, 2004). As a result, raw scores for the two test forms yielded by the same ability estimate are assumed to be equivalent. A test taker's equated raw score on the new form, then, is the raw score on the reference form. Conversion of test takers' equated raw scores to the base scale is accomplished by applying the raw-to-scale score conversion table originally obtained from the base form.

Dimensionality Analysis and Scale Anchoring Study

Pilot testing of the TEFT and the TPK was conducted in targeted countries in Fall 2012 as reported by Freeman et al. (2013). Data from 1,303 examinees who took the TEFT and 1,091 examinees who took the TPK in that pilot administration were used in a dimensionality study in order to evaluate the underlying structure of these assessments. In this study, item-level data were analyzed for both assessments in LISREL using confirmatory factor analysis with diagonally weighted least squares. Note that running confirmatory factor analysis using different statistical software packages can and does occasionally lead to different results. For the TEFT, several latent factor models were tested: one factor, two factors based on oral and written constructs, three factors based on different content areas (managing the classroom; understanding and communicating lesson content; providing feedback), and four factors based on language modality. There were no appreciable gains in model fit beyond the one-factor model; although not all of the latent factor correlations exceeded 0.90, all were reasonably close to this criterion.[6] For the TPK, the initial hypothesis was that a single latent factor fit the data as well as a two-factor model based on content areas (foundations of PK; core teaching skills). In the two-factor model, the correlation of the latent factors was 0.97, indicating that a one-factor model was appropriate. In addition, there was no improvement in the comparative fit index, which was well above the 0.95 threshold for acceptability (Byrne, 2006). Thus, for each assessment, reporting one overall score seemed to be well supported by the pilot data.

A scale anchoring study was conducted to define score bands and to develop performance descriptors. For this study, we implemented the direct scale anchoring method, which uses the proportions of successful responses to items at different score levels (Beaton & Allen, 1992). The scale anchoring process was similar for each assessment and involved several steps: examining the distribution of raw scores, mapping each item to the raw score scale, grouping raw scores into bands, providing average item scores for examinees within each score band, and creating score descriptors from the items mapped into each band. Items were mapped to raw score points using 65% of the maximum score points for constructed-response items and 70% of the maximum for multiple-choice items. A team of assessment developers, psychometricians, and researchers created three score ranges for each assessment and verbal descriptors for the score bands. These verbal descriptors are the result of a thorough analysis of test-taker performance on each of the test tasks and the relation this has to the tests' overall claims and subclaims. For example, the highest of the TPK score bands reads as follows:

> A typical test taker in Band Three **demonstrates comprehensive knowledge** of **all** content areas in the curriculum. A test taker in Band Three **consistently** identifies concepts and connects them to examples of ELT in classroom situations presented in the Professional Knowledge for ELT course.[7]

Automated Scoring for the Test of English-for-Teaching (TEFT) Assessment

One of the design goals of the TEFT was to produce an assessment design within which the speaking and writing tasks would be amenable to automated scoring. This goal is consistent with the intention to design the ELTeach program to be suitable for professional development purposes and not as a factor in higher-stakes decisions such as teacher certification. To support this goal, the TEFT assessment development team consulted regularly with ETS natural language processing research scientists throughout the process of designing the assessment task types. The goal was to develop speaking and writing task types that maximize the likelihood that valid and reliable models for automated scoring can be built while ensuring appropriate measurement of the construct, regardless of the approach used for scoring.

ETS is developing a new automated scoring system for TEFT writing responses while building customized models for spoken responses, leveraging its existing engine, *SpeechRater*SM. For written responses, features based on string edit distance metrics and context-free grammars that measure content and grammar accuracy in the responses are used to predict scores; whereas, for spoken responses, features related to fluency, pronunciation, prosody, vocabulary, grammar, and content accuracy are used to predict scores for each item. Furthermore, SpeechRater employs various filtering models to flag responses that should not be scored automatically (e.g., due to a high noise level), and such responses are routed to and scored by human raters during operational scoring. The TEFT was launched operationally using human scoring, but the plan is to transition to automated scoring with appropriate checks and controls when that occurs.

Score Reporting

In designing the score reporting materials for ELTeach, two basic goals were set: to ensure that the score reports and associated materials would be appropriate for the purposes described in the introduction to this report and to ensure that they would reflect the organization and values of the ELTeach program as it had been designed. The ELTeach test-taker score reports, summary score reports, and various supporting documents were all designed with these two goals in mind, as well as to provide this information in a format that is easy to access and comprehend.[8]

Test-Taker Score Reports

As described above, one major purpose of the ELTeach program is to support teachers in improving their command of the language and skills needed to teach English in English (through the English-for-Teaching course) and/or their PK (through the Professional Knowledge in ELT course). The most prominent score information on these reports is the total scaled score and the associated bands and band descriptors. The overall purpose of the total scaled score, the bands, and the band descriptors is to provide statistically reliable performance information and also descriptive information that allows teachers and other score users to meaningfully interpret the scores.

The total scaled score for the TEFT ranges from 400 to 700 in increments of 10; the total scaled score for the TPK ranges from 100 to 300 in increments of 10. For each assessment, this total scaled score is the score information that is most psychometrically reliable and thus is the most appropriate basis for any decisions to be made based on ELTeach results. (ELTeach is designed to support low-stakes decisions such as those related to professional development.)

In order to provide meaning to these scores, scale anchoring studies were conducted as described above. These scale anchoring studies produced the bands and band descriptors which accompany the total scaled scores. The bands themselves help score users to understand the total scaled score by providing a numerical context for each score. Of greater interest and value to score users, however, are the band descriptors, which complement the numerical score information with descriptive information.

The language of the descriptors themselves gives teachers clear and detailed information about how their performance on the assessment relates to the ECD claims (see the "Construct Definitions" section in this paper). For the TEFT, the band descriptors for the total scaled scores are presented as a table, providing information about how teachers in Band One, Band Two, and Band Three typically performed in relation to the three categories around which the claims, the learning materials, and the TEFT are organized: managing the classroom, understanding and communicating lesson content, and providing feedback. For the TPK, the band descriptors reflect the TPK claim, describing three levels of command over the knowledge presented in the PK learning materials and included on the TPK assessment.

It is worth noting that the bands do not cover the entire scale; instead, there is a gap between each band and the one above and/or below it. This is because bands with such gaps proved to be better defined for providing detailed descriptive information of teacher performance than broader bands without gaps would have been. A note is provided on each score report to clarify where the teacher's performance locates them. (For example, a teacher receiving a total scaled score of 550 on the TEFT receives a note saying, "Your score of 550 is between Band Two and Band Three. This score indicates that your performance shares the characteristics of Band Two and may have one or more of the characteristics of Band Three.")

In designing the score reports, a conscious choice was made to provide the text of all band descriptors to each test taker, regardless of the individual's score. This enables teachers receiving the score report to review all of the descriptors, giving

them a sense of the overall scale rather than just the band in which they scored and providing a context for them to set targets for continuing professional improvement.

Some additional score information is also included on the test-taker score reports. The nature of this information is somewhat different for the TEFT and the TPK, reflecting differences in the frameworks, the learning materials, and the assessments for each. For both assessments, this additional information is given less prominence on the score report, as it is not as psychometrically robust as the total scaled scores. While the total scaled scores are appropriate to be used for the purposes of the assessments identified above, this additional information is provided primarily to complement the professional development uses of the scaled scores.

The TEFT test-taker score report contains two categories of additional information. Nonequated scaled scores are provided based on items that assess skills in listening, speaking, reading, and writing. These scales range from 40 to 70 in 5-point increments. A note is included to clarify that these scores refer to skills only as used to complete the tasks on the TEFT and do not provide information about general proficiency in these skill areas. Raw scores are also provided reflecting student performance on the three content areas of the assessment (managing the classroom, understanding and communicating lesson content, and providing feedback). In addition to the teacher's raw score obtained, the 25th and 75th percentiles for each category are marked to provide reference points.

The TPK test-taker score report contains additional score information for the two major categories into which TPK items are organized: foundations of PK and core teaching practices. As with the TEFT content areas, the teacher's raw score obtained is reported, with the 25th and 75th percentiles marked for reference.

Summary Score Reports

A second major purpose of the ELTeach program (also described above), is to provide MOEs, LEAs, or other institutional score users with analytic information about learning and performance. Specifically, the ELTeach program is designed to provide institutional users and educational authorities with information

- to document the overall PK and language skills of groups of teachers in the content of the ELTeach program curriculum;
- to document the degree to which candidates within preservice teacher preparation or training programs have developed the PK and language skills presented in the ELTeach curricula; and
- to help meet policy and accountability goals and benchmarks in professional development planning in ELT.

To satisfy these goals, summary score reports are provided to institutional score users. These summary score reports provide a combination of performance summaries and information about the performance of individuals for teachers taking the assessments during a given administration window.

The first section of the summary score report provides summary information related to the total scaled scores and bands overall number of test takers, their mean scores, and how many teachers scored in each of the seven positions relative to the bands (above Band Three, within Band Three, between Band Three and Band Two, etc.). This is followed by text of the band descriptors themselves.

The next section of the summary score report provides mean scores for those categories listed as additional information on the test-taker score report. Next, a graph is provided showing the distribution of teachers into the seven positions relative to the band. The final section of the summary report contains roster information, providing total scaled score, band achieved, and other high-level information for each teacher in the administration.

Supporting Documents

On the public ELTeach.com website, information is provided to help teachers and other score users interpret their scores. These materials include annotated versions of each score report, providing further explanations of each section of each score report.

Research Agenda

This section outlines the broad organization of a research agenda for the ELTeach program. As described above, the ELTeach program has two overarching purposes: (a) to support teachers in improving their command of the language

and skills needed to teach English in English (through the English-for-Teaching course) and/or their PK (through the Professional Knowledge in ELT course), while (b) providing institutional users with analytic information about learning and performance. Therefore, in this section we acknowledge that the program presents a unique research ecology that combines data on teachers' use of learning materials with data on their performance on the aligned assessments. At the same time, it is appropriate for assessment framework documents to consider current thinking on test validation in the fields of educational measurement and second language assessment in order to point to directions for research to support the use of scores from the TEFT and the TPK.

Recent validity frameworks (e.g., M. Kane, 2013, 2006) draw on Toulmin's argument structure (Toulmin, 2003), which consists of making claims based on data. M. Kane's seminal paper (1992) presented earlier approaches to argument-based validation (Cronbach, 1988; House, 1980) by employing two kinds of arguments. First, the *interpretive argument* specifies the proposed interpretations and uses of assessment results. This is done through a network of inferences and assumptions from the observed performances to the conclusions and decisions based on the assessment scores. More recently, M. Kane (2013) referred to an *interpretation/use argument* (IUA) to emphasize that the interpretive argument is about both interpretations and uses of test scores. Second, the *validity argument* evaluates the interpretive argument's coherence and the plausibility of the inferences and assumptions (M. Kane, 1992, 2006, 2013).

Building on Kane's work, the *assessment use argument* (AUA; Bachman, 2005; Bachman & Palmer, 2010) provides a framework for test developers and decision makers to justify the intended uses of a particular assessment. According to Kenyon (2012), the AUA framework made Toulmin's approach widely known in the field of language assessment. When an AUA is used for the selection of an assessment or the development of a new one, claims about the intended consequences, decisions, and interpretations will need to be stated and supported by warrants so that test developers and decision makers are accountable to stakeholders (Bachman & Palmer, 2010, p. 92). Although the inclusion of consequences in argument-based validity has been challenged in the educational measurement literature (Cizek, 2012), the AUA reflects the attention of the language testing community to the consequences of test uses, primarily because of Messick's (1989, 1996) influential work, as McNamara (2006) points out, and also because of the impact of tests on learning and teaching through washback investigations (Wall, 2000; Wall & Alderson, 1993).

This section provides the general parameters of a potential research agenda within the ecology discussed earlier: that is, that assessments are aligned and inextricably linked to the learning materials. At the same time, and in accordance with the recent approaches to validity described above (IUA or AUA), a future research agenda should consider the core validation questions that should be addressed by a testing program in order to justify and support claims about

- the interpretations of test scores;
- the decisions to be made on the basis of test scores; and
- the intended consequences of the use of the assessment.

One possibility for the organization of the proposed research agenda is to not only consider but also fully adopt the categorization of an IUA or an AUA. However, because the ELTeach program brings together two components—learning materials and assessments—the proposed agenda is structured along two broad foci for research to emphasize the designed interrelationship of these components: Those studies that focus within each of the components, called here *intracomponent studies*, and those studies that focus on the relations between components or between the program and its use environments, called here *intercomponent studies*. Intracomponent studies examine how each component functions in relation to how it was designed and how it is implemented. When stating research questions to support claims about the TEFT and TPK assessments, these two research foci might not be mutually exclusive. Therefore, it may be more accurate to talk about *primarily* intracomponent studies and primarily intercomponent studies. To support the usefulness of the assessments, both foci are important for future studies. The assessments first have to be of high psychometric quality; otherwise, they cannot support the aims of the ELTeach program. The psychometric quality is assessed by studies with an intracomponent focus. Studies with an intercomponent focus will need to examine whether the claimed relationship between assessments and the learning materials is upheld; without evidence of the relationship between the two components, the claim of usefulness of the ELTeach program is not warranted.

The purposes discussed in the beginning of this section suggest a set of high-level core research questions, some of which focus on claims about the learning outcomes of the program, while others focus on the learning processes designed into both the learning materials and the assessments. Some research topics for future studies with intracomponent and

intercomponent foci are presented below. This list of topics is not exhaustive, and it is intended as a guide for studies whose results can support the claims made about the ELTeach program.

Possible Studies With an Intracomponent Focus

- How well can the TEFT and TPK assessments predict performance in relation to other measures of the same ability?
- To what extent do the assessments document teachers' knowledge or performance in a useful way?
- To what extent do the assessments engage relevant teacher knowledge as specified in the test design claims?

Possible Studies With an Intercomponent Focus

Learning Outcomes

- To what extent does the ELTeach program (learning materials coupled with assessments) help teachers improve their functional command of classroom language and/or their PK?
- To what extent do the learning materials and the assessments for each course accomplish the outcomes claimed in their respective design frameworks?
- To what extent have teachers improved the quality of their teaching after successfully completing the ELTeach program?
- To what extent do teachers feel more confident in teaching EFL in English after successfully completing the ELTeach program?

Learning Processes

- How does each course accomplish the outcomes claimed in its design framework? How do the various activities support teachers' learning and performance according to the claims?
- How do the learning materials support teachers' performance on the aligned assessments?

Summary

In summary, the purpose of this document was to describe the design and development of the assessments associated with the ELTeach integrated program of learning—the TEFT and the TPK. We described the major aspects of the two assessments: (a) the purposes and intended uses for test scores/certificates, (b) the target populations, (c) the construct definitions, (d) the program frameworks and domain descriptions, (e) test designs, (f) scoring systems, and (g) directions for research in support of the program. In documenting the design and development of these assessments, we seek to ensure that the assessment development process meets the professional standards for the intended uses of the assessments and that the assessments are appropriate for their intended uses. Additionally, this document serves as a reference for future investigations of test validity that may be required to support the intended test uses as part of the ELTeach program.

Acknowledgments

The authors would like to acknowledge the contributions to this report from their ETS and NGL colleagues: Anne Katz, Anne Burns, Harold Van Hise, Sultan Turkan, Eric Steinhauer, Maria Konkel, Rick Morgan, Tsung-Han Ho, Yoko Futagi, and Derrick Higgins.

Notes

1 With specific regard to teacher language proficiency, current conceptions of such proficiency in assessment include contexts of language use and communicative language competence, which encompasses all the other underlying competencies (strategic, sociolinguistic, and grammatical; Butler, 2004). However, Elder (2001) questioned how the tests of such proficiency could elicit a

sample of language use that essentially represents "teacherliness" differently from the ways in which general proficiency tests would. In response, Elder implicitly noted that the construct of teacher language proficiency, though quite complex, could be differentiated from general language proficiency through defining the participants, channel of the input, expected response, and the nature of the input-response relationship.

2 The design and sequencing of the learning materials embeds concepts and tasks relevant to professional commitment. Therefore, as teachers work through the course they are both developing and expanding their understanding of this area; in the TPK, these areas are documented in a number of items.

3 The raw scores on the TEFT range from 0 to 189.

4 An additional indication of the generally strong performance is that fewer than 1% of test takers received a raw score below 80.

5 The raw scores on the TPK range from 0 to 62.

6 During the design process, several approaches to defining scaled scores for the TEFT were explored, including having two scaled scores (e.g., one for oral language and one for written language) and having separate scores for each of the four skills. Once pilot data had been analyzed, the decision to report a single scaled score was made because (a) it was in line with the construct (the general focus was on the command of English to perform specific, identified tasks, rather than language skills per se); (b) statistical analysis and dimensionality studies provided strong support for it; (c) it allowed us to provide feedback on the three content areas as ranges in the total scale and with descriptive information in the band descriptors; and (d) it was parallel to the reporting approach taken for the TPK.

7 Performance descriptors for both TEFT and TPK can be found in the sample score reports in Appendix B.

8 Sample score reports can be found in Appendix B.

References

Abell, S. K., Rogers, M. P., Hanuscin, D. L., Lee, M. H., & Gagnon, M. J. (2009). Preparing the next generation of science teacher educators: A model for developing PCK for teaching science teachers. *Journal of Science Teacher Education*, *20*(9), 77 – 93.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, *2*(1), 1 – 34.

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

Ball, D. L., Hill, H. C., & Bass, H. (2005). Knowing mathematics for teaching: Who knows mathematics well enough to teach third grade, and how can we decide? *American Educator*, *29*, p. 4 – 17, 20 – 22, 43 – 46.

Ball, D. L., Thames, M., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*(5), 389 – 407.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational and Behavioral Statistics*, *17*, 191 – 204.

Ben-Peretz, M. (2011). Teacher knowledge: What is it? How do we uncover it? What are its implications for schooling? *Teaching and Teacher Education*, *27*, 3 – 9.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397 – 472). Reading, MA: Addison-Wesley.

Butler, Y. G. (2004). What level of English proficiency do elementary school teachers need to attain to teach EFL? Case studies from Korea, Taiwan, and Japan. *TESOL Quarterly*, *38*(2), 245 – 278.

Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Mahwah, NJ: Erlbaum.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, *17*(1), 31 – 43.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, *25*(2), 119 – 142.

Connelly, M., & Clandinin, D. J. (1988). *Teachers as curriculum planners: Narratives of experience*. New York, NY: Teachers College Press.

Consolo, D. A. (2006). On a (re)definition of oral language proficiency for EFL teachers: Perspectives and contributions from current research. *Melbourne Papers in Language Testing*, *11*(1), 1 – 28.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3 – 17). Hillsdale, NJ: Lawrence Erlbaum.

Dewey, J. (1920). *Reconstruction in philosophy*. New York, NY: Henry Holt.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge, England: Cambridge University Press.

Dudley-Evans, T., & St John, M. J. (1998). *Developments in English for specific purposes*. Cambridge, England: Cambridge University Press.

Dudley-Marling, C., Abt-Perkins, D., Sato, K., & Selfe, R. (2006). Teacher quality: The perspective of NCTE members. *English Education*, *38*, 167 – 193.

Elbaz, F. (1983). *Teacher thinking: A study of practical knowledge*. New York, NY: Nichols.

Elder, C. (2001). Assessing the language proficiency of teachers: Are there any border controls? *Language Testing*, *18*(2), 149 – 170.

Engestrom, Y. (2001). Expansive learning at work: Toward an activity theoretical reconceptualization. *Journal of Education and Work*, *14*(1), 133 – 156.

Freeman, D. (1996). The 'unstudied problem': Research on teacher learning in language teaching. In D. Freeman & J. C. Richards (Eds.), *Teacher learning in language teaching* (pp. 351 – 377). New York, NY: Cambridge University Press.

Freeman, D., & Johnson, K. E. (1998). Reconceptualizing the knowledge-base of language teacher education. *TESOL Quarterly*, *32*(3), 397 – 417.

Freeman, D., Katz, A., Le Dréan, L., Burns, A., & Hauck, M. (2013). *ELTeach global pilot report 2012*. Retrieved from http://elteach.com/ELTeach/media/Documents/ELTeach_GPR_9-20-13.pdf

Graddol, D. (2006). *English next: Why global English may mean the end of 'English as a foreign language.'* London, England: British Council.

Graddol, D. (2010). *English next: India*. London, England: British Council.

Grossman, P. L. (1990). *The making of a teacher: Teacher knowledge and teacher education*. New York, NY: Teachers College Press.

Haberman, S. J. (2006). *An elementary test of the normal 2PL model against the normal 3PL alternative* (Research Report No. RR-06-14). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2006.tb02020.x

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. London, England: Sage.

Harris, L., & Bain, R. (2011). Pedagogical content knowledge for world history teachers: What is it? How might prospective teachers develop it? *The Social Studies*, *102*(1), 9 – 17.

House, E. R. (1980). *Evaluating with validity*. Beverly Hills, CA: Sage.

Hutchinson, T., & Waters, A. (1987). *English for specific purposes: A learning-centered approach*. New York, NY: Cambridge University Press.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527 – 535.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1 – 73.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17 – 64). Westport, CT: American Council on Education/Praeger.

Kelly, L. G. (1969). *Twenty-five centuries of language teaching*. Rowley, MA: Newbury House.

Kenyon, D. M. (2012). Using Bachman's assessment use argument as a tool in conceptualizing the issues surrounding linking ACTFL and CEFR. In E. Tschirner (Ed.), *Aligning frameworks of reference in language testing: The ACTFL proficiency guidelines and the common european framework of reference for languages* (pp. 23 – 34). Tübingen, Germany: Stauffenburg Verlag.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, linking, and scaling: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Larsen-Freeman, D., & Freeman, D. (2008). Language moves: The place of "foreign" languages in classroom teaching and learning. *Review of Research in Education*, *32*(1), 147 – 186.

Magnusson, S., Krajcik, J., & Borko, H. (1999). Nature, sources, and development of pedagogical content knowledge for science teaching. In J. Gess-Newsome & N. G. Lederman (Eds.), *Examining pedagogical content knowledge: The construct and its implication for science education* (pp. 95 – 132). Dordrecht, The Netherlands: Kluwer Academic.

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly*, *3*(1), 31 – 51.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13 – 103). New York, NY: Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(4), 241 – 256.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika*, *59*, 439 – 483.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence centered design* (Research Report No. RR-03-16). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2003.tb01908.x

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *Evidence-centered assessment design*. Princeton, NJ: Educational Testing Service.

Monte-Sano, C., & Budano, C. (2013). Developing and enacting pedagogical content knowledge for teaching history: An exploration of two novice teachers' growth over three years. *The Journal of the Learning Sciences*, *22*(2), 171 – 211.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159 – 176.

Sešek, U. (2007). English for teachers of EFL — Toward a holistic description. *English for Specific Purposes*, *26*(4), 411 – 425.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1 – 21.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201 – 210.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge, England: Cambridge University Press.

Turner-Bisset, R. (1999). The knowledge bases of the expert teacher. *British Educational Research Journal*, *25*(1), 39 – 55.

Van Driel, J. H., Verloop, N., & de Vos, W. (1998). Developing science teachers' pedagogical content knowledge. *Journal of Research in Science Teaching*, *35*, 673 – 695.

Wall, D. (2000). The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled? *System*, *28*(4), 499 – 509.

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing*, *10*(1), 41 – 69.

Wilson, S., Shulman, L., & Richert, A. (1987). '150 different ways' of knowing: Representations of knowledge in teaching. In J. Calderhead (Ed.), *Exploring teachers' thinking* (pp. 104 – 124). London, England: Cassel.

Zakeri, A., & Alavi, M. (2011). English language teachers' knowledge and their self-efficacy. *Journal of Language Teaching and Research*, *2*, 413 – 419.

Appendix A
Representative Examples of English-for-Teaching (EFT) Teacher Tasks and Associated Subclaims

Table A1 Tasks and Associated Subclaims of English-for-Teaching

Task	Subclaim	Functional Area	Skill
Reading Task 1 — Reads student content in instructional materials, (e.g., reading passages, instructions for students)	Reading Subclaim 1 — Can comprehend written instructional materials for the purpose of preparing and enacting a lesson • instructions, explanations, exemplifications, exercises, and answer keys in students' textbooks • texts in students' textbooks (written for different purposes, such as letters, articles, signs, posters, stories, menus, etc.)	Understanding and communicating lesson content	Reading
Speaking Task 3 — Delivers instructions in limited formulaic language to organize and manage different types of classroom activities	Speaking Subclaim 2 — Can produce formulaic language when enacting a lesson for the purposes of managing the classroom (giving instructions, starting activities, organizing students into groups, announcing lesson goals, etc.)	Managing the classroom	Speaking
Writing Task 4 — Corrects word form and usage errors in student-written text based on instructional materials	Writing Subclaim 4 — Can provide written feedback on student output by using formulaic language	Providing feedback	Writing
Listening Task 2 — Listens to instructional materials within the curriculum in preparation for classroom instruction, and whenever necessary, follows transcript while listening to support comprehension	Listening Subclaim 1 — Can comprehend spoken instructional materials for the purpose of preparing and enacting a lesson • exemplifications, instructions, explanations, and exercises in students' audio materials • audio components in students' materials (recorded for different purposes, such as dialogues, telephone conversations, stories, narratives, songs, news reports, etc.)	Understanding and communicating lesson content	Listening

Appendix B
Sample Score Reports for Test of English-for-Teaching (TEFT) and Test of Professional Knowledge (TPK)

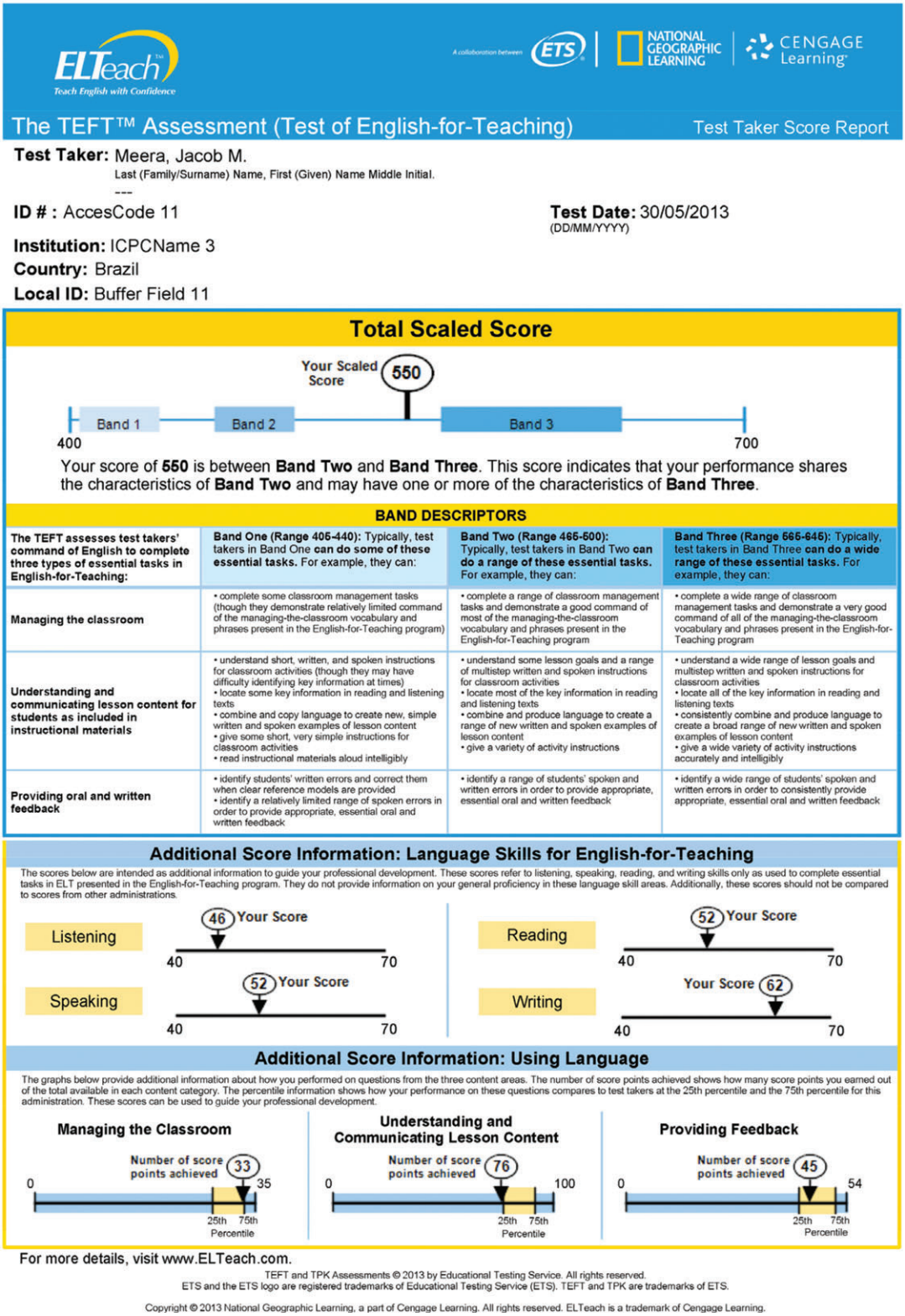
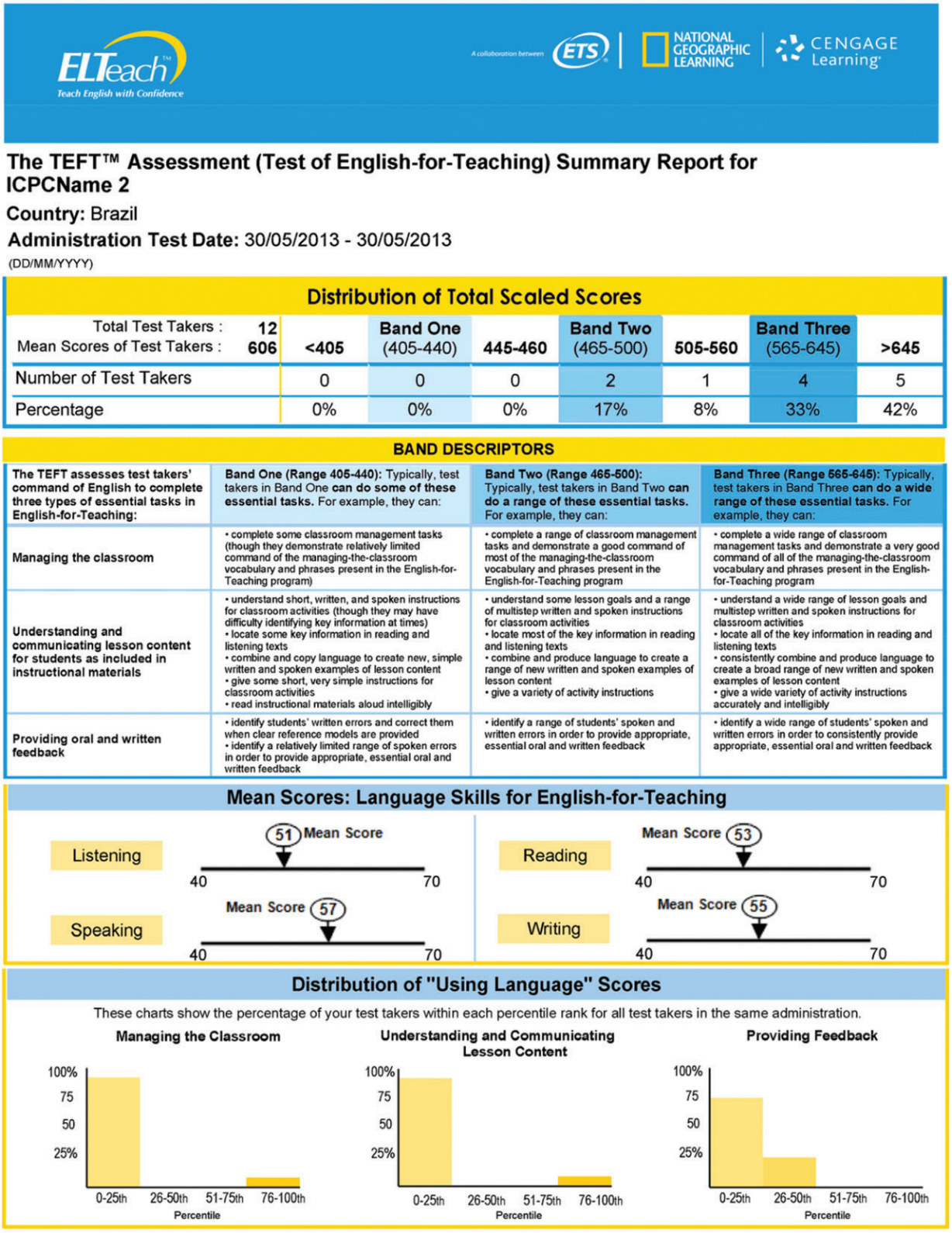


Figure B1 Test of English-for-Teaching (TEFT) test-taker score report.



ELTeach™ Teach English with Confidence

A collaboration between ETS | NATIONAL GEOGRAPHIC LEARNING | CENGAGE Learning

The TEFT™ Assessment (Test of English-for-Teaching) Summary Report for ICPCName 2

Country: Brazil

Administration Test Date: 30/05/2013 - 30/05/2013

(DD/MM/YYYY)

Distribution of Total Scaled Scores

Total Test Takers : 12 Mean Scores of Test Takers : 606	<405	Band One (405-440)	445-460	Band Two (465-500)	505-560	Band Three (565-645)	>645
Number of Test Takers	0	0	0	2	1	4	5
Percentage	0%	0%	0%	17%	8%	33%	42%

BAND DESCRIPTORS

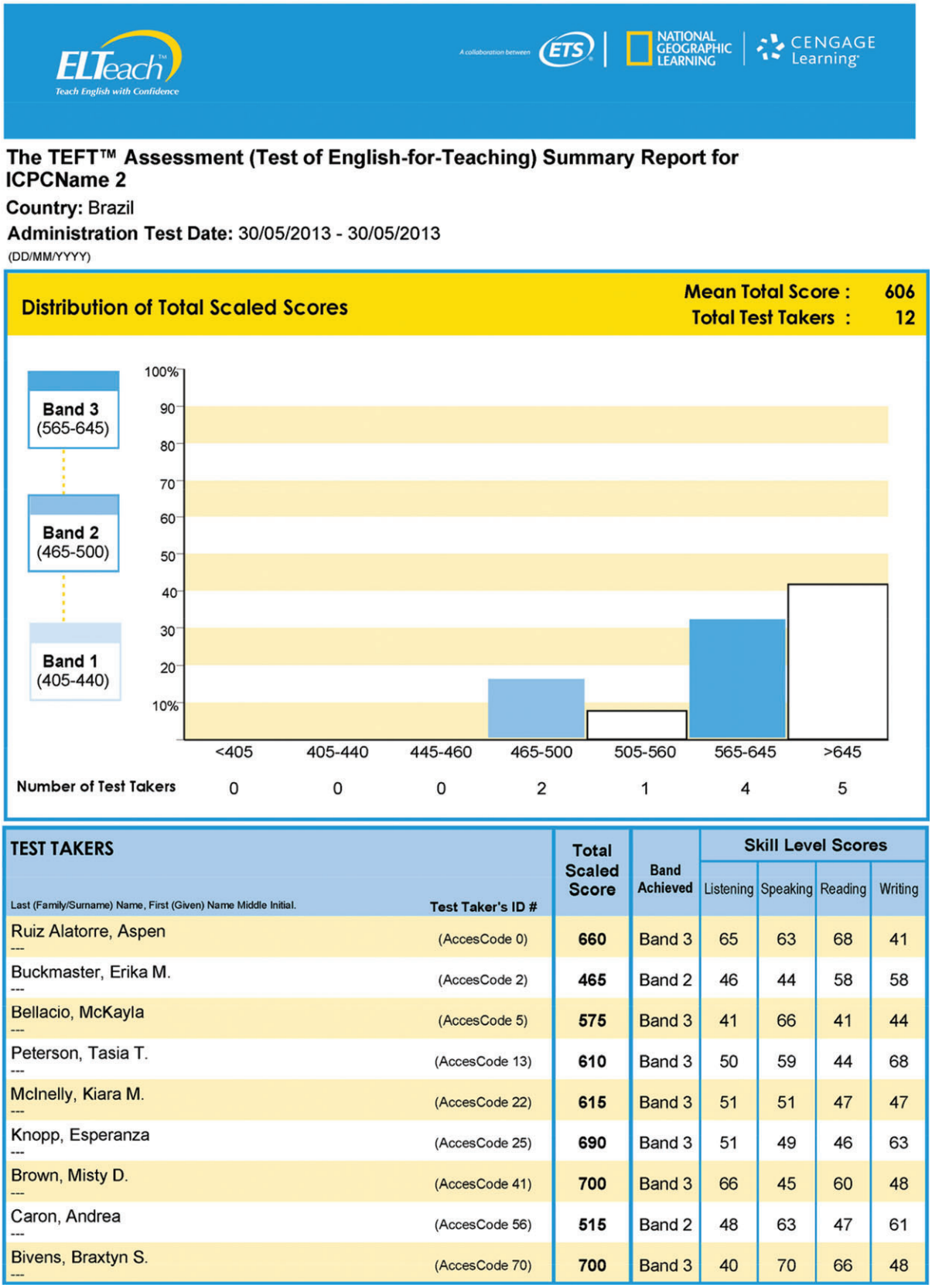
The TEFT assesses test takers' command of English to complete three types of essential tasks in English-for-Teaching:	**Band One (Range 405-440):** Typically, test takers in Band One **can do some of these essential tasks.** For example, they can:	**Band Two (Range 465-500):** Typically, test takers in Band Two **can do a range of these essential tasks.** For example, they can:	**Band Three (Range 565-645):** Typically, test takers in Band Three **can do a wide range of these essential tasks.** For example, they can:
Managing the classroom	• complete some classroom management tasks (though they demonstrate relatively limited command of the managing-the-classroom vocabulary and phrases present in the English-for-Teaching program)	• complete a range of classroom management tasks and demonstrate a good command of most of the managing-the-classroom vocabulary and phrases present in the English-for-Teaching program	• complete a wide range of classroom management tasks and demonstrate a very good command of all of the managing-the-classroom vocabulary and phrases present in the English-for-Teaching program
Understanding and communicating lesson content for students as included in instructional materials	• understand short, written, and spoken instructions for classroom activities (though they may have difficulty identifying key information at times) • locate some key information in reading and listening texts • combine and copy language to create new, simple written and spoken examples of lesson content • give some short, very simple instructions for classroom activities • read instructional materials aloud intelligibly	• understand some lesson goals and a range of multistep written and spoken instructions for classroom activities • locate most of the key information in reading and listening texts • combine and produce language to create a range of new written and spoken examples of lesson content • give a variety of activity instructions	• understand a wide range of lesson goals and multistep written and spoken instructions for classroom activities • locate all of the key information in reading and listening texts • consistently combine and produce language to create a broad range of new written and spoken examples of lesson content • give a wide variety of activity instructions accurately and intelligibly
Providing oral and written feedback	• identify students' written errors and correct them when clear reference models are provided • identify a relatively limited range of spoken errors in order to provide appropriate, essential oral and written feedback	• identify a range of students' spoken and written errors in order to provide appropriate, essential oral and written feedback	• identify a wide range of students' spoken and written errors in order to consistently provide appropriate, essential oral and written feedback

Mean Scores: Language Skills for English-for-Teaching

Distribution of "Using Language" Scores

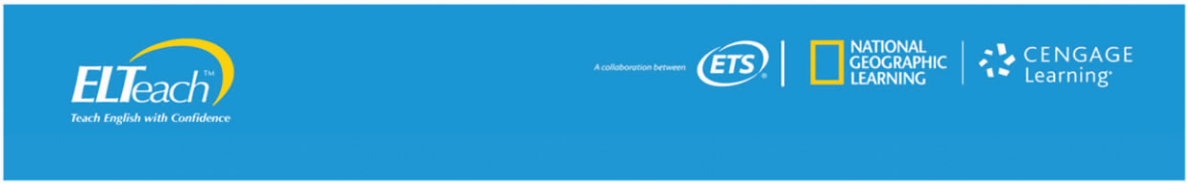
These charts show the percentage of your test takers within each percentile rank for all test takers in the same administration.

Figure B2 Test of English-for-Teaching (TEFT) summary score report, page 1.



TEST TAKERS Last (Family/Surname) Name, First (Given) Name Middle Initial.	Test Taker's ID #	Total Scaled Score	Band Achieved	Skill Level Scores			
				Listening	Speaking	Reading	Writing
Ruiz Alatorre, Aspen ---	(AccesCode 0)	**660**	Band 3	65	63	68	41
Buckmaster, Erika M. ---	(AccesCode 2)	**465**	Band 2	46	44	58	58
Bellacio, McKayla ---	(AccesCode 5)	**575**	Band 3	41	66	41	44
Peterson, Tasia T. ---	(AccesCode 13)	**610**	Band 3	50	59	44	68
McInelly, Kiara M. ---	(AccesCode 22)	**615**	Band 3	51	51	47	47
Knopp, Esperanza ---	(AccesCode 25)	**690**	Band 3	51	49	46	63
Brown, Misty D. ---	(AccesCode 41)	**700**	Band 3	66	45	60	48
Caron, Andrea ---	(AccesCode 56)	**515**	Band 2	48	63	47	61
Bivens, Braxtyn S. ---	(AccesCode 70)	**700**	Band 3	40	70	66	48

Figure B3 Test of English-for-Teaching (TEFT) summary score report, page 2.



The TEFT™ Assessment (Test of English-for-Teaching) Summary Report for ICPCName 2

Country: Brazil

Administration Test Date: 30/05/2013 - 30/05/2013

(DD/MM/YYYY)

TEST TAKERS		Total Scaled Score	Band Achieved	Skill Level Scores			
Last (Family/Surname) Name, First (Given) Name Middle Initial.	Test Taker's ID #			Listening	Speaking	Reading	Writing
Ronald, Bill M. ---	(AccesCode 74)	**570**	Band 3	45	60	46	58
Reagan, Clinton M. ---	(AccesCode 75)	**690**	Band 3	51	59	44	62
Armstrong, Meera M. ---	(AccesCode 92)	**485**	Band 2	54	50	66	61

Figure B4 Test of English-for-Teaching (TEFT) summary score report, page 3.

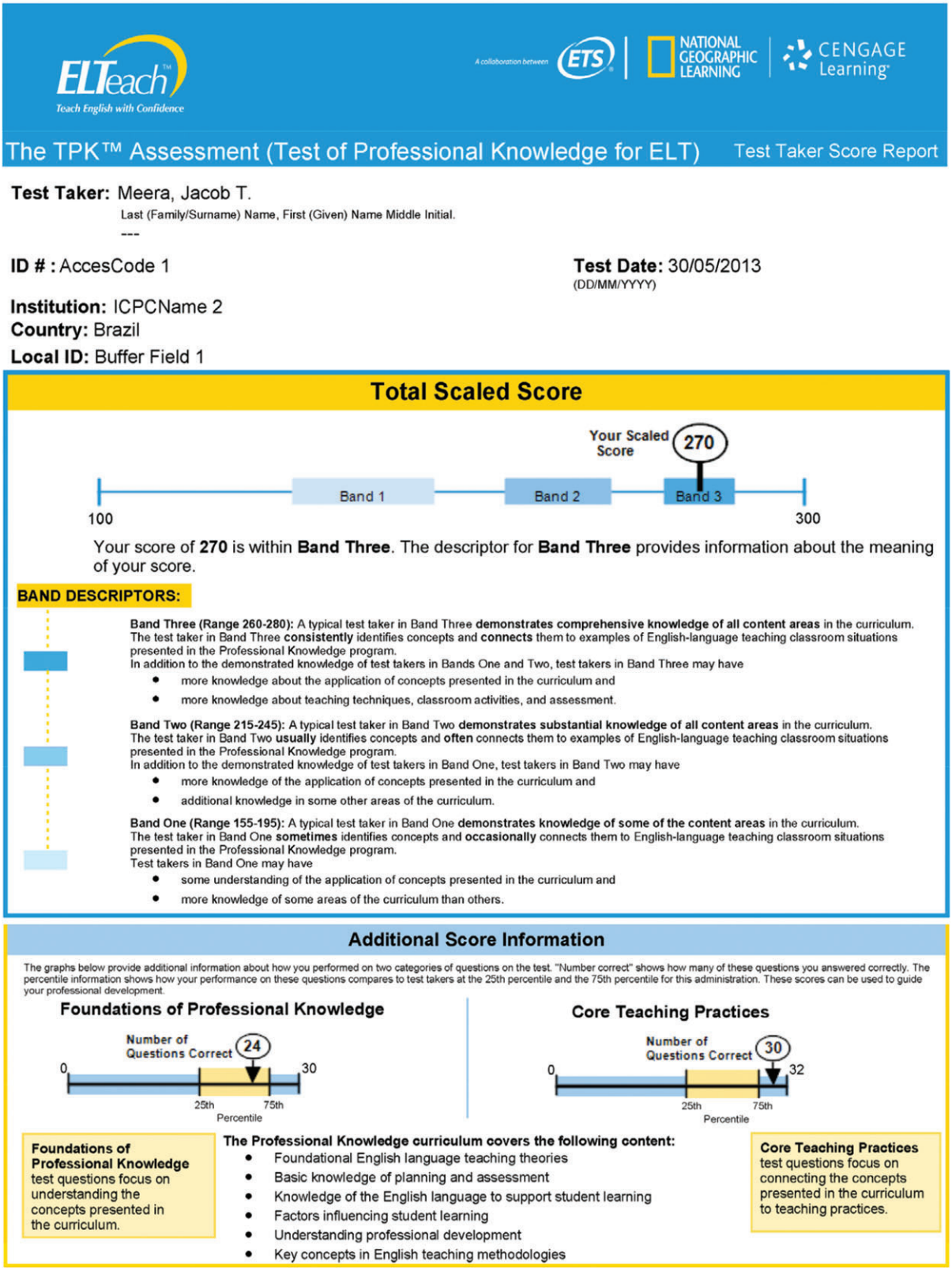
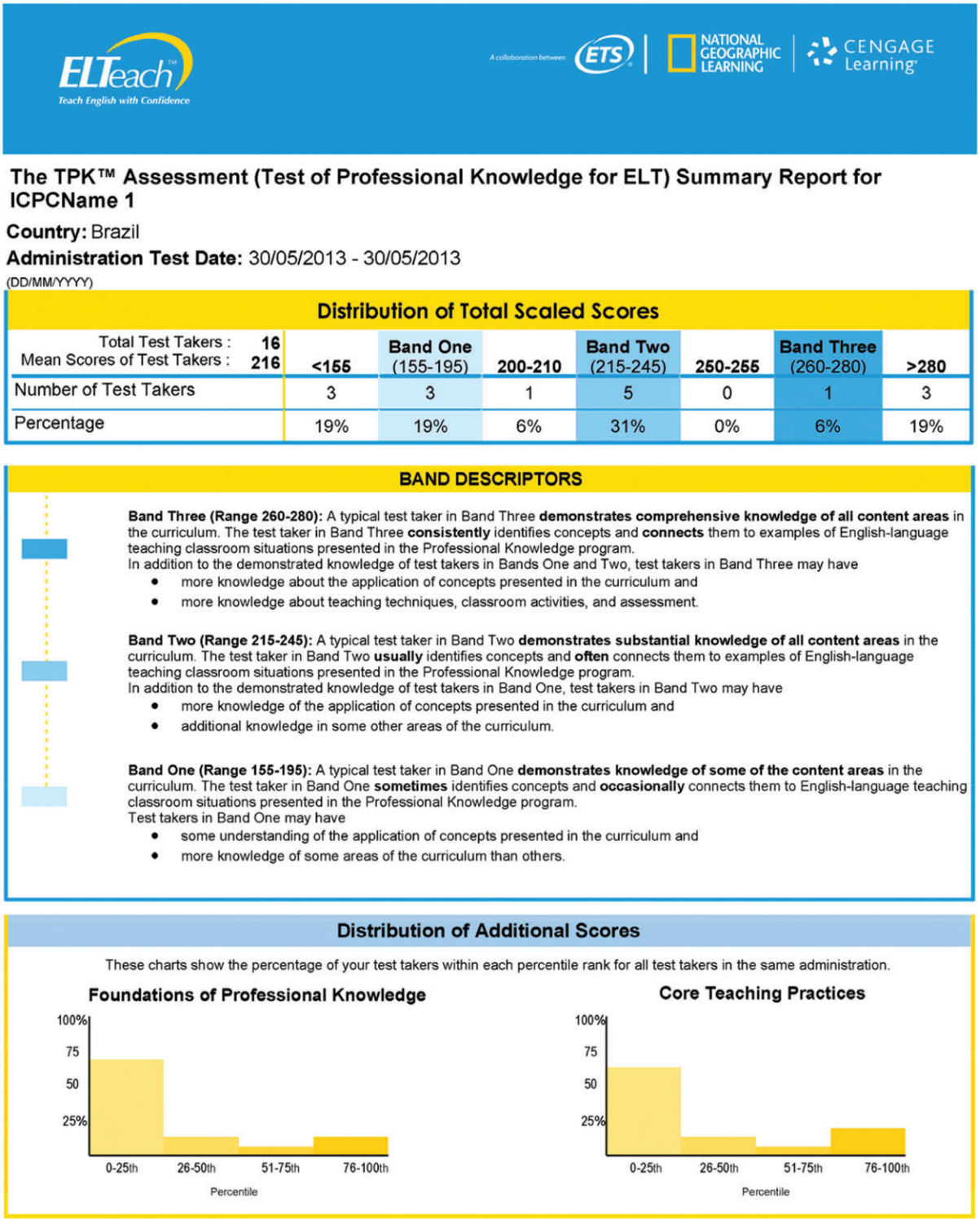


Figure B5 Test of Professional Knowledge (TPK) test-taker score report.



ELTeach™ Teach English with Confidence

A collaboration between ETS | NATIONAL GEOGRAPHIC LEARNING | CENGAGE Learning

The TPK™ Assessment (Test of Professional Knowledge for ELT) Summary Report for ICPCName 1

Country: Brazil

Administration Test Date: 30/05/2013 - 30/05/2013

(DD/MM/YYYY)

Distribution of Total Scaled Scores

Total Test Takers : 16 Mean Scores of Test Takers : 216	<155	Band One (155-195)	200-210	Band Two (215-245)	250-255	Band Three (260-280)	>280
Number of Test Takers	3	3	1	5	0	1	3
Percentage	19%	19%	6%	31%	0%	6%	19%

BAND DESCRIPTORS

Band Three (Range 260-280): A typical test taker in Band Three **demonstrates comprehensive knowledge of all content areas** in the curriculum. The test taker in Band Three **consistently** identifies concepts and **connects** them to examples of English-language teaching classroom situations presented in the Professional Knowledge program.
In addition to the demonstrated knowledge of test takers in Bands One and Two, test takers in Band Three may have

- more knowledge about the application of concepts presented in the curriculum and
- more knowledge about teaching techniques, classroom activities, and assessment.

Band Two (Range 215-245): A typical test taker in Band Two **demonstrates substantial knowledge of all content areas** in the curriculum. The test taker in Band Two **usually** identifies concepts and **often** connects them to examples of English-language teaching classroom situations presented in the Professional Knowledge program.
In addition to the demonstrated knowledge of test takers in Band One, test takers in Band Two may have

- more knowledge of the application of concepts presented in the curriculum and
- additional knowledge in some other areas of the curriculum.

Band One (Range 155-195): A typical test taker in Band One **demonstrates knowledge of some of the content areas** in the curriculum. The test taker in Band One **sometimes** identifies concepts and **occasionally** connects them to English-language teaching classroom situations presented in the Professional Knowledge program.
Test takers in Band One may have

- some understanding of the application of concepts presented in the curriculum and
- more knowledge of some areas of the curriculum than others.

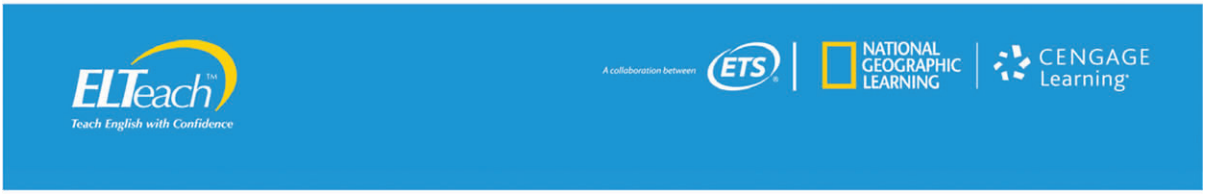
Distribution of Additional Scores

These charts show the percentage of your test takers within each percentile rank for all test takers in the same administration.

Foundations of Professional Knowledge

Core Teaching Practices

Figure B6 Test of Professional Knowledge (TPK) summary score report, page 1.



The TPK™ Assessment (Test of Professional Knowledge for ELT) Summary Report for ICPCName 1

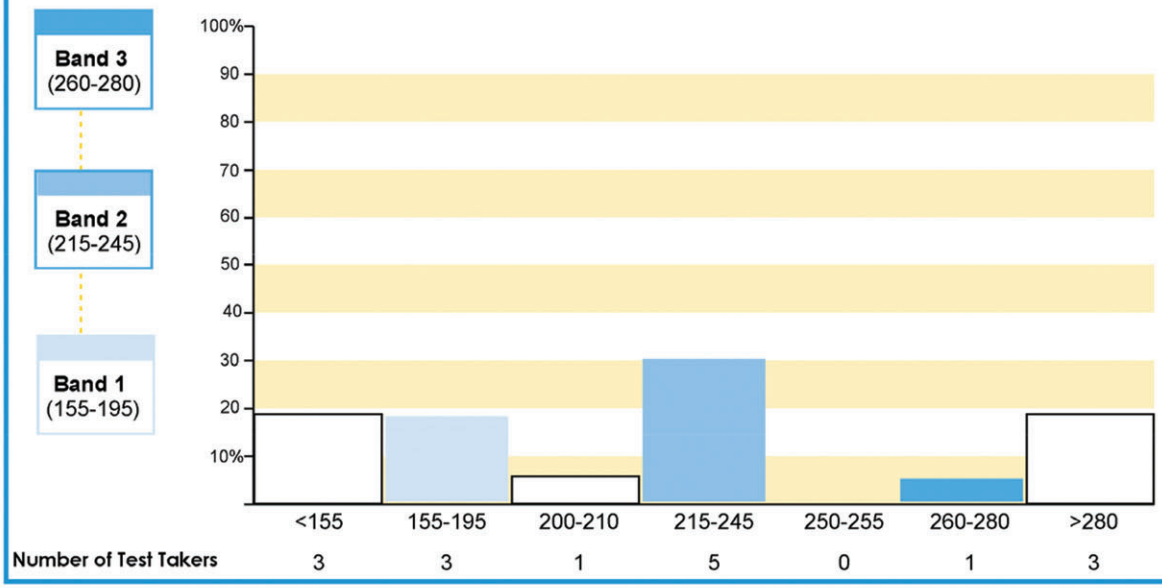
Country: Brazil

Administration Test Date: 30/05/2013 - 30/05/2013

(DD/MM/YYYY)

Distribution of Total Scaled Scores

Mean Total Score : 216
Total Test Takers : 16



TEST TAKERS Last (Family/Surname) Name, First (Given) Name Middle Initial .	Test Taker's ID #	Total Scaled Score	Band Achieved
Ackerman, Berkli J. ---	(AccesCode 10)	285	Band 3
Peterson, Sydney E. ---	(AccesCode 12)	115	---
Betzold, Sierra G. ---	(AccesCode 19)	160	Band 1
Mollett, Zendiah ---	(AccesCode 28)	150	---
Bosch, Kaitlyn M. ---	(AccesCode 34)	175	Band 1
Stanley, Nuha N. ---	(AccesCode 41)	230	Band 2
Rochlitz, Laynee ---	(AccesCode 48)	260	Band 3
Poston, Rachel T. ---	(AccesCode 54)	285	Band 3
Smedley, Piper ---	(AccesCode 57)	210	Band 1

Figure B7 Test of Professional Knowledge (TPK) summary score report, page 2.

The TPK™ Assessment (Test of Professional Knowledge for ELT) Summary Report for ICPCName 1

Country: Brazil

Administration Test Date: 30/05/2013 - 30/05/2013

(DD/MM/YYYY)

TEST TAKERS Last (Family/Surname) Name, First (Given) Name Middle Initial	Test Taker's ID #	Total Scaled Score	Band Achieved
Wolfe, Madelynn	(AccesCode 62)	140	---
Blair, Celeste D.	(AccesCode 68)	195	Band 1
Morales, Shaylee S.	(AccesCode 72)	245	Band 2
Rauch, Taylor J.	(AccesCode 74)	230	Band 2
Buffaloe, Ashley	(AccesCode 77)	240	Band 2
Nelson, Lindsey M.	(AccesCode 97)	230	Band 2
Wiesenhahn, Tehilah M.	(AccesCode 99)	300	Band 3

Figure B8 Test of Professional Knowledge (TPK) summary score report, page 3.

Suggested citation:

Young, J. W., Freeman, D., Hauck, M. C., Garcia Gomez, P., & Papageorgiou, S. (2014). *A design framework for the ELTeach program assessments* (ETS Research Report No. RR-14-36). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12036

Action Editor: Donald Powers

Reviewers: Lin Gu and Gary Ockey