1. What is standardized testing?

A standardized test is one that meets the following characteristics, at least in theory:

- ✓ A rigorous development, **trialling** and revision process, determining the measurement properties of the test. This process includes principled **sampling** from the **population** of interest; defining the measurement scale; establishing adequate reliability and validity for the test's intended purpose; creating **norms** for the population on the basis of the score **distribution** of the sample under test; and statistical equating of all forms of the test, so that reported scores always represent the same level of ability.

- ✓ Standard procedures for administration and scoring of the test. Raw scores obtained on the test are often transformed into **percentiles** or **z scores** for reporting.

- ✓ Standard **test content** in all its versions. This content is based on a set of test specifications which may reflect a theory of language proficiency or a view of candidates' anticipated needs. **Alternative forms** of the test are examined for content equivalence.

Standardized tests place emphasis on reliability. This reliability is seen as making the test suitable for the purposes of comparability across large groups of test takers. Because their measurement properties have been established, **standardized tests are often used as a control for conducting research into the effects of other variables**. For example, a researcher trying to determine the relative value of two methods of language instruction might use a suitable standardized test as a means of comparing learner performance before and after the course of instruction.

However, the attention to reliability may not be matched by a similar concern for validity. For example, a) the emphasis on reliability in standardized tests often leads to reliance on objectively scored test items, which may not represent a valid method of judging communicative ability; b) the validity of test content may not be properly established for the entire test population, since it is extremely hard to generate material which is suitable for large populations which may have very disparate learning backgrounds and experiences.

2. Standardized testing and objective questions

Multiple choice and short answer questions became the conventional format for standardized language test items, given their objectivity. While in principle there is nothing to prevent a standard test from using essays or extended reasoning, reliable scoring of extended answers and essays requires careful training and monitoring of test evaluators and substantially more effort. Rushed and inept evaluation of extended answers can be at least as troublesome as restricting testing to multiple choice and short answer formats. The preference for MC items lies in the fact that these provide the test-designer with an objective means for determining correct and incorrect responses, and therefore it is the preferred mode for large-scale tests. However, standards are equally involved in certain human-scored tests of oral production and writing, such as the Test of Spoken English (TSE) and the Test of Written English (TWE), both produced by ETS.

3. Advantages and disadvantages of standardized testing

Advantages of standardized tests include:

a) A ready-made previously validated product that frees the teacher from having to spend hours creating a test.

b) Administration to large groups (large-scale testing) can be done within reasonable time limits.

c) If MC in format, scoring procedures are streamlined for fast turnaround time.

d) There is an air of face validity to such authoritative-looking instruments.

Disadvantages of standardized tests include:

a) The possibility of an inappropriate use of these tests. For example, using an overall proficiency test as an achievement test simply because of the convenience of standardization.

   A case in point: a course director administered a MC grammar achievement test for placement purposes, even though the curriculum of the course in question was mostly listening and speaking and involved few of the grammar points tested. The test had the appearance and face validity of a good test but it lacked content validity.

b) The potential misunderstanding of the difference between direct and indirect testing. Some standardized tests include tasks that do not directly specify performance in the target objective. For example, prior to 1996, the TOEFL did not include written nor oral production sections, yet statistics showed a high correlation between performance on the TOEFL, ie. the results obtained by the test-takers, and their written and (to a lesser degree) oral production. This means that the TOEFL, a comprehension-based test then, could be claimed to be an indirect test of production in English. By the same token, a test of reading comprehension that proposes to measure ability to read extensively and that engages test-takers in reading only short one- or two-paragraph texts is an indirect measure of extensive reading.

4. Standardized tests of English language ability

Examples of language tests which are generally considered to be standardized include the following:

KET (Key English Test)	**KET** is Cambridge ESOL's exam which recognizes the ability to deal with everyday written and spoken English at a basic level.
PET (Preliminary English Test)	**PET** is an exam for people who can use everyday written and spoken English at an intermediate level. It covers all four language skills — reading, writing, listening and speaking.
FCE (First Certificate in English)	**FCE** is an exam for people who can use everyday written and spoken English at an upper-intermediate level. It is an ideal exam for people who want to use English for work or study purposes.
CAE (Certificate in Advanced English)	**CAE** is an exam for advanced users of English. This exam is aimed at people who can use written and spoken English for most professional and social

	purposes. It is widely recognised for work or study purposes.
CPE (Certificate of Proficiency in English)	**CPE** is Cambridge ESOL's most advanced exam. It is aimed at people who use English for professional or study purposes.
TOEFL (Test of English as a Foreign Language)	The **TOEFL** measures your ability to communicate in English in colleges and universities. It is the most widely accepted English-language test in the world. It is currently produced by the ETS in the US and/or its British counterpart, the IELTS, in affiliation with the UCLES.
MTELP (Michigan Test of English Language Proficiency)	The **MTELP** is a language certificate measuring a student's ability in English as a second or foreign language. Its primary purpose is to access a learner's English language ability at an academic or advanced business level. It's an 100-item MC test of grammar, vocabulary and reading comprehension for advanced level speakers of English as a second language
MELAB (Michigan English Language Assessment Battery)	The **MELAB** assesses advanced-level English language competence of adult nonnative speakers of English such as students applying to United States, Canadian, British, and other educational institutions where the language of instruction is English; professionals who need English for work or training purposes; anyone interested in obtaining a general assessment of their English language proficiency for educational or employment opportunities. It's accepted as an alternative to the TOEFL and IELTS tests. It includes a written composition, a listening test, a grammar/cloze/vocabulary/reading test and an optional speaking test.
TOEIC (Test of English for International Communication)	**TOEIC** is mainly intended for students of business and administration, and largely taken up in the Pacific region. It operates only in the receptive modalities of reading and listening, though investigations are currently in train to determine the feasibility of developing a speaking test as well.
TSE (Test of Spoken English)	The TSE is defined as the most widely used assessment of spoken English worldwide, measuring the ability of nonnative speakers of English to communicate effectively. The test is used for employment, graduate assistantships, and certification purposes.
ESLPT (English as a Second Language Placement Test)	

5. Developing a standardized test

1. Determine the purpose and objectives of the test

a) The **TOEFL**: The purpose of the TOEFL is to evaluate the English proficiency of people whose native language is not English. More specifically, the TOEFL is designed to help institutions of higher learning make valid decisions concerning English language proficiency in terms of their own requirements. Most colleges and universities in the US use TOEFL scores to admit or refuse international applicants. Thus, **the high-stakes, gate-keeping nature of the TOEFL is evident**.

b) The **ESLPT**: The ESLPT is locally designed and administered at San Francisco State University so as to place already admitted students in an appropriate course in academic writing, with the secondary goal of placing students into courses in oral production and grammar-editing. While the test's primary purpose is to make placements, another desirable objective is to provide teachers with some diagnostic information about their students on the first day or two of classes.

c) The **GET**: Also designed at SFSU, it is given to prospective graduate students –both native and non-native speakers – in all disciplines to determine whether their writing ability is sufficient to allow them to enter graduate-level courses in their programs. Students who fail or marginally pass the GET are ineligible to take graduate courses in their disciplines.

☛ As can be seen, the objectives of these three tests are very specific. The content of each test must therefore be designed to achieve those particular ends. Each test has also a specific gate-keeping function to perform. Therefore, the criteria for entering those gates must be accurately specified.

2. Design test specifications

2.1.- Because the **TOEFL** is a proficiency test, the first step in its development is to **define the construct of language proficiency**. How you view language will make a difference in how you assess language proficiency. After breaking language competence down into subsets of listening, speaking, reading, and writing, each performance mode can be examined on a continuum of linguistic units: phonology (pronuniciation) and orthography (spelling), words (lexicon), sentences (grammar), discourse, and pragmatic (sociolinguistic, contextual, functional, cultural) features of language. **How will the TOEFL sample from all these possibilities?**

E.g.: Oral production tests can be tests of overall conversational fluency or pronunciation of a particular subset of phonology, and can take the form of imitation, structured responses, or free responses. Etc.

In short, **from the sea of potential performance modes that could be sampled in a test, the test developer must select a subset on some systematic basis**. Thus, the TOEFL had for many years included three types of performance in its organizational specifications: listening, structure, and reading, all of which tested comprehension through standard MC tasks. In 1996, written production was included in the computer-based TOEFL by adding a slightly modified version of the already existing Test of Written English (TWE). In doing so, some validity (face and content) were improved for this test (but it became increasingly more expensive to administer, so its practicality diminished).

TOEFL specs in Brown 2004:72. Give them the photocopy.

2.2.- Designing the specs for the **ESLPT** was much easier. Because the purpose of the test is placement, **its construct validation consisted mainly in the close study of the content of the ESL courses it gave access to**. Thus, having established the importance of designing ESLPT tasks that simulated classroom tasks used in the courses, the designers ultimately specified two writing production tasks (one a response to an essay the students read, and the other a summary of another essay) and one multiple-choice grammar-editing task. These specifications mirrored the reading-based, process-writing approach used in the courses.

2.3.- The specifications of the **GET** (Graduate Essay Test) arouse out of the perceived need to provide a threshold of acceptable writing ability for all prospective graduates at SFSU, native and non-native speakers of English. The specifications of the GET are the skills of writing grammatically and rhetorically acceptable prose on a topic of some interest, with clear organization of ideas and logical development. The GET is a direct test of writing ability in which test-takers must, in two hours, write an essay on a given topic.

3. Design, select, and arrange test tasks/items

The specifications serve as a blueprint in determining the types and number of items to be created.

TOEFL design, selection and arrangement of test items/tasks: Brown 2004:74. (A)

Items are then designed by a team that select and adapt items solicited from a bank of items provided by free-lance writers and ETS staff.

TOEFL design, selection and arrangement of test items/tasks: Brown 2004:74. (Consider…)

Before items can be released into a form of the TOEFL (or other validated standardized tests), they are piloted and scientifically selected to meet difficulty specifications within each subsection, section, and the test overall.

ESLPT & GET design, selection and arrangement of test items/tasks: Brown 2004:76 (B & C).

4. Make appropriate evaluations of different kinds of items

4.1.- For MCIs:

IF (Item facility, or difficulty, index)
ID (Item discrimination index)
Distractor analysis

4.2.- For other types of responses (namely, production responses):

Practicality issues: clarity of directions, timing of the test, ease of administration, and how much time is required to score responses.

Reliability issues: one or more than one scorer?

Facility issues: are directions clear? Is the language unnecessarily complex? Are the topics rather obscure? Are there any fuzzy data provided in the test material or test questions? Is there any culturally-biased information? All these could lead to a higher level of difficulty than the one desired.

Brown 2004:78 for decisions concerning the evaluation of items in the TOEFL, ESLPT, and GET.

5. Specify scoring procedures and reporting formats
6. Perform ongoing construct validation studies

Percentile: A conversion of a raw score which provides a means of locating the score in relation to the distribution of all scores. Percentiles generally relate the score obtained by a particular candidate to the proportion of scores at or below it. Thus, if a score is said to be at the 68th percentile, this means that this candidate did as well as or better than 68% of the whole cohort. The 50th percentile represents the median score.

Z score: A way of placing an individual score in the whole distribution of scores on a test; it expresses how many standard deviation units lie above or below the mean. Scores above the mean are positive; scores below the mean are negative.

Test content: The skills and components of language ability which are measured in a test, and the manner in which they are measured. The test content may be based on a theory of language proficiency, as in a proficiency test, or on a specified syllabus, as in an achievement test. In either case, the content should be a representative sample of the domain to be tested. For example, in a language aptitude test the test content will be based on a sampling of the abilities deemed to be associated with successful language learning.

The content is normally defined in the test specifications in terms of a range of features of the test. These features may include: skills, vocabulary, grammar, communicative functions; topics, type and length of texts, task or item types and format used; and form of responses required.

Alternative forms of a test: Two or more tests designed to the same specifications. Each form should correspond in terms of: number of test items, type of items, content of items, difficulty level of item, as well as in instructions, time allowed, etc.

Norms: Language tests are doubly normative. First because they are necessarily based on a standard of speaking or writing (often the native speaker); and second because they impose that standard on test-takers and, through them, on the community.

Item facility/difficulty index (IF): The degree of difficulty of a test item which is calculated on the basis of a group's test performance. Davies et al. 1999:95-96.

Item discrimination index (ID): The capacity of test items to differentiate among candidates possessing more or less of the trait that the test is designed to measure. A test with consistently high levels of ID is considered to be reliable.