Teacher-made vs standardized tests

- Teacher made (classroom) tests: in terms of criteria

I. Preparation and construction: the same person as instructor, test writer, and evaluator

II. Administration: no uniform procedures

III. Content and objectives coverage: those determined by the teacher in the classroom

IV. Scoring: subjective and usually biased and judgment evaluative

V. Purpose and use: measures particular objectives and is used to make intraclass comparisons
I. Preparation and construction: a team of experts
II. Administration: standard uniform procedures
III. Content and objectives coverage: determined by ministry of education, existing curricula and syllabi
IV. Scoring: objective, usually machine-scored
V. Purpose and use: measures broad objectives and is used to make interclass, school, and national comparisons
Formative vs summative tests

Formative assessment: in terms of two criteria

I. Purpose of data gathering: to provide feedback and improve learning
   1. Delivery of information by teacher
   2. Internalization of information by students
   3. Achieving competence or learning
   4. Not part of a final grade or mark
   5. More experimentation, challenging of preconceived ideas, and development of higher cognitive skills
   6. Looking forward

II. Time of assessment: during the course of instruction: informal assessment as formative assessment
Summative assessment

I. Purpose of data gathering: to determine the success of learning and teaching
   A. Measure or summarize the amount of learning
   B. Achieving limited assigned material or knowledge
   C. Less experimentation
   D. Final grade or mark
   E. Looking back

II. Time of assessment: the end of the course or unit: all quizzes, midterms, final test, and proficiency exams are summative assessment
To bring the discussion down to earth

Read the following two anonymous quotes to figure out the distinction between formative and summative assessment.

1. When the cook tastes the soup, that's formative; when the guests taste the soup, that's summative (Stake).

2. When the chef tastes the sauce it is formative assessment; when the customer tastes sauce, it is summative (Anon)
**Norm-referenced vs criterion referenced tests**

- Norm-referenced assessment: in terms of criteria
  
  A. Type of interpretation: relative and in percentile
  B. Type of measurement: global abilities
  C. Purpose of testing: spread of scores along a continuum to reveal existing differences
  D. Distribution of scores: normally distributed, heterogeneous
  E. Tests structure: long and comprehensive in content coverage;
  F. Knowledge of questions: unknown content but known test format
A. Type of interpretation: absolute and in percentage
B. Type of measurement: well-defined learning points
C. Purpose of testing: the amount of knowledge and skill learned by learners
D. Distribution of scores: homogeneous
E. Test structure: short and less comprehensive in content coverage
F. Knowledge of questions: known content and test formats
High stakes vs low-stakes tests

High-stakes tests:
A. Major, life-threatening decisions
B. Difficulty of correcting decision errors
C. High costs of making wrong decisions

Low-stakes tests
A. Minor decisions
B. Easiness of decision errors to correct
C. Low costs of making wrong decisions
Performance-based testing:

A. Definition:
A communicative test in which learners must perform tasks as authentic as possible, the outcome of which must be rated by qualified judges.

B. Features:
1. Learners make a constructed response.
2. They engage in higher-order thinking, with open-ended tasks.
3. Tasks are meaningful, engaging, and authentic.
4. Tasks call for the integration of language skills.
5. Both process and product are assessed.
A. Tests usage rather than use.
B. Tests competence than performance.
C. Tests underlying ability to perform a skill than the skill itself
Traditional vs alternative tests

Traditional testing: main features
A. one-shot, standardized exams or tests
B. Timed, multiple-choice format
C. Decontextualised test items
D. Scores suffice to feedback
E. Norm-referenced scores
F. Focus on the right answer
G. Summative
H. Oriented to product
I. Non-integrative performance
J. Fosters extrinsic motivation
Alternative assessment

A. Continuous long-term assessment
B. Untimed, free-response format
C. Contextualized communicative tasks
D. Individualized feedback and washback
E. Criterion-referenced scores
F. Open-ended, creative answers
G. Formative
H. Oriented to process
I. Interactive performance
J. Fosters intrinsic motivation
Defining alternative assessment

Procedures and techniques which can be used within the context of instruction and can be easily incorporated into the daily activities of the school or classroom.

Self-assessment and peer-assessment are two well-known examples of alternative assessment.
Direct vs indirect tests

A. Direct tests
1. Performance based in nature
2. Testing format and procedure duplicating the setting and operation of real-life situations
3. Performance of the skill at hand
4. Assessment of performance or response to a naturally evolving situation
5. More suited to productive skills of speaking and writing
6. Easy to create
7. Beneficial backwash effect
Indirect tests

A. Measurement of abilities underlying skills than the skills proper such as a multiple-choice test to measure writing
B. No relationship between performance on them and performance of the skills
C. Possibility of measuring a large number of abilities
D. Provision of negative backwash
Discrete-point vs integrative tests

A. Discrete-point tests
1. Test language bit by bit one at a time
2. Separate testing of individual skills and components as isolated phenomena
3. Measure knowledge or performance in very restricted areas of language such as listening
4. Individual test item taken together as a sign of overall language proficiency
5. Diagnosis oriented in nature and indirect and usage based
B. Integrative tests

1. Concurrent tapping of a greater variety of language abilities
2. Measure overall language proficiency
3. Integration of language skills and components
4. Still usage based, but less diagnosis oriented in nature
5. Cloze and dictation as standard types of integrative tests
Speed vs power tests

- **Speed tests:**
  1. Easy items
  2. Limited amount of time

- **Power tests**
  1. Difficult items
  2. Enough time
Other main test terms

Practice effect or rehearsal effect:

1. Two administrations (test-retest or pre and post)
2. Higher scores on the second test due to experience or practice of the first test
3. Depresses reliability of a test
4. To minimize practice effect:
   a. Time interval
   b. Counterbalancing
1. Test taking experience of testees, than knowledge of the subject matter
   a. Familiarity with and easy comprehension of test format characteristics, directions, testing situations; guessing; and knowledge of strategies
   b. Flaws in construction of test items, stems, and options
2. closely related objective tests, namely, multiple choice item tests
3. Overestimated test scores
4. Correlated with all age groups, test anxiety, culture, but not gender
5. Heavily researched in the United States of America, Canada, and European countries including Sweden
6. Use of metacognition in answering test items
7. Solutions: provision of clear instructions and minimization of clues or absurd options
Coaching or teaching to the test

A. Special efforts to prepare students to score higher on target language examinations
   1. Item-teaching: instruction targeted at actual items or clone items
   2. Curriculum-teaching: instruction targeted at knowledge and skills or objectives

B. Either productive or counterproductive

C. Concerns both low-stakes and high-stakes tests

D. As a special type of testwiseness

E. Depressed consistency, appropriacy and genuineness
I. Backwash effect
A. The effect of testing on teaching
B. Empirically investigated existence of backwash
C. Bidirectional: Either beneficial or harmful

II. Impact
A. Effect of testing on teaching, learning, society and educational systems
B. Much broader and more encompassing than backwash
Test compromise

- Acquisition of prior knowledge of test content
- Solutions
  1. Infrequent test administration—once or twice a year with new forms of a test
  2. A sufficient number of test forms
  3. Counterbalancing of test forms
  4. Securing the tests, answer keys, and answer sheets
Ceiling effect or lack of headroom (# basement effect or floor effect)

A. Majority of scores near the top of the scale
B. Lack of adequate discrimination among higher ability students
C. Implications:
   1. Negatively skewed distribution
   2. Impossibility of statistics to interpret
C. Causes:
   1. Subjective ratings
   2. Easy test or tasks
D. Solutions:
   1. Rater training
   2. Provision of more detailed descriptors
Testing bias

A. favorable or unfavorable inclinations toward subgroups or individual students

B. Causes of bias:
   1. Administration procedures
   2. Test directions
   3. Test content
   4. Test knowledge selection
   5. Testing method
   6. Rating/scoring

C. Solutions
   1. Provision of careful guidelines for item writers and test designers
   2. Making results public
Features for classifying different types of tests

A. Intended use
B. Content
C. Frame of reference
D. Scoring procedure
E. Testing method