DEPARTMENT OF
OPERATIONAL
SUPPORT

FOUNDATIONS OF LANGUAGE ASSESMENT:

AN INTRODUCTION TO THE WHYS, WHATS AND HOWS

Neus Figueras, 24 February 2023 8:30-11:30AM

Session Dossier

The main aim of the webinar was to raise awareness of the issues involved in the development, administration and use(s) of language assessments so that they are “honest, reliable, valid, transparent and portable” (John Trim, 2011). The session combined theory and practice. Reflection slots and self-correcting activities were included to facilitate participation and to clarify and discuss issues raised.

The main referent of the session was the UN Language Framework (UNLF), which states that

Language in use is a very complex phenomenon which calls on a large number of different skills or competences. It is important to start a testing project with an explicit model of these competences and how they relate to each other. The role of such a model is to identify significant aspects of competence for our consideration. It is a starting point for deciding which aspects of use or competence can or should be tested, and helps to ensure that the test results will be interpretable and useful. The mental characteristic identified by the model is also called a CONSTRUCT. https://hr.un.org/page/harmonization-language-learning-and-assessment

These were the contents addressed:

1. The “Why” of Assessment:
   - Assessment vs. testing: similarities & differences
   - Why and when are they necessary?
2. The “What” of Language Assessment in the UN:
   - The UN Language Framework (UNLF) and its uses
3. The “How”: From Purpose to Results
   - Elements of quality, fair and reliable assessment
4. Hands-on analysis:
   - Scoring of sample tasks and language performances
5. Concluding remarks

This dossier contains a selection of the documentation presented in the webinar with some supplementary texts and references for further study.

INTRODUCTION

The past 10 years have seen an increased interest in assessment, understood as a very broad term, and have blurred the line between testing and assessment, traditionally opposed terms. Research into how assessment/testing can contribute to learning, and media debates on the fairness of certificate examinations have also resulted in new (sometimes confusing) terminology. Practitioners today often struggle to see the differences and the similarities between assessment, evaluation and testing; between continuous assessment, formative assessment and summative assessment or proficiency assessment; between assessment for learning (AFL), assessment of learning (AOL), and learning-oriented assessment (LOA), which sometimes represent similar concepts or approaches for different authors. The sections that follow outline the main principles to consider when developing, administering and grading any assessment, which can be defined as ***a systematic collection of information with the purpose of making decisions, and making one or more judgements based on the data collected in relation to reference values.***

1. THE "WHY" OF ASSESSMENT

Assessment vs. Testing: Similarities & Differences; Why and When are they Necessary?

The first consideration when starting to plan any assessment is why it is needed and what purpose it will serve. Different purposes need different approaches to assessment, with different types of text. In addition to purpose, context also needs to be considered as both will impact the function of the test and its impact or stakes. Although techniques for writing test tasks with different purposes, for different contexts, etc. may be very specific, the same principles, concepts and recommendations (e.g. on checking usefulness, on careful planning) are applicable in all assessment contexts, whether they use standardized exams or not. For a glossary of testing terms, check the list of references at the end of the dossier.

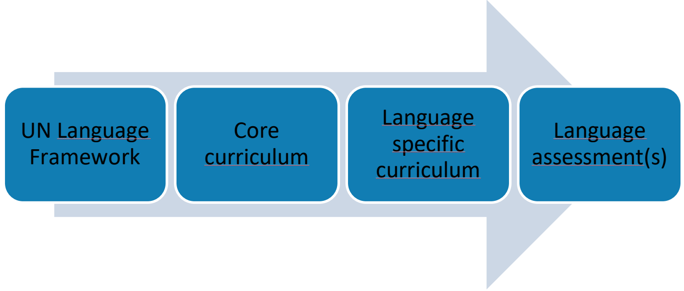
PURPOSE	TYPE OF TEST	CONTEXT	CONTENT	FUNCTION	STAKES
1. Place learners in the right level/class at the beginning of a course.	Placement	School*	School curriculum; Wide focus	Pedagogic	Medium
2. Find out how learners are improving and what difficulties they may be having.	Progress	Classroom	Course objectives; narrow focus	Pedagogic	Low
3. Identify prior knowledge of the contents of a lesson.	Diagnostic	Classroom	Course objectives; narrow focus	Pedagogic	Low
4. Find out whether course objectives have been fulfilled.	Achievement	Classroom	Course objectives; wide focus	Pedagogic and social	Medium/ High
5. Check language ability in the world beyond the test.	Proficiency	School* Classroom Society	"Real World"; Wide Focus	Social	High

*In this context, school refers to any language teaching centre or institution.

2. THE "WHAT" OF LANGUAGE ASSESSMENT:

The UNLF and its Uses

The webinar briefly outlined the main components of the United Nations Language Framework (UNLF) and its related documents, highlighting the fact that different users would be focusing on and using different levels and different sections of the different documents in the UNLF, which is organized from the more general to the more specific.



For a detailed description of the UNLF and its applications, please visit the UNLF training modules, available at: http://elounge.unssc.org/login/index/php

As a summary of the UNLF level labels and descriptors, the following table may serve as a reminder of what different learners/speakers at different levels can do, how well and under which conditions and limitations.

UNLF	WHAT	HOW WELL	CONDITIONS & LIMITATIONS
I	Use the language in a simple manner, in non-demanding everyday contexts and situations, when dealing with routine or predictable matters in the personal, public and professional domains, throughout the Organization.	Show basic linguistic competence and use a restricted range of social language conventions to meet simple communication needs. Show limited facility in understanding if an action or response is required and some autonomy to respond	Usually require reference resources and models, templates or external help to prepare in advance, check understanding or repair communication.
II	Use the language with moderate fluency and accuracy, in everyday contexts and situations, when dealing with ordinary or general matters in the personal, public and professional domains, throughout the Organization.	Show an appropriate command of a moderate range of linguistic and pragmatic competences and of social language conventions to meet ordinary general communication needs. Understand if any action or response is required and show adequate autonomy to respond.	Often require reference resources and models or external help to prepare in advance, check understanding and improve or support communication

III	Use the language efficiently, with a high degree of fluency and accuracy, in a variety of contexts and situations, when dealing with a wide variety of general matters in the personal, public and professional domains, throughout the Organization.	Show a good command of a range of linguistic and pragmatic competences and of social language conventions to meet most communication needs. Respond autonomously and sufficiently to most required actions.	Use reference resources to confirm and refine interpretation, and to improve communication.
IV	Use the language efficiently and flexibly, consistently maintaining a high degree of fluency, accuracy and precision. Function in a large variety of demanding contexts and situations, even adverse or unpredictable, when dealing with a wide range of matters, even highly specific or sensitive, in the personal, public and professional domains, throughout the Organization.	Show an excellent command of a wide range of linguistic and pragmatic competences and of social language conventions to meet any communication need. Respond to and follow up on any required action appropriately and without hesitation.	Use reference resources to enhance communication with sophisticated precision.

3. THE "HOW": FROM PURPOSE TO RESULTS

Quality, Fair and Reliable Assessment

QUALITY AND USEFULNESS

Bachman and Palmer (1996) use the term **usefulness** to summarize test quality (1996:16-42). When developing a test it is key to **strive for fairness**, making sure that the tasks to complete/perform by the test takers match the defined construct, and that administration, marking and scoring procedures are adequately standardized.

The notion of **usefulness** can be described as a function of several different qualities, all of which contribute in unique but interrelated ways to the overall usefulness of a given test. Usefulness cannot be evaluated in the abstract, but in relation to the context and purpose of a test in particular. Three principles should guide the consideration of usefulness:

1. It is the overall usefulness of the test that is to be maximized, rather than the individual qualities that affect usefulness.
2. The individual test qualities cannot be evaluated independently, but must be evaluated in terms of their combined effect on the overall usefulness of the test.
3. Test usefulness and the appropriate balance among the different qualities cannot be prescribed in general, but must be determined for each specific situation.

- **Validity**: this concept pertains to the meaningfulness and appropriateness of the interpretations that we make on the basis of scores. Does the exam assess what it claims to assess (the construct)? To what extent can we justify the interpretations?
- **Reliability**: this concept is often defined as consistency of measurement. Can we trust the results?
- **Authenticity**: this concept is defined as the degree of correspondence of the characteristics of a given language test task to the features of a real-life task. Does the performance on the language test correspond to language use in specific domains other than the test itself?
- **Interactiveness**: this concept is defined as the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task. Are the test taker's language knowledge, metacognitive strategies, topical knowledge and affective schemata engaged by the test tasks?
- **Impact:** all tests imply certain values and goals, and they all have consequences for or impact on, both the individuals and the system involved. The influence of a test in the learning process is often referred to as "washback".
- **Practicality**: this concept pertains to the ways in which the test is developed and implemented. If the resources required for implementing the test exceed the resources available, the test will be impractical.

THE TEST DEVELOPMENT CYCLE

Effective test development requires a systematic, detail-oriented approach based on sound theoretical educational measurement principles. Steven M. Downing and Thomas M. Haladyna edited a seminal book in 2006, *Handbook of Test Development,* with 32 chapters by experts on assessment. Their chapter on Twelve Steps for Effective Test Development discusses 12 discrete test development procedures or steps that typically must be accomplished in the development of most tests. Following these 12 steps tends to maximize validity evidence for the intended test score interpretation.

STEPS	**EXAMPLE TEST DEVELOPMENT TASKS**
1. Overall Plan	Systematic guidance for all test development activities: construct; desired test interpretations; test format(s); major sources of validity evidence; clear purpose; desired inferences; psychometric model; timelines; security; quality control.
2. Content Definition	Sampling plan for domain/universe; various methods related to purpose of assessment; essential source of content-related validity evidence; delineations of construct.
3. Test specifications	Operational definitions of content; framework for validity evidence related to systematic, defensible sampling of content domain; norm or criterion referenced; desired item characteristics.
4. Item development	Development of effective stimuli; formats; validity evidence related to adherence to evidence-based principles; training of item writers, reviewers; effective item editing; CIV (construct-irrelevant variance) owing to flaws.
5. Test design & assembly	Designing and creating test forms; selecting items for specified test forms; operational sampling by planned blueprint; pretesting considerations.

6. Test production	Publishing activities; printing or CBT (computer- based testing) packaging; security issues; validity issues concerned with quality control.
7. Test administration	Validity issues concerned with standardization; disability issues; proctoring; security issues; timing issues.
8. Scoring test responses	Validity issues: quality control; key validation; item analysis.
9. Passing scores	Establishing defensible passing scores; relative vs. absolute; validity issues concerning cut scores; comparability of standards: maintaining constancy of score scale (equating, linking).
10. Reporting test results	Validity issues: accuracy, quality control; timely; meaningful; misuse issues; challenges; retakes.
11. Item banking	Security issues; usefulness, flexibility; principles for effective item banking.
12. Test technical report	Systematic, thorough, detailed documentation of validity evidence; 12-step organization; recommendations.

WRITING TEST TASKS: TEXT SELECTION AND ITEM WRITING

Reception: Text selection and item writing

Selecting a text should be regarded as the first step to the successful development of a test task. Whenever possible, authentic texts requiring minimum editing should be prioritized. It is important that texts (both written and spoken) have a clear logical structure and contain sufficient information. Texts published online need to be carefully checked for coherence and cohesion.

Although characteristics such as text type, text length, domain, topic, content or language complexity should inform a first step in the selection of a text (written or spoken), it is important to establish, on a principled basis, the content (questions or items) that should be extracted in line with the established purpose for reading or listening. A common practical, “utilization- focused” procedure, referred to as **text mapping** or diagramming should be used (Weir et al., 2000).

1. Individual first reading/listening. Checking of test adequacy for the purpose of the set task and identification of main ideas.
2. Second individual reading/listening to identify detail and supporting ideas.
3. Exchange with the group. Identification of consensus on 1 and 2.
4. Does the text allow the production of sufficient items to assess the cognitive processes expected? With what type of reading/listening?
5. Which ideas/information can yield questions?

When writing items, both form and content are important, and item writers need to take into consideration a) the objective(s) of the task and b) the operations the task is expected to assess. Item writers also need to consider the following recommendations:

- construct each item to assess a single objective
- always write items for listening tasks on the basis of the oral text, not the transcript.

- focus on relevant concepts and detail
- avoid tricky or excessively complex items
- the items should follow the order of the text
- only ONE option should be correct
- the language used in the items cannot be more difficult than the language in the text.
- use correct grammar, punctuation and spelling
- write clear and simple instructions

Production and interaction

Production tasks are often described as constructed response tasks, as opposed to the selected response tasks often used in reception tasks and they offer the candidate the opportunity to:

- generate or create a response
- go beyond the requirements of selected response items

Given the interaction between prompt (input text, rubric) and candidate, which impacts the performance obtained and, consequently, the score given, the selection of the topic of the task, the way the task is presented, its contextualization and its specific wording is very important. Prompts for productive tasks should:

- be accessible to all candidates regardless of gender, culture, background or position
- provide sufficient guidance to the candidate whilst allowing for some freedom of performance within the necessary standardization (length, expected response format, register, etc.)
- be sufficiently complex in content to elicit the best possible performance and contain all the information necessary, but worded in clear language and avoid unnecessary reading time
- foster authentic, communicative language
- aim at tapping a variety of language functions and linguistic features (grammar & vocab)

RATING SCALES

Rating scales are necessary to guarantee that the grading focus corresponds to the purpose of the assessment and its characteristics (validity), but they are also very important to standardize grades (reliability), helping different graders be systematic in their work.

A rating scale is a set of descriptors which describe performances at different levels, showing which grade each performance level should receive. In many contexts, the term rubric or marking scheme is often used with the same meaning.

Rating scales reduce the variation inherent in the subjectivity of human judgements and are key for reliability.

There is a range of options to consider:

- ➢ **Holistic scales:** a single mark based on a single scale describing a range of features in each level of performance.
- ➢ **Analytic scales:** a mark is given for each of a range of criteria (accuracy, task fulfillment, etc.)
- ➢ **Checklists**: a grade is given based on a list of yes/no judgements as to whether a performance fulfills specific requirements or not.

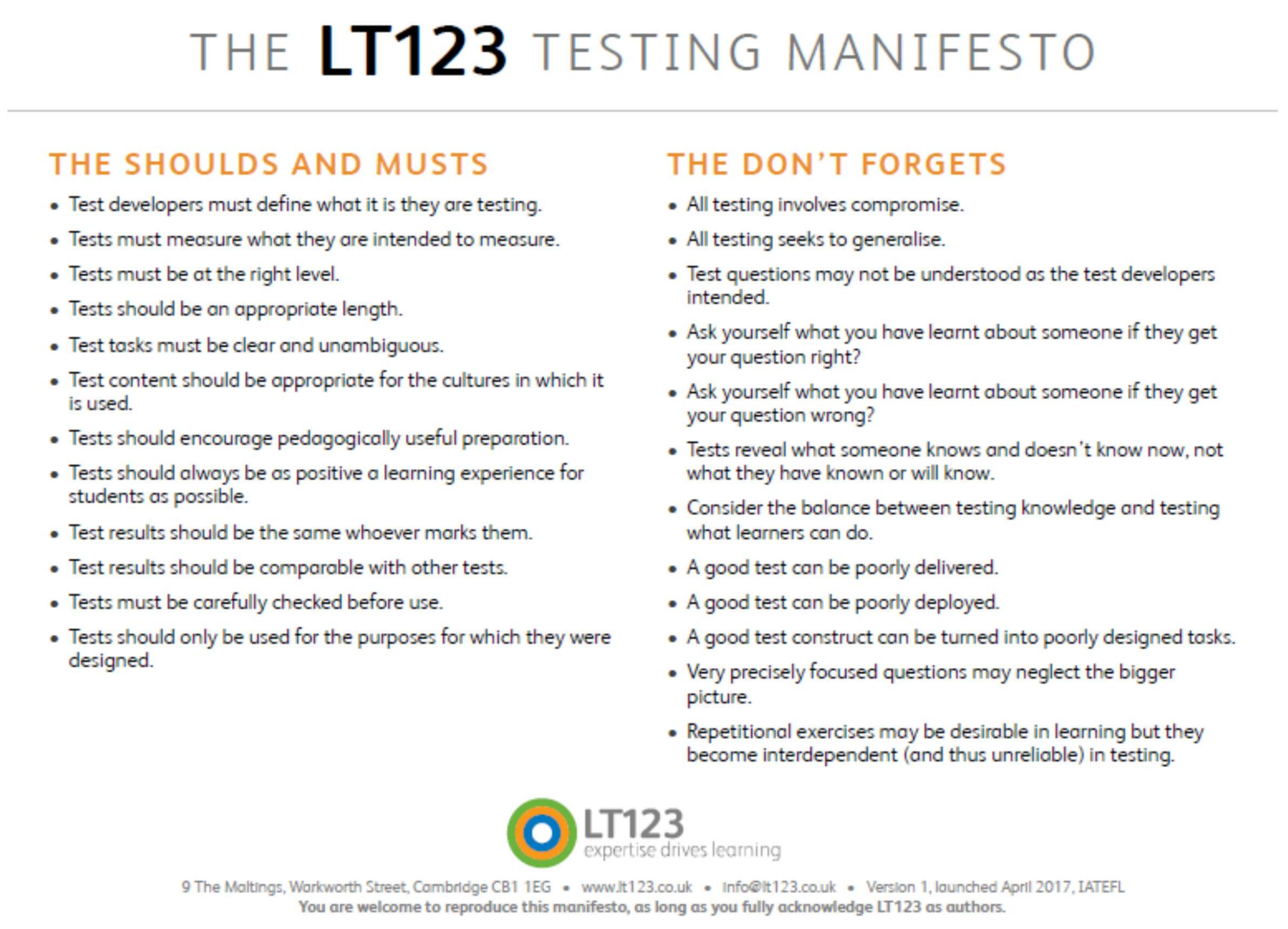
- **Generic or task-specific** scales: providing a generic scale or checklist for all tasks or provide criteria for each specific task.

Whatever approach you choose to grade performances, the options above share similar underlying principles and require that all graders using the same scale have attended norming sessions and use it appropriately and reliably. The statements below need to be remembered:

- All rating depends on the raters understanding the levels.
- Exemplars are essential to defining and communicating this understanding, typically at norming sessions.
- The test tasks used to generate the rated performance are critically important to working with scales.

GUIDELINES, STANDARDS, CODES OF ETHICS

Most assessment organizations and exam boards have published codes of practice, codes of ethics, standards or guidelines for good practice to guide the development and evaluation of tests. Such documents vary in length, thoroughness and complexity. The Manifesto below, by a private consultancy, represents a brief, user-friendly and sufficient set of principles to start with. Those interested in the topic can access other documents listed in the list of references, such as the Guidelines of Good Practice for Language Testing and Assessment by EALTA, the ALTE Principles of Good Practice or the Standards of Quality by SICELE.



THE LT123 TESTING MANIFESTO

THE SHOULDS AND MUSTS

- Test developers must define what it is they are testing.
- Tests must measure what they are intended to measure.
- Tests must be at the right level.
- Tests should be an appropriate length.
- Test tasks must be clear and unambiguous.
- Test content should be appropriate for the cultures in which it is used.
- Tests should encourage pedagogically useful preparation.
- Tests should always be as positive a learning experience for students as possible.
- Test results should be the same whoever marks them.
- Test results should be comparable with other tests.
- Tests must be carefully checked before use.
- Tests should only be used for the purposes for which they were designed.

THE DON'T FORGETS

- All testing involves compromise.
- All testing seeks to generalise.
- Test questions may not be understood as the test developers intended.
- Ask yourself what you have learnt about someone if they get your question right?
- Ask yourself what you have learnt about someone if they get your question wrong?
- Tests reveal what someone knows and doesn't know now, not what they have known or will know.
- Consider the balance between testing knowledge and testing what learners can do.
- A good test can be poorly delivered.
- A good test can be poorly deployed.
- A good test construct can be turned into poorly designed tasks.
- Very precisely focused questions may neglect the bigger picture.
- Repetitional exercises may be desirable in learning but they become interdependent (and thus unreliable) in testing.

LT123
expertise drives learning

9 The Maltings, Workworth Street, Cambridge CB1 1EG • www.lt123.co.uk • info@lt123.co.uk • Version 1, launched April 2017, IATEFL

4. Hands-on Analysis: Scoring of Sample Tasks and Language Performances

Once a production/ interaction task has been developed and is considered adequate, a decision needs to be made on how to grade it, the type of rating scale to use and the necessary procedures to ensure that all quality elements are taken into consideration during the administration and grading phases. The choice of rating scale is related – again – to the purpose of the assessment, but also to the degree of detail needed and the availability of human and financial resources. A simple way to get started is by using a generic checklist like the one below, as it does not require the development of level descriptors – a difficult task as described by Brophy (2014)- and only requires judgements related to the levels being referenced. The use of a checklist, however, or of any marking scheme, is closely related to the framework informing the assessment asks (in this case the UNLF), and all users need to be familiarized with the contents of the framework in question. In this case, and depending on the level of detail/granularity required, graders will refer only to the overall descriptors for each of the four levels or will also consider the descriptors for each activity/skill within each level.

	YES	NO	NOT REALLY	UNLF LEVEL
TASK				
Does the text/performance respond effectively to the prompt? Length?				
Is the style/ register/tone performance adequate?				
Is it coherent and well organized?				
Does it contain specific details and examples?				
LANGUAGE				
Has the writer/speaker appropriately used a wide range of language structures & vocabulary?				
Is the text/performance accurate and largely free of mistakes?				
Do errors affect comprehension?				
Is the pronunciation intelligible? Is it comprehensible?				

USEFUL REFERENCES

You will find a lot of assessment/testing related materials on the web, but the animated videos (4-6 minutes each) that the British Council has prepared (some with worksheets) are really useful to get a better understanding of what developing a language exam entails: https://www.britishcouncil.org/exam/aptis/research/assessment-literacy

Alderson, Ch., Clapham, C. & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

ALTE- Council of Europe. (2009). *Manual for language test development and examining.* Available from https://rm.coe.int/manual-for-language-test-development-and-examining-for-use-with-the-ce/1680667a2b

Bachman, L. and Palmer, A. (1996). *Language testing in practice.* Oxford University Press.

Bachman, L. and Palmer, A. (2010) L*anguage assessment in practice*. Oxford University Press.

Downing, S.M. and Haladyna, Th.M. (eds). (2006). *Handbook of test development*. Lawrence Erlbaum Associates.

ELT Glossary of testing terms. Available from https://www.eltconcourse.com/training/glossaries/ELT_Concourse_glossary_testing.pdf

Mc.Namara, T. (1996) *Measuring second language performance*. Longman.

Weir, C. (1993). *Understanding and developing language tests*. Prentice Hall.

Guidelines and codes

Principles of good practice. Available from https://www.alte.org/resources/Documents/ALTE%20Principles%20of%20Good%20Practice%20Online%20(Final).pdf

European Association for language testing and Assessment – EALTA. *EALTA Guidelines for good practice in language testing and assessment .* Available from www.ealta.eu.org

Sicele: Estándares de calidad. Available from https://asociac
Association of Language testers in Europe – ALTE . *ALTE* ionsicele.org/es/node/9

Rubrics

TELC – German institution providing language tests in Arabic, English, French, Spanish, Russian, which has online mock examinations including their rating scales https://www.telc.net/

Brophy, T. (2014). *Writing effective rubrics*. University of Ohio.
Available from http://web.cse.ohio-state.edu/~soundarajan.1/abet/writing_effective_rubrics_guide_v2.pdf