

Introduction to Language Testing & Assessment: An Overview

Ahmet Dursun

Director, Office of Language Assessment

Preview

In this talk:

- Review key concepts, terms and principles of language assessment

- Explain the principles of designing and developing useful language test

Perceptions of language tests

How is the language testing perceived by students?

SCAN ME



OR

Join at
slido.com
#3469 887

Perceptions of language tests

How do you perceive the language testing?

SCAN ME



OR

Join at
slido.com
#3469 887

Justification

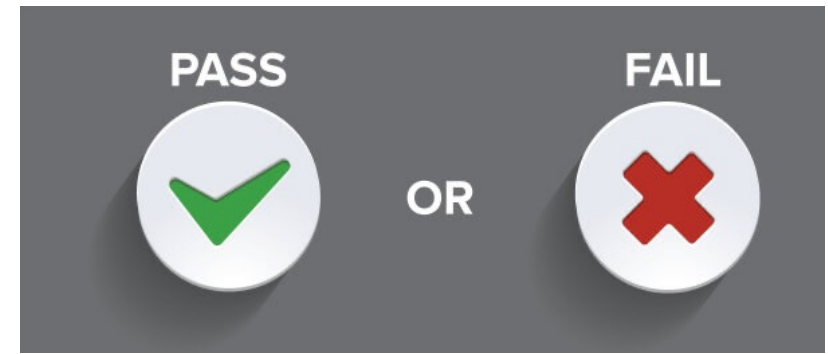
Negative washback on teachers, students, and language instruction

Socio-cultural and socio-economic factors

The role of assessment in language instruction

Teachers' beliefs about the relationship between pedagogy and assessment

**When testing only follows
teaching...**



Key Terms & Principles



What is a test?

A test is method of measuring a person's ability, knowledge, or performance in a given domain (Brown, 2004).

- A test has **a method**. It is **an instrument**. There is a set of techniques, procedures, or items.
- A test must measure (individual's ability, knowledge, or performance)
- A test measures a given domain (overall proficiency in a language in a language vs. a limited set of phonemic minimal pairs)

What is assessment?

Assessment is an ongoing process of measuring a person's ability in a wider domain.

Tests are a subset of assessment

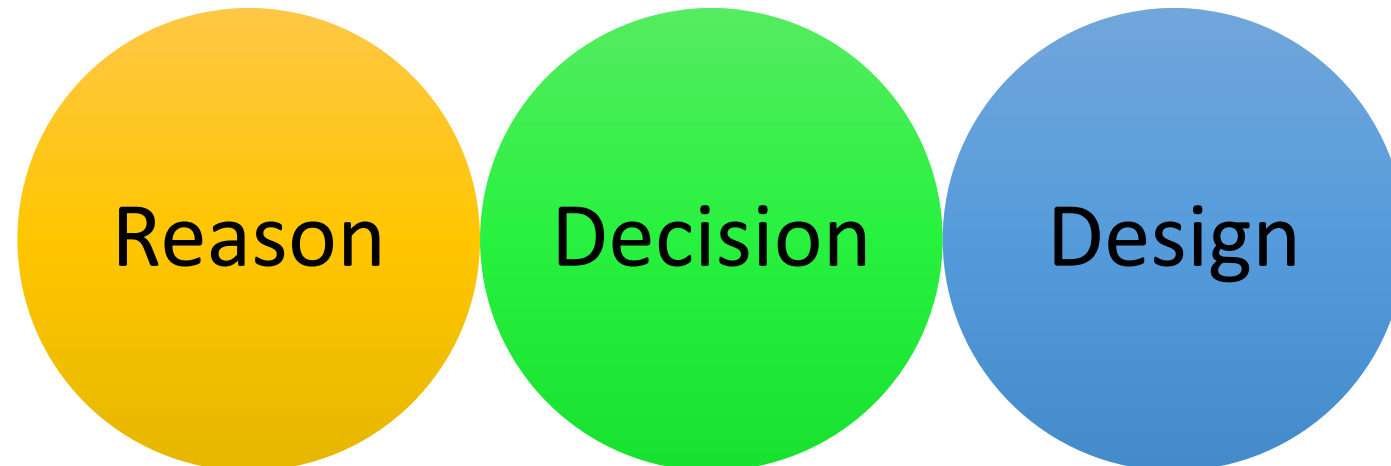
- **Informal Assessment:** Incidental, unplanned, impromptu feedback
- **Formal Assessment:** Purposefully designed systematic exercise and procedures

Tests & Assessments are Tools

The ultimate goal is **decision-making!**

Therefore before we even start writing a test, we should:

Identify & define the reason/purpose



A particular tool is better for some tasks than for others...

TASK



TOOLS



KEY POINTS:

- Certain tests might work quite well for one purpose, but not so well for something else.
- Some tools are poorly made, and are not useful for much of anything; so are some tests.

Test Purposes & Types



Test Purposes & Types

Tests are referred by **types of decisions** they are used to make:

Curriculum-related decisions	Non-curriculum-related decisions
• Admission • Placement • Diagnostic • Progress • Achievement	• Proficiency • Screening

Curriculum-Related Tests

- **Admission Test:** Used to decide whether or not a student should be admitted to the program at all.

- **Placement Test:** Used to decide at which level in the language program a student should study.

- **Diagnostic Test:** Used to identify learners' areas of strength and weakness.

Curriculum-Related Tests

- **Progress Test:** Used to assess how well students are doing in terms of mastering course content and meeting course objectives. The purpose is to decide what or how fast to teach.
-
- **Achievement Test:** Used to identify how well students have met course objectives or mastered course content. The purpose is to decide the extent to which individual students have learned what they were supposed to.

Non-curriculum-related Tests

- **Proficiency Tests:** Used to assess an examinee's level of language ability without respect to a particular curriculum. (This typically involves assessing more than one narrow aspect/element of language).
-
- **Screening Tests:** Used to decide whether someone is sufficiently proficient (based on a proficiency test) in the target language to be qualified for a particular job.

Activity: Match the types of decisions to the appropriate test type

Test Types

Achievement

Admission

Progress

Placement

Diagnostic

Proficiency

Decisions

Identify learners’ areas of strength and weakness

How well students are doing in terms of mastering course content and meeting course objectives

To identify the level in the language program a student should study

To identify an examinee’s level of language ability

Whether or not a student should be admitted to the program at all

To identify how well students have met course objectives or mastered course content

slido



To identify how well students have met course objectives or mastered course content

① Start presenting to display the poll results on this slide.

slido



Identify learners' areas of strength and weakness

① Start presenting to display the poll results on this slide.

slido



Whether or not a student should be admitted to the program at all

① Start presenting to display the poll results on this slide.

slido



To identify an examinee's level of language ability

① Start presenting to display the poll results on this slide.

slido



How well students are doing in terms of mastering course content and meeting course objectives

① Start presenting to display the poll results on this slide.

slido



To identify the level in the language program a student should study

① Start presenting to display the poll results on this slide.

Activity: Match the types of decisions to the appropriate test type

Test Types

Achievement

Admission

Progress

Placement

Diagnostic

Proficiency

Decisions

Identify learners’ areas of strength and weakness

How well students are doing in terms of mastering course content and meeting course objectives

To identify the level in the language program a student should study

To identify an examinee’s level of language ability

Whether or not a student should be admitted to the program at all

To identify how well students have met course objectives or mastered course content

Test Type Classifications



Other ways of Test Classifications

Tests can be classified/categorized in multiple other ways, according to:

- How we **interpret the results**
- The things that examinees have to do during the test
- The **ways that the tests are scored**

Other ways of Test Classifications-1

With regard to interpreting test scores

Norm-referenced tests (NRT):

- Test takers' results are interpreted by comparing them to how well others did on the test
- Scores are usually reported in percentile scores (NOT percentage correct)
- NRTs tell test users how a particular examinee performed with respect to other examinees, NOT how well that person did in absolute terms.
- NRTs do not tell us how much ability or knowledge examinees demonstrated

Other ways of Test Classifications-1

With regard to interpreting test scores

Criterion-referenced tests (CRT):

- Test takers' results are interpreted in terms of how much learners know in “absolute” terms in relation to one or more standards, objectives, or other criteria
- For CRTs, there is no impact of how much other learners know
- Scores are generally reported in terms of percentage correct (NOT percentile), scoring rubric, or rating scale
- CRTs have at least one (pass vs. fail or mastery vs. non-mastery) or more than one cut scores

Other ways of Test Classifications-2

With regard to when tests are administered and the purposes the results are used for

Summative Assessments:

- They are typically given at end of a unit, course, program etc.
- They provide information about how much students learned (i.e. to sum up how well someone did or can do)

Other ways of Test Classifications-2

With regard to when tests are administered and the purposes the results are used for:

Formative Assessments:

- They take place while students are still in the process of learning something
- They are used to monitor how well that learning is processing (i.e. to shape or form what is being taught)
- They are closely related to language curricula

Other ways of Test Classifications-3

With regard to the testing approach:

Discrete-Point Tests:

- They use a series of separate, unrelated tasks (usually test questions) to assess one “bit” of language ability
- Typically done with multiple choice or other types of selected-response items (e.g., matching, fill-in the blanks)
- They hardly make for life-like language use (NOT authentic: that is, VERY seldom in real life does anyone use language one discrete point at a time)
- They are useful for testing very specific areas of language (e.g. grammar points that have been covered in a unit)

Therefore, discrete-point tests provide an incomplete picture of what learners can actually do with the language

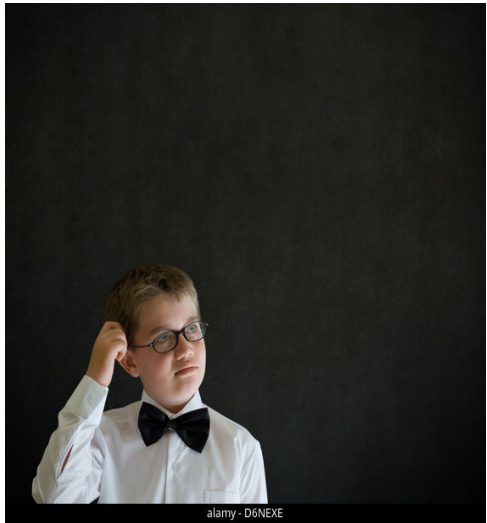
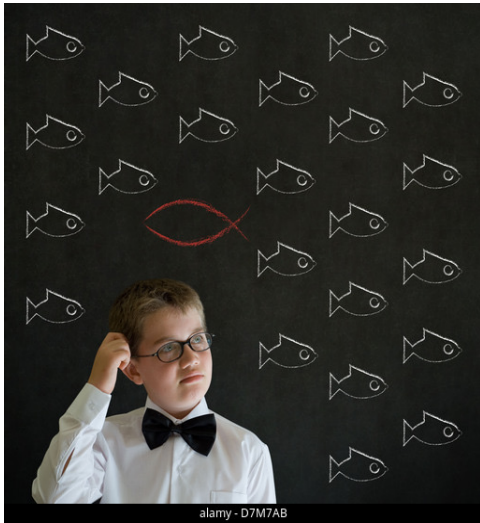
Other ways of Test Classifications-3

With regard to the testing approach:

(Integrated) Performance Tests:

- They require examinees to use multiple aspects of language ability, typically to perform more life-like tasks
- Tasks closely resemble real-life language use tasks, thus require more communicative language use
- They are more difficult to score than discrete-point selected-response questions/items
- The focus is typically on the level of the language used in performing the task

Some skills (because of their target domain) are naturally performance-oriented!



Activity: Choose an appropriate item from each category to measure the following construct

Functional Outcome: Be able to talk about daily routines using simple present tense

Category 1

Achievement

Admission

Progress

Placement

Diagnostic

Proficiency

Category 2

Formative Assessment

Summative Assessment

Category 3

Discrete-point Test

Integrated Performance Test

Category 4

Norm-referenced Test

Criterion-referenced Test

Activity: Choose an appropriate item from each category to measure the following construct

Functional Outcome: The ability to start, maintain and end a conversation on a variety of topics with greater ease and accuracy

Category 1

Achievement

Admission

Progress

Placement

Diagnostic

Proficiency

Category 2

Formative Assessment

Summative Assessment

Category 3

Discrete-point Test

Integrated Performance Test

Category 4

Norm-referenced Test

Criterion-referenced Test

In the SEALC workshop this year, you are going to design a:

proficiency-oriented

performance-based

criterion-referenced

summative

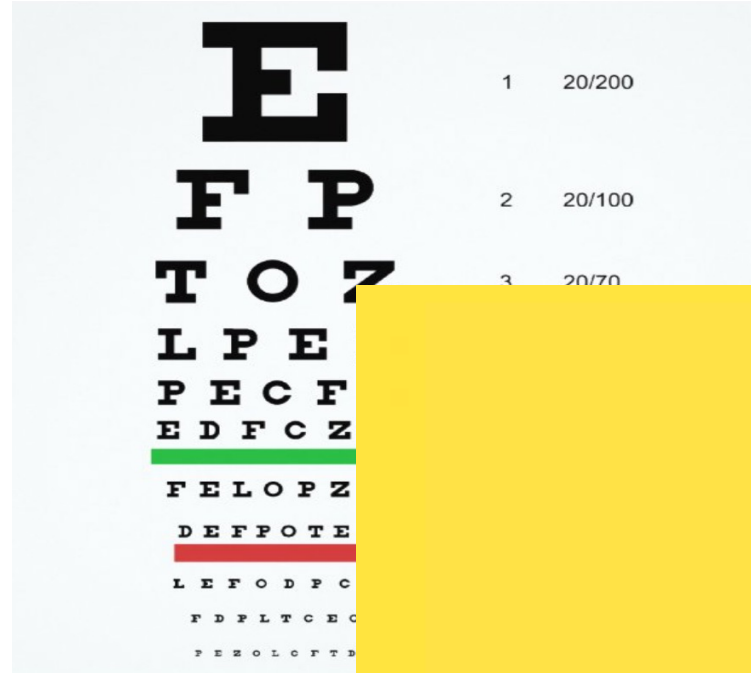
LISTENING test.

Testing & Making Inferences



Testing & Making Inferences

Eye Examination



DECISION



What we Want Tests to Tell us

Language tests act as an operational definition of the construct

What we Want Tests to Tell us

Why defining an intended construct is so important?

We make inferences on the basis of learner performance(s) in a test that acts as an operational definition of the construct.

Misrepresented or underrepresented constructs will NOT tell us what we want tests to tell us!

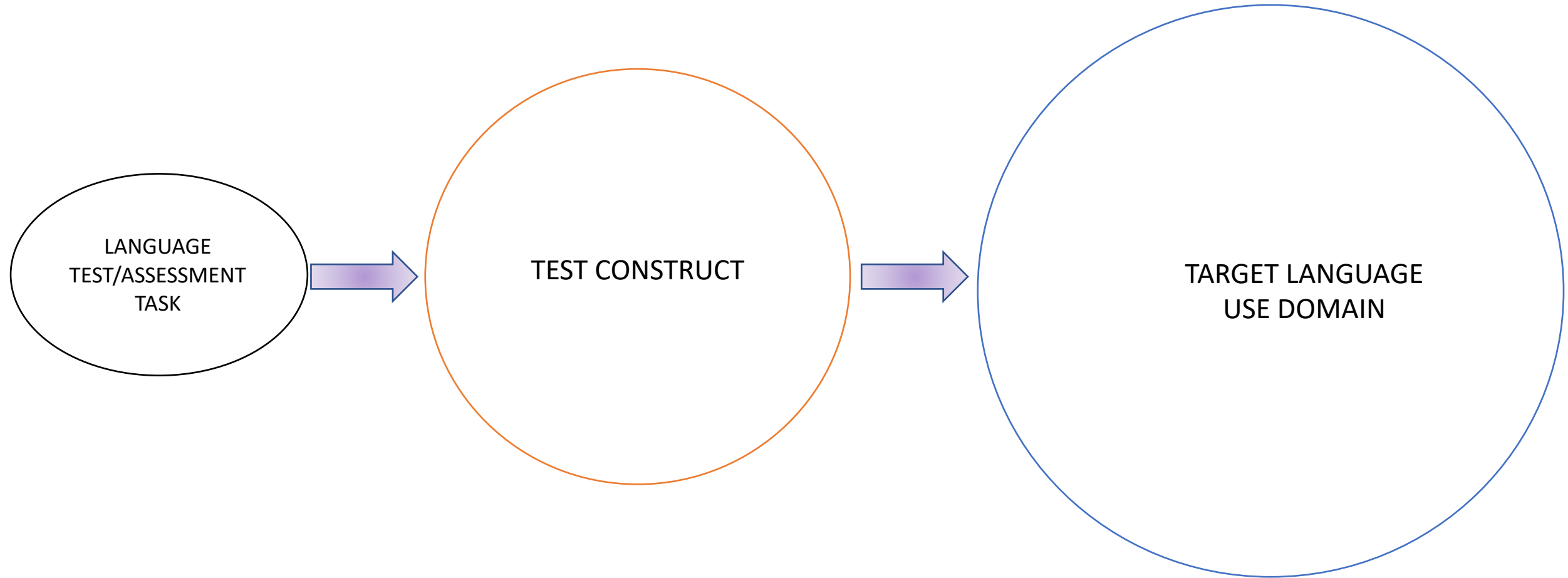
Designing & Developing Useful Tests



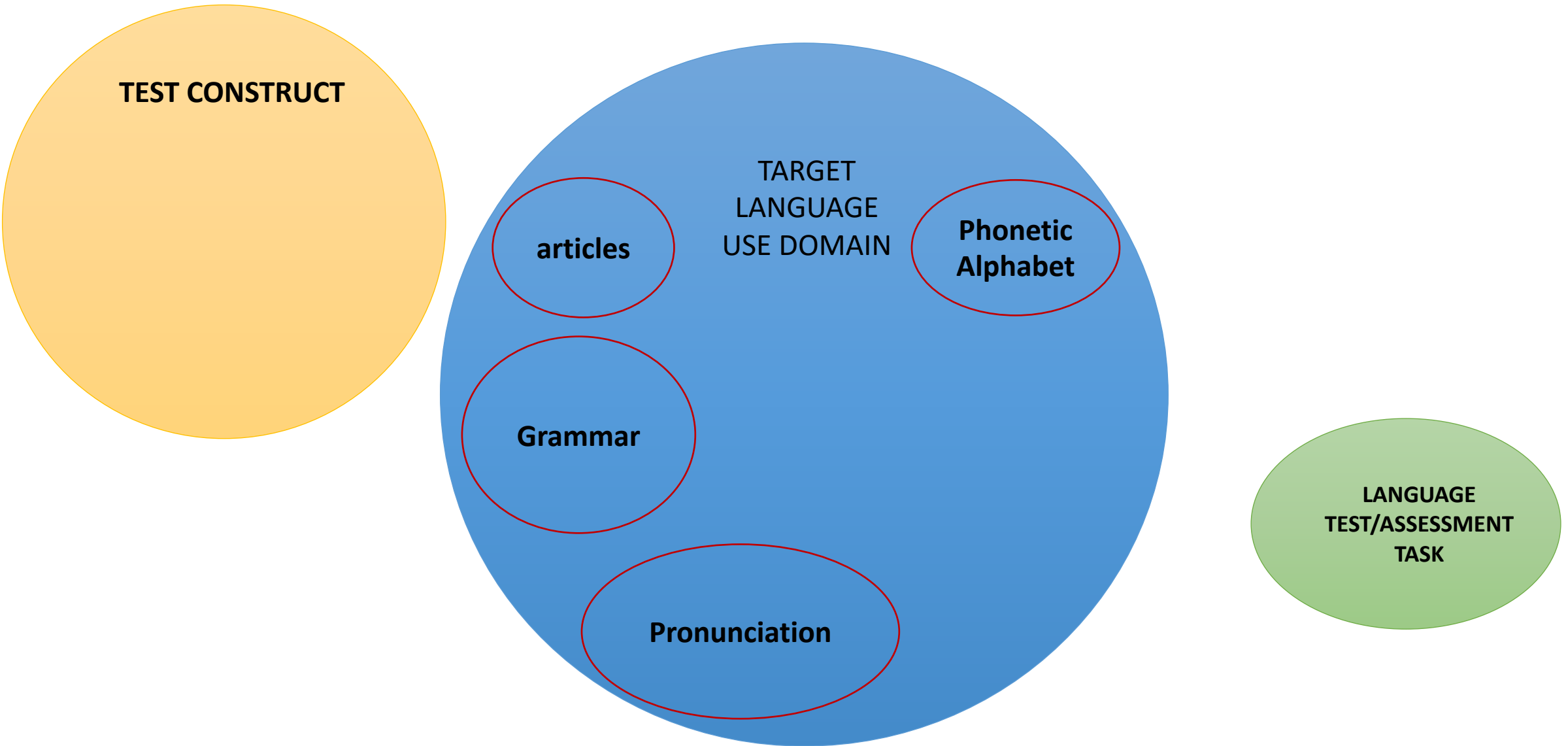
The Key Goal in Testing

In testing or assessing one's language ability or performance, the key goal is to provide **an accurate and valid interpretation of a person's language skills in the real-life context** for which the test is developed (Dursun, 2019).

Designing & Developing Useful Language Tests



Designing & Developing Useful Language Tests



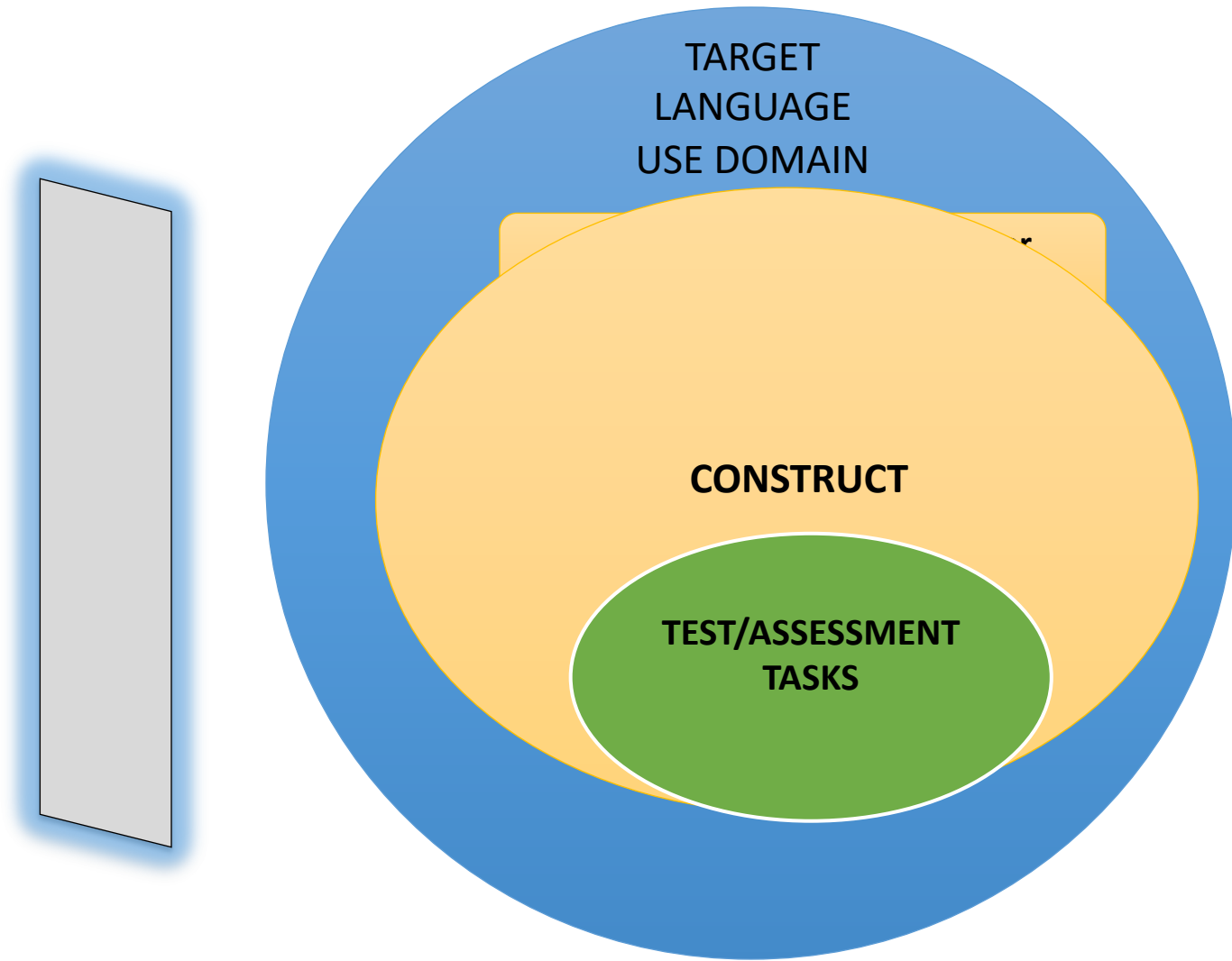
Designing & Developing Useful Language Tests

What it means to know a language

Underlying factors relating to the ability to use language

How we understand specific instances of language use

Fulcher & Davidson, 2007



The diagram consists of two circles. The left circle is orange and contains the text 'CONSTRUCT' and 'The ability to start, maintain and end a conversation on a variety of topics with greater ease and accuracy'. The right circle is blue and contains the text 'Target Domain: Speaking ability in second or foreign in conversational discourse'. The blue circle is significantly larger than the orange circle.

CONSTRUCT

The ability to start, maintain and
end a conversation on a variety
of topics with greater ease and
accuracy

**Target Domain: Speaking ability in
second or foreign in conversational
discourse**

Test Task: Option 1

GOING SHOPPING



Fill in the blanks with an appropriate word or expression and write the numbers.

AT A CLOTHES SHOP

Salesperson: Excuse me. Can I help you?

Customer: Oh, yes, please. *I'm looking for* a jacket to wear at a party I'm attending tonight.

Salesperson: We have this nice selection of leather jackets *on sale*. *Would you like to* take a look?

Customer: Yes, sure. [...] this black one?

Salesperson: It's \$..... (one hundred fifty-nine ninety-nine).

Customer: Wow! It's *really* Can I try it on?

Salesperson: *Of course*.

Customer: [...] I think it's small. Do you have it large?

Salesperson: Yes, *here you are*. The large one is \$..... (two hundred nine ninety-nine).

Customer: Oh, okay. [...] This one is much better. I will it.

Salesperson: Excellent choice!

AT A GIFT SHOP

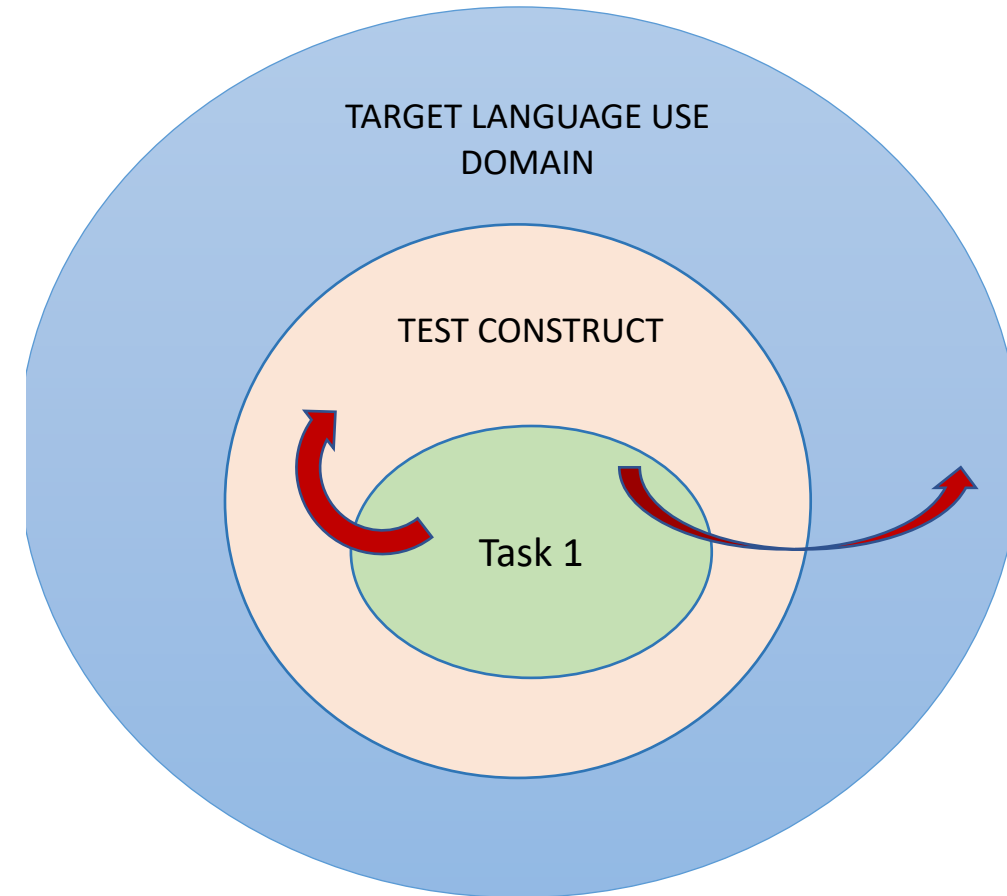
Salesperson: Good evening. How can I help you?

Customer: Hi, good evening. a nice *gift* for someone's birthday.

Salesperson: I see. Is it a *child* or an *adult*?

Customer: It's an adult.

What can we infer?



Test Task: Option 2

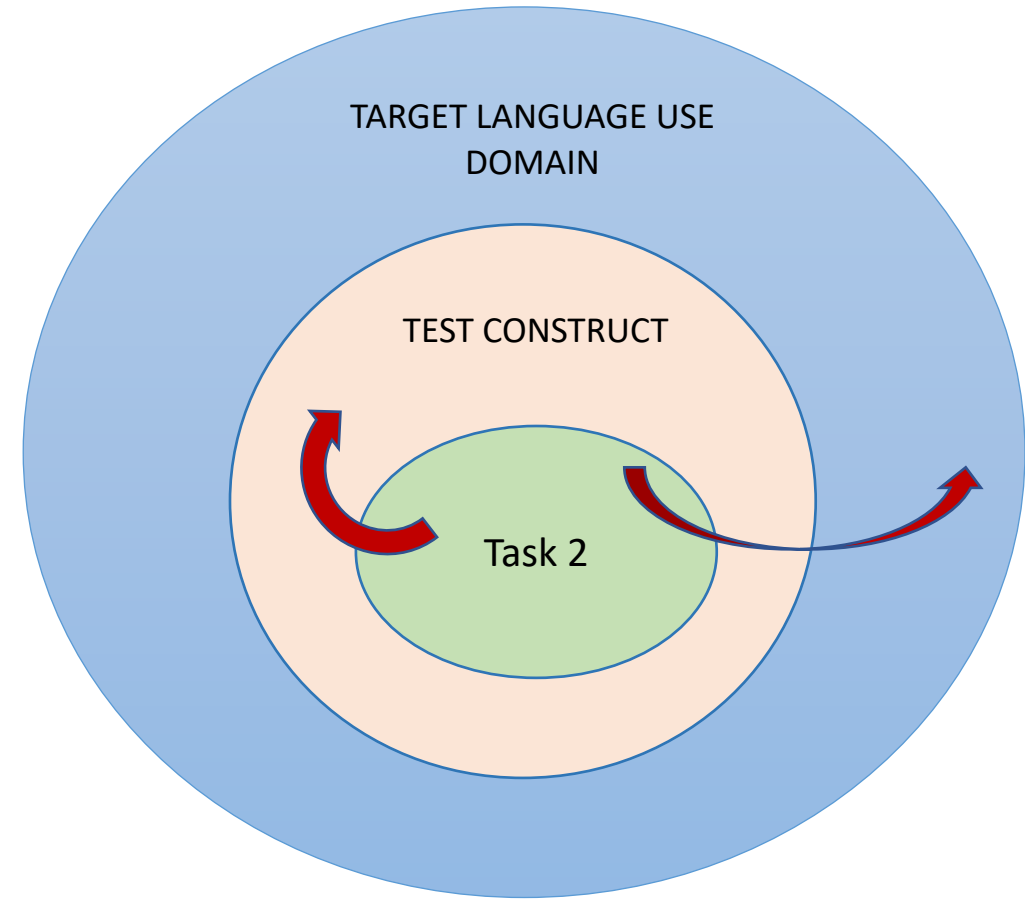
Conducting a mini OPI:

An authentic, highly-structured one-on-one conversation lasting 5-10 minutes.

- Your goal is to elicit the test taker's best possible sample of functional speaking ability, rather than only accuracy and grammar.
- It will include a role play where the test taker and the language specialist act out a scenario that the test taker could encounter in a region where the language is spoken.

.

What can we infer?



Designing & Developing Useful Language Tests

Key Takeaways:

- The test should be useful for its intended purposes

Therefore, before we even start developing a test, we should define:

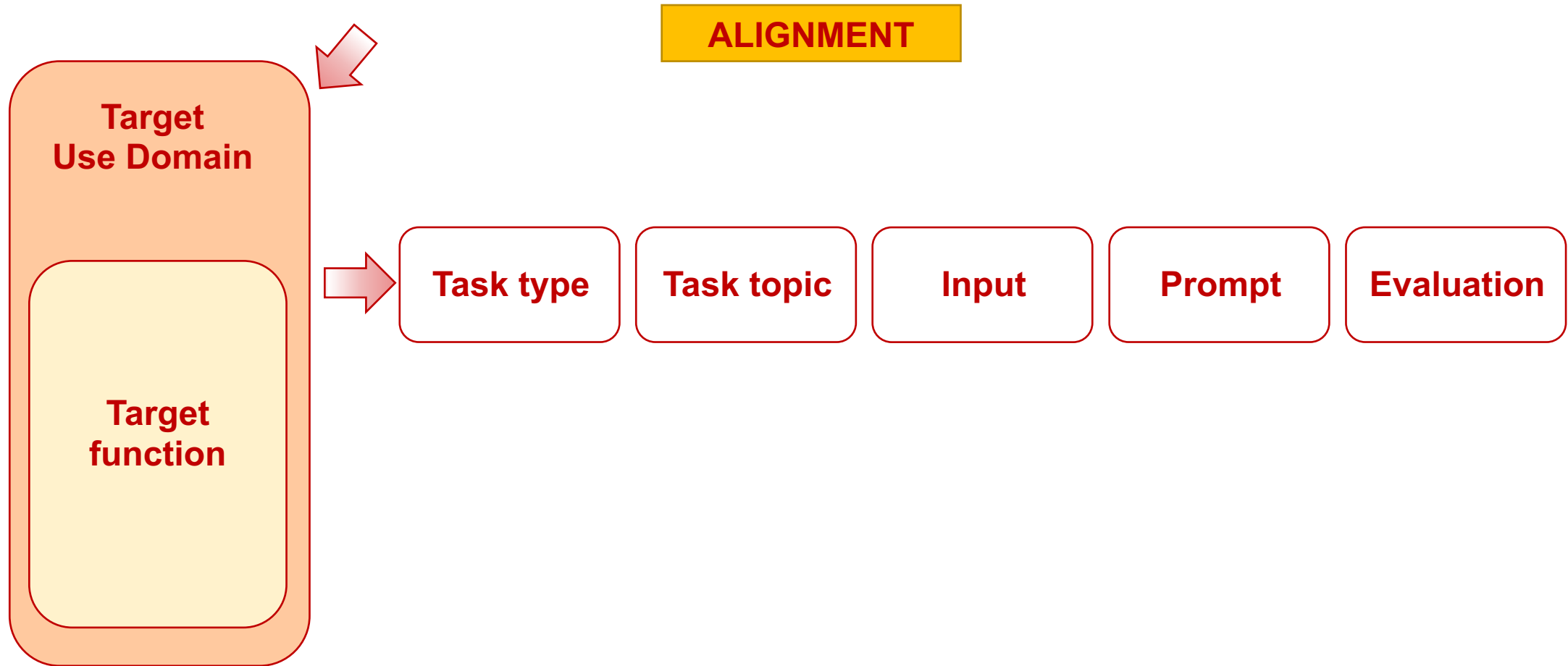
- the intended purpose(s)
 - target language use domain for the intended purpose(s)
 - the construct(s)
-
- The construct(s) must be representative of target language use domain
 - The test tasks must well represent/measure the construct(s) of interest (with correct and adequate number of test tasks)
 - The test should do its job with fairness
 - The decisions made based on a test should be based on the intended purposes

Test Design Framework



How do we measure Language Proficiency?

- Proficiency-oriented performance-based assessment design



Everything from **designing to developing an assessment task**, and then
from **grading the task to reporting the its results** **MUST work in harmony**
to elicit the evidence that the learner can function in the **TARGET**
LANGUAGE USE DOMAIN.

Proficiency-oriented performance-based assessment design

Design Matrix: A blue print

Target Skill & Level	Target Functions	Topics / Inputs	Assessment Task Types	Response Types	Evaluation Criteria	Notes
	Functions	Context Content	Tasks	Elicitation	Performance Accuracy Comprehensibility	

Assessment Design Matrix – LISTENING

	Target Functions	Topics / Text Types	Assessment Task Types	Response Types	Evaluation Criteria	Notes
LISTENING NOVICE	- Can understand key words, true aural cognates, and formulaic expressions that are highly contextualized and highly predictable - Can understand words and phrases from simple questions, statements, and high frequency commands - Can recognize rather than fully comprehend	Topics: familiar, everyday (greetings, demographic info, numbers, family, hobbies, likes/dislikes, weather, clothing, simple descriptions, food, common objects, etc.) Text types: Messages or announcements (voice messages, airport announcements), exchanges or transactions, conversations, monologues	- Tables or charts, <i>e.g. Fill in the table with the foods Marie and John like, don't like, and plan to buy.</i> - Open-ended questions that require comprehending a word or list of words, phrases or numbers, <i>e.g. List Joe's family members and their ages.</i>	- Words and phrases, especially lists (e.g. lists of nouns, verbs, numbers) - Responses in English to prevent interaction with writing or spelling ability.	- Accurate content (i.e. should answer the question) - Accurate words and phrases	- Prompts and/or column headings should be in English so that tasks are clear. - Prompts should always provide context, <i>"You'll hear a weather report for Barcelona."</i> - Prompts should elicit lists whenever possible, avoid prompts that elicit a one-word or one-phrase response - Filling in tables and charts help listeners organize information.
LISTENING INTERMEDIATE	- Can understand speech that conveys basic information simply, with minimal connection, containing high-frequency vocabulary - Can understand adequately, both the overall message as well as some details, despite occasional gaps and misunderstandings	Topics: a broad range of familiar topics pertaining to daily life; personal interactions and social transactions (demographic info, family, hobbies, likes/dislikes, travel, weather, clothing, simple descriptions, living situations, travel, etc.) Text types: - Messages/ announcements (voice messages) - Exchanges or transactions - Conversations in person or on the phone - Monologues / Interviews	- Recall protocols - Open-ended questions that require comprehension of simple sentences or phrases, <i>e.g. What are all the things Ali will do on his day off?</i> - Summary tables, <i>e.g. Fill in the table with the things Elena and Vlad plan to do on vacation, don't want to do on vacation, can't afford to do on vacation</i> - Outlining (using diagrams that fit the organization of the text)	Responses in English to prevent interaction with writing or spelling ability. NOTE: The expected response should be evaluated in terms of the extent to which they need to LISTEN and comprehend in order to complete the task or respond to the question.	- Have understood the information from the text [Comprehension is evidentiary through complete idea units.] - Provide accurate information NOTE: Information not contained in the original text should earn no points and does not count as an error or cause the student to lose points	- An idea unit is usually a verb + complement <i>e.g. Elena wants to swim, go to museums, eat new foods</i> - Prompts and/or table headings should be in English and always provide context, <i>"You'll hear Elena and Vlad talking about travel plans."</i> - Listeners relay heavily on redundancy, restatement, paraphrasing and contextual clues.

Task Design Framework



Developing Performance-based Tasks Scenarios

- Proficiency-Oriented Performance Tasks¹:

- ✓ Present test takers with **a problem, key challenging situation**

A realistic situation/problem

- ✓ Push test takers to develop **a concrete product** or **performance** for an intended audience

To “do” the subject

- ✓ Set **evaluation criteria** and **performance standards**

Articulation of evaluation criteria

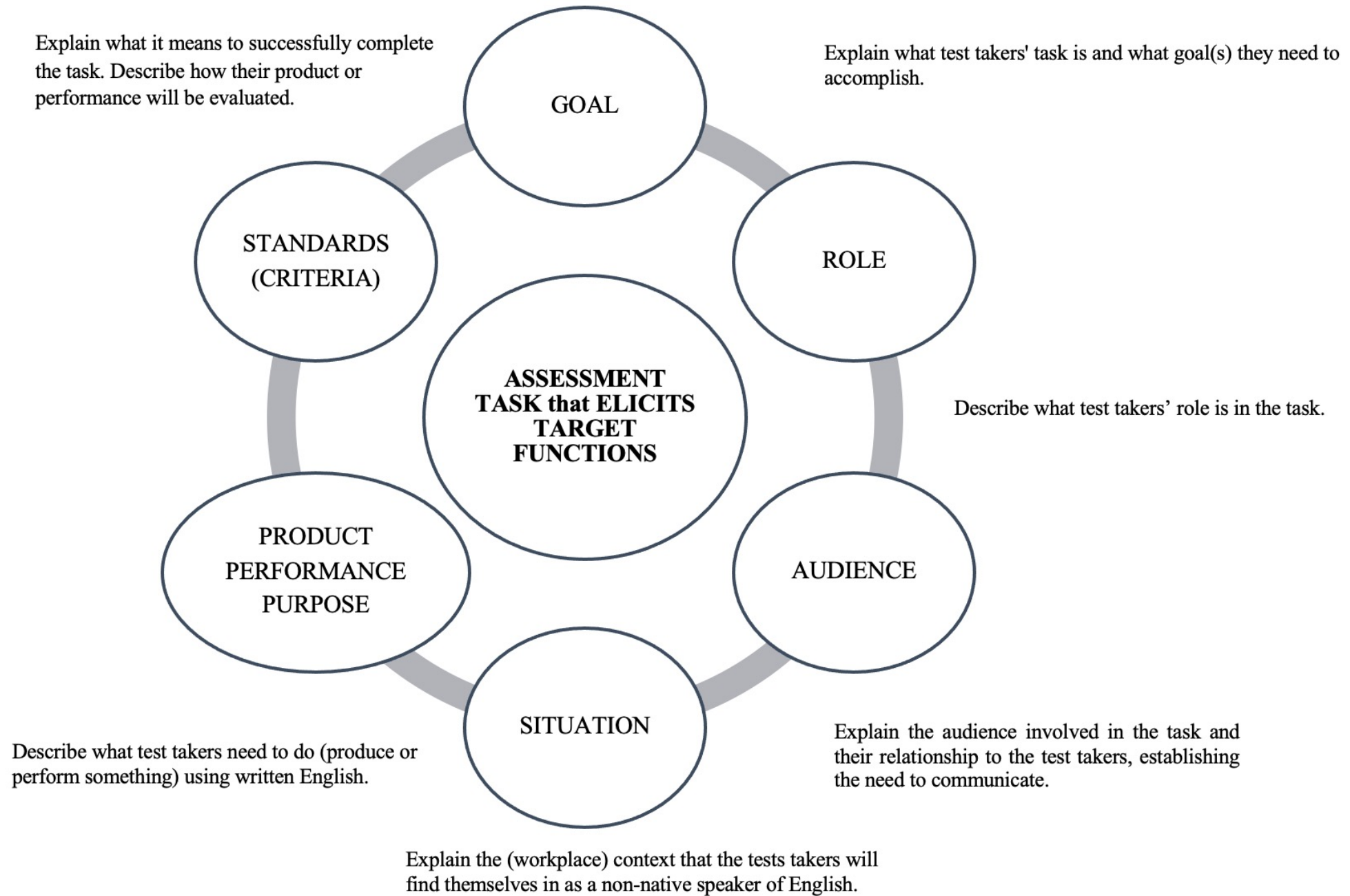
Checklist

CONTEXTUALIZED
AUTHENTIC, RELEVANT

HOW TO PERFORM vs.
WHAT TO PERFORM

SENSE OF TASK COMPLETION

¹ Adapted from Wiggins, G., & McTighe, J. (2008). *Understanding by Design Professional Development Workbook*. Alexandria, VA: Association for Supervision and Curriculum Development.



This model was selected for a few reasons:

- It defines and sets criteria to indicate successful completion of authentic proficiency-oriented, performance-based tasks
- It defines clear parameters to contextualize the tasks
- It shifts the task completion criteria from quantity-based standards to quality-based, real-world language use criteria
- It allows the measurement of **sociocultural competency** and **communication strategies** in a task-based assessment model by specifying the minimum characteristics needed to complete an assessment task, including:
 - communicative context
 - a specific audience and the relationship to that audience, and a clear purpose towards a realistic and relevant goal

Oxford Handbooks

for Language Teachers

Language Assessment for Classroom Teachers



Lyle Bachman | Barbara Damböck

OXFORD

Thank You!

Questions?

Evaluating Tests



Qualities of Usefulness

How do we evaluate the usefulness of a test?

Classroom-based Assessments:

- Content match
- The opportunity to notice a gap (scoring reports/ results summary)
- Suggestions for improvement (feedback)

Qualities of Usefulness

A Possible Framework Evaluating Classroom-based Assessments

First Step: Analysis of content in the learning program (curriculum, syllabi, course book, etc.)

Second Step: Analysis & Evaluation of content coverage in assessments

Is the content in the text/materials adequately covered?

Third Step: Evaluation of the perception of the gap (feedback) in assessments.

To what extent assessments give students the opportunity to notice the gaps in their knowledge/learning?

Fourth Step: The evaluation of closing the gap in assessments (evaluation of the content of the feedback provided after the assessment).

Would feedback (real-time or delayed) help a student and his/her teacher to address strengths and weaknesses?

Qualities of Usefulness

How do we evaluate the usefulness of a test?

High-Stakes Assessments (Bachman & Palmer, 1996)

QUALITY	DEFINITION
Reliability	Consistency of scoring (i.e. to determine how much effect has error has on test scores)
Authenticity	The degree to which test tasks resemble TLU tasks (both situationally and interactionally)
Construct Validity	The degree to which it is appropriate to interpret a test score as an indicator of the construct (i.e., the ability) of interest
Impact	Effects of tests on people, institutions, including (but not limited to) washback-the effect of a test on teaching and learning
Practicality	The extent to which the demands of the particular test specifications can be met within the limits of existing resources.

Table 1.2 Bachman and Palmer's Qualities of Usefulness (Carr, 2011, p. 20).

Activity: Qualities of Usefulness

QUALITY	DEFINITION
Reliability	Consistency of scoring (i.e. to determine how much effect has error has on test scores)
Authenticity	The degree to which test tasks resemble TLU tasks (both situationally and interactionally)
Construct Validity	The degree to which it is appropriate to interpret a test score as an indicator of the construct (i.e., the ability) of interest
Impact	Effects of tests on people, institutions, including (but not limited to) washback-the effect of a test on teaching and learning
Practicality	The extent to which the demands of the particular test specifications can be met within the limits of existing resources.