University of Jendouba

Higher Institute of Applied Languages and Computer Science (ISLAIB) Beja- TUNISIA

Empirical Reading Test Construct Validity: Some Emerging *Ethical* Issues
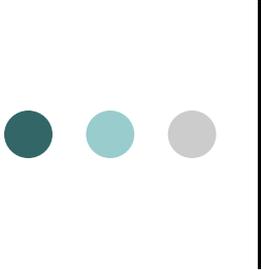
Asma MAAOUI GAALOUL

The 8th Annual EALTA Conference, 2011
Ethics in Language Testing and Assessment
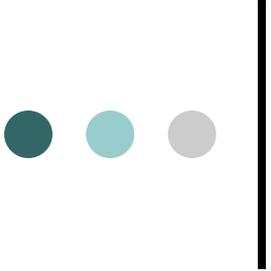
Outline

- Background and Rationale
- The study
- Methodology
- Findings (preliminary)
- Conclusions and Implications

Background and Rationale

○ Confirming validity claims of language assessment tools is a multifaceted and complex process in which the test designer's and test taker's perspectives should be both investigated.

○ From a social and ethical perspective, testing is "an important but potentially dangerous component of language teaching" (Spolsky, 2000:537).

○ Validity is "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (Messick, 1989:13).
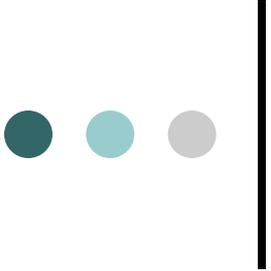
- "Tests are not developed and used in a value-free psychometric test-tube" (Bachman, 1990:279)

- Validity research and arguments have been inevitably oriented towards the social and political spheres relating to testing purposes (Davidson & Lynch, 2002; Shohamy, 2001).

- "Validation of the evidential basis of test-score interpretation involves thinking about the *logic* of the test, particularly its design and its *intentions*, and also looking at the empirical evidence" (McNamara, 2006:33; emphasis mine).
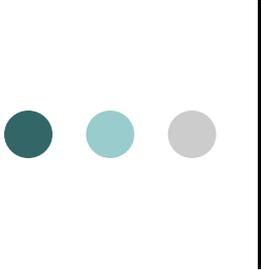
A "useful" (Bachman & Palmer, 1996) and valid reading comprehension test should be designed to measure the examinee's reading ability in the target language.

➡️ Do the strategies used during item solving reflect what is intended to be measured?

"Our conceptions of what it is that we are measuring and the things we prioritize in measurement, will reflect *values*, which we can assume will be social and cultural in origin […]"
(McNamara & Roever, 2006, emphasis mine).

The study generally aims at empirically confirming:

- the degree of validity of this EFL reading test through short answer questions (SAQ);

- the nature of "strategic competence" (Bachman & Palmer,1996) in language test performance;

- what distinguishes more high from low-achievers in language tests.
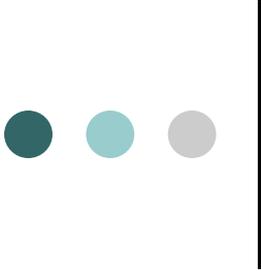
The study

## The research site and study participants

- During an end-of-term (high stakes, achievement) reading comprehension test (duration=1 hour).
- the Department of English at the Faculty of Human and Social Sciences in Tunis, Tunisia.

Primary participants	Secondary participants
Regular First Year students of English as a Foreign Language (EFL) at the tertiary level **(N=571)**	**Three** teachers of EFL reading and writing (the test developers and raters)
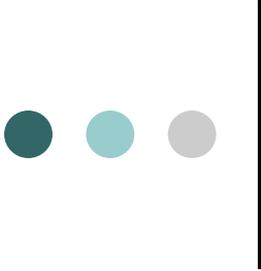
The paper-and-pencil RC test

	Question type	Item format	Nb of items
Section 1	Open-ended/ short answer	open-ended format	3
Section 2	True/ false with justification	Fixed format/ open-ended format	3
Section 3	Paraphrasing statements	open-ended format	2
Section 4	Summary of specific text ideas	open-ended format	1
Section 5	Vocabulary	open-ended format	4
Total	5		13
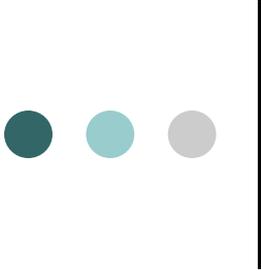
Research Questions

RQ1: What are EFL teachers' perceptions about the reading test and to what extent does this relate to the test takers' actual strategy use?
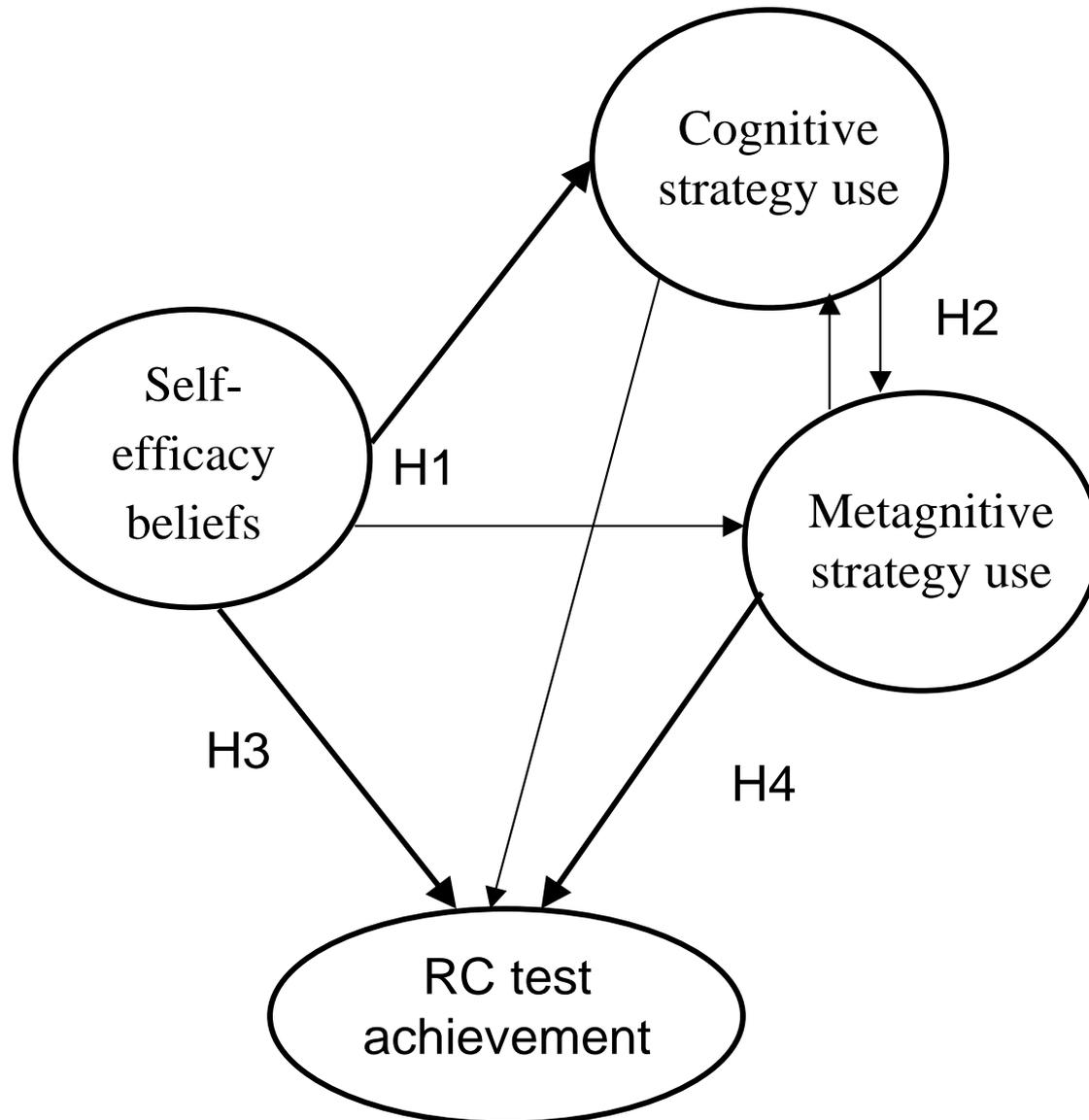
RQ2: What are the cognitive, metacognitive and test taking strategies (TTS) that the examinees use during item solving?
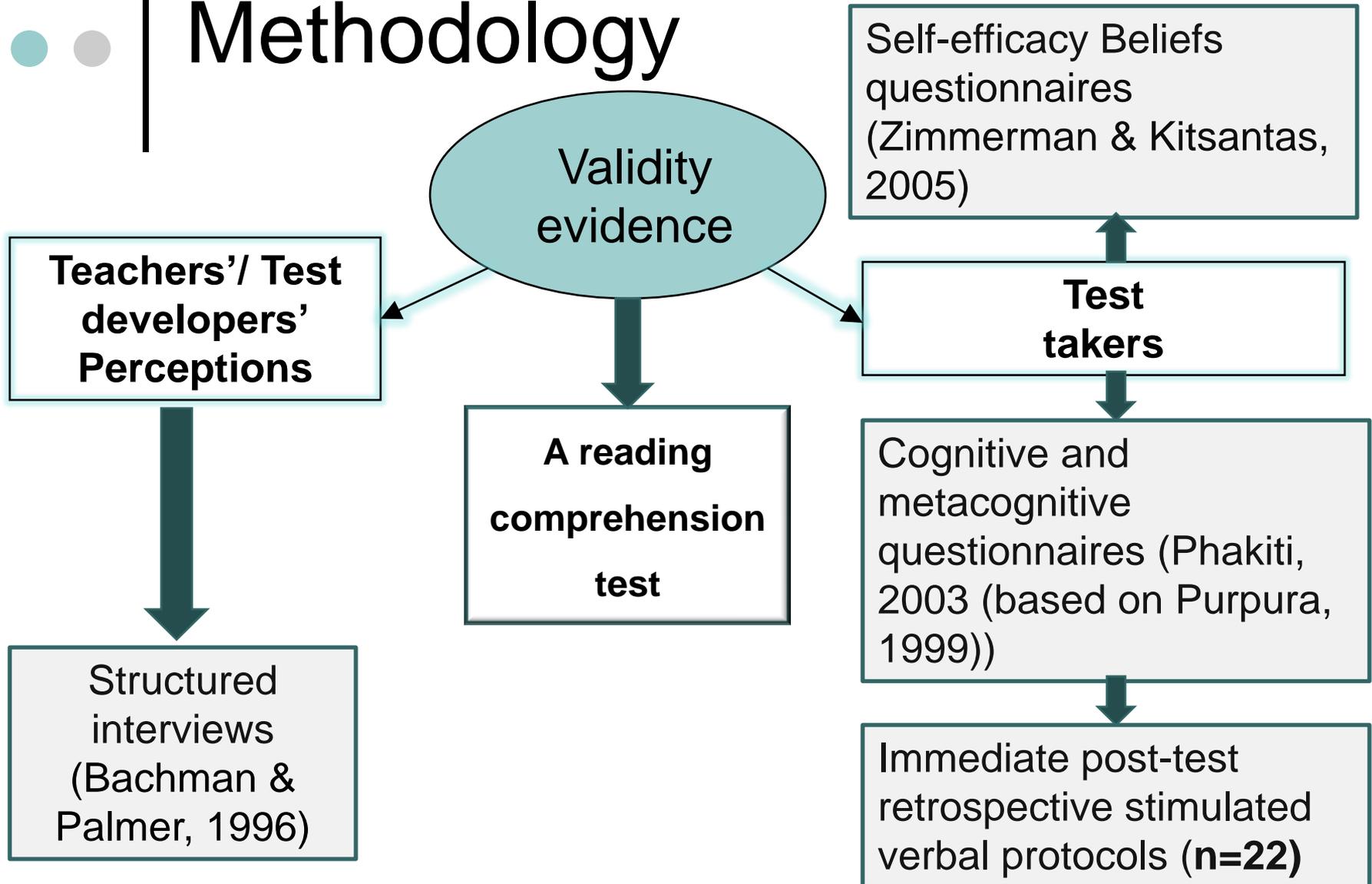
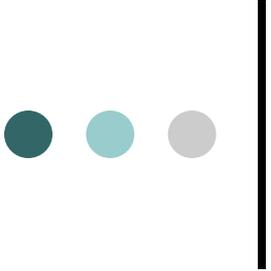RQ3: What is the relationship between the test takers' perceived strategy use, TTS and achievement?

RQ4: How do individual test takers differ in the use of cognitive and metacognitive strategies?

The Hypothesized Structural Model

Methodology

Validity evidence

Teachers'/ Test developers' Perceptions

Test takers

Self-efficacy Beliefs questionnaires (Zimmerman & Kitsantas, 2005)

A reading comprehension test

Structured interviews (Bachman & Palmer, 1996)

Cognitive and metacognitive questionnaires (Phakiti, 2003 (based on Purpura, 1999))

Immediate post-test retrospective stimulated verbal protocols (**n=22**)

Findings (1)

Descriptive Statistics

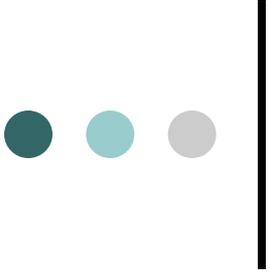|	N	Mean	SD	Variance	Kurtosis	Skewness
Reading achievement	571	7.7	3.2	10.2	-0.4	0.2

Kurtosis (Standard error)= 0.2
Skewness (Standard error)= 0.1
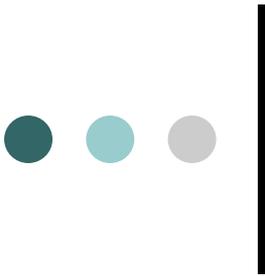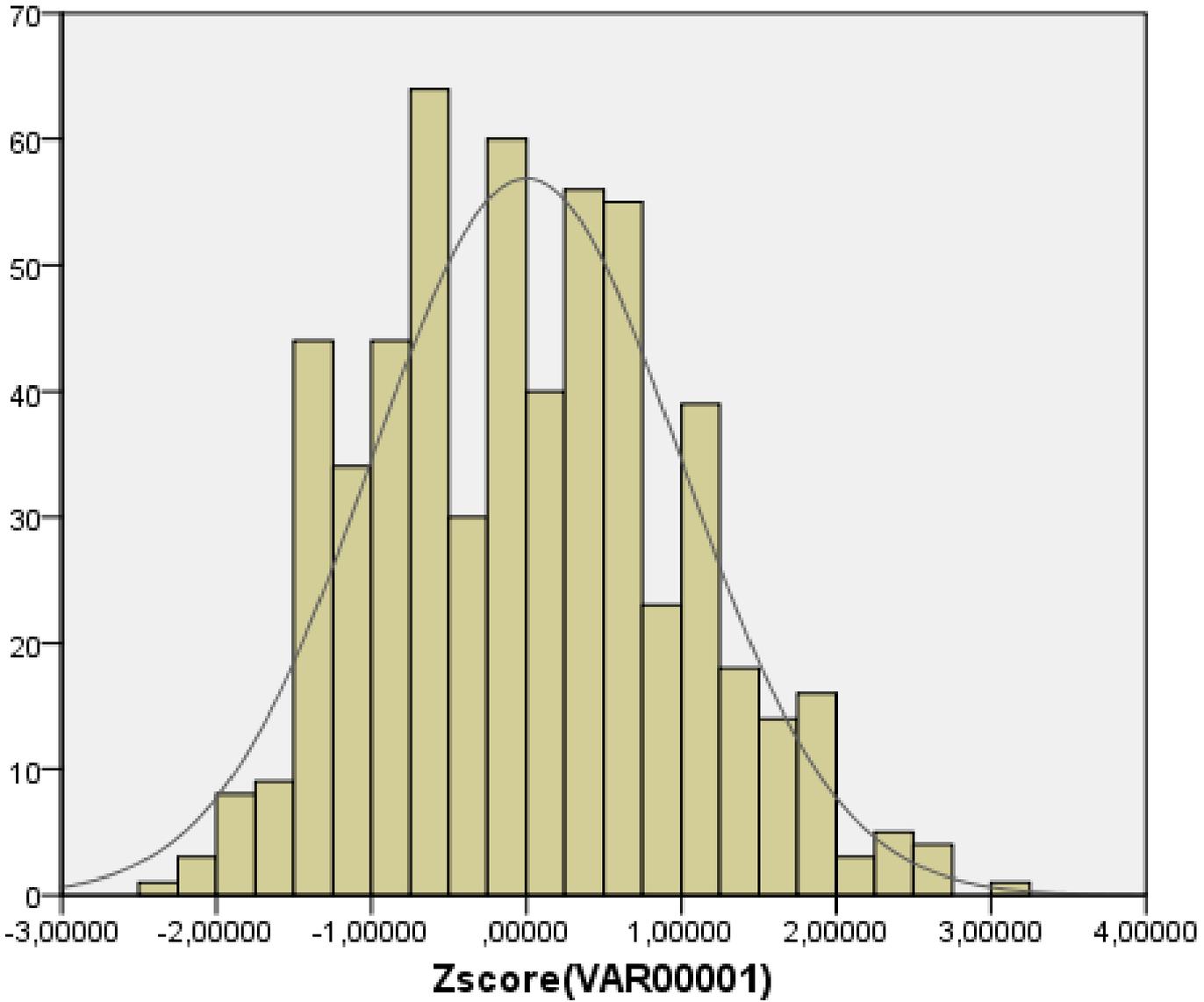
SD = 3.2
Variance= 10.2

Findings (2)

The test developers' profiles versus their perceptions

- The test developers exhibited limited knowledge of test design and only one of them had a limited training in testing;

- the test developers' professional background might be a major cause of how they perceived and defined the measured reading ability (two of them were specializing in literature);

- The test takers' strategy use reflected reading-irrelevant difficulties during item processing (writing and paraphrasing/ vocabulary knowledge problems)
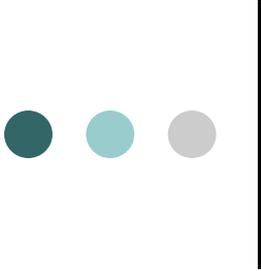
Findings (3)

The participants' views of the strategies required for item solving

	Teacher 1	Teacher 2	Teacher 3
Open-ended/ short answer	Skimming and scanning	Understanding discourse/ non-linear reading	Paraphrasing the text ideas
True/ false with justification	Being able to read and understand detailed ideas	Being able to justify what is correct and understand that an idea is false	Make the difference between correct and incorrect information
Paraphrasing statements	Getting the text meaning from the context Understanding keywords and finding equivalents to them	Expressing the original text ideas in a different way	Paraphrasing by using their words and by changing the sentence structure

| Summary of specific text ideas	Summarize the text ideas	Provide a coherent and well written summary	Sum up the text using their own words
Vocabulary	Finding equivalents	Reading the text and findings equivalents	Understanding word meaning by using the context

- Intuitively developed test items;
- no clear agreement concerning *what* was being tested;
- there was almost a total absence of any planning during item construction;
- the measured construct could not be explicitly defined by the test designers.
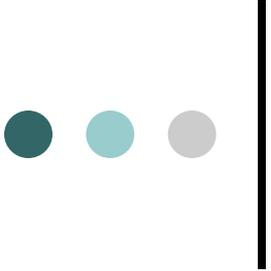
Conclusions and implications

- The measured reading ability was relatively underrepresented;

- variance in test performance was not necessarily dependant on different levels of reading ability;

- inferences and scores were probably"muddied" by a knowledge-related problem of *what* and *how* to measure EFL reading;

- the construct validity of this teacher-produced test might have been affected by a lack of test planning.
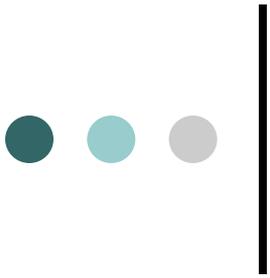
▶ The study of the test designer's validity claims versus actual test performance is extremely important in different educational contexts;

▶ Whatever the paradigm in which an EFL reading course is framed, it is important for teachers to be aware of the practical demands of valid reading assessment.
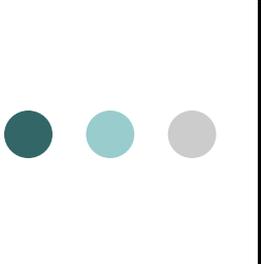
Benefits of the study

○ Combining qualitative (immediate post-test retrospection) with Structural Equation Modeling (SEM)

○ Relying on **multiple sources of validity evidence**

○ Adopting a cognitive psychology framework in the interpretation of the test performance data

○ Studying the test developer's underlying perceptions of test design (intended and unintended)

Ethical and responsible assessment is a "shared responsibility" (Shohamy, 2001).
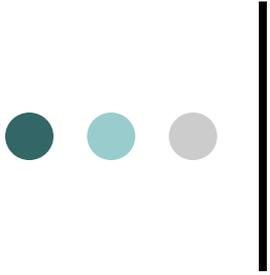
A top-down power-based testing culture

The political context

The educational context

The institution

The test designer

The test taker
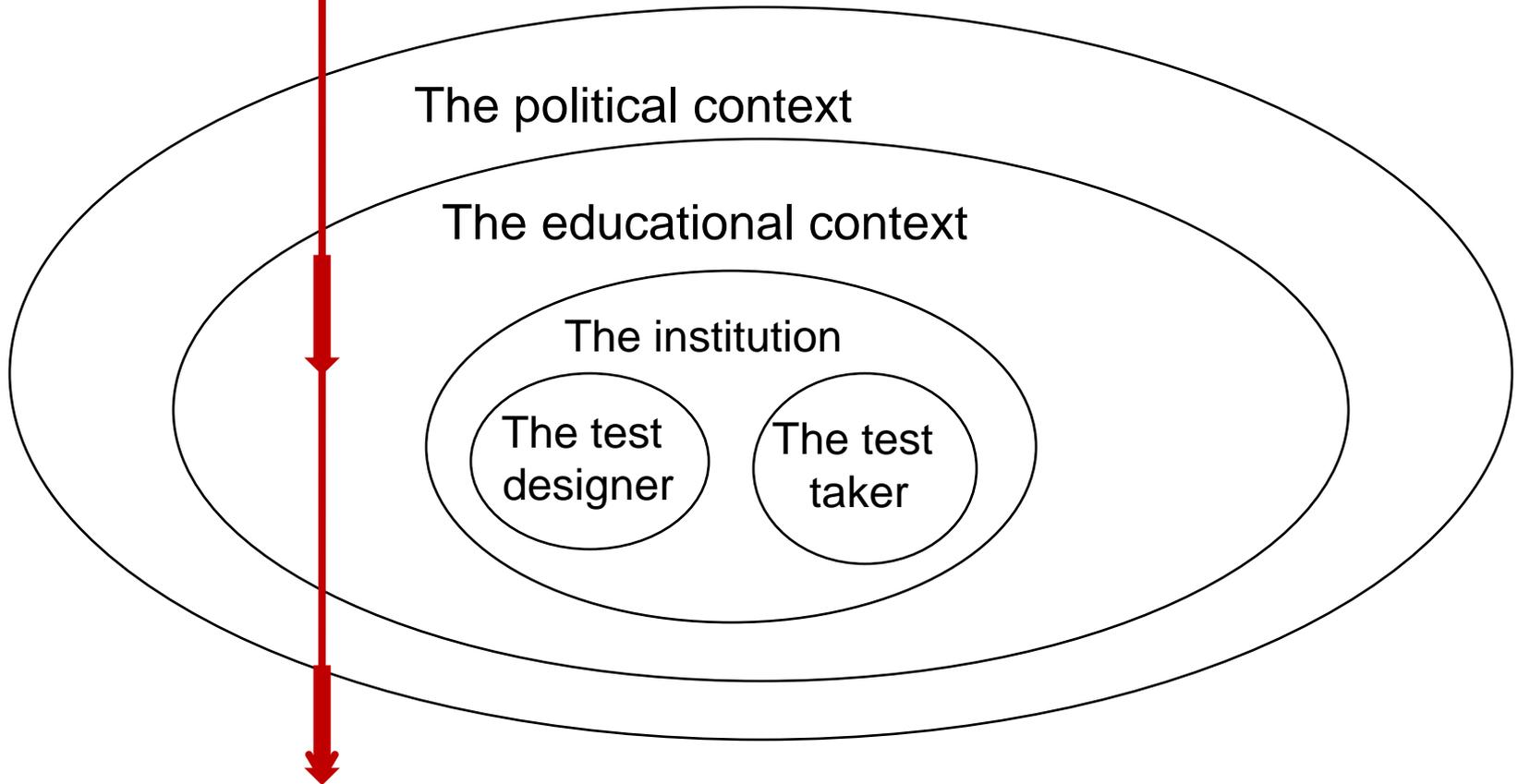
Issues of professional responsibility and ethicality are raised!

References

- Bachman, L. F. & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

- Davidson, F. & Lynch, B. K. (2002). *Testcraft. A Teacher's Guide to Writing and Using Language Test Specifications.* New Haven and London: Yale University Press.

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

- McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing,* 18(4), 333-350.

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

- Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing,* 20(1), 26–56.

- Purpura, J. E. (1999). *Learner Strategy Use and Performance on Language Tests: A Structural Equation Modeling Approach*. Cambridge: Cambridge University Press.

- Roever, C. & McNamara, T. (2006). Language testing: the social dimension. *International Journal of Applied Linguistics,* 16(2), 242-258.

- Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing,* 18(4), 373–391.

- Spolsky, B. (2000). Language testing in *The Modern Language Journal*. *The Modern Language Journal,* 84, 536-552.

- Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Houndmills: Macmillan.

Thank you