BRITISH COUNCIL

Tensions and synergies in standardized testing: making the numbers meaningful

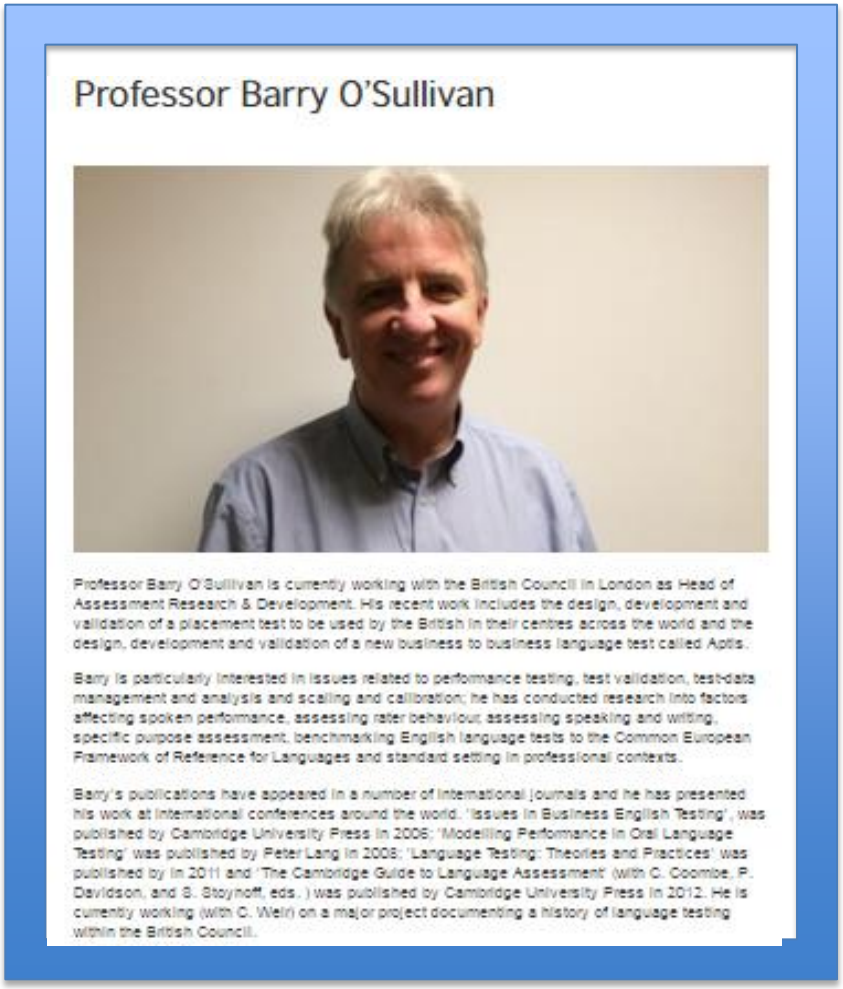
New Directions Hanoi 2016

Dr Jamie Dunlea

Assessment Research Group

Research and test expertise

- Assessment Advisory Board
- Assessment Research Awards and Grants
- Assessment Research Group
- Publications
- Research projects
- The Davies Lecture



Professor Barry O'Sullivan

Professor Barry O'Sullivan is currently working with the British Council in London as Head of Assessment Research & Development. His recent work includes the design, development and validation of a placement test to be used by the British in their centres across the world and the design, development and validation of a new business to business language test called Aptis.

Barry is particularly interested in issues related to performance testing, test validation, test-data management and analysis and scaling and calibration; he has conducted research into factors affecting spoken performance, assessing rater behaviour, assessing speaking and writing, specific purpose assessment, benchmarking English language tests to the Common European Framework of Reference for Languages and standard setting in professional contexts.

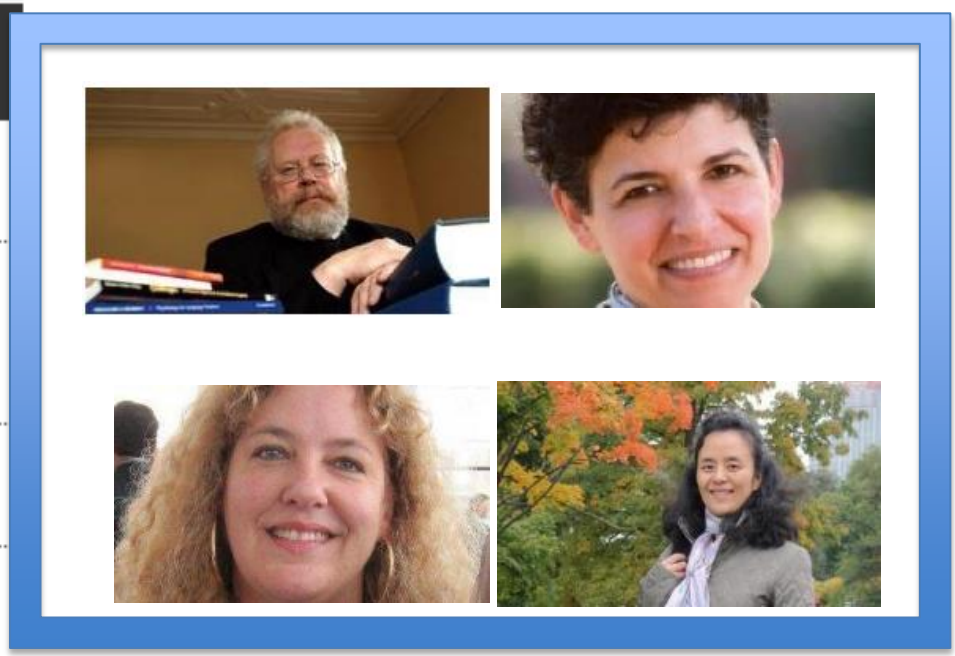
Barry's publications have appeared in a number of international journals and he has presented his work at international conferences around the world. 'Issues in Business English Testing', was published by Cambridge University Press in 2006; 'Modelling Performance in Oral Language Testing' was published by Peter Lang in 2008; 'Language Testing: Theories and Practices' was published by in 2011 and 'The Cambridge Guide to Language Assessment' (with C. Coombe, P. Davidson, and B. Stoynoff, eds.) was published by Cambridge University Press in 2012. He is currently working (with C. Weir) on a major project documenting a history of language testing within the British Council.

http://www.britishcouncil.org/exam/aptis/research

Assessment Research Group

Research and test expertise

- Assessment Advisory Board
- Assessment Research Awards and Grants
- Assessment Research Group
- Publications
- Research projects
- The Davies Lecture



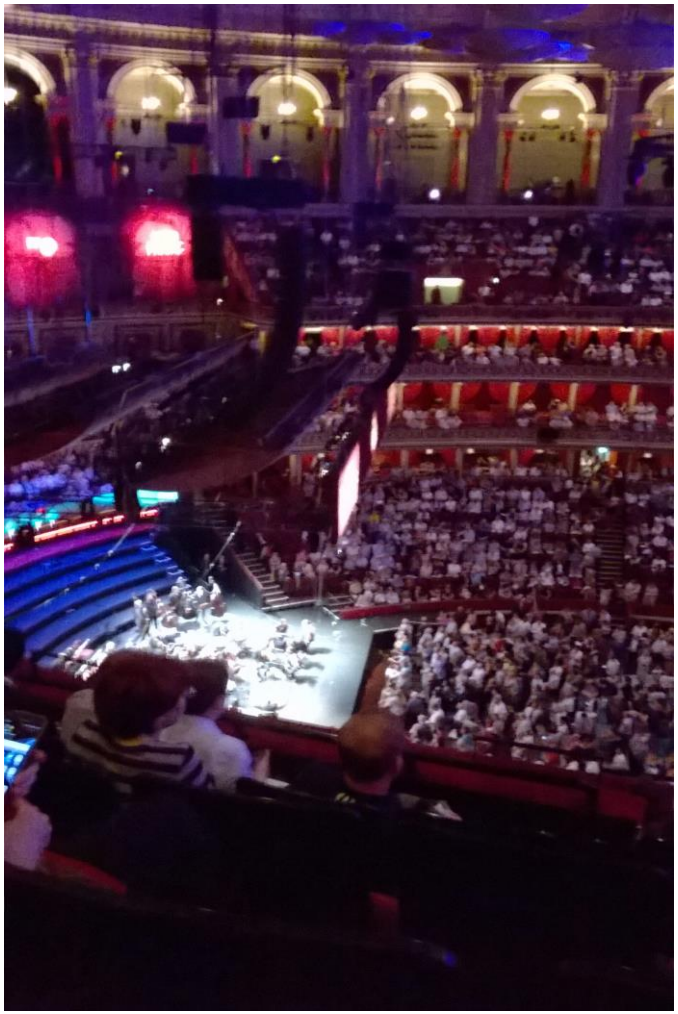
http://www.britishcouncil.org/exam/aptis/research

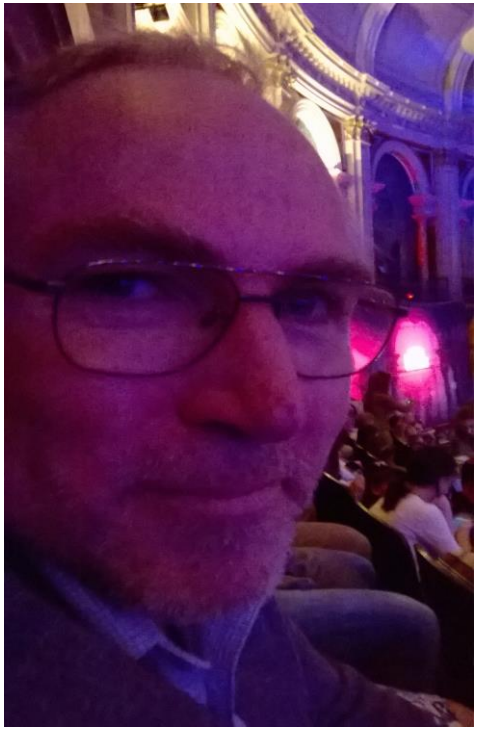
Assessment Research Group

Research and test expertise

- Assessment Advisory Board
- Assessment Research Awards and Grants
- Assessment Research Group
- Publications
- Research projects
- The Davies Lecture

Tensions and synergies





Tensions and synergies

- (Age-old) validity / reliability tensions
- Measurement ideals and practical realities
- Test users' demands and needs and the limits of reliable, meaningful measurement
- Between tests with wide applicability/usability and the localized needs of each context of use
- Between feedback which is interpretable and comparable across contexts and meaningful for individuals and individual contexts

Tensions and synergies

- Why do we assess / test / evaluate?
- Is the feedback we provide / get from assessment really informing learning and teaching?
- Is what we teach (and test) relevant to what our students will need to do with the language in the future?
- Is it realistic to envisage change in practice without systemic change in our working/learning/living environments?

Validation and validity

- **Messick, 1986, p. 13 (also republished in Wainer & Braun (Eds), 2015)**
  - *One recommendation is to contrast the potential social consequences of the proposed testing with those of alternative procedures and even of procedures antagonistic to testing, such as not testing at all*
  - *(Ebel, 1964)* .

Validation and validity

- **Messick, 1986, p. 13 (also republished in Wainer & Braun (Eds), 2015)**
  - *the construct meaning of measures plays a central role. Just as the construct meaning of the test provided a rational basis for hypothesizing predictive relationships to criteria, construct meaning also provides a rational basis for hypothesizing potential outcomes and for anticipating possible side effects.*

Validation and validity

- First explicit categorization of validity evidence to include construct validity was presented by the American Psychological Association in 1954
- The taxonomy was presented as a four-way distinction: ***predictive validity, concurrent validity, content validity*** and ***construct validity.***
- Cronbach and Meehl (1955, pp. 281-282) suggested that predictive and concurrent approaches could be subsumed under the umbrella of ***criterion validity*** evidence, and this **tripartite distinction** became the defacto standard for validity for then next 30 years

Validation and validity

- The importance of defining the construct of interest for a test has become a well-established part of the general tenets of the unified approach to validity.
- The understanding in the field of what that means in practice, however, has changed considerably from the early presentations of the concept of construct validity.

Validation and validity

Cronbach and Meehl (1955) recognized that the state of knowledge regarding the constructs underlying most psychological tests was far from the ideal , noting that rather than empirically supported, well defined theories, “psychology works with crude, half-explicit formulations” (p. 294).

Validation and validity

- Messick's definition of construct validation:
  - an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment (1989, p. 13).

Validation and validity

- The field of language testing and assessment has been faced with the same issues regarding construct definition.
- While a number of models of second language proficiency have been proposed, there remains no consensus model with universal support
- Language testers have accepted a looser interpretation of construct which encompasses both ***descriptions of the underlying abilities relevant to language use*** for particular purposes but also ***clear descriptions of the contextual features of tasks relevant to the target language use domain which is the target of testing***.

Validation and validity

- Messick listed six aspects of a validity which must all be considered. He called this "touching all the bases"
- If time or resources aren't available to investigate all, the test developer must still explain why, and "touch all the bases"
- Messick included the importance of **consequences and values** in his six categories

A model of validity

Validation and validity

- Messick (1989) remains the “touchstone” for discussions of validity in educational measurement
- But the 1990s and 2000s saw growing criticism of the difficulty of operationalizing the model
- Kane (1992, 2001, 2013) promoted the argument-based approach. Applied in language testing by Chapelle et al (2008)
- Bachman (2005) and Bachman and Palmer (2010) promoted the assessment use argument
- Mislevy et al (2003) proposed the evidence-centred design approach

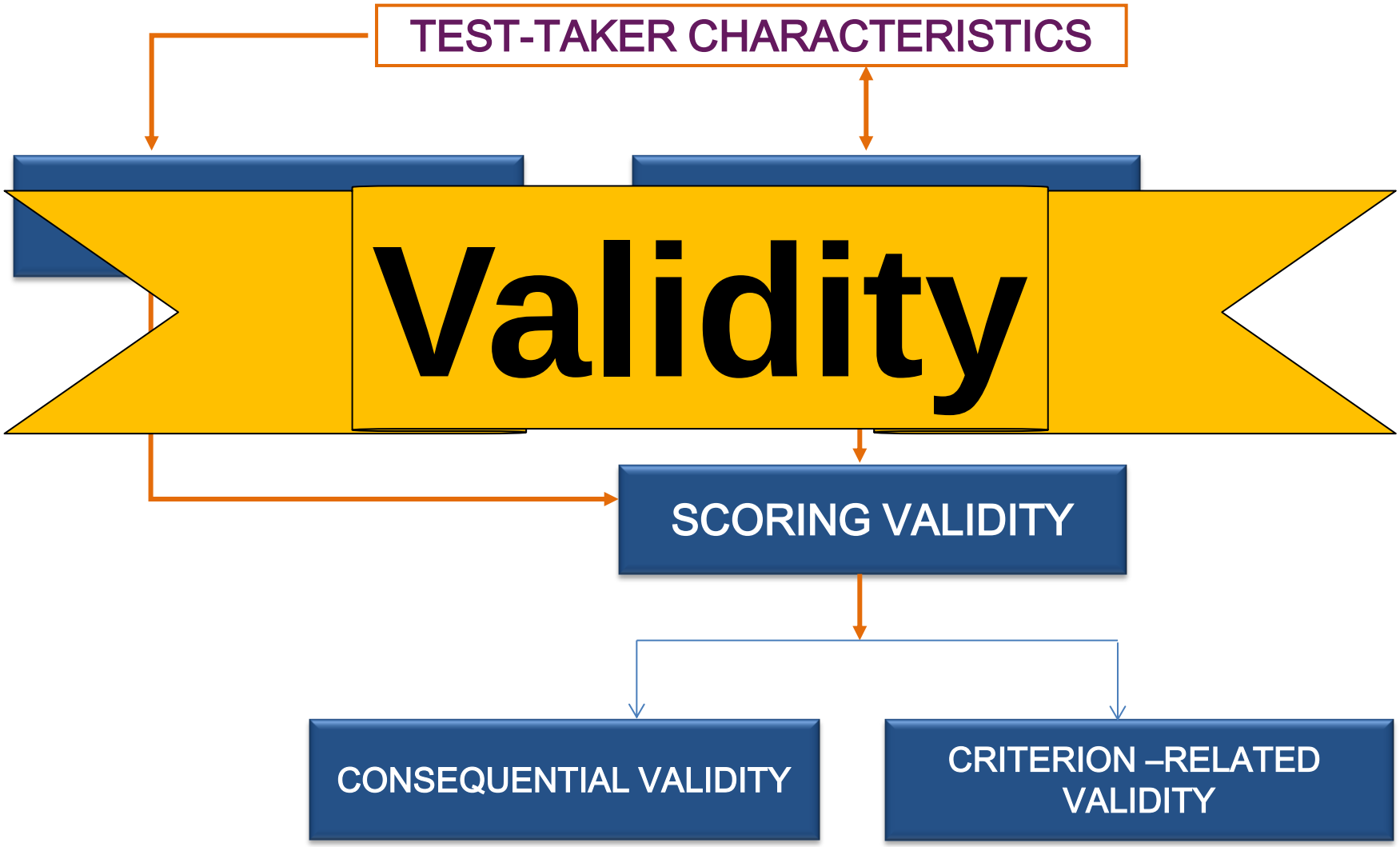
Validation and validity

- Chalhoub-Deville (2003) suggests models still fail to capture the dynamic relationship between context and underlying ability, with neither being fixed but impacting on and influencing the other.
- Chalhoub-Deville (2003, p. 380) calls on language testing researchers to “develop local theories that detail the L2 ‘ability – in language user – in context’ interactions.”
- Weir et al (2013, pp. 99-100) suggest that “testing researchers in the future will need to explore these interrelationships further and determine more closely if and how individual ability and contextual factors interact, and whether and how the ability changes as a result of that interaction.”

Validation and validity

- These models are by design general and do not try to contain taxonomies of evidence relevant to justifying the uses and interpretations of langauge tests, or to help us define the construct underlying our language tests.
- They do not help us find answers to the question ***"how much of what kind of evidence to we need to be confident that our tests are useful and work in the way intended?"***

Socio-cognitive model of language test development and validation



Socio-cognitive model of language test development and validation

What is validity?

Does the test measure what we want it to measure?

- CONTEXT VALIDITY
- COGNITIVE VALIDITY

Are the scores from the test accurate, reliable, meaningful?

- SCORING VALIDITY

Are the scores useful for test users to make decisions?

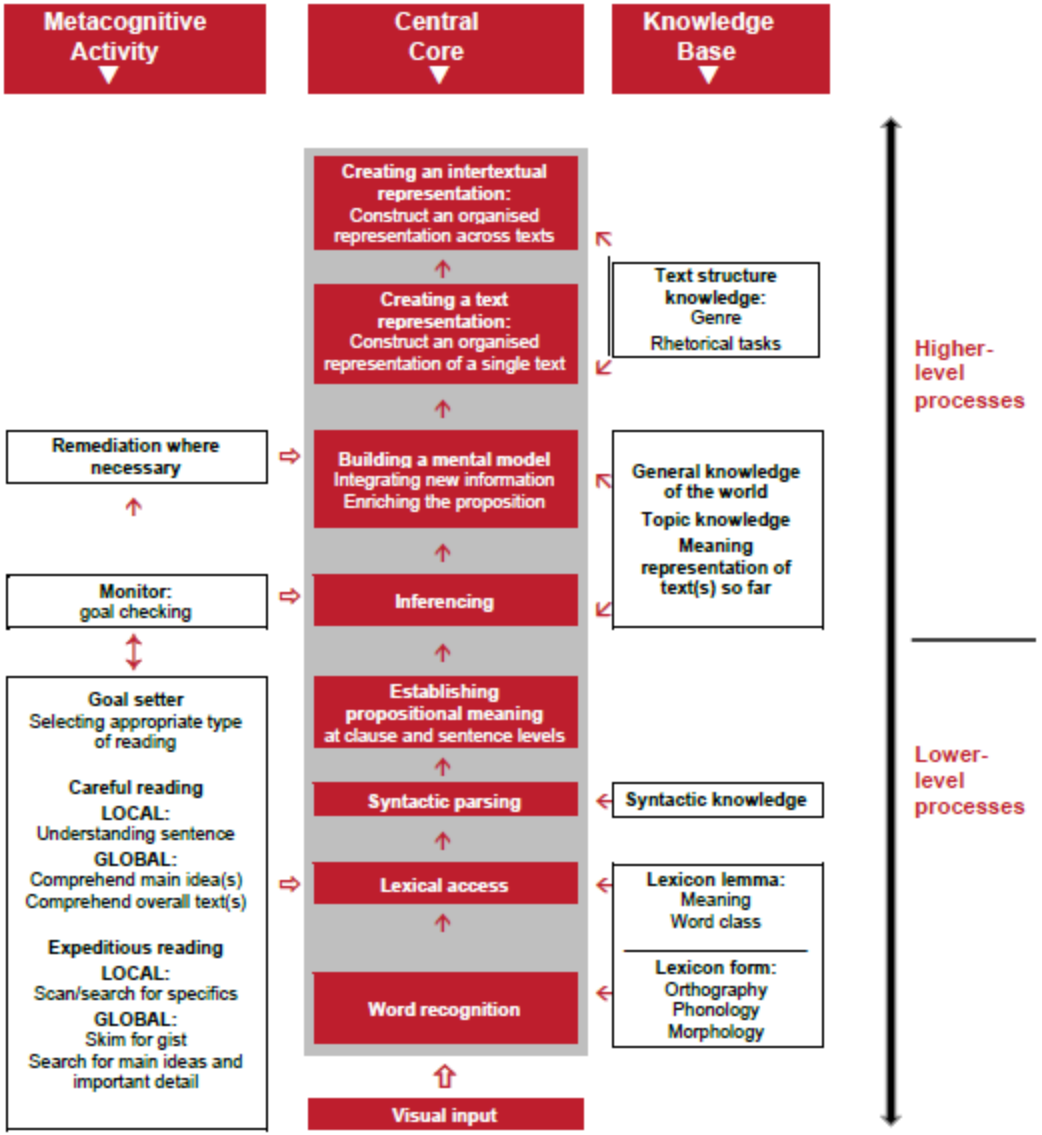
- CONSEQUENTIAL VALIDITY
- CRITERION –RELATED VALIDITY

Contextual and Cognitive parameters: Reading

Context validity	
Task Setting • Response method • Weighting • Knowledge of criteria • Order of items • Channel of presentation • Text length • Time constraints **Setting: administration** • Physical conditions • Uniformity of administration • Security	**Linguistic Demands: Task Input & Output** • Overall Text purpose • Writer reader relationship • Discourse mode • Functional resources • Grammatical resources • Lexical resources • Nature of information • Content knowledge

Cognitive validity
Cognitive Processes • Goal setting • Word recognition • Lexical access • Syntactic parsing • Establish propositional meaning • Inferencing • Building a mental model • Creating a text level representation • Creating an inter-textual representation • Monitoring comprehension

Cognitive processing model: reading



A cognitive processing model of reading based on Khalifa & Weir (2009)

Figure taken from Brunfaut & McCray, 2015

Cognitive processing model: Reading

Types of reading (goal setting)	Expeditious reading: local	Careful reading: local
	Expeditious reading: global	Careful reading: global

Levels of reading	

Cognitive processing model: Reading

Types of reading (goal setting)	Expeditious reading: local	Careful reading: local
	Expeditious reading: global	Careful reading: global

Levels of reading	Word recognition
	Lexical access
	Syntactic parsing

Cognitive processing model: Reading

Types of reading (goal setting)	Expeditious reading: local	Careful reading: local
	Expeditious reading: global	Careful reading: global

Levels of reading	Word recognition
	Lexical access
	Syntactic parsing
	Establishing propositional meaning
	Inferencing

Cognitive processing model: Reading

Types of reading (goal setting)	Expeditious reading: local	Careful reading: local
	Expeditious reading: global	Careful reading: global

Levels of reading	Word recognition
	Lexical access
	Syntactic parsing
	Establishing propositional meaning
	Inferencing
	Building a mental model
	Creating a text level representation

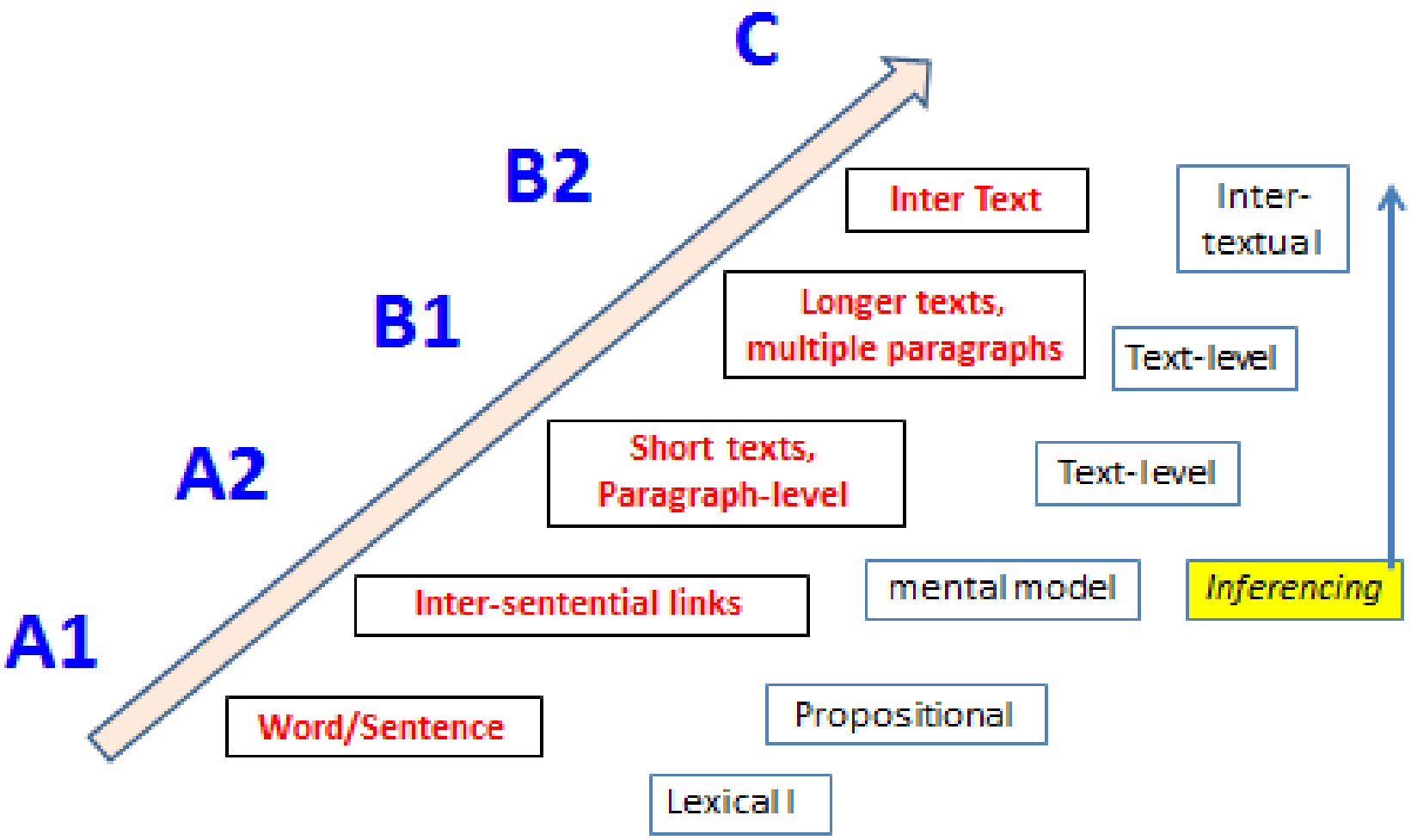
Cognitive processing model: Reading

Types of reading	Expeditious reading: local	Careful reading: local
	Expeditious reading: global	Careful reading: global

Levels of reading	Word recognition
	Lexical access
	Syntactic parsing
	Establishing propositional meaning
	Inferencing
	Building a mental model
	Creating a text level representation
	Creating an intertextual representation

From Khalifa & Weir (2009)

Operationalizing the Model



Task specs: an example

Test	Aptis General	Component	Reading	Task	Multiple Choice Gap-Fill

Features of the Task						
Skill focus	Reading comprehension up to the sentence level					
Task Level	A1	A2	B1	B2	C1	C2
task description	Multiple-choice gap fill. A short text of 6 sentences is presented. Each sentence contains one gap. Test takers choose the best option from a pull-down menu for each gap to complete the sentence. The first sentence is an example with the gap completed. Each gap can be filled by reading within the sentence.					
Cognitive processing Goal setting	**Expeditious reading: local** (scan/search for specifics)			**Careful reading: local (understanding sentence)**		
	Expeditious reading: global (skim for gist/search for key ideas/detail)			**Careful reading: global** (comprehend main idea(s)/overall text(s))		
Cognitive processing Levels of reading	**Word recognition**					
	Lexical access					
	Syntactic parsing					
	Establishing propositional meaning (cl./sent. level)					
	Inferencing					
	Building a mental model					
	Creating a text level representation (disc. structure)					
	Creating an intertextual representation (multi-text)					

Task specs: an example

Features of the Input Text											
Words	40-50 words (including target words for gaps)										
Domain	Public		Occupational		Educational		Personal				
Discourse mode	Descriptive		Narrative		Expository		Argumentative		Instructive		
Content knowledge	General						Specific				
Cultural specificity	Neutral						Specific				
Nature of information	Only concrete		Mostly concrete		Fairly abstract		Mainly abstract				
Lexical Level	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	
Text genre	E-mails, letters, notes, postcards										
Features of the Response											
Target	**Length**	1 word	**Lexical**	K1	**Part of Speech**	Noun, verb, adjective					
Distractors	**Length**	1 word	**Lexical**	K1	**Part of Speech**	Noun, verb, adjective					
Key	Within sentence		Across sentences		Across paragraphs						

Task specs: an example

Choose one word from the list for each gap. The first one is done for you.

Dear Morgan,

Thank you for a wonderful weekend. I had a really great time with you and Becky. Your wife is a good cook and she [ ] a very nice dinner. I am writing this note in my hotel room and I can [ ] the park from my window. My plane leaves tomorrow and I will take a taxi to the airport [ ] breakfast. I hope you and Becky will come and [ ] with me in Rome next summer. I am feeling a little tired now and I [ ] to have a sleep.

Thanks again and see you soon,

James

Task specs: an example

Test	Aptis General	Component	Reading	Task	Matching headings to text

Features of the Task						
Skill focus	*Expeditious global reading of longer text, integrating propositions across a longer text into a discourse-level representation.*					
Task Level	**A1**	**A2**	**B1**	**B2**	**C1**	**C2**
task description	*Matching headings to paragraphs within a longer text. Candidates read through a longer text consisting of 7 paragraphs, identifying the best heading for each paragraph from a bank of 8 options.*					
Cognitive processing Goal setting	**Expeditious reading: local** (scan/search for specifics)			**Careful reading: local (understanding sentence)**		
	Expeditious reading: global (skim for gist/search for key ideas/detail)			**Careful reading: global** (comprehend main idea(s)/overall text(s))		
Cognitive processing Levels of reading	**Word recognition**					
	Lexical access					
	Syntactic parsing					
	Establishing propositional meaning (cl./sent. level)					
	Inferencing					
	Building a mental model					
	Creating a text level representation (disc. structure)					
	Creating an intertextual representation (multi-text)					

Task specs: an example

Features of the Input Text										
Words	*700-750 words*									
Domain	**Public**	Occupational	**Educational**	Personal						
Discourse mode	Descriptive	Narrative	**Expository**	**Argumentative**	Instructive					
Content knowledge	**General**				Specific					
Cultural specificity	**Neutral**				Specific					
Nature information	Only concrete	Mostly concrete	**Fairly abstract**	Mainly abstract						
Lexical Level	**K1**	**K2**	**K3**	**K4**	**K5**	K6	K7	K8	K9	K10
Readability	*Flesch-Kincaid Grade Level 9-12*									
Grammar	*A1-B2 Exponents*	Average sentence length	*18-20 words*							
Text genre	*Magazines, newspapers, instructional materials (such as extracts from undergraduate textbooks describing important events and ideas, etc).*									

Task specs: an example

Features of the Response						
Target	**Length**	Up to 10 words	**Lexical**	K1-K5	**Grammar**	A1 – B2
Distractors	**Length**	Up to 10 words	**Lexical**	K1-K5	**Grammar**	
Key	Within sentence		Across sentences	Across paragraphs		

Task specs: an example

READING

Read the passage quickly. Choose a heading for each numbered paragraph (1-7) from the drop-down box. There is one more heading than you need.

1
2
3
4
5
6
7

Bone Wars

In the summer of 1868 a train carrying a group of American scientists made its way through the western frontier state of Wyoming. On board was O.C. Marsh, an expert in geology and the first person in the country to hold the position of professor of palaeontology at the University of Yale. Like his fellow passengers, Marsh was impressed by the enormous landscapes of dry rock, and he knew that the ancient stones must hold evidence of prehistoric life. It was during this journey that he made a decision that was to have a lasting impact not only on his own professional career but on the American scientific community.

1. In 1800 the French naturalist Georges Cuvier identified a fossil [old bone] as the remains of a small flying reptile. This was the first recorded example of a species that later became known as the dinosaur. Although these creatures no longer existed, Cuvier showed that they could be studied through an examination of fossil records, buried and preserved in rock. So the science of palaeontology – the study of prehistoric life – began.

2. Over the next two decades some spectacular finds were made by English

Aptis Reading test spec

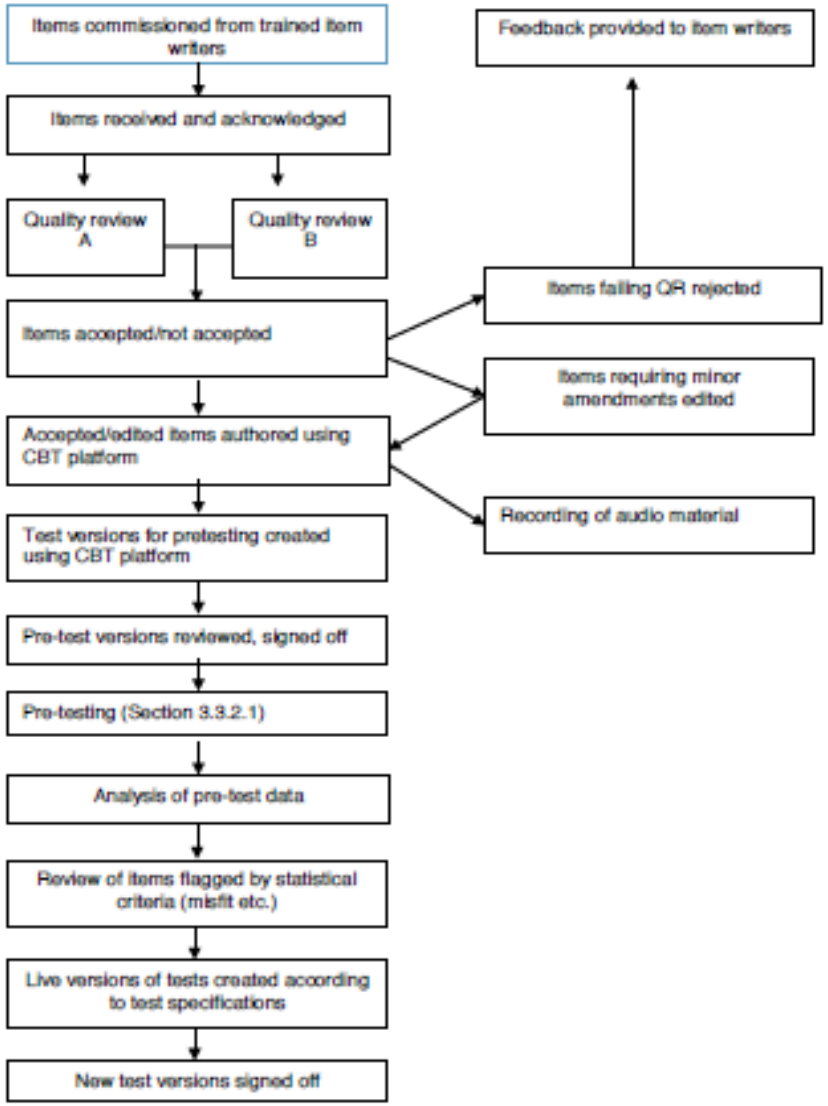
Skill focus	Lvl	Task description	Cognitive processes
Sentence level meaning	**A1**	A short text with 5 gaps. Filling each gap only requires comprehension of the sentence containing the gap. Text-level comprehension is not required.	• Careful local reading • Syntactic parsing • Understanding propositional meaning
Inter-sentence cohesion	**A2**	Reorder jumbled sentences to form a cohesive text	• Careful global reading • Inferencing • Building a mental model
Text-level comprehension of short texts	**B1**	A short text with 7 gaps. Requires comprehension of text across sentences.	• Careful global reading • Building a mental model
Integrating macro-propositions and understanding important ideas in longer texts	**B2**	Matching the most appropriate heading to paragraphs. Requires integration of micro- and macro-propositions within and across paragraphs, and comprehension of discourse structure of more complex and abstract texts.	• Expeditious global reading • Creating a text level representation

Putting it all together

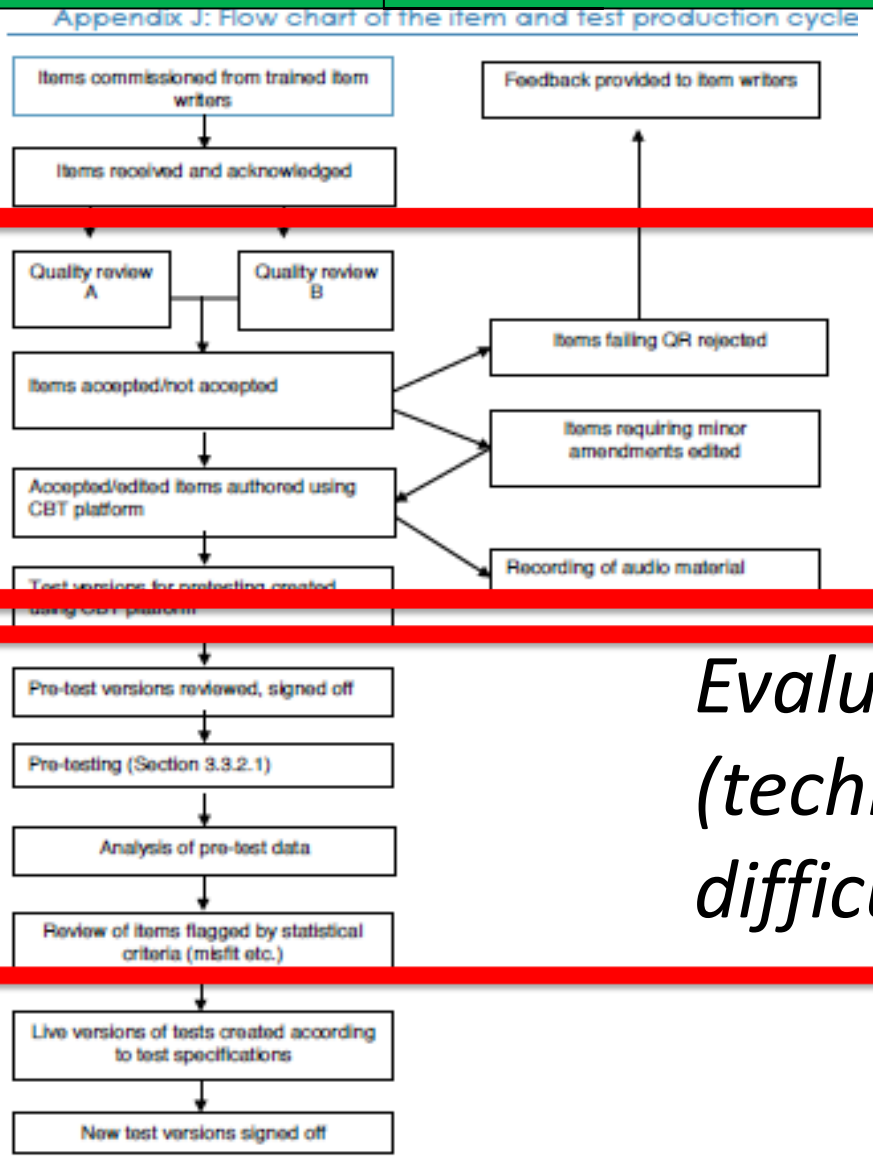
- Synergy between contextual, cognitive and scoring aspects of validity
- Model underpinning specs allows for a cycle ***of test design, development, validation, evaluation and revision.***
- ***Illustrate with an example of ongoing evaluation of the Aptis Reading test***

The life of an item (Aptis)

Appendix J: Flow chart of the item and test production cycle



Synergy



Evaluate cognitive and contextual features

Evaluate scoring validity (technical properties and difficulty)

Synergy: Cognitive, contextual, scoring validity

Looking into test-takers cognitive processes

Looking into test-takers cognitive processes while completing reading tasks - Brunfaut and McCray

- Empirical difficulty estimated through the Rasch model confirmed impressions that B1 reading tasks needed revision
- The studies carried out to investigate cognitive processing also confirmed that the B1 reading task was not eliciting the processes the model calls for

Resolving tension

- Separate empirical validation of the cognitive processing model identified that the B1 task, while working as a measurement instrument was not eliciting the “across sentences” reading intended
- Other tasks conformed to the model
- So there was a synergy between the construct representation and cognitive processing and the Rasch model empirical difficulty

Happy ending?

- On-going adjustment is necessary and to be expected
- Adjustments will be necessary to the measurement instrument but also to our understanding of the construct
- We can't expect to be perfect, but there is a tension between how confident we can be that our constructed measures are plausible and useful, and the caveat that we know we will learn more as we go and need to change
- Communicating the need to expect change to test users, while still meeting the needs for meaningful, reliable measurement outcomes, and comparable interpretable measures is a challenge.

Some final thoughts…

- The socio-cognitive model provides a coherent methodology for collating, organizing and evaluating the evidence gathered through a validation research agenda,
- It allows us to "touch all the bases" in Messick's terms.
- The model nonetheless clearly identifies a road map for designing and carrying out such a research agenda to help design an agenda to answer the question of **how much of what is needed to justify the uses and interpretations of a language test?**

Some final thoughts…

- *To summarize there is no gold standard, there is no true cut-off score, there is no best standard setting method, there is no perfect training, there is no flawless implementation of any standard setting method on any occasion and there is never sufficiently strong validity evidence. In three words, nothing is perfect. (Kaftandjieva, 2004)*