NCME, New Orleans
April 11, 2011

Symposium:
Innovations in the automated scoring of spoken responses

Evaluating the constructs and automated scoring performance for speaking tasks in the Versant Tests and PTE Academic

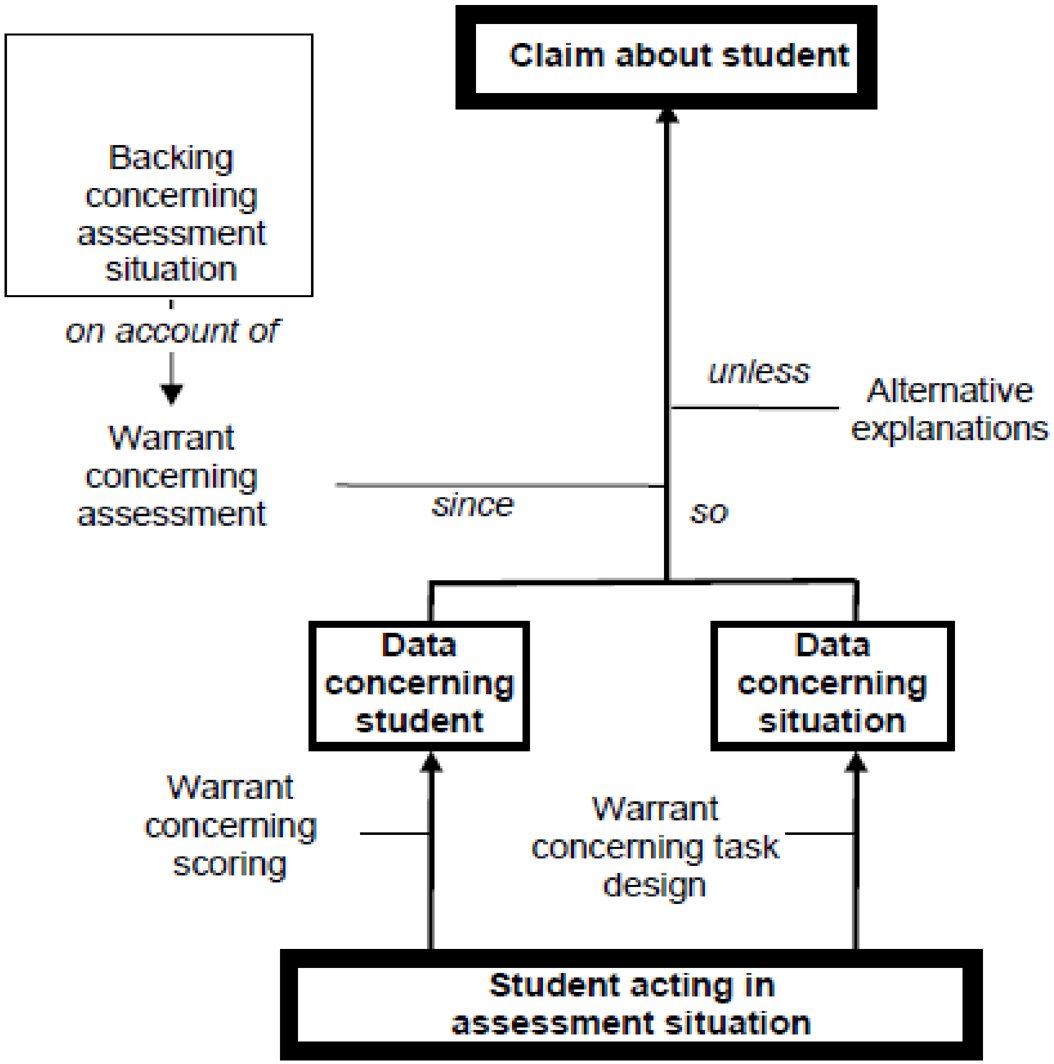
Alistair Van Moere
Knowledge Technologies
Pearson

Automated Tests of Spoken Proficiency

- **Versant Test**
  - Listening-Speaking test
  - Uses: Job recruitment, placement, progress monitoring
  - Available in English, Spanish, Arabic, Dutch, (French, Chinese)

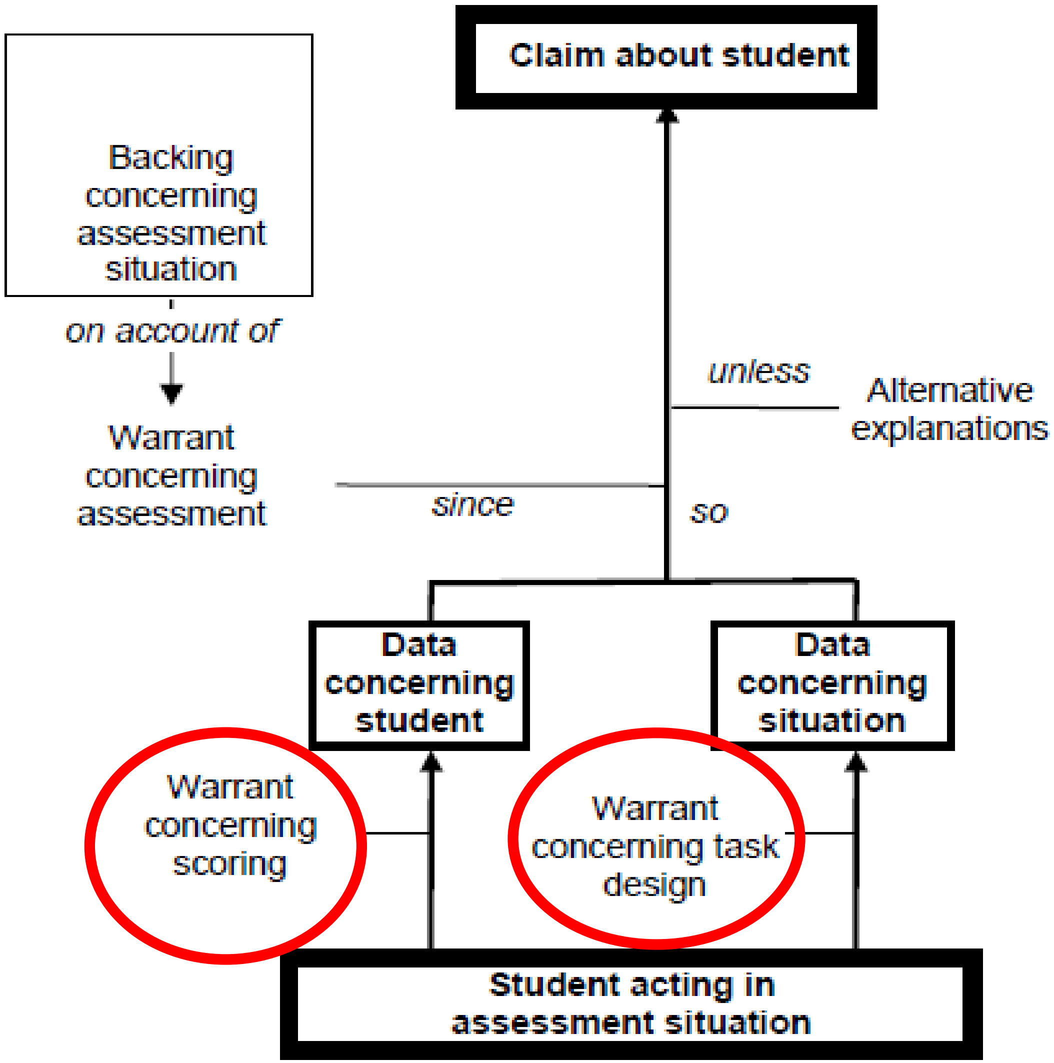
- **PTE Academic**
  - 4-skills language proficiency test
  - Uses: Entrance into English-speaking universities

Assessment argument



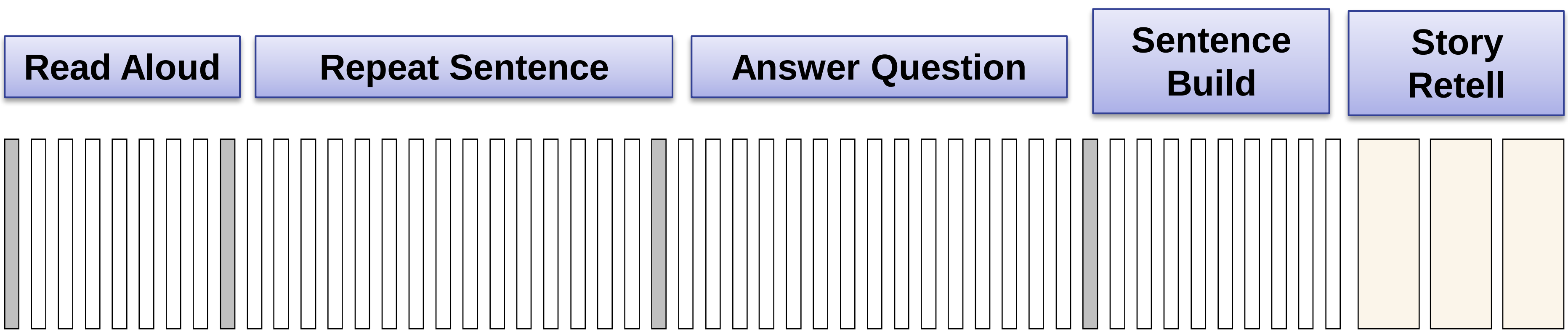
(Mislevy 2005)

Assessment argument

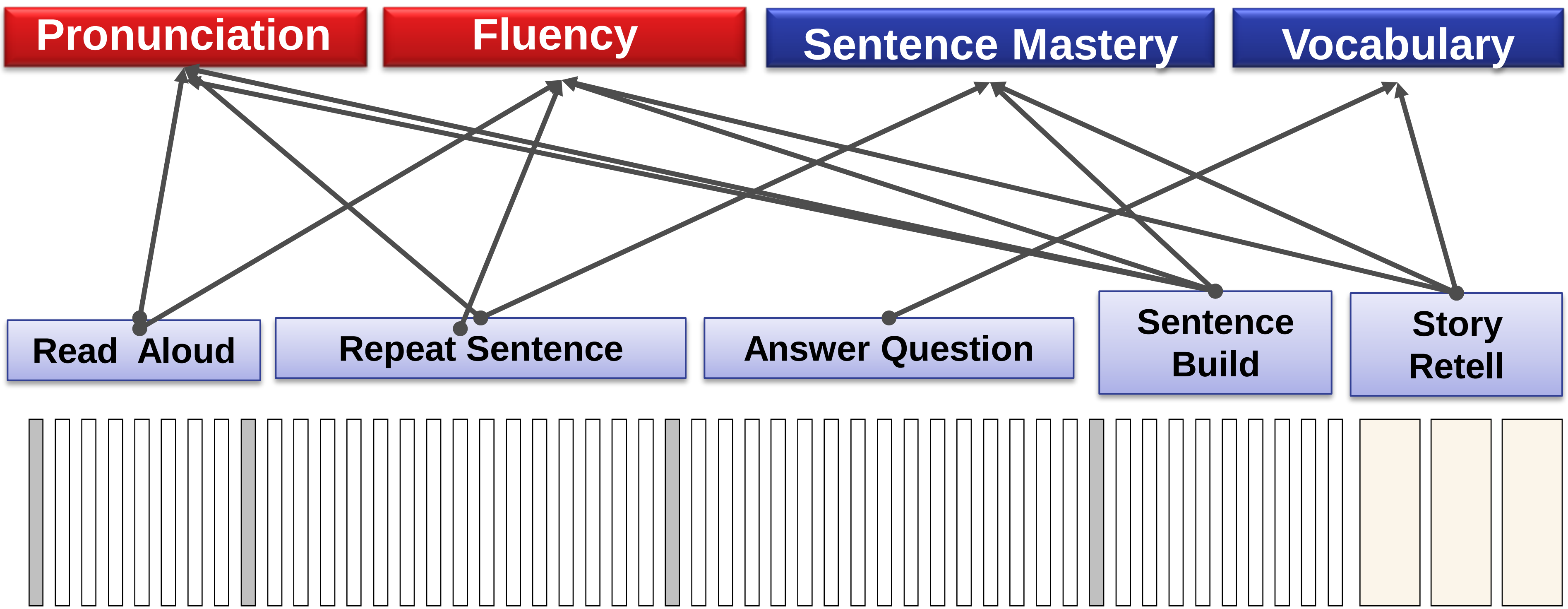


(Mislevy 2005)

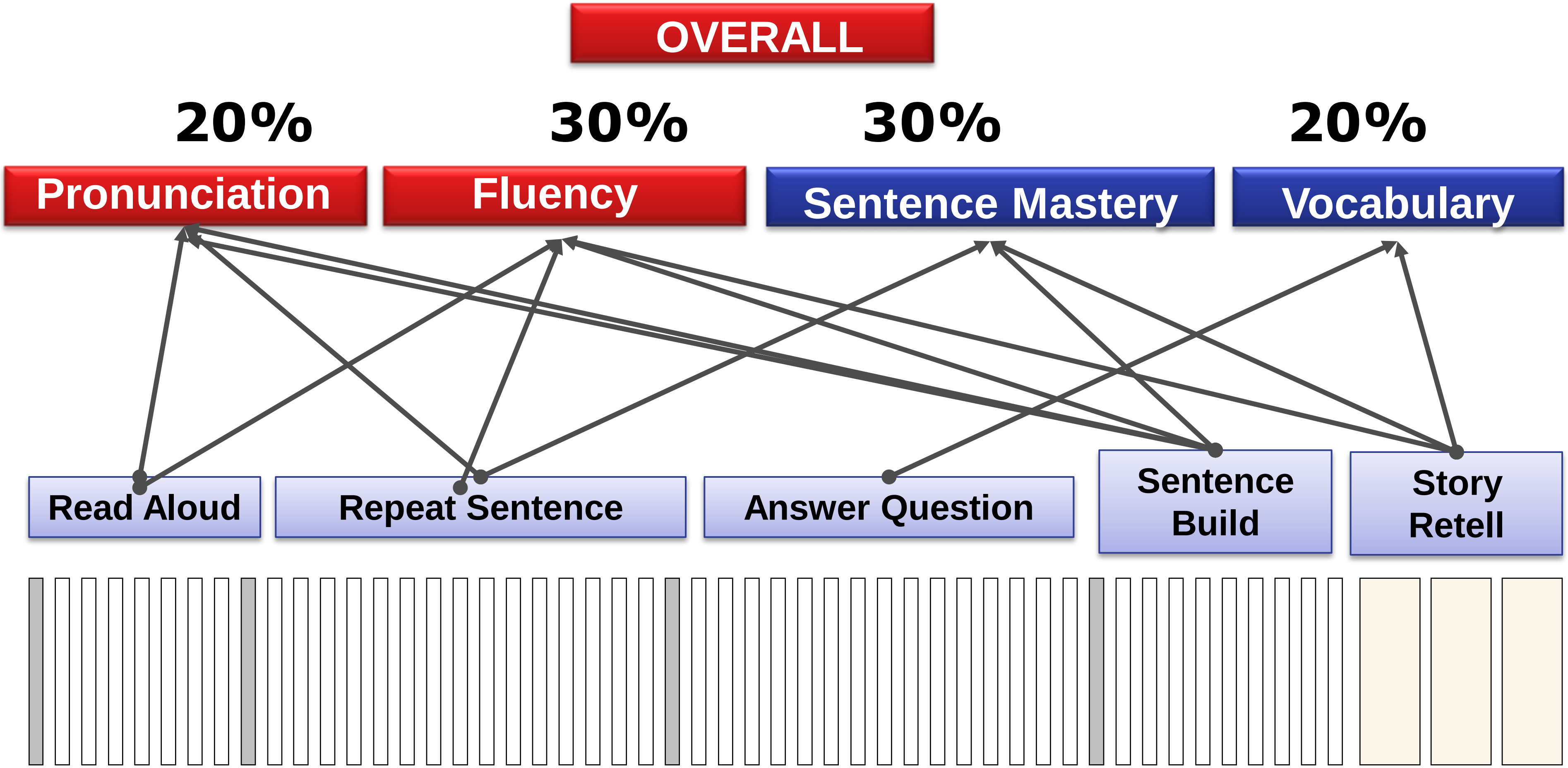
Versant Tasks and Scoring



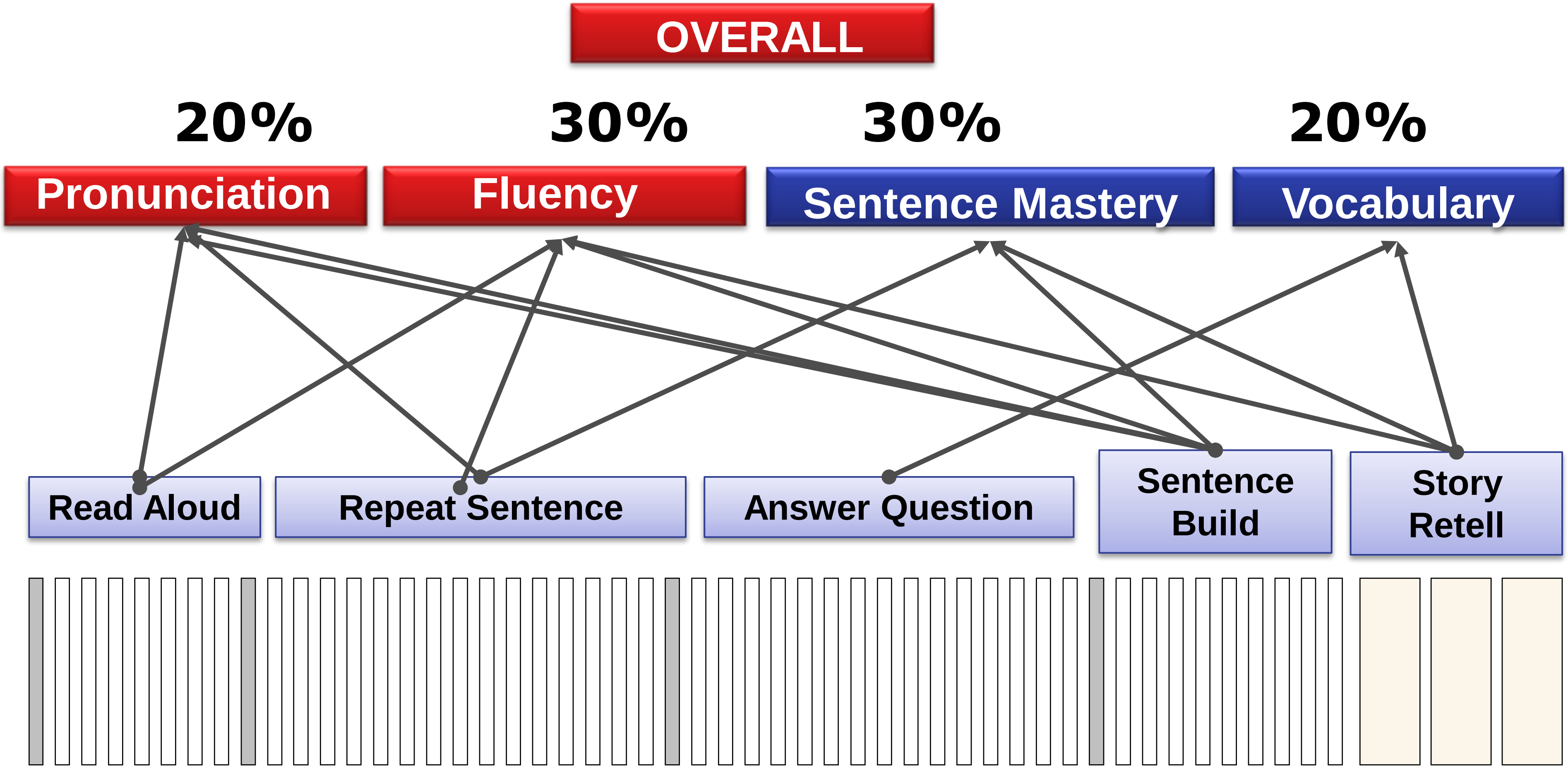
Versant Tasks and Scoring



Versant Tasks and Scoring



Versant Tasks and Scoring



63 responses , 3’30 mins speech

Versant Test Scoring

Trait	Scoring
Fluency	Temporal features of speech predict expert human judgments
Pronunciation	Spectral properties and segmental aspects predict human judgments
Vocabulary	i) Rasch-based ability measures from dichotomous-scored vocabulary items; ii) LSA-based measures on constructed responses predict human judgments
Sentence Mastery	Rasch-based ability measures from word errors on increasingly complex sentences

Versant Test Scoring

Trait	Scoring	Machine-Human, r	Human split-half	Machine split-half
Fluency	Temporal features of speech predict expert human judgments	.94	.99	.97
Pronunciation	Spectral properties and segmental aspects predict human judgments	.88	.99	.97
Vocabulary	i) Rasch-based ability measures from dichotomous-scored vocabulary items; ii) LSA-based measures on constructed responses predict human judgments	.96	.93	.92
Sentence Mastery	Rasch-based ability measures from word errors on increasingly complex sentences	.97	.95	.92
	Overall	**.97**	**.99**	**.97**

Validation sample, n=143, flat score distribution

Scoring Model for Sentence Mastery

Repeat Sentence:

Scoring Model for Sentence Mastery

<u>Repeat Sentence:</u>

I'll catch up with you soon.

"uh .. I'll catch up you … I don't know"

Security wouldn't let him in because he didn't have a pass.

"Security wouldn't help him pass"

Scoring Model for Sentence Mastery

<u>Repeat Sentence:</u>

I'll catch up <u>with</u> you <u>soon</u>.

"uh .. I'll catch up you … I don't know" **= 2 word errors**

Security wouldn't <u>let</u> him <u>in because he didn't have a</u> pass.

"Security wouldn't help him pass" = 7 word errors

Scoring Model for Sentence Mastery

Repeat Sentence:

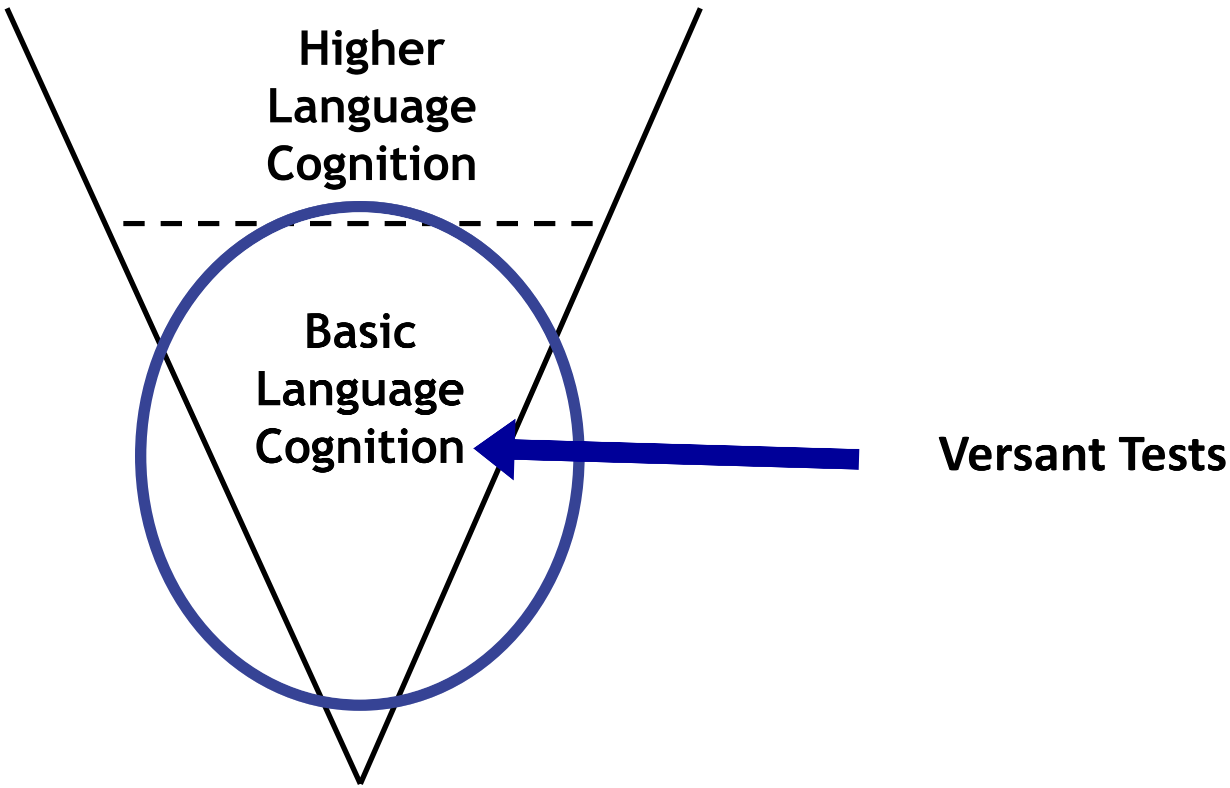
I'll catch up with you soon.

"uh .. I'll catch up you … I don't know" **= 2 word errors**

Security wouldn't let him in because he didn't have a pass.

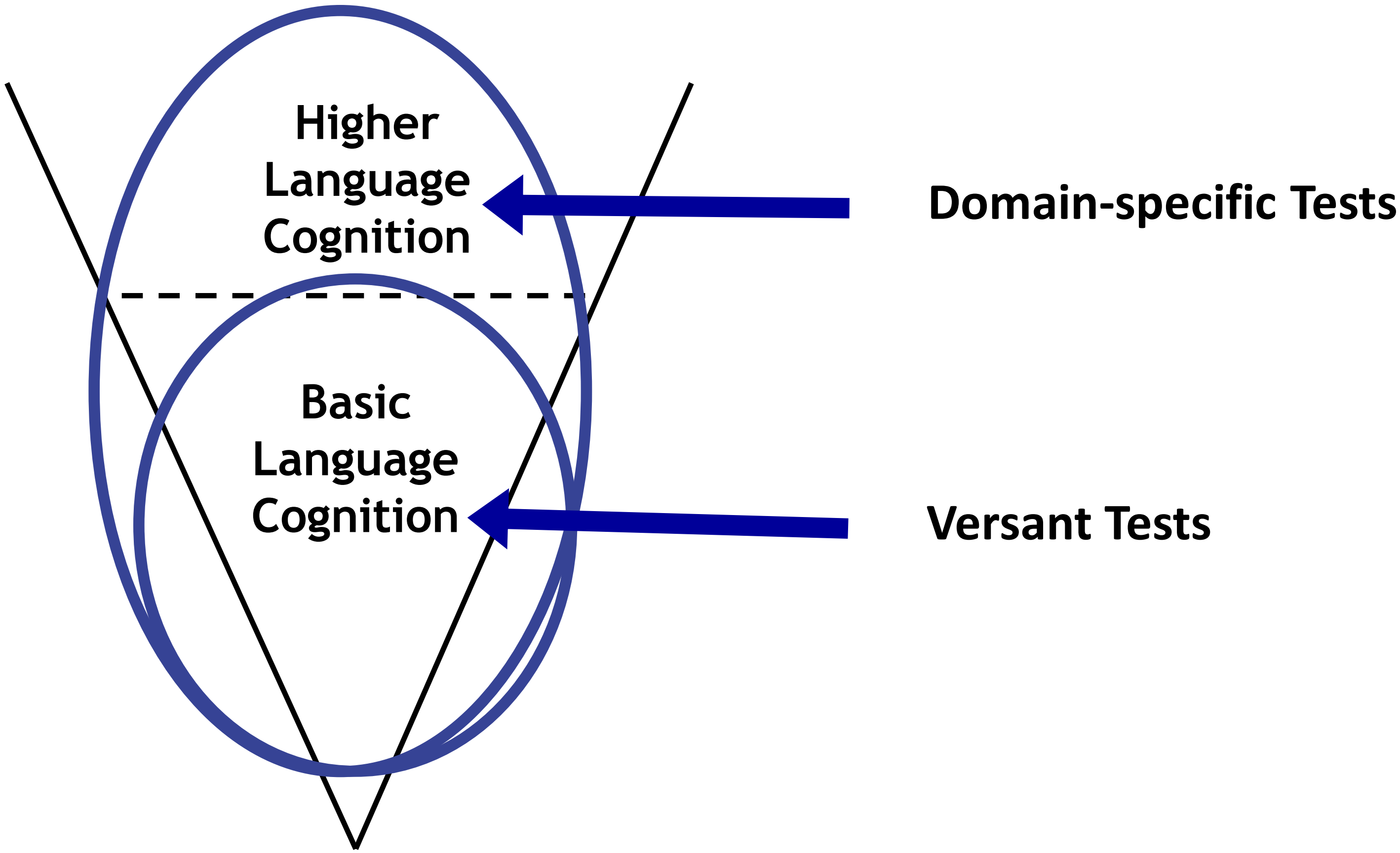
"Security wouldn't help him pass" = 7 word errors

Versant's Domain of Use



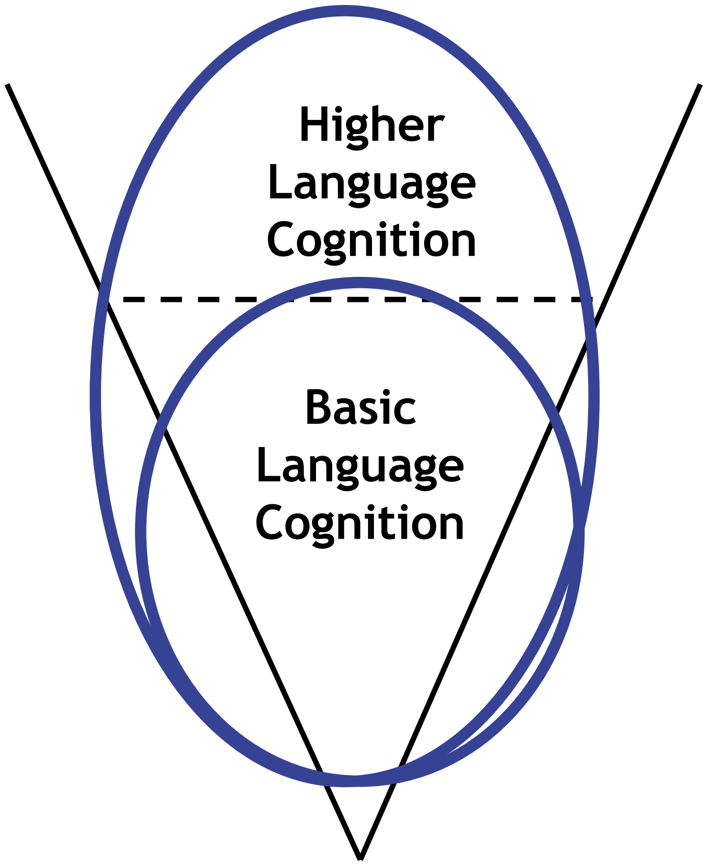
Hulstijn (2010)

Versant's Domain of Use



Hulstijn (2010)

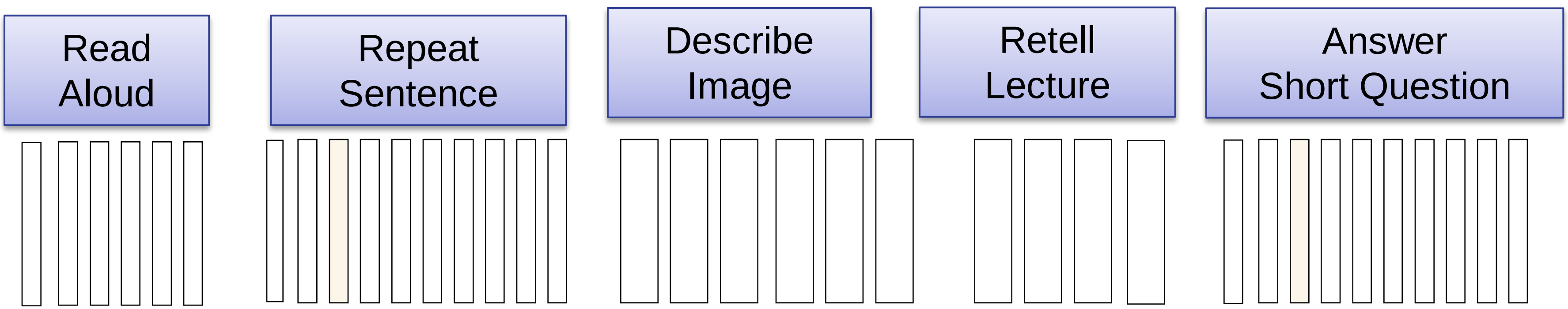
Versant's Domain of Use



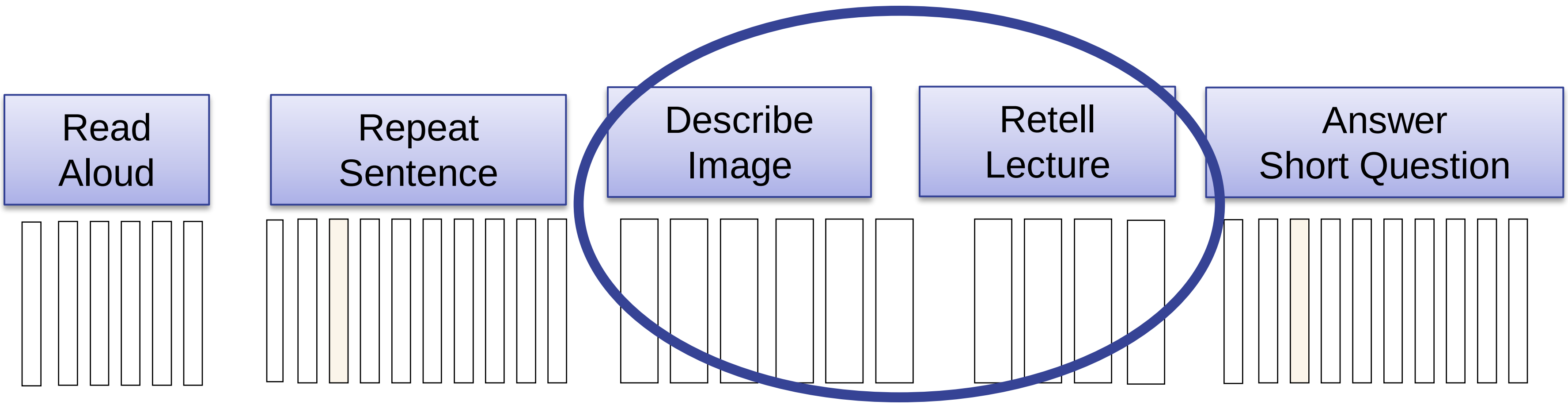
Versant test score correlations with communicative tests

Communicative test	r	n
Test of Spoken English (TSE)	0.88	58
New TOEFL Speaking	0.84	321
BEST Plus interview	0.86	151
IELTS interview test	0.76	130

PTE Academic: Broader construct

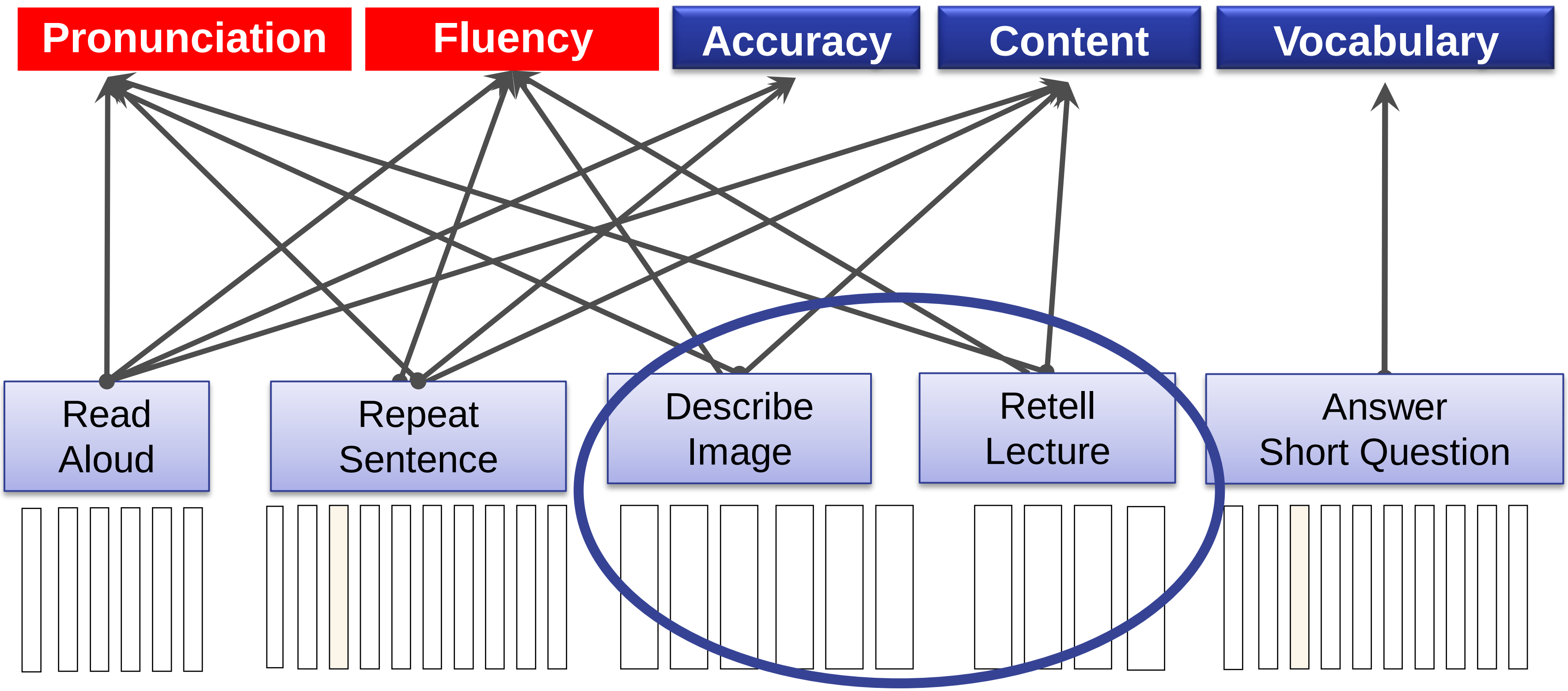


PTE Academic: Broader construct



	Describe Image	Retell Lecture
Preparation time	**25 secs**	**40 secs**
Response time	**40 secs**	**40 secs**

PTE Academic: Broader construct



	Describe Image	Retell Lecture
Preparation time	**25 secs**	**40 secs**
Response time	**40 secs**	**40 secs**

PTE Academic: Sampling Academic Domain

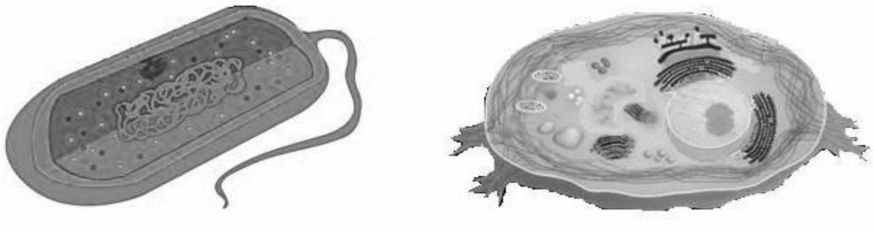
- **5 tasks:**
  - ~ 36 responses
  - ~ 8 minutes of speech
- **Input:**
  - Reading texts
  - Listening texts
  - Visual (non-linguistic)
- **Output:**
  - Prepared monologues
  - Short, real-time responses

Content Scoring of Constructed Responses

- Word choice (Latent Semantic Analysis)
- Content relevance
- Lexical measures
- Words in sequence; collocations

Example item:

Retell Lecture

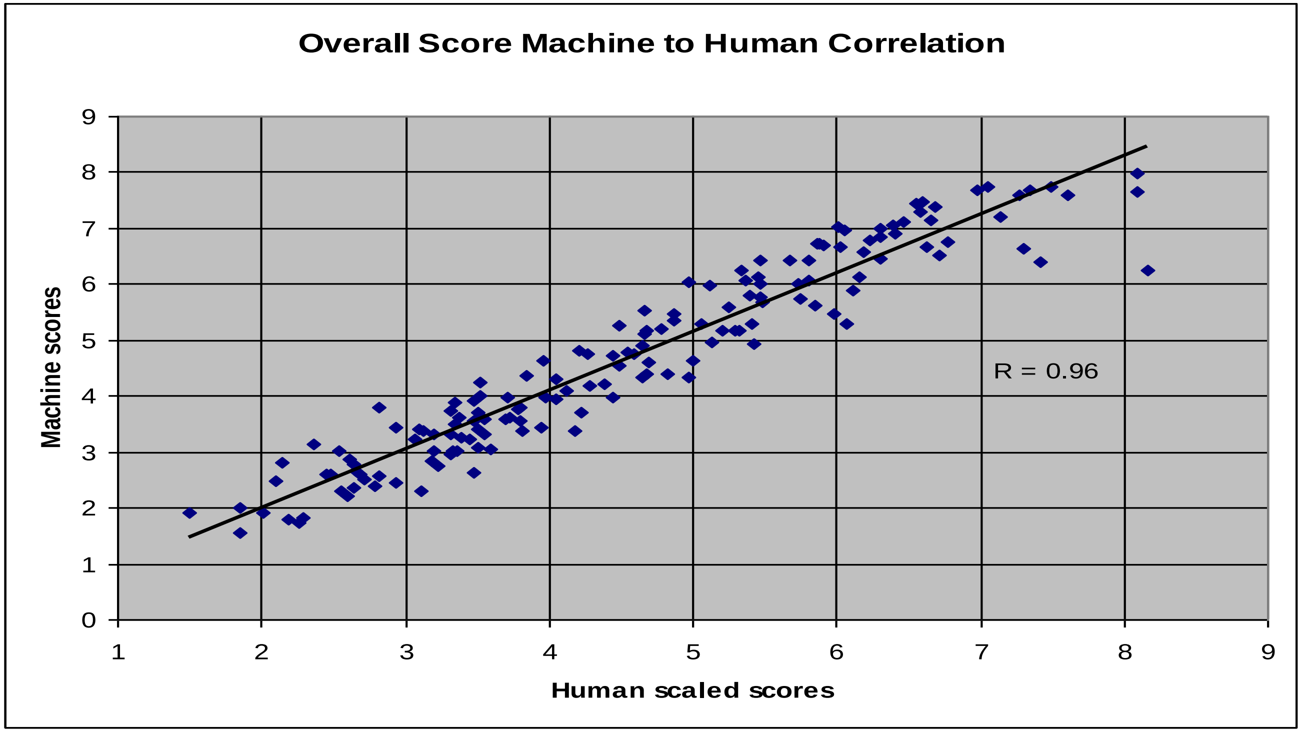


Prokaryotic cell Eukaryotic cell

Sample response

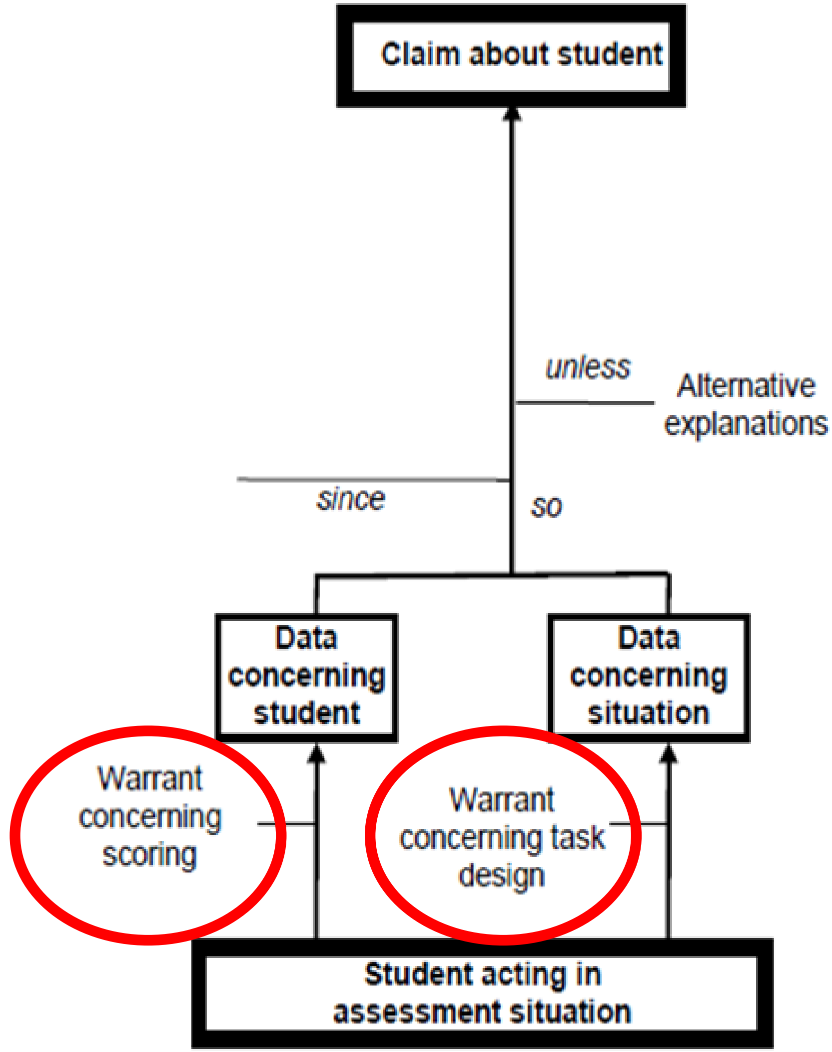
“the lecture was given about biotic cells prokaryotic cell was first described and eukaryotic cell was secondly ref uh described uh it was said eukaryotic cells are more complicated than prokaryotic cell eukaryotic cell is microorganisms where it is it has one single cell and multi cell organisms are also present in eukaryotic cell this more complicated than prokaryotic cell which is placed in right side of the screen”

PTE Academic: Reliability



Validation sample
n=158

Scoring	Machine-Human, r	Human split-half	Machine split-half
Overall	.97	.97	.96


Claim about student
unless
Alternative explanations
since
so
Data concerning student
Data concerning situation
Warrant concerning scoring
Warrant concerning task design
Student acting in assessment situation

TASKS

SCORING

REBUTTAL

↑

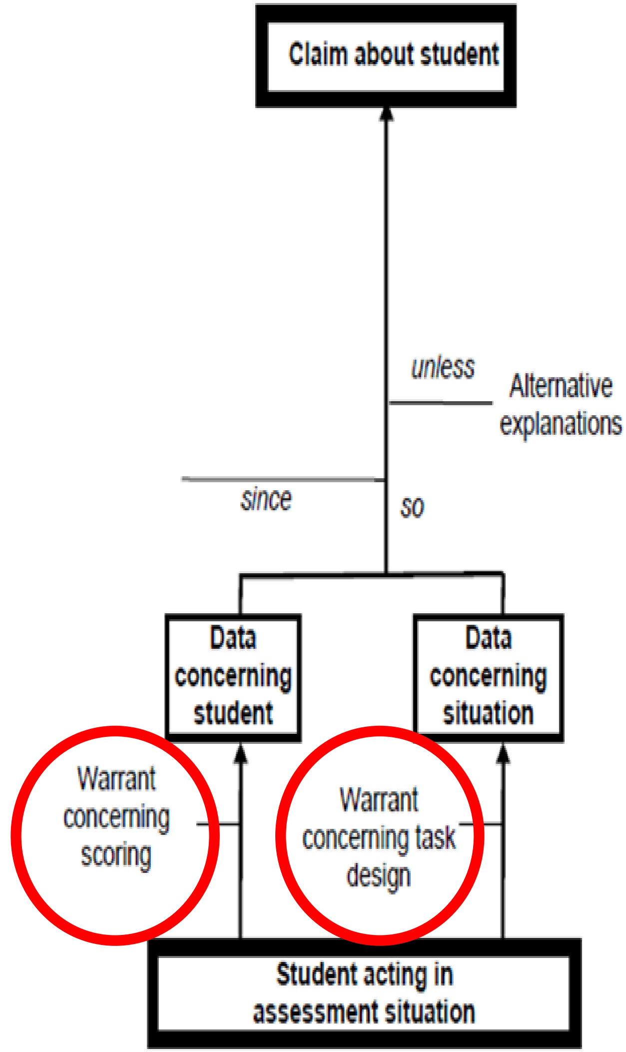
COUNTER

↑

BACKING

↑

CLAIM



TASKS

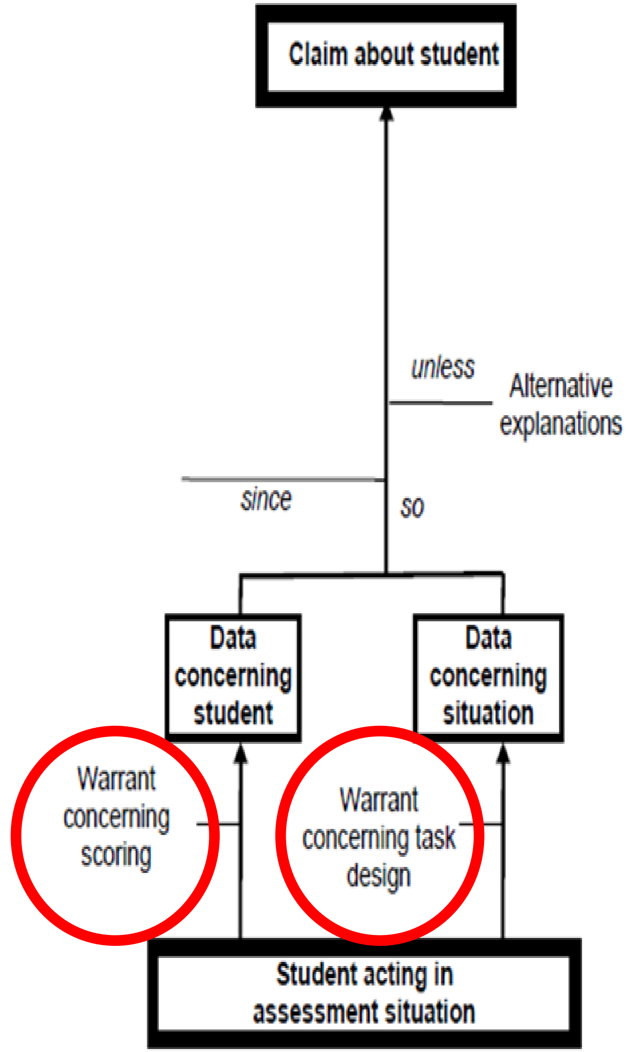
SCORING

REBUTTAL

COUNTER

BACKING

CLAIM



The tasks are valid for assessing spoken language proficiency

TASKS

SCORING

REBUTTAL

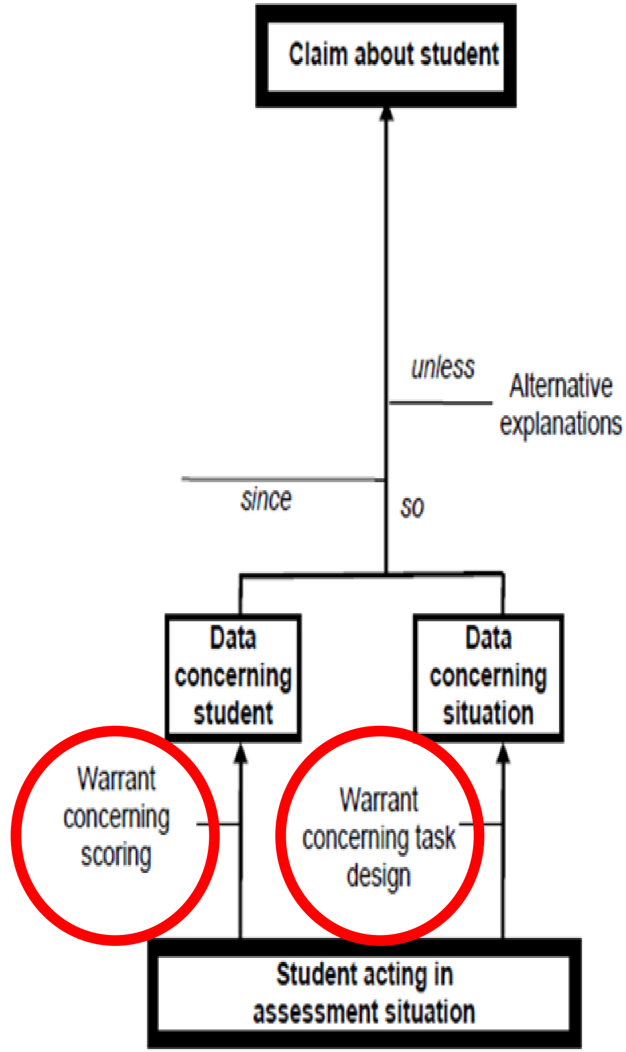
COUNTER

BACKING

The tasks tap real-time automatic processes, and sample academic language & domain interactions

CLAIM

The tasks are valid for assessing spoken language proficiency



TASKS

SCORING

REBUTTAL

COUNTER

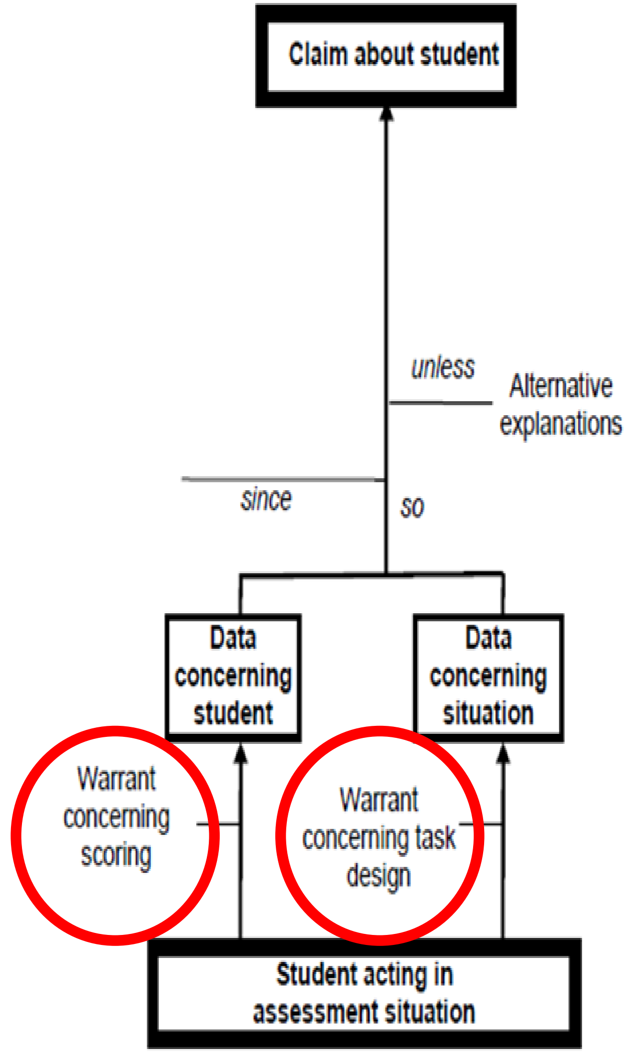
Some tasks are not authentic; the interactions are too constrained

BACKING

The tasks tap real-time automatic processes, and sample academic language & domain interactions

CLAIM

The tasks are valid for assessing spoken language proficiency



TASKS

SCORING

REBUTTAL

Many concurrent validation correlations with interview tests > 0.80 (different tasks and different performances)

COUNTER

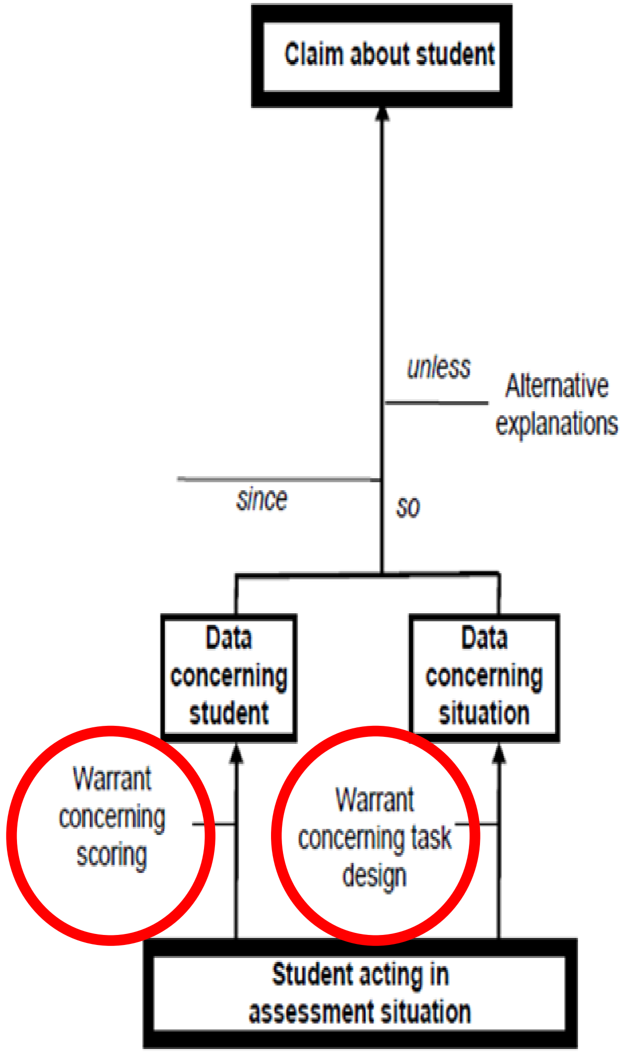
Some tasks are not authentic; the interactions are too constrained

BACKING

The tasks tap real-time automatic processes, and sample academic language & domain interactions

CLAIM

The tasks are valid for assessing spoken language proficiency



TASKS

SCORING

REBUTTAL

Many concurrent validation correlations with interview tests > 0.80 (different tasks and different performances)

COUNTER

Some tasks are not authentic; the interactions are too constrained

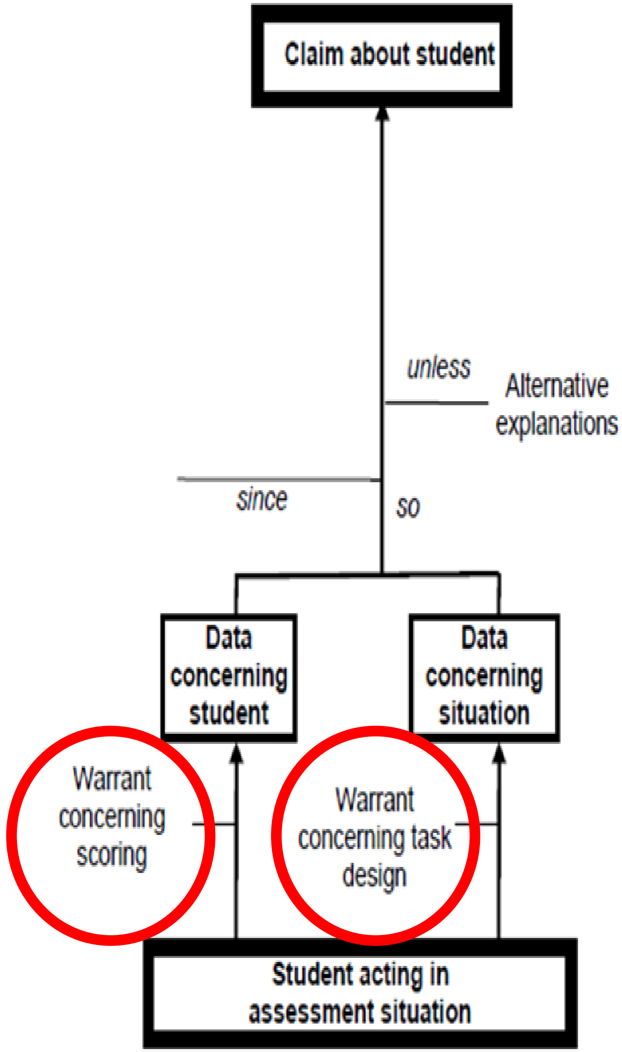
BACKING

The tasks tap real-time automatic processes, and sample academic language & domain interactions

CLAIM

The tasks are valid for assessing spoken language proficiency

The scoring is sufficiently accurate to replace humans



TASKS

SCORING

REBUTTAL

Many concurrent validation correlations with interview tests > 0.80 (different tasks and different performances)

COUNTER

Some tasks are not authentic; the interactions are too constrained

BACKING

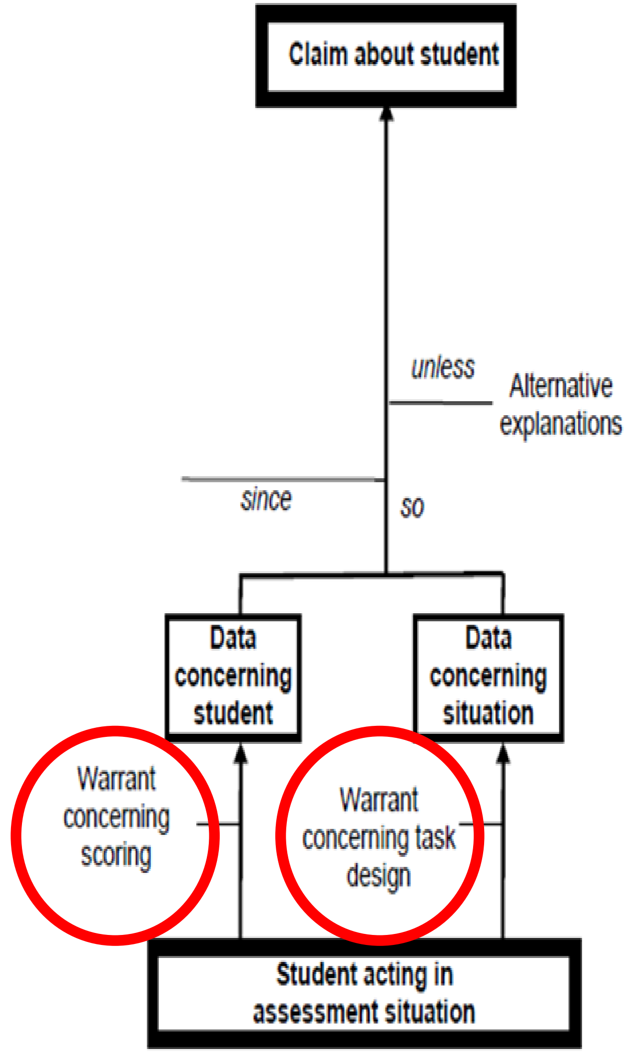
The tasks tap real-time automatic processes, and sample academic language & domain interactions

Machine-to-human score correlations ~0.97 (same tasks, same performance instance)

CLAIM

The tasks are valid for assessing spoken language proficiency

The scoring is sufficiently accurate to replace humans



TASKS

SCORING

REBUTTAL

Many concurrent validation correlations with interview tests > 0.80 (different tasks and different performances)

COUNTER

Some tasks are not authentic; the interactions are too constrained

Machines are notoriously error prone; scores may be triple counting poor pronunciation

BACKING

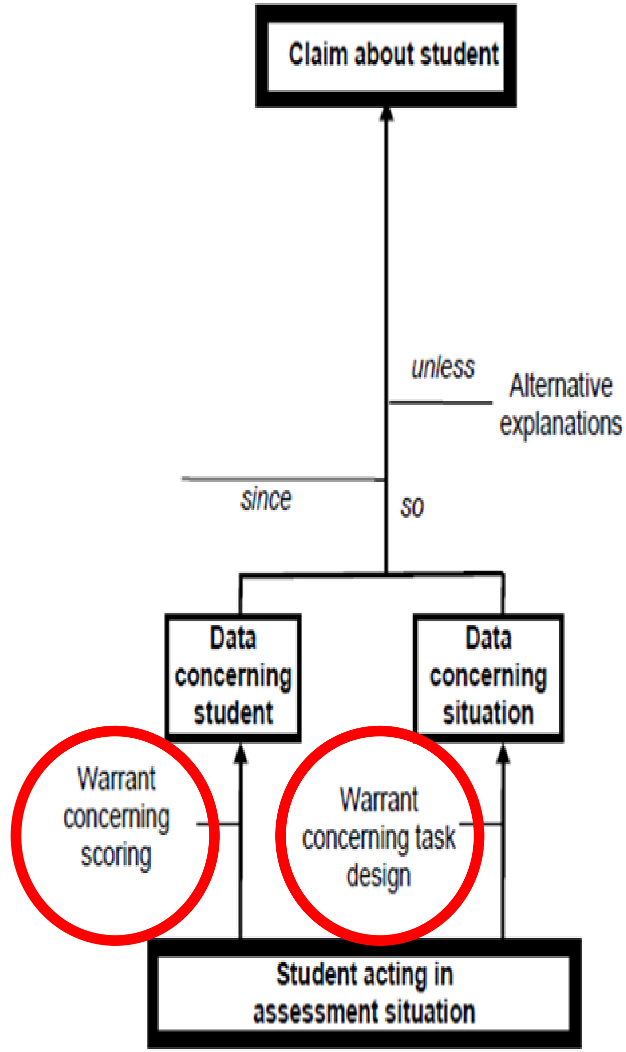
The tasks tap real-time automatic processes, and sample academic language & domain interactions

Machine-to-human score correlations ~0.97 (same tasks, same performance instance)

CLAIM

The tasks are valid for assessing spoken language proficiency

The scoring is sufficiently accurate to replace humans



TASKS

SCORING

REBUTTAL

Many concurrent validation correlations with interview tests > 0.80 (different tasks and different performances)

Scores are relatively insensitive to simulations of worse recognition; systems should be optimized for score accuracy.

COUNTER

Some tasks are not authentic; the interactions are too constrained

Machines are notoriously error prone; scores may be double counting poor pronunciation

BACKING

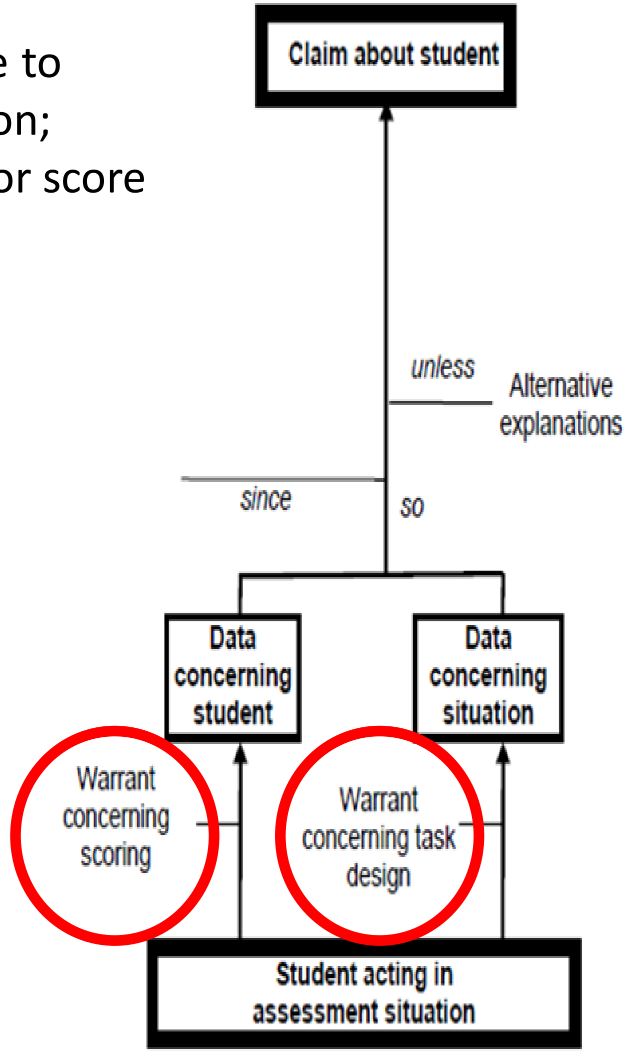
The tasks tap real-time automatic processes, and sample academic language & domain interactions

Machine-to-human score correlations ~0.97 (same tasks, same performance instance)

CLAIM

The tasks are valid for assessing spoken language proficiency

The scoring is sufficiently accurate to replace humans



Acknowledgements

Dr Jared Bernstein,
Consulting Scientist, Knowledge Technologies, Pearson

Prof John De Jong,
SVP Global Strategy & Business Development, Pearson