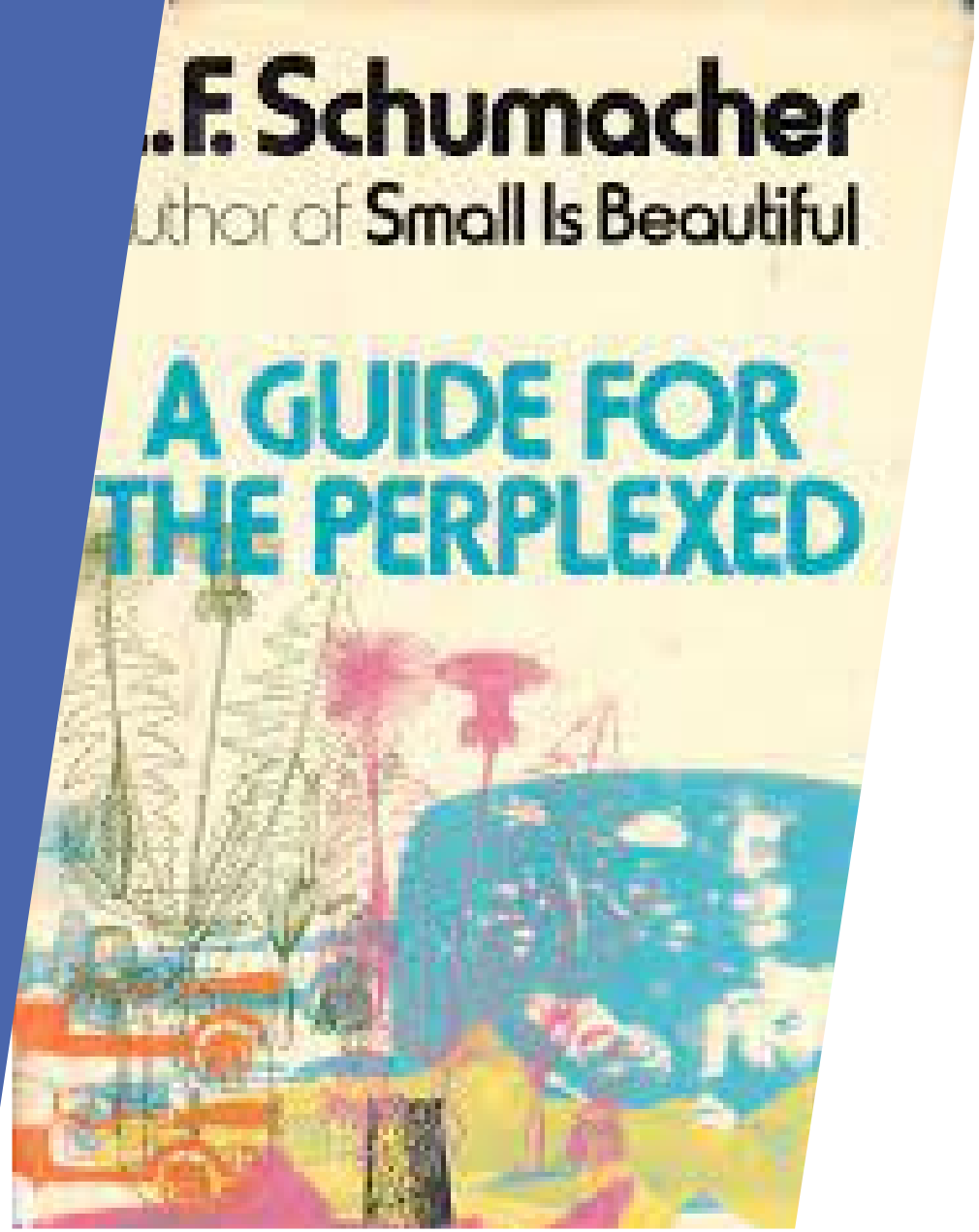
Language Courses and Tests

GROW MEASURE INTERACT

The practicality of theory: assessment and applied linguistics

Albert Weideman



Do we need theory?

"One way of looking at the world as a whole is by means of a map… some sort of … outline that shows … the things that are most prominent… for orientation - outstanding landmarks … Mapmaking … employs a high degree of abstraction, but nonetheless clings to reality with something akin to self-abandonment." (Schumacher 1977: 7)

Starting in1973…

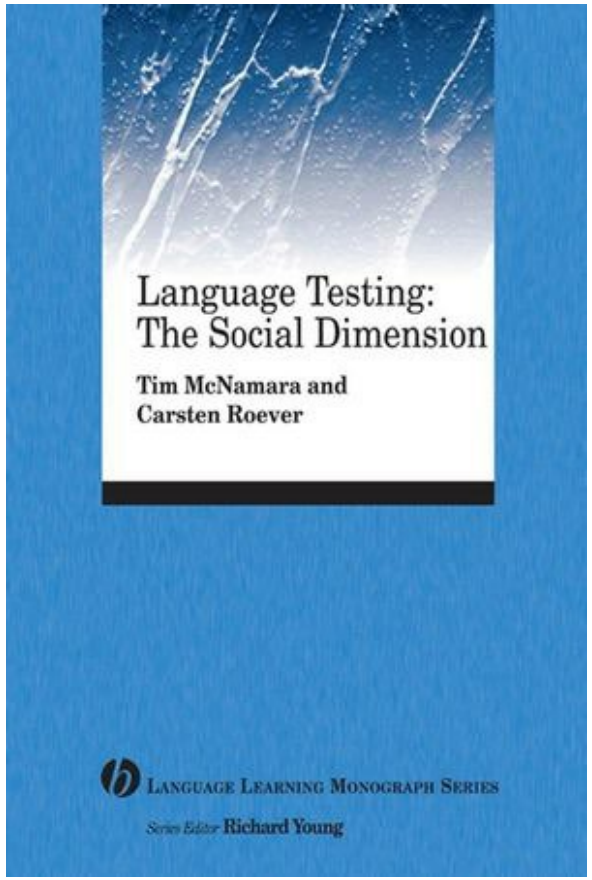
Designing language tests “is an activity of the applied linguist” (Pit Corder 1973: 353).

Everyone agrees, but there is neither argument nor evidence…

… applied linguistics is prepared in its curricula and its assessments and in its planning… to be accountable…”, [accomplishing that] “by theorising practice”

(Davies 2008: 298).

And fast forward from 1973…



Language "testing is … a central area of applied linguistics" (McNamara and Roever 2006: 255).

Language assessment and applied linguistics

- Language assessment is a subfield of applied linguistics …
- … and there are two crucial arguments that have not yet been made to demonstrate that.

Language assessment …

… is a subfield of applied linguistics because it …

1. shares in the elementary concepts / primitives of other applied linguistics subfields:
   - language policy development and
   - language course design;
2. demonstrates that **design** is key.

Language assessment is one of three subfields of applied linguistics

Normative level Blueprint (conditions for)	Factual level User interface artefact
language curriculum or development plan	language course and language learning
test construct and specifications	language test or battery of assessments
institutional language policy	language management plans and strategies
Levels of prime applied linguistic intervention designs and artefacts	

Where does one start to demonstrate conceptual convergence?

- with a theoretical clarification of what makes applied linguistic conceptualization possible
- In short: with a theory of applied linguistics

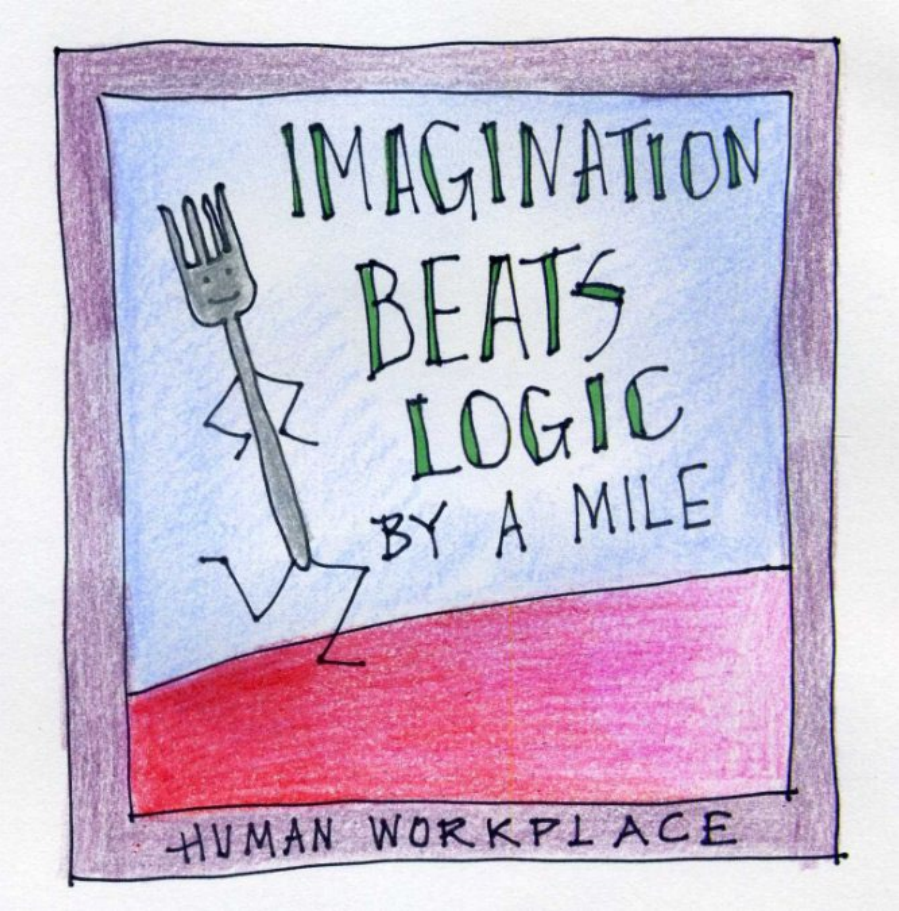
A theory of applied linguistics

- Identifies the *terminal* (qualifying and founding) functions of the three instruments/interventions:

  leading function: the **technical**

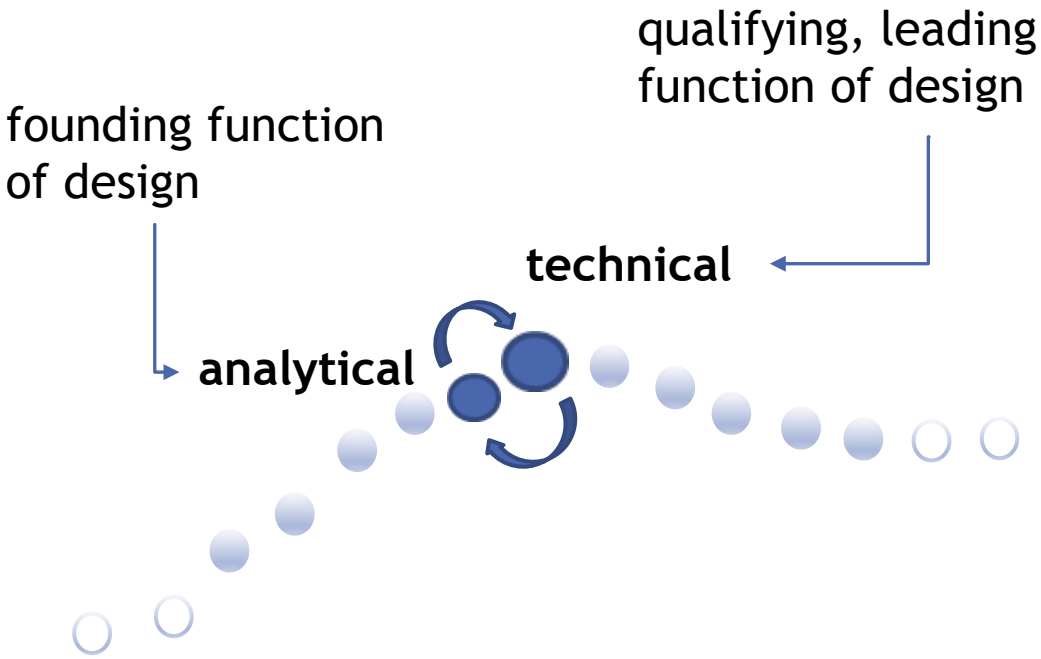
  founding function: the **analytical**

- Then focusses on the nuclear meaning of the technical: **designing,** shaping, planning, forming, styling, arranging, influencing

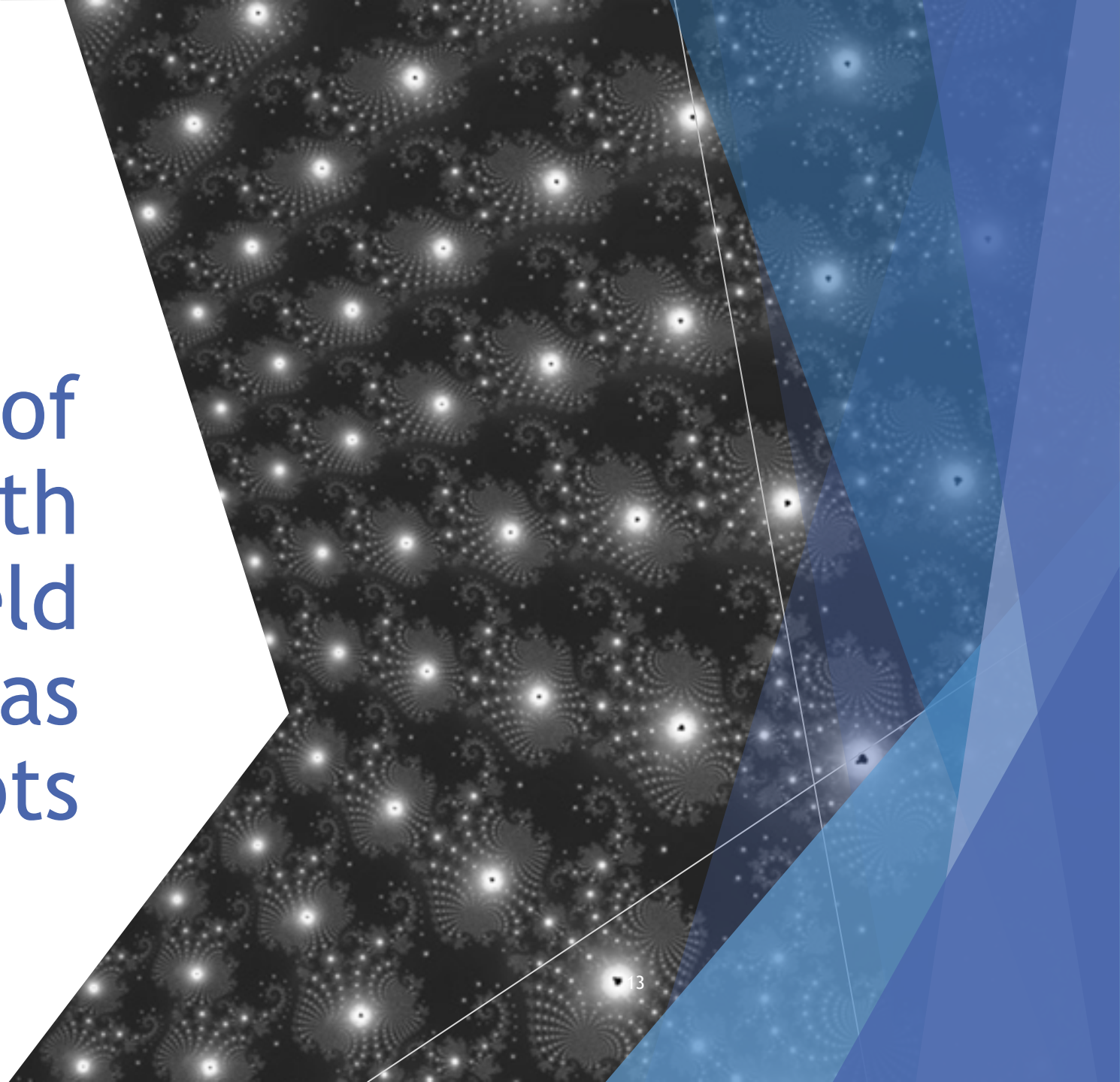


… in each course, test, or policy, the technical imagination leads the design

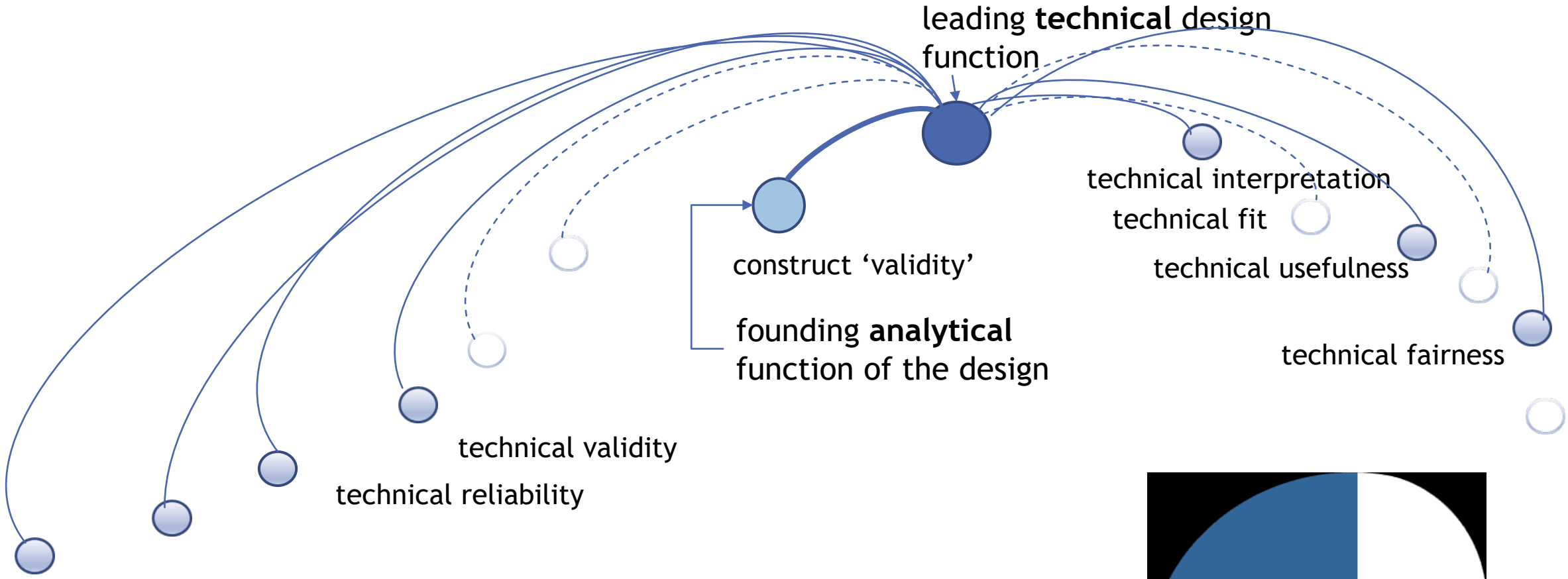
Theory conceptualizes interconnections that yield analogical concepts and ideas



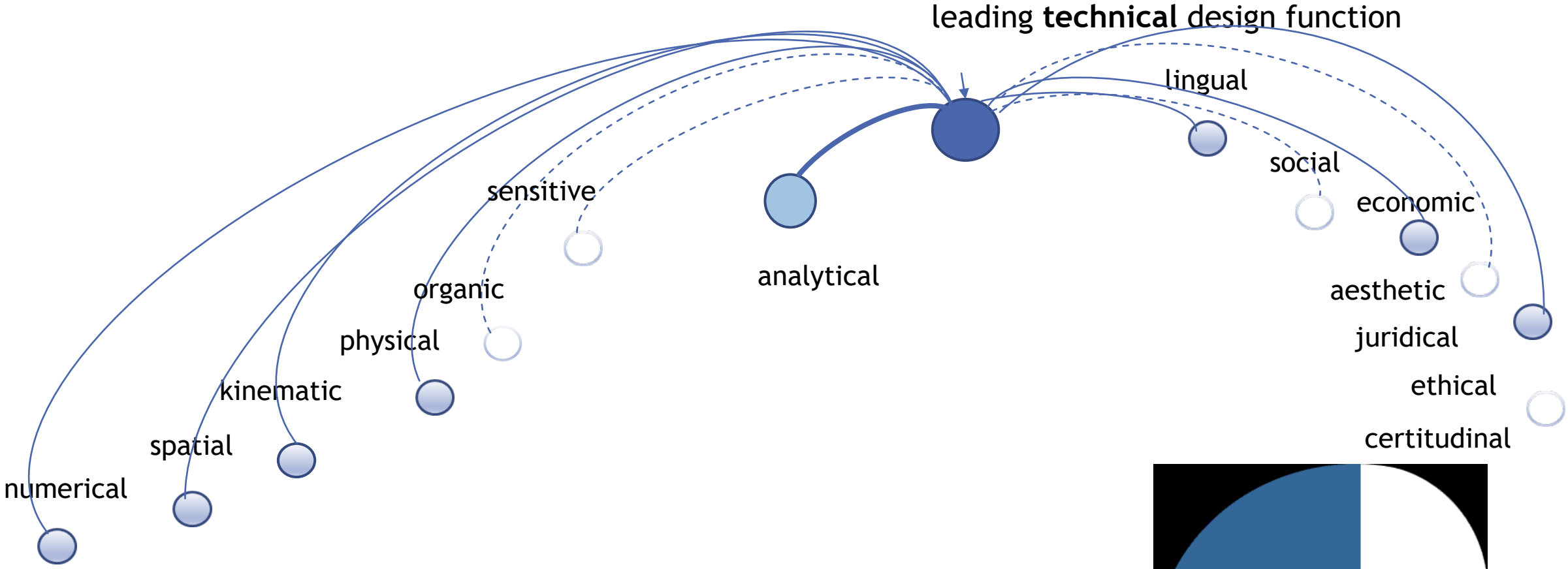
Interconnections of the technical with others yield fundamental ideas and concepts



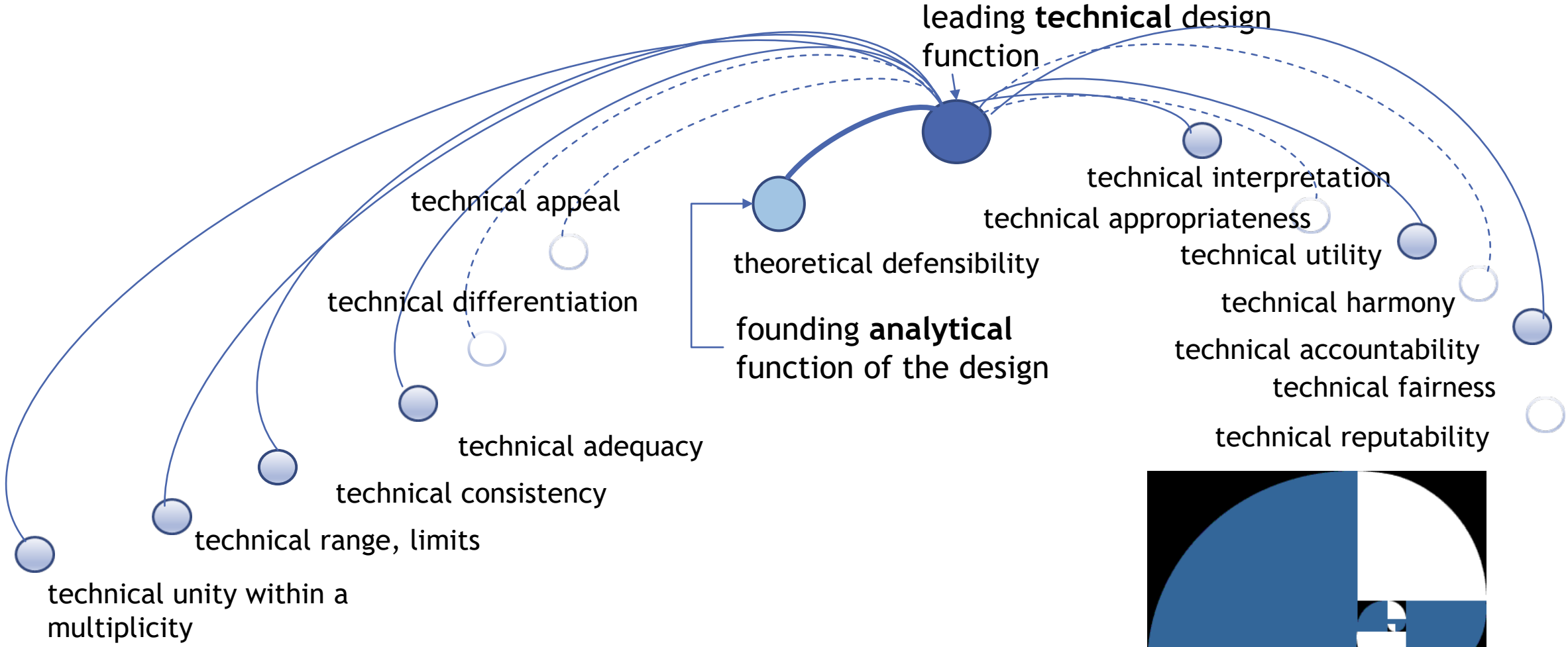
A range of conceptual primitives



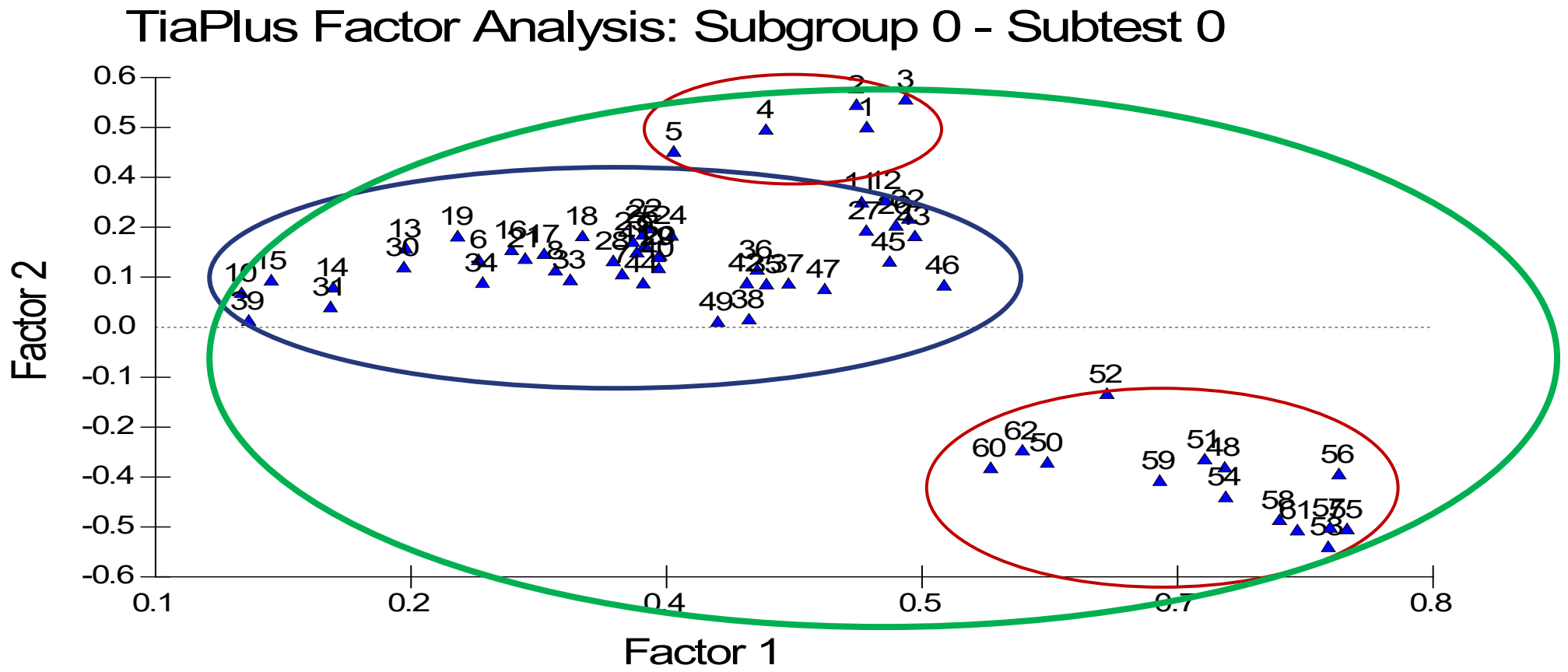
… deriving from analogical technical concepts and ideas …



Pivotal concepts and ideas generate design principles

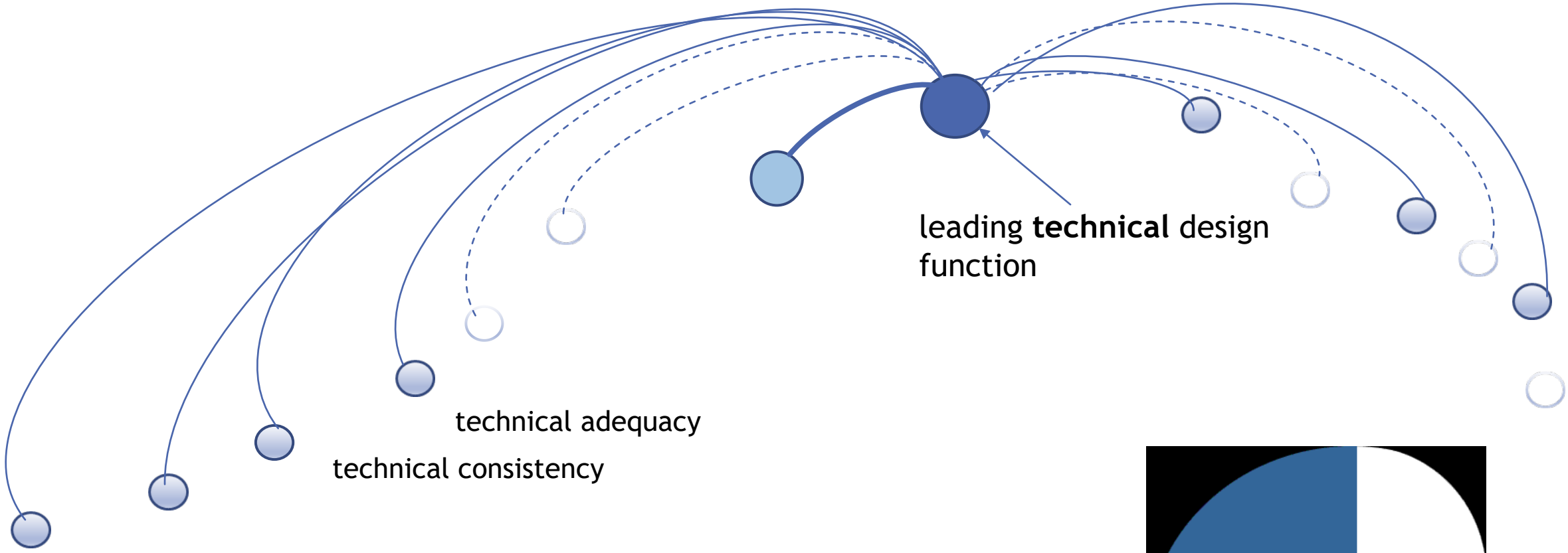


An first primitive: unity within multiplicity



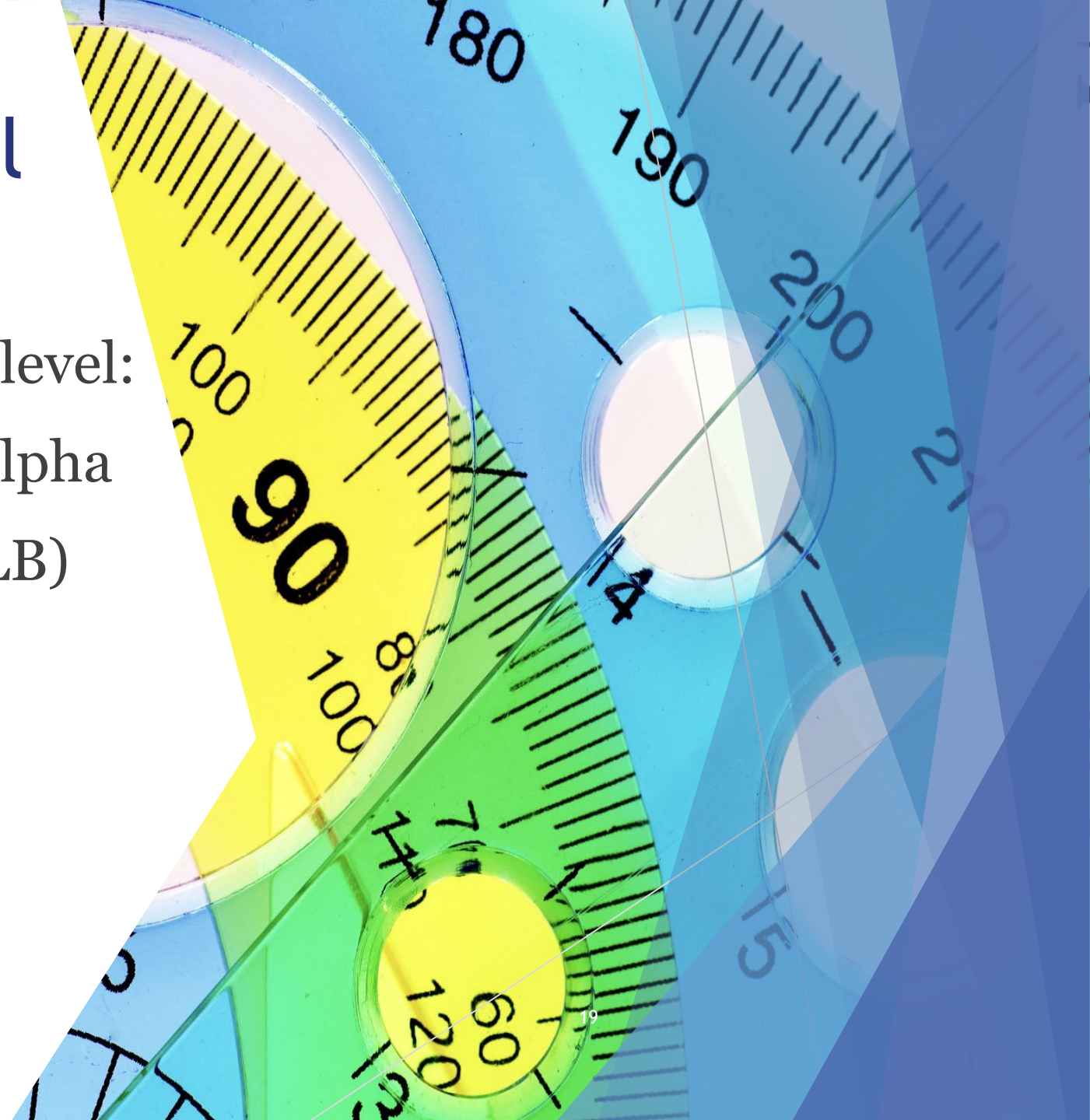
Factor analysis of TALL 2006

How do language interventions demonstrate consistency?



Measures of technical consistency

- From CTT, at test and item level:
  - (Cronbach's) coefficient alpha
  - Greatest lower bound (GLB)
- From Rasch
  - Person reliability
  - 'Item' reliability



An illustration of technical consistency

Year	University (campus)	Number of candidates	Reliability
2004	NWU (Potchefstroom)	124	0.92
	Pretoria (main)	3833	0.95
2005	NWU (Potchefstroom)	135	0.94
	Pretoria (main)	3310	0.93
2006	Pretoria (main)	3652	0.94
	NWU (Potchefstroom)	143	0.93
	Stellenbosch	2952	0.91
2007	Pretoria (main)	3905	0.94
2008	Pretoria (main)	4325	0.94
	NWU (Potchefstroom)	140	0.94
	Stellenbosch	4219	0.91
2009	Pretoria (main)	5191	0.94
Total and average		31929	**0.93**

Reliability indices for TALL (2004-2009)

Technical consistency relating to persons

SUMMARY OF 677 MEASURED (NON-EXTREME) PERSON

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	58.6	80.0	1.42	.34	1.00	.10	.99	.09
SEM	.6	.0	.05	.01	.00	.03	.01	.03
P.SD	15.4	.0	1.30	.14	.09	.70	.27	.90
S.SD	15.5	.0	1.30	.14	.09	.71	.27	.90
MAX.	79.0	80.0	4.68	1.01	1.31	3.82	2.79	4.51
MIN.	7.0	80.0	-2.58	.24	.75	-2.07	.20	-2.00

REAL RMSE	.37	TRUE SD	1.25	SEPARATION 3.36	PERSON RELIABILITY .92
MODEL RMSE	.37	TRUE SD	1.25	SEPARATION 3.42	PERSON RELIABILITY .92
S.E. OF PERSON MEAN = .05					

MAXIMUM EXTREME SCORE: 4 PERSON .6%

Person reliability and infit measures: ALLT

Differentiated design for technical viability

Subtest		Test	1	2	3	4
Skommelteks (Scrambled text)	1	0.53				
Woordeskat (Vocabulary)	2	0.62	0.11			
Grafiese & visuele informasie (Graphic & visual information)	3	0.60	0.18	0.48		
Teksbegrip (Text comprehension)	4	0.83	0.35	0.47	0.40	
Grammatika & teksverband (Grammar & text relations)	5	0.67	0.16	0.24	0.26	0.32

Subtest intercorrelations of “Apparaatjies en oulike uitvindsels” [Gadgets and freaky inventions]

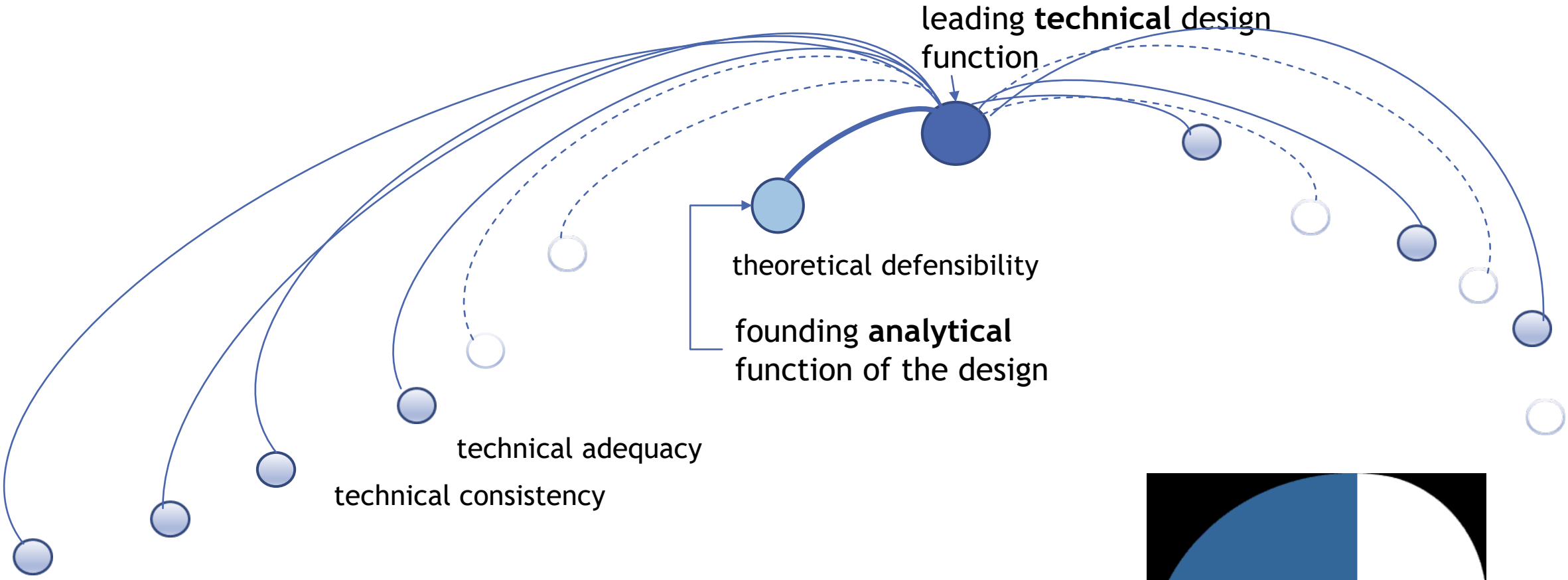
Differentiated and organized from easy to difficult

	Test	1	2	3	4	5	6
Number of testees	446	446	446	446	446	446	446
Number of items	80	18	20	12	5	5	20
Average test score	40.5	11.5	10.8	5.5	2.5	1.5	5.6
Average P-value	50.6	63.8	54.0	46.0	49.9	37.8	41.4
Standard deviation	12.34	3.77	3.61	2.29	1.54	1.31	5.55
SEM	2.35	1.32	1.74	1.48	0.75	0.71	1.22
Coefficient Alpha	0.90	0.75	0.69	0.60	0.58	0.51	0.90

Facility values of ALEF: test and subtests

Differentiation applies not only to tests but also to course design

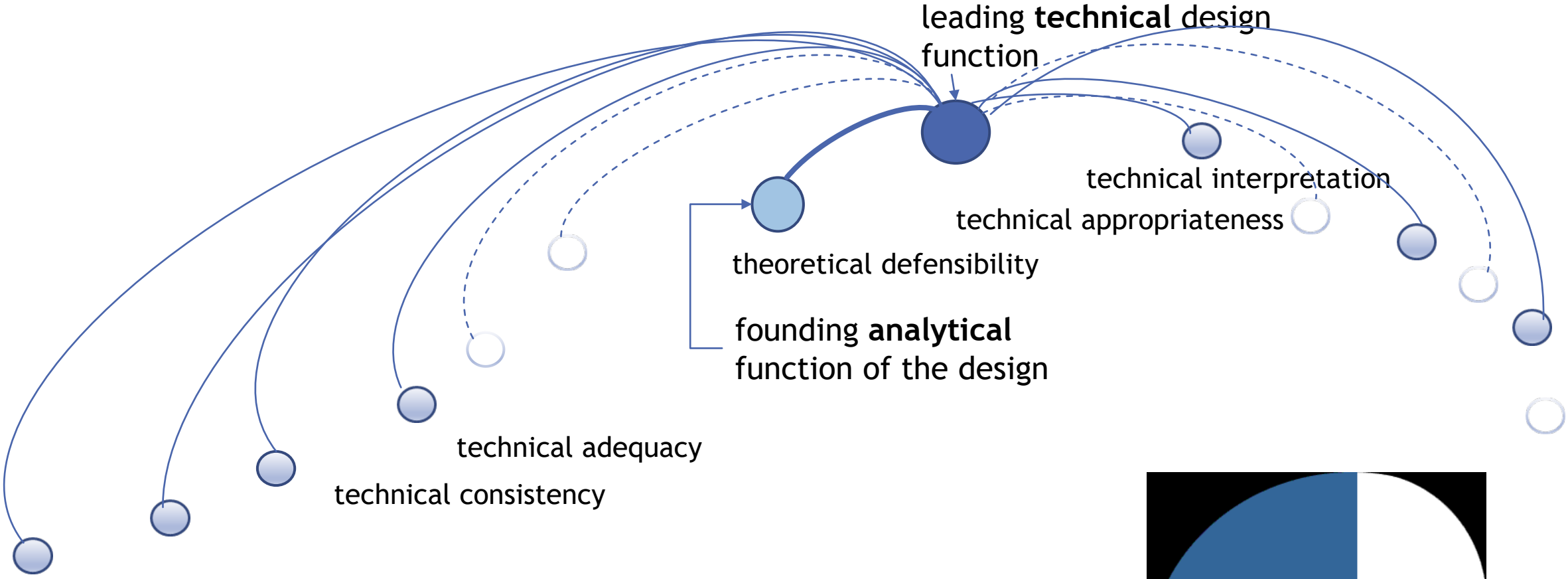
Construct validity = theoretical rationale



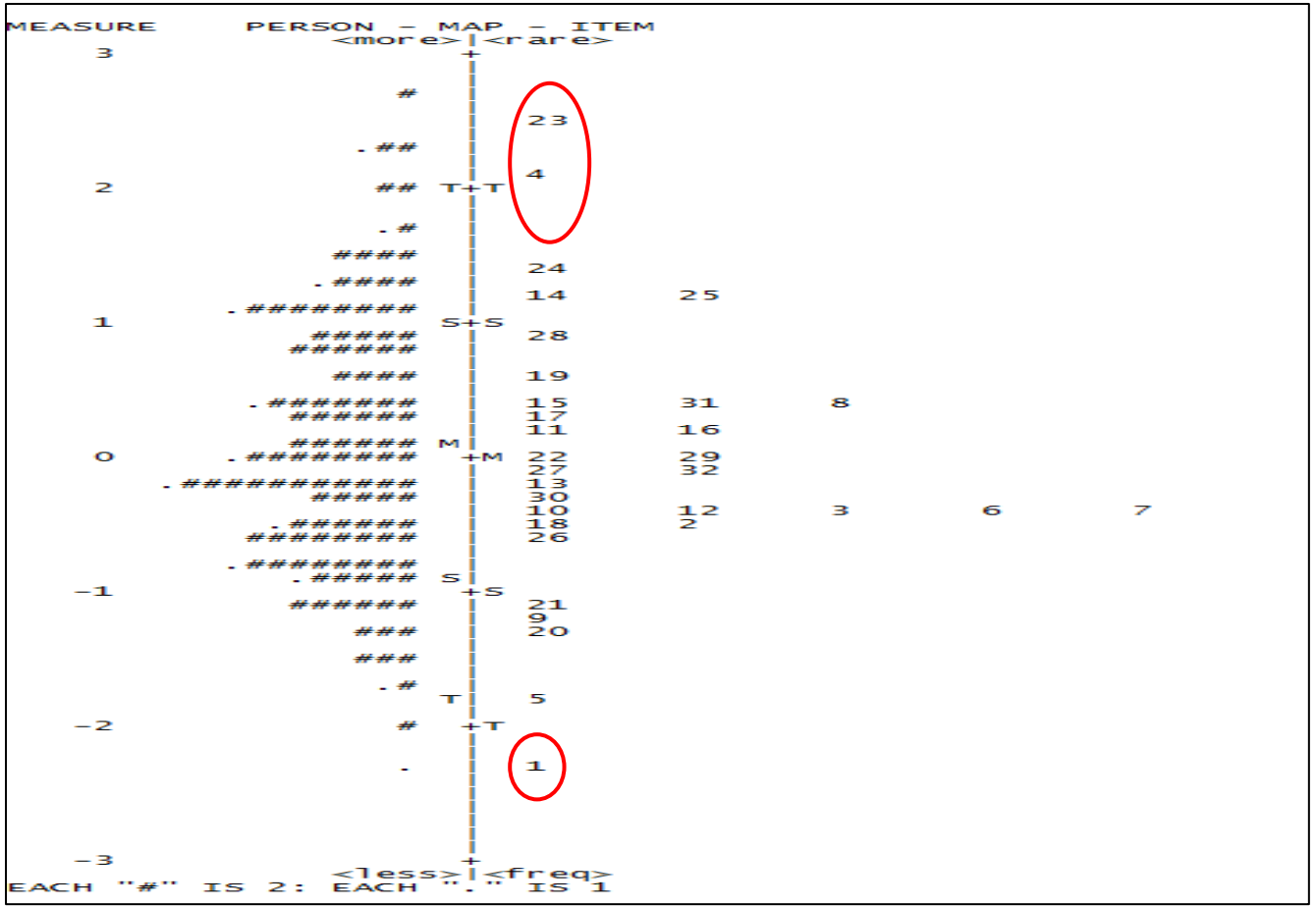
The construct of academic literacy

… the mastery of academic discourse, in which distinction-making and analytically qualified argument building are central

Appropriate and implementable: the social sides of language intervention design



Is there fit between ability and difficulty?

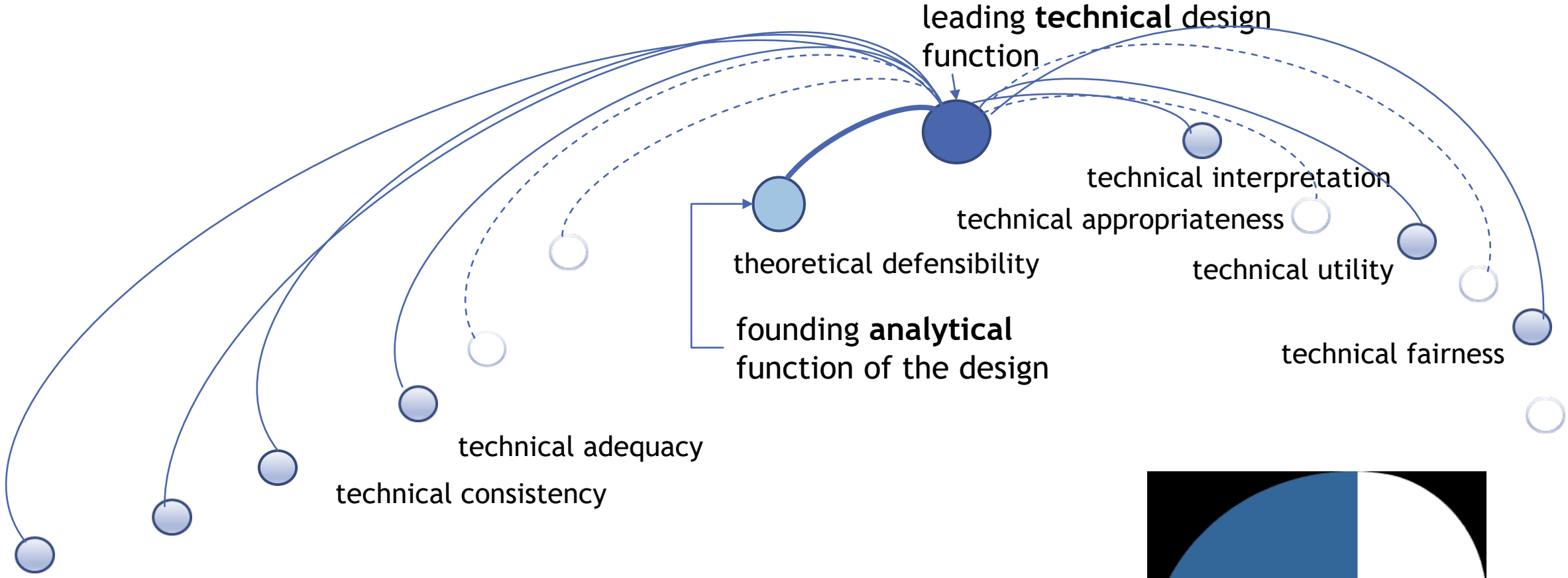


Person-item fit: TAGNaS Section 2

Further disclosures of design

- Usefulness: analogically *economic* dimension of language intervention design and implementation, such as efficiency and utility.
- What about “consequential validity”, the impact of a design? How fairly does a test measure? How compassionately does it treat those taking it? Was it designed with care and concern for them?
- *Technically* stamped *ethical* dimension: fairness, compassion and care

Designing for efficiency, and with care and compassion

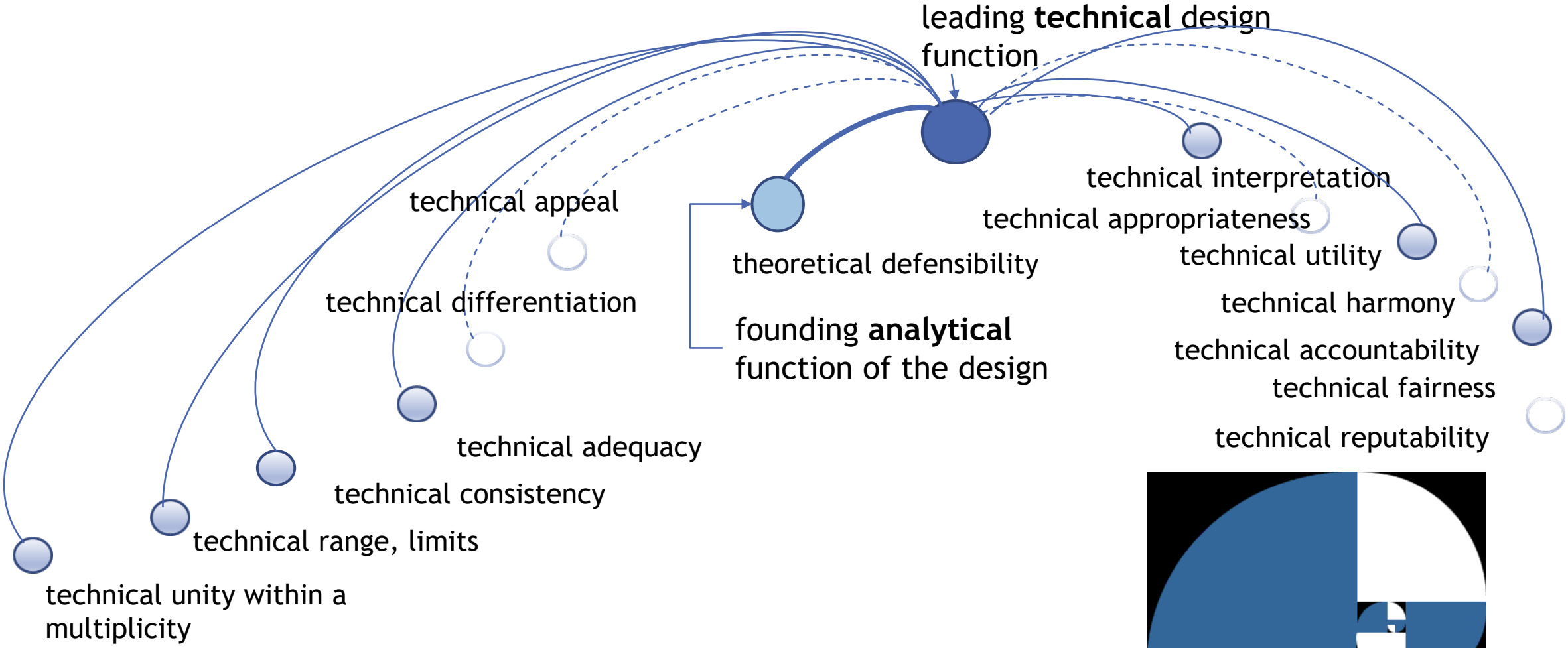


How fairly do tests treat those measured?

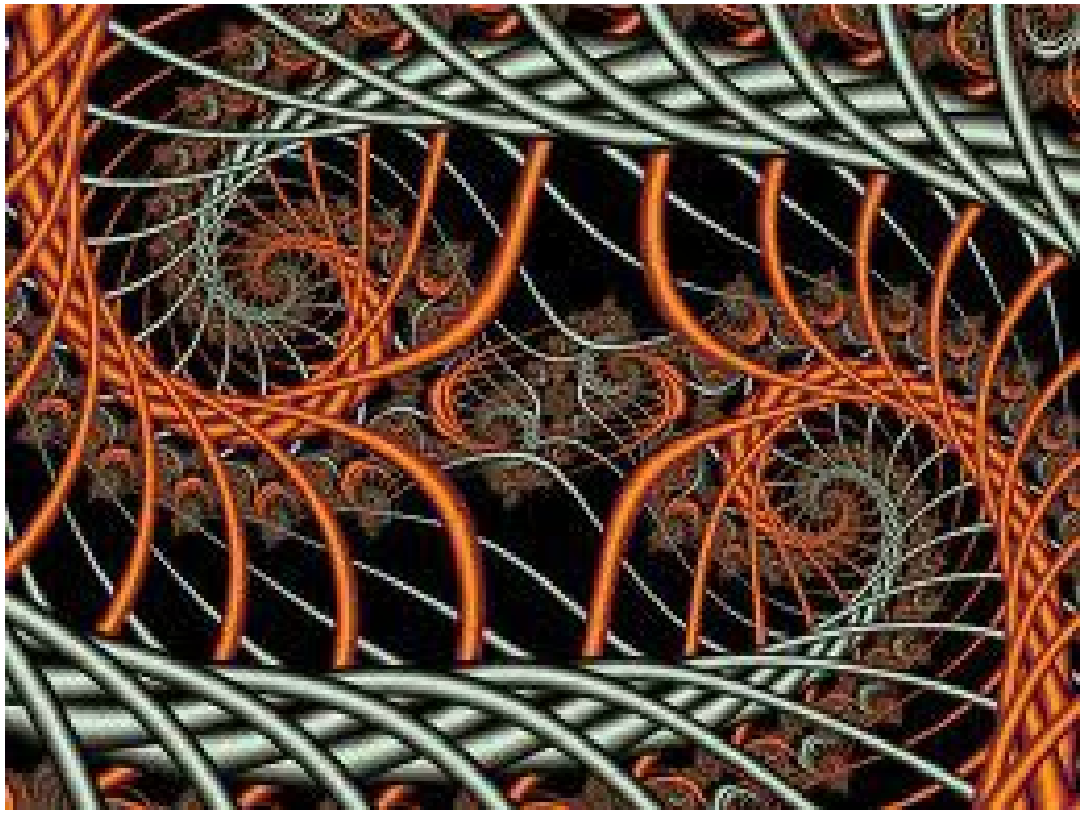
Misclassifications (out of a total of 5383 test takers)			
Alpha based		GLB based	
- Rxx' case: Percentage	3.8%	Percentage	3.3%
Number	204	Number	178
- Rxt case: Percentage	2.7%	Percentage	2.4%
Number	148	Number	128

Potential misclassifications for TALL (2011)

Primitives generate common principles of responsible design



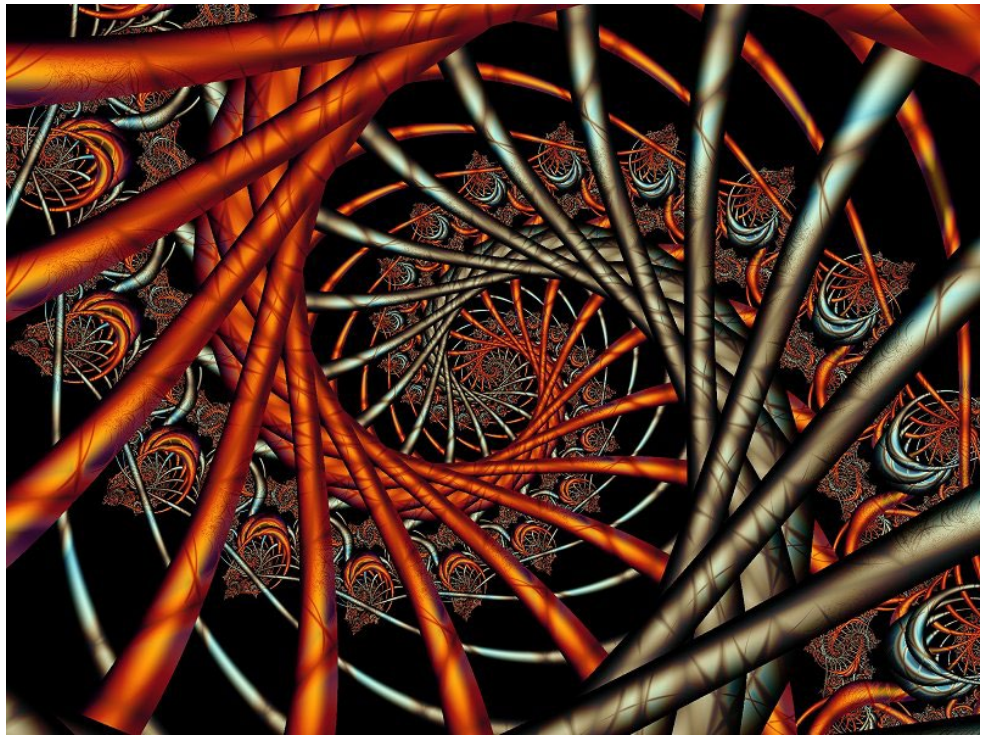
Learning lessons from design



- As diligent a validation of curricula and courses?
- Can our tests be as field specific as courses?
- How would policies become more transparent and accountable?

A robust, philosophically grounded theory of applied linguistics will …

- encompass all three prime language interventions (policies; assessments; courses)
- recognize both typical and general design principles
- discover these principles in factual designs, and coherently conceptualize them
- promote their responsible design



LCaT Language Courses and Tests

GROW MEASURE INTERACT

https://lcat.design

The ‘Blog’ tab on:

https://albertweideman.com

Albert Weideman

albert@lcat.design

albert.weideman48@gmail.com