UNIVERSITY ADMISSION L2 TESTS:
Knowing what you need to measure

Download the presentation at https://tinyurl.com/y882jpne

Or the dissertation at https://tinyurl.com/yct784dq

UNIVERSITY ADMISSION L2 TESTS:
Knowing what you need to measure

(And actually measuring it)

Bart Deygers, University of Leuven

ALTE 51st Meeting and Conference, Cluj-Napoca, 11th - 13th April 2018

Introduction
Who actually read the abstract?

Who actually read the abstract?

University admission L2 tests: Knowing what you need to measure, and actually measuring it.

During this workshop we will consider the assessment of languages for academic purposes (LAP), which has a lot of similarities with LSP testing. Nevertheless, there are a few important differences in terms of constructs and task selection, which we will discuss in this entry-level workshop.
The workshop will begin by determining the purpose of university admission testing for international L2 students and by listing which conditions might make a university admission policy effective, fair, valid, and just. We will compare different ways in which to measure students' language proficiency, and introduce methodologies that can be used to determine an adequate threshold language level.

Participants can expect to gain knowledge and insights regarding the construct and measurement of LAP, regarding admission policy effectiveness, and regarding responsible use of tests. During the workshop, we will use real-world test data collected in a four-year research project in Flanders, Belgium.

Who actually read the abstract?

University admission L2 tests: Knowing what you need to measure, and actually measuring it.

During this workshop we will consider the assessment of languages for academic purposes (LAP), which has a lot of similarities with LSP testing. Nevertheless, there are a few important differences in terms of constructs and task selection, which we will discuss in this entry-level workshop.
The workshop will begin by determining the purpose of university admission testing for international L2 students and by listing which conditions might make a university admission policy effective, fair, valid, and just. We will compare different ways in which to measure students' language proficiency, and introduce methodologies that can be used to determine an adequate threshold language level.

Participants can expect to gain knowledge and insights regarding the construct and measurement of LAP, regarding admission policy effectiveness, and regarding responsible use of tests. During the workshop, we will use real-world test data collected in a four-year research project in Flanders, Belgium.

"Everyone has the right to education [...] higher education shall be equally accessible to all on the basis of merit"

UN General Assembly. (1948, December 10). Universal Declaration of Human Rights. UN General Assembly. (Art. 26, §1)

"Everyone has the right to education […] higher education shall be equally accessible to all on the basis of merit"

Construct-relevant measurement

A lot of people, a long time

325,000 Study abroad students
1,700,000 Erasmus students
3,300,000 International students
+_____

5,325,000 (> 10,000,000 candidates?)

European Commission. (2015). *Erasmus. Facts, figures & trends*. Brussels: European Commission.
NAFSA. (2017). Trends in U.S. Study Abroad. Retrieved February 6, 2018, from nafsa.org.
OECD. (2017). *Education at a glance 2017: OECD indicators.* Paris: OECD Publishing.
Spolsky, B. (1995). *Measured Words: The Development of Objective Language Testing*. Oxford University Press (Sd).
Stein, Z. (2016). Social Justice and Educational Measurement. Oxon and New York: Routledge.

"When students are excluded from college because they do not secure a certain grade in a written examination, we assume a serious responsibility. The least we can do is to make a scientific study of our methods and results"

Cattell, J. M. (1905). Examinations, Grades and Credits. *Popular Science Monthly*, *66*, 367–378.

1 What are university admission tests for?

I need two volunteers*

(* it's about chocolate)

I need two volunteers

Volunteer 1:

I need two volunteers

Volunteer 2: Fill the gaps

Belgium's association with chocolate goes back as far as 1635 when the country was under Spanish occupation. The composition ____ Belgian chocolate has been ____ by law since 1894 ____ a minimum level of 35 ____ pure cocoa was imposed. ____ this day, many Belgian firms ____ chocolates by hand, ____ is laborious and explains the ____ of small, independent chocolate ____. Most chocolate companies ____ traditional recipes for their products.

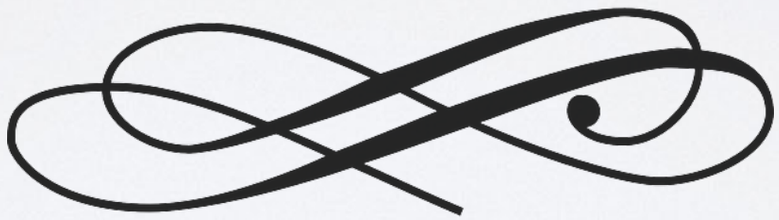
Q | Who are the primary stakeholder groups in a university admission testing policy?

Q | In one sentence, what is the purpose of university admission language tests?

3'

positive formulation

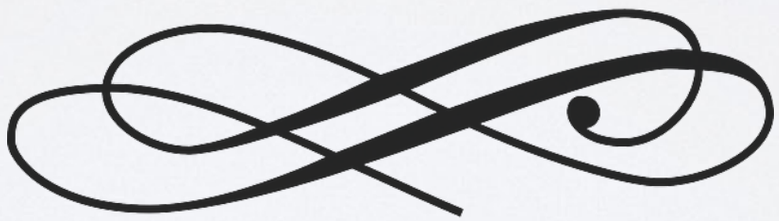
To select students who have a sufficient level of L2 proficiency to be able to attend university in that language



Deygers, B. & Malone, M. (accepted). Language assessment literacy in university admission policies, or the dialogue that isn't. *Language Testing*.

negative formulation

To identify students who do not have a sufficient level of L2 proficiency to be able to attend university in that language



McNamara, T., & Ryan, K. (2011). Fairness Versus Justice in Language Testing: The Place of English Literacy in the Australian Citizenship Test. *Language Assessment Quarterly*, *8*(2), 161–178.

cynical formulation

To control the flow of incoming students and generate revenue

Deygers, B. & Malone, M. (accepted). Language assessment literacy in university admission policies, or the dialogue that isn't. *Language Testing*.
Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2018). One Framework to Unite Them All? Use of the CEFR in European University Entrance Policies. *Language Assessment Quarterly*, *15*(1), 3–15.

Claims are made for testing

Kane (2013)

> Validation is empirically verifying the claims made on the basis of test scores
>
> Whoever makes / relies on a claim, is responsible for proving it
>
> Claims need to be supported by robust evidence
>
> The more unlikely a claim is, or the more impact it has, the more convincing the evidence should be

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Claims are made for testing

Kane

⇡

Messick (1989)

> Validity is a unitary concept containing five aspects (content, substantive, structural, external, consequential) that can be used when making a validity argument

Messick, S. (1989). Validity. In *Educational Measurement* (3rd ed., pp. 13–103). Washington, D.C.: American Council on Education / Macmillan.

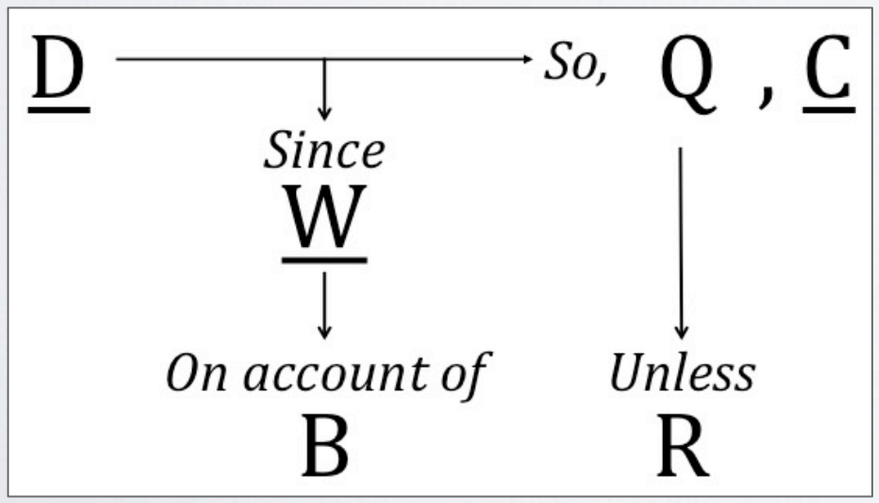
Claims are made for testing

Kane

Messick (1989)

Toulmin (1958, 2003)

You can't just make any claim, if you don't have the data to support it



Toulmin, S. (2003). *The Uses of Argument. Updated Edition.* (Updated edition). Cambridge, U.K.; New York: Cambridge University Press.

Claims are made for testing

Kane

Messick (1989)

Toulmin (1958, 2003)

↑

Laplace (1812)

"Plus un fait est extraordinaire, plus il a besoin d'être appuyé de fortes preuves"

"The weight of evidence for an extraordinary claim must be proportioned to its strangeness."

Laplace, P.S. (1812). T*héorie analytique des probabilités*. Paris: Courcier.

Q

What are some of the major claims made in university admission test policies?

Which stakeholders are responsible for which claims?

5’

Today's framework

Quality criteria for credentialing test programs:

- The test content needs to reflect the target domain

Today's framework

Quality criteria for credentialing test programs:

- The test content needs to reflect the target domain
- Operationalization should provide an appropriate basis for decisions

Today's framework

Quality criteria for credentialing test programs:

- The test content needs to reflect the target domain
- Operationalization should provide an appropriate basis for decisions
- The testing program should be of high psychometric quality

Today's framework

Quality criteria for credentialing test programs:

- The test content needs to reflect the target domain
- Operationalization should provide an appropriate basis for decisions
- The testing program should be of high psychometric quality
- The scores should be free of extraneous sources of variance (i.e. bias)

Today's framework

Quality criteria for credentialing test programs:

- The test content needs to reflect the target domain
- Operationalization should provide an appropriate basis for decisions
- The testing program should be of high psychometric quality
- The scores should be free of extraneous sources of variance
- If the testing program has an adverse impact, this impact should reflect real differences in the populations rather than defects in the testing program

Today's framework

Quality criteria for credentialing test programs:

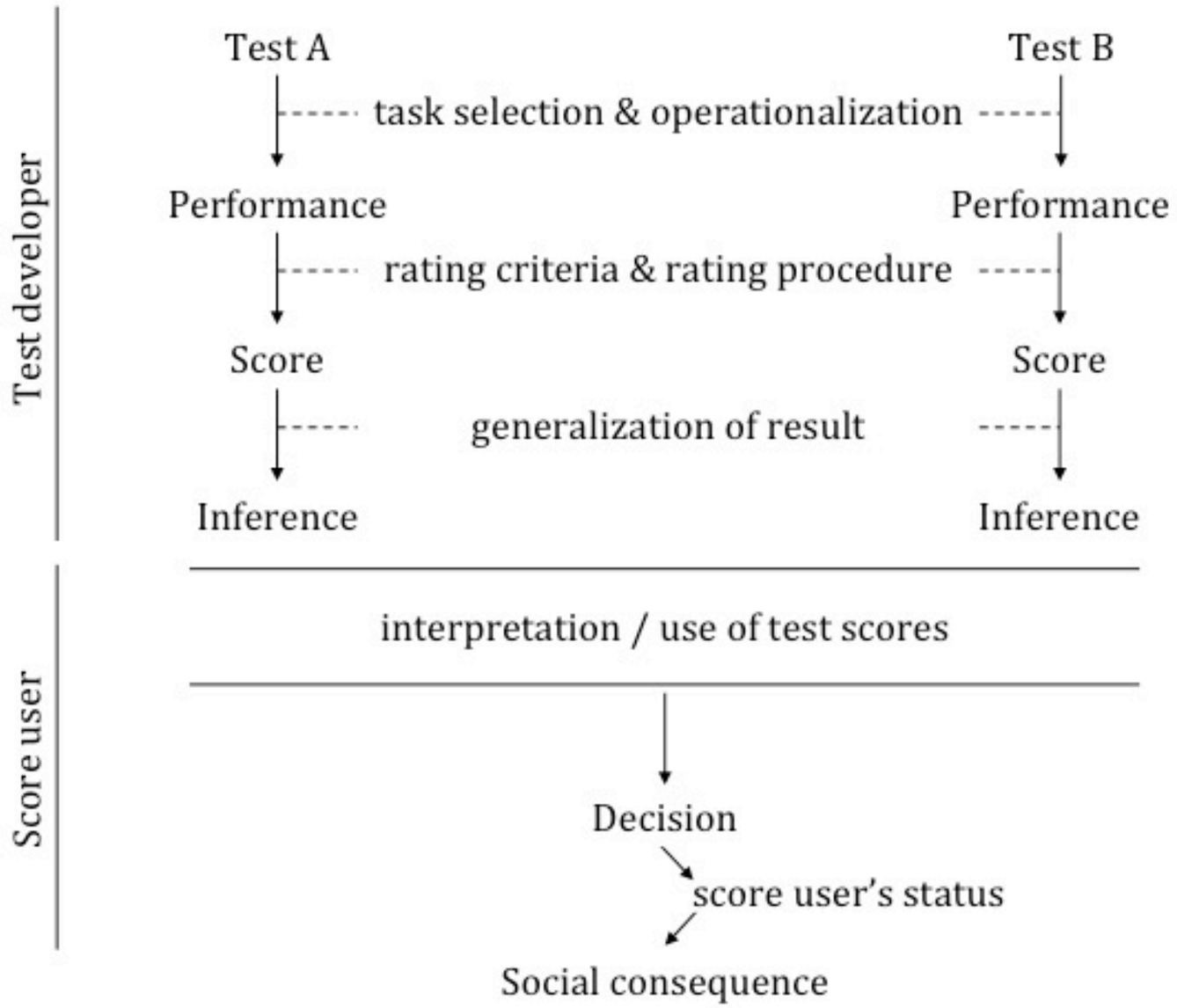
- The test content needs to reflect the target domain
- Operationalization should provide an appropriate basis for decisions
- The testing program should be of high psychometric quality
- The scores should be free of extraneous sources of variance
- If the testing program has an adverse impact, this impact should reflect real differences in the populations rather than defects in the testing program
- The cut scores should not be too high or too low

Today's framework

Quality criteria for credentialing test programs:

- The test content needs to reflect the target domain
- Operationalization should provide an appropriate basis for decisions
- The testing program should be of high psychometric quality
- The scores should be free of extraneous sources of variance
- If the testing program has an adverse impact, this impact should reflect real differences in the populations rather than defects in the testing program
- The cut scores should not be too high or too low
- The program should be as transparent as possible

Kane, M. T., Kane, J., & Clauser, B. E. (2017). A validation framework for credentialing tests. In C. W. Buckendahl & S. Davis-Becker (Eds.), *Testing in the Professions: Credentialing Polices and Practice* (pp. 20–41). New York, Routledge.



Deygers, B. (2017). *Assessing high-stakes assumptions. A longitudinal mixed-methods study of university entrance language tests, and of the policy that relies on them*. Leuven: Acco.

II

Claims

Q

Pick a topic, join a group

1. Task selection
2. Difficulty level
3. Psychometrics, cut scores & equivalence
4. Bias & DIF
5. Openness & communication

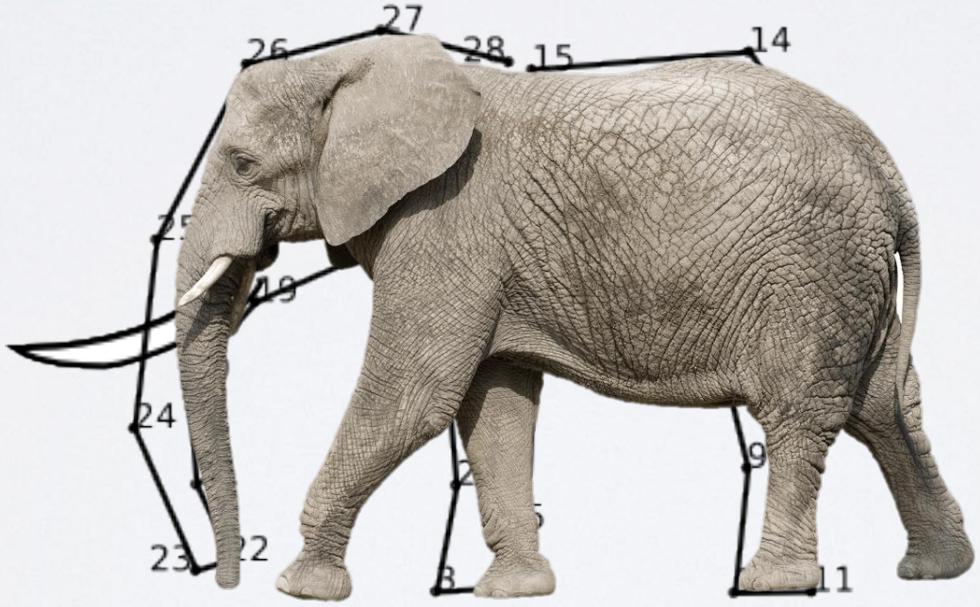
Check your assignment

Please do run around, mingle and join the discussion!
(it's ok to switch groups mid-session)

20'

1 The tasks are representative

The test content needs to reflect the target domain
KSJ operationalized in the test need to provide an appropriate basis for decisions



Continuum of opinions

"for face validity reasons, the stimuli in such tests will be field related"'

(Davies, 2001)

"we must go as far as we can"

(Hyland, 2002)

Davies, A. (2001). The logic of testing Languages for Specific Purposes. Language Testing, 18(2), 133–147.
Hyland, K. (2002). Specificity Revisited: How Far Should We Go Now? *English for Specific Purposes*, *21*(4), 385–395.

LAP skills & tests

- Are all test tasks essential for the target context?
- Would the absence of certain test tasks make a substantial difference in real-world practice?

LAP skills

Compose a logical argumentation
Describe graphs & tables
Express ideas accurately
Give a presentation
Grammatical accuracy
Look up information
Summarize long text
Summarize multiple sources
Take class notes
Understand coherence & cohesion
Understand general academic lexis
Understand implicit message
Understand scientific text as a whole
Understand scientific text in detail
…

Deygers, B., Van den Branden, K., & Van Gorp, K. (2017). University entrance language tests: A matter of justice. *Language Testing*.
Hulstijn, J. (2015). Language Proficiency in Native and Non-native Speakers: Theory and research. Amsterdam; Philadelphia: John Benjamins
Snow, C. (2010). Academic Language and the Challenge of Reading for Learning About Science. Science, 328(5977), 450–452.

LAP skills & tests

	T1						T2						
	Argumentative writing from audio	Summarize lecture	Argumentative writing from text	Summarize paper	Oral argumentation	Presentation	Language-in-use, vocabulary	Language-in-use, grammar	Reading comprehension	Listening comprehension	Listening comprehension, dictation	Oral argumentation	Presentation
Compose a logical argumentation	★		★		★							★	★
Take class notes	★	★										★	★
Express ideas accurately	★	★	★	★	★	★							
Grammatical accuracy	★	★	★	★	★	★		★				★	★
Understand general academic lexis	★	★	★	★	★	★	★		★	★		★	★
Understand coherence & cohesion	★	★	★	★	★	★			★			★	★
Understand implicit message	★	★		★						★			
Understand scientific text as a whole				★					★				
Look up information													
Summarize long text				★									
Summarize multiple sources													
Understand scientific text in detail				★									
Describe graphs & tables						★							★
Give a presentation						★							★

2 The difficulty level is appropriate

Content representation = task selection + task operationalization

Necessary vs effective

"Goldilocks Criterion": neither too high, nor too low

Balancing necessary performance (what is minimally required) and effective performance (what should ideally be mastered)

Carlsen, C. H. (2018). The Adequacy of the B2 Level as University Entrance Requirement. *Language Assessment Quarterly*, 15(1), 75–89.
Kane, M. T., Kane, J., & Clauser, B. E. (2017). A validation framework for credentialing tests. In C. W. Buckendahl & S. Davis-Becker (Eds.), *Testing in the Professions: Credentialing Polices and Practice* (pp. 20–41). New York, Routledge.
Markus, K. A., & Borsboom, D. (2013). *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning*. New York: Routledge.

Level requirements

S W L R

Listening

Lexis - Speech fluency - Background noise – Structure - Accent

		1K-2K*	5K-7K	7K+	w/m#
Test (*N* = 8)	*M*	5.93	2.33	6.50	148.33
	SD	4.43	2.52	2.78	18.04
Les (*N* = 12)	*M*	6.67	1.23	10.37	103.86
	SD	3.79	1.08	5.82	18.60

* % of words used in frequency band

#mean words/minute

Field, J. (2011). Into the mind of the academic listener. Journal of English for Academic Purposes, 10(2), 102–112.

3 The psychometrics are sound

The testing program should be of high psychometric quality

Typically: Raw score > IRT model > latent score
Latent score is stable over replications of testing procedure

Also, see ALTE's minimum standards!

The cut scores are just right

Flemish context: No significant relationship between language test scores & academic success

T1: $W = 46$, $p = .625$, $r = -.115$;
T2: $W = 51$, $p = .599$, $r = -.120$

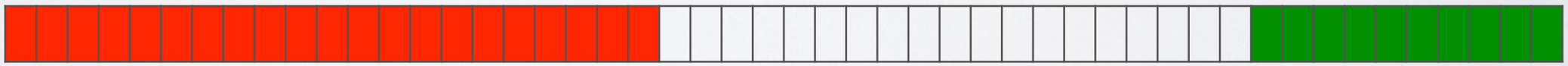
Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT® scores to academic performance: Some evidence from American universities. Language Testing, 29(3), 421–442.
Deygers, B., Van den Branden, K., & Van Gorp, K. (2017). University entrance language tests: A matter of justice. *Language Testing*.

Cut scores

Predictive validity, false positives & false negatives

B2
False positives: 41%
False negatives: 6%

C1
False positives: 21%
False negatives: 42%



Deygers, B. (2017). *Assessing high-stakes assumptions. A longitudinal mixed-methods study of university entrance language tests, and of the policy that relies on them*. Leuven: Acco.

Lee, Y.-J., & Greene, J. (2007). The Predictive Validity of an ESL Placement Test A Mixed Methods Approach. *Journal of Mixed Methods Research, 1*(4), 366–389.

Tests are equivalent

Correlations & CEFR levels do not guarantee equivalence

Overall high correlation

r = .767**; writing r = .694**; oral τ = .387**

But T-tests: significant differences mean scores ($p < 0.001$)

d = -0.53 (writing) d = -1.41 (speaking)

And significantly different pass probability

T1 (P = .50), T2 (P = .35) (p = .02)

Deygers, B., Van Gorp, K., & Demeester, T. (2018). The B2 Level and the Dream of a Common Standard. *Language Assessment Quarterly*, 15(1), 44–58.

Green, A. (2018). Linking Tests of English for Academic Purposes to the CEFR: The Score User's Perspective. Language Assessment Quarterly, 15(1), 59–74.

4 The scores are bias-free

The scores should be free of extraneous sources of variance

Differences in scores should reflect real differences in the populations

Bias, & DIF

Differential Item Functioning (DIF):

- certain test items function differently for different populations
- not necessarily problematic: certain groups could systematically underperform for construct-relevant reasons

Bias:

- Items with DIF systematically (dis)advantage specific populations on construct-irrelevant grounds

Fairness, in the narrow sense:

- Absence of bias

McNamara, T., & Roever, C. (2006). Language Testing: The Social Dimension. New York: John Wiley & Sons.

Fairness / justice

MFRA for facet “Group”

	Measure	Model SE	Infit	% below cut-off
Flemish	0.57	0.02	1.12	11
$L2_I$	-0.15	0.01	0.82	30
$L2_F$	-0.43	0.01	1.29	57

A testing policy is likely to be *unjust* if it wilfully and avoidably restricts test takers’ freedom without an empirically sound or reasonable motivation.

Deygers, B. (2017). Just testing. Applying theories of justice to high-stakes language tests. *ITL – International Journal of Applied Linguistics*, *168*(2), 143–162.

III

Test developers & policy makers

5 Openness & communication

We know about:

- LAL literature
- MS 17
- ...

Testers - society

The importance of public justification

Click to open expanded view

Zorvo

Zorvo Mini USB Fridge Cooler Beverage Drink Cans Cooler/Warmer Refrigerator Laptop PC Office Car Refrigerator

4 customer reviews

Price: $22.59

Get $50 off instantly: Pay $0.00 upon approval for the Amazon Rewards Visa Card.

Only 20 left in stock - order soon.

This item does not ship to **Gentbrugge, Belgium**. Please check other sellers who may ship internationally.
Learn more

Sold by SimSimCar and Fulfilled by Amazon. Gift-wrap available.

Color: **Black**

$22.59 | $20.99

- ✔ Portable mini fridge, apply to all PC or other USB interface, no batteries or driver required, plug and play
- ✔ Switch Control: "I" for heating, "II" for cooling, O for off
- ✔ Volume: 7.6*3.5*3.5 inches, Powered by 4 feet USB cable, suitable for car office or home to keep beverage, coca-cola, cans, coffee and other drinks
- ✔ Temperature Range:8.5 deg C for cooling; 70deg C for heating,(Note: it is without Cooling $ Heating function as the actual fridge)

Kunnan, A. J. (2018). Evaluating language assessments. New York & London: Routledge.

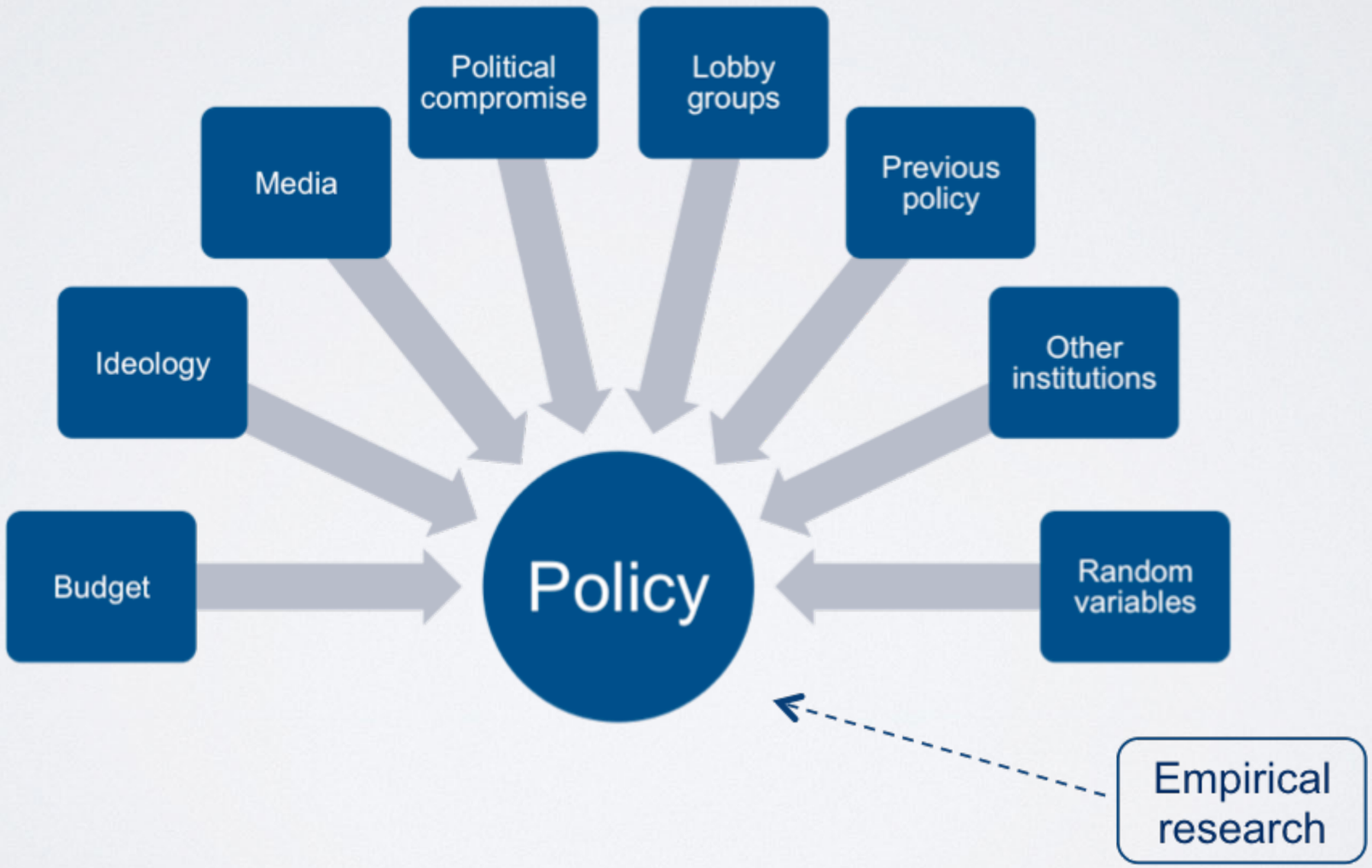
Testers - society

assessment institutions ought to be asked … to submit their assessments and assessment practices to formal evaluations …test reviewers do not have …assessment performance data supplied to them for secondary analyses…Thus, unlike consumer reports of products or car reviews that are based on test trials, assessment reviewers write reviews without access to the actual assessment instrument and assessment performance data.

Kunnan, A. J. (2018). Evaluating language assessments. New York & London: Routledge.

Testers - policy

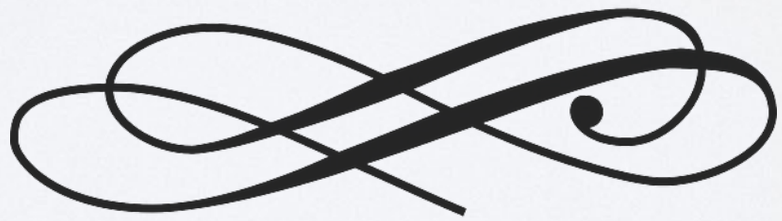
Policy as a pragmatic compromise



Jann, W., & Wegrich, K. (2007). Theories of the Policy Cycle. In F. Fischer & G. J. Miller (Eds.), *Handbook of Public Policy Analysis: Theory, Politics, and Methods* (pp. 43–62). Boca Raton: CRC Press.

Testers - policy

"Actually, research is often used when it can help to prove a point we like to make. Research is often used a little selectively, like when it can support our policy"



Deygers, B. & Malone, M. (accepted). Language assessment literacy in university admission policies, or the dialogue that isn't. *Language Testing*.

Testers - policy

Acceptance into university is a clear example of a decision affecting the lives of individuals. It is therefore necessary to continue to explore the process by which language assessment scores are used for this purpose. Any way forward must continue to foster productive collaboration between language test score users and language assessment specialists.

Baker, B. A., Tsushima, R., & Wang, S. (2014). Investigating language assessment literacy: Collaboration between assessment specialists and canadian university admissions officers. *Language Learning in Higher Education*, *4*(1), 137–157.

Testers - policy

Need for testers to

- increase policy literacy
- end top-down thinking about LAL
- interact with policy makers

Baker, B. A. (2016). Language assessment literacy as professional competence: The case of Canadian admissions decision makers. *Canadian Journal of Applied Linguistics*, *19*(1), 63–83.
Lo Bianco, J. (2014). Dialogue between ELF and the field of language policy and planning. *Journal of English as a Lingua Franca*, *3*(1), 197–213.
McNamara, T., & Shohamy, E. (2008). Language tests and human rights. *International Journal of Applied Linguistics*, *18*(1), 89–95.

IV

Concluding

Determine construct

Without overly relying on outcomes of earlier research conducted in different contexts

Deygers, B., Van den Branden, K., & Van Gorp, K. (2017). University entrance language tests: A matter of justice. *Language Testing*.
Fløttum, K., Gedde-Dahl, T., & Kinn, T. (2006). *Academic Voices: Across Languages and Disciplines*. Amsterdam: John Benjamins Publishing.

…measure it

Evidence for every inference!
Using suitable data!
Supported by robust backing!

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73.
Kane, M. T., Kane, J., & Clauser, B. E. (2017). A validation framework for credentialing tests. In C. W. Buckendahl & S. Davis-Becker (Eds.), *Testing in the Professions: Credentialing Polices and Practice* (pp. 20–41). New York, Routledge.

…and communicate it well

Speak the language of policy makers & interact!
Be clear to candidates & teachers

Lo Bianco, J. (2014). Dialogue between ELF and the field of language policy and planning. *Journal of English as a Lingua Franca*, *3*(1), 197–213.
O'Sullivan, B. (2016, April). A Story to Tell, a Lesson to Learn: The Testing Industry and Validation. Presented at the ALTE 48th Conference Day, Stockholm.

Do get in touch!

bart.deygers@kuleuven.be

References

Baker, B. A. (2016). Language assessment literacy as professional competence: The case of Canadian admissions decision makers. *Canadian Journal of Applied Linguistics*, *19*(1), 63–83.

Baker, B. A., Tsushima, R., & Wang, S. (2014). Investigating language assessment literacy: Collaboration between assessment specialists and canadian university admissions officers. *Language Learning in Higher Education*, *4*(1), 137–157.

Carlsen, C. H. (2018). The Adequacy of the B2 Level as University Entrance Requirement. *Language Assessment Quarterly*, 15(1), 75–89.

Cattell, J. M. (1905). Examinations, Grades and Credits. *Popular Science Monthly*, *66*, 367–378.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT® scores to academic performance: Some evidence from American universities. Language Testing, 29(3), 421–442.

Cumming, A. (2013). Assessing Integrated Writing Tasks for Academic Purposes: Promises and Perils. *Language Assessment Quarterly*, 10(1)

Deygers, B. (2017). Just testing. Applying theories of justice to high-stakes language tests. *ITL – International Journal of Applied Linguistics*, *168*(2), 143–162.

Deygers, B. & Malone, M. (accepted). Language assessment literacy in university admission policies, or the dialogue that isn't. *Language Testing*.

Deygers, B., Van den Branden, K., & Van Gorp, K. (2017). University entrance language tests: A matter of justice. *Language Testing*.

Deygers, B., Van Gorp, K., & Demeester, T. (2018). The B2 Level and the Dream of a Common Standard. *Language Assessment Quarterly*, 15(1), 44–58.

Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2018). One Framework to Unite Them All? Use of the CEFR in European University Entrance Policies. *Language Assessment Quarterly*, *15*(1), 3–15.

European Commission. (2015). *Erasmus. Facts, figures & trends*. Brussels: European Commission.

Field, J. (2011). Into the mind of the academic listener. Journal of English for Academic Purposes, 10(2), 102–112.

Fløttum, K., Gedde-Dahl, T., & Kinn, T. (2006). *Academic Voices: Across Languages and Disciplines*. Amsterdam: John Benjamins Publishing.

Green, A. (2018). Linking Tests of English for Academic Purposes to the CEFR: The Score User's Perspective. Language Assessment Quarterly, 15(1), 59–74.

Hulstijn, J. (2015). Language Proficiency in Native and Non-native Speakers: Theory and research. Amsterdam; Philadelphia: John Benjamins

References

Jann, W., & Wegrich, K. (2007). Theories of the Policy Cycle. In F. Fischer & G. J. Miller (Eds.), *Handbook of Public Policy Analysis: Theory, Politics, and Methods* (pp. 43–62). Boca Raton: CRC Press.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kane, M. T., Kane, J., & Clauser, B. E. (2017). A validation framework for credentialing tests. In C. W. Buckendahl & S. Davis-Becker (Eds.), *Testing in the Professions: Credentialing Polices and Practice* (pp. 20–41). New York, Routledge.

Krekeler, C. (2006). Language for special academic purposes (LSAP) testing: the effect of background knowledge revisited. Language Testing, 23(1), 99–130.

Kunnan, A. J. (2018). Evaluating language assessments. New York & London: Routledge.

Laplace, P.S. (1812). T*héorie analytique des probabilités*. Paris: Courcier.

Lee, Y.-J., & Greene, J. (2007). The Predictive Validity of an ESL Placement Test A Mixed Methods Approach. *Journal of Mixed Methods Research*, *1*(4), 366–389.

Lo Bianco, J. (2014). Dialogue between ELF and the field of language policy and planning. *Journal of English as a Lingua Franca*, *3*(1), 197–213.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning*. New York: Routledge.

McNamara, T., & Roever, C. (2006). Language Testing: The Social Dimension. New York: John Wiley & Sons.

McNamara, T., & Ryan, K. (2011). Fairness Versus Justice in Language Testing: The Place of English Literacy in the Australian Citizenship Test. *Language Assessment Quarterly*, *8*(2), 161–178.

McNamara, T., & Shohamy, E. (2008). Language tests and human rights. *International Journal of Applied Linguistics*, *18*(1), 89–95.

Messick, S. (1989). Validity. In *Educational Measurement* (3rd ed., pp. 13–103). Washington, D.C.: American Council on Education / Macmillan.

NAFSA. (2017). Trends in U.S. Study Abroad. Retrieved February 6, 2018, from nafsa.org.

OECD. (2017). *Education at a glance 2017: OECD indicators.* Paris: OECD Publishing.

O'Sullivan, B. (2016, April). A Story to Tell, a Lesson to Learn: The Testing Industry and Validation. Presented at the ALTE 48th Conference Day, Stockholm.

Sasayama, S. (2016). Is a "Complex" Task Really Complex? Validating the Assumption of Cognitive Task Complexity. The Modern Language Journal, 100(1), 231–254.

Snow, C. (2010). Academic Language and the Challenge of Reading for Learning About Science. Science, 328(5977), 450–452.

Spolsky, B. (1995). *Measured Words: The Development of Objetive Language Testing*. Oxford University Press (Sd).

References

Stein, Z. (2016). Social Justice and Educational Measurement. Oxon and New York: Routledge.

Toulmin, S. (2003). *The Uses of Argument. Updated Edition.* (Updated edition). Cambridge, U.K.; New York: Cambridge University Press.

UN General Assembly. (1948, December 10). Universal Declaration of Human Rights. UN General Assembly. (Art. 26, §1)

Xi, X. (2010). Aspects of performance on line graph description tasks: influenced by graph familiarity and different task features. *Language Testing*, *27*(1), 73–100.