Swansea University
Prifysgol Abertawe
School of Culture and Communication
Ysgol Diwylliant a Chyfathrebu

The new LLAMA tests (v.3): some initial thoughts on reliability and overlap with memory?

Vivienne Rogers, Paul Meara &
Brian Rogers
Swansea University

1

Background

Swansea University
Prifysgol Abertawe
School of Culture and Communication
Ysgol Diwylliant a Chyfathrebu

- LLAMA tests are widely used for language learning aptitude.
- Developed by Paul Meara (2005) as a tool for teaching research methods.
- Freely available and language neutral.

- Number of criticisms:
  - Reliability (Bokander & Bylund 2020 with v.1).
  - Aptitude is working memory (Wen, 2016).
  - Aptitude is Long Term memory (Buffington & Morgan-Short, 2019).

2

Test development

Swansea University
Prifysgol Abertawe
School of Culture and Communication
Ysgol Diwylliant a Chyfathrebu

- Since 2013/14, two parallel strands of development

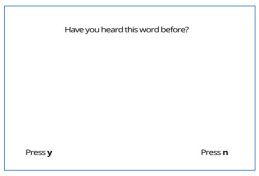
1. Creating a web-based, cross-platform version of the tests.
   a) 2016: LLAMA B (vocabulary test) made available online (v.2)
   b) 2018/19: other tests available online
   c) Beta version (ALPACAA) created (presented at EUROSLA 2019)
   d) 2019: version 3 online – major changes to various subcomponents (next slide)
2. Making the tests more reliable.
   a) Rogers, Meara et al (2016 & 2017) on factors that can influence the test scores.
   b) Bokander & Bylund (2020)
   c) Upcoming paper with Rogers, Bokander, Meara & Rogers
      a) (sneak preview: reliability is much better)

3

LLAMA tests (v.3): Rogers, Meara & Rogers (forthcoming)

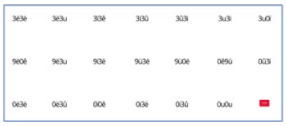
Swansea University
Prifysgol Abertawe
School of Culture and Communication
Ysgol Diwylliant a Chyfathrebu

LLAMA_D Sound Recognition

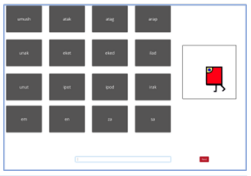


LLAMA_B Vocabulary Test

LLAMA_E Sound Symbol Correspondence



LLAMA_F Grammatical Inferencing



- Standardised input screen and instructions (English)
- LLAMA B: unchanged
- LLAMA D: no separate learning phase. All scoring items included.
- LLAMA E: test phase with 20 possibilities.
- LLAMA F: 10 items, each scored twice according to different target rules.

4

Research questions

Swansea University
Prifysgol Abertawe
School of Culture and Communication
Ysgol Diwylliant a Chyfathrebu

1. Are the LLAMA v.3 tests (more) reliable?
2. Do the LLAMA v.3 tests measure the same thing as common WM and LTM tests?

5

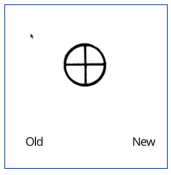
Methodology

Swansea University
Prifysgol Abertawe
School of Culture and Communication
Ysgol Diwylliant a Chyfathrebu

Participants

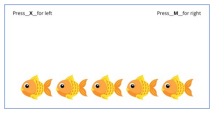
- n= 210
- F=145, M=56, NB = 9
- Mean age = 23,
  - (SD = 8.629, Mode= 20)
- L1 English = 47,
- L1 German = 110,
- bilingual L1 = 18,
- other L1/missing = 35

www.gorilla.sc



Tasks

- 4 LLAMA tests
- Auditory/ oral digits forwards and backwards
- Flanker



- Towers of Hanoi
  - procedural

- CVMT (Trahan & Larrabee, 1988, Morgan-Short et al., 2014)
  - declarative

6

Results RQ1

Are the LLAMA v.3 tests (more) reliable?

Swansea University Prifysgol Abertawe
School of Culture and Communication
Ysgol Diwylliant a Chyfathrebu

Cronbach's alpha scores.
NB: no change to LLAMA D – newer version on website with improved alpha scores

LLAMA Subtest	Original tests Bokander & Bylund (2020, table 4)	New LLAMA tests v.3 Rogers & Meara (2019)
LLAMA B	.81	.855
LLAMA E	.74	.902
LLAMA F	.60	.854
LLAMA D	.54	.562

7

Results RQ2

Do the LLAMA v.3 tests measure the same thing as common WM and LTM tests?

Swansea University Prifysgol Abertawe
School of Culture and Communication
Ysgol Diwylliant a Chyfathrebu

Descriptive Statistics for LLAMA tests

	LLAMA D - Total	LLAMA D –yes only	LLAMA B	LLAMA E	LLAMA F
Valid	135	135	210	210	210
Mean	29.644	15.133	9.652	8.505	8.905
Std. Deviation	3.920	2.788	4.777	5.655	5.060
Minimum	16.000	6.000	0.000	0.000	0.000
Maximum	39.000	20.000	20.000	20.000	20.000

8

Results RQ2

Do the LLAMA v.3 tests measure the same thing as common WM and LTM tests?

Swansea University Prifysgol Abertawe
School of Culture and Communication
Ysgol Diwylliant a Chyfathrebu

Descriptive Statistics for memory tests

	Flanker Difference incon-con	CVMT_Total	DF	DB	Hanoi 3-4	Hanoi 3-5
Valid	210	193	178	174	188	131
Missing	0	17	32	36	22	79
Mean	8580.845	77.793	6.466	5.787	420.945	536.746
Std. Deviation	119576.896	12.111	1.823	1.579	512.601	449.626
Minimum	-33470.980	3.000	0.000	2.000	-910.313	-488.376
Maximum	1.732e+6	99.000	10.000	9.000	3375.899	1910.654

9

Results RQ2

Do the LLAMA v.3 tests measure the same thing as common WM and LTM tests?

Swansea University Prifysgol Abertawe
School of Culture and Communication
Ysgol Diwylliant a Chyfathrebu

- All raw scores converted to z-scores
- Exploratory Factor Analysis (rotation: promax, Eigenvalues >1)
- Model:
- χ^2 (17, n=210) = 32.301, p=.014

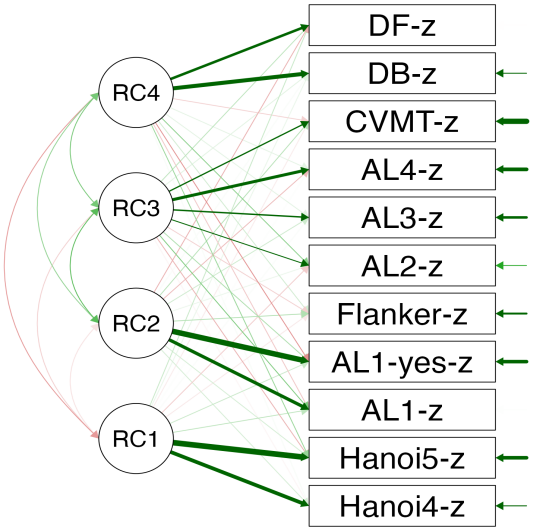
Factor Loadings

	Factor 1	Factor 2	Factor 3	Factor 4	Unique
Hanoi5-z	1.014				-0.001
Hanoi4-z	0.768				0.413
D-yes-z		0.991			0.004
D-z		0.712			0.421
F-z			0.678		0.544
E-z			0.479		0.745
CVMT-z			0.455		0.770
B-z			0.414		0.755
DB-z				0.807	0.344
DF-z				0.620	0.621
Flanker-z					0.972

Note. Applied rotation method is promax.

10

Swansea University
Prifysgol Abertawe
School of Culture and Communication
Ysgol Diwylliant a Chyfathrebu



11

Discussion

Swansea University
Prifysgol Abertawe
School of Culture and Communication
Ysgol Diwylliant a Chyfathrebu

1. Are the LLAMA v.3 tests (more) reliable?
   a) yes
2. Do the LLAMA v.3 tests measure the same thing as common WM and LTM tests?
   a) LLAMA scores split between D and others – implicit vs explicit? (Granena, 2013)
   b) LLAMA D and CVMT are same task but in different modalities – yet load on different factors.
   c) LLAMA B, E, F same factor as declarative memory
      a) Role of DM in L2 (Ullman 2015)
      b) Vocab and grammar in DC
   d) WM measures not on same factors as LLAMA tests (contra Wen, 2016)
      a) Same as in previous studies by Rogers et al presentations (2017, 2018, 2019)

12

Conclusion

Swansea University
Prifysgol Abertawe
School of Culture and Communication
Ysgol Diwylliant a Chyfathrebu

- New LLAMA tests are more reliable and test something different to WM measures
- Language learning aptitude is domain specific?
- Further research needed re: LTM.
- Please use version 3 not original download version.

- **Limitations**
  - Scoring of Hanoi
  - missing data: LLAMA D n=135, CVMT n=193, DB/DF n=174
  - Trustworthiness of participants recording themselves in DF/DB task

13

Conclusion

Swansea University
Prifysgol Abertawe
School of Culture and Communication
Ysgol Diwylliant a Chyfathrebu

- **Thanks for help with data collection/transcription**
  - Thomas Wagner, Dieter Thoma (and their students)
  - BA students: Richard Reed & Alex Torry
  - Martin Lee-Paterson & Tesni Galvin

- **Next steps**
  - Rogers, Bokander, Meara & Rogers (in prep) on LLAMA website data (n=640)

14

Swansea University
Prifysgol Abertawe
School of Culture and Communication
Ysgol Diwylliant a Chyfathrebu

Thank You!

v.e.rogers@swansea.ac.uk
https://viviennerogers.info
https://llamatests.org (coming soon)

15

16

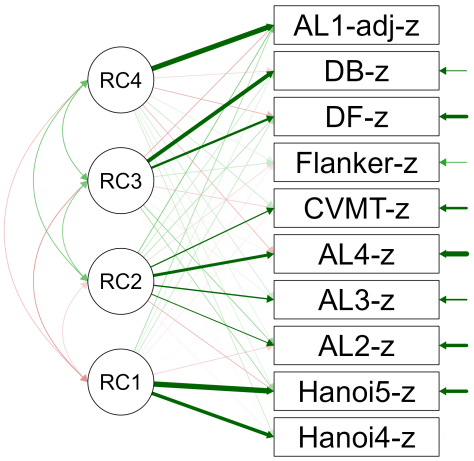
Swansea University
Prifysgol Abertawe

LLAMA D scoring options: 20 “old” and 20 “new

- Total /40 (/2)
- Yes for “old” only
- Yes for “old” +1, yes for “new” = -1

Descriptive Statistics

	D (all)	D-yes	D-adj
Mean	29.644	15.133	9.644
Std. Deviation	3.920	2.788	3.920
Minimum	16.000	6.000	-4.000
Maximum	39.000	20.000	19.000

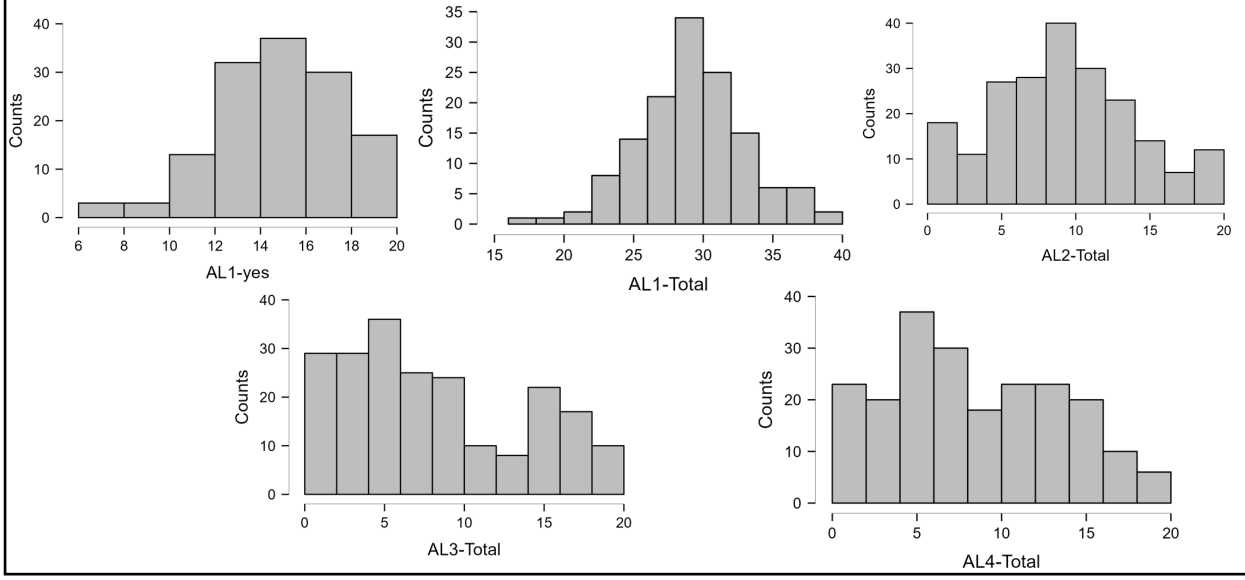


17

Results RQ2

Do the LLAMA v.3 tests measure the same thing as common WM and LTM tests?

Swansea University
Prifysgol Abertawe
School of Culture and Communication
Ysgol Diwylliant a Chyfathrebu



18

Variable		AL1-z	AL2-z	AL1-yes-z	AL3-z	AL4-z	DB-z	DF-z	CVMT-z	Flanker-z	Hanoi4-z	Hanoi
1. AL1-z	Pearson's r	—										
	p-value	—										
2. AL2-z	Pearson's r	0.182 *	—									
	p-value	0.035	—									
3. AL1-yes-z	Pearson's r	0.749 ***	0.116	—								
	p-value	< .001	0.180	—								
4. AL3-z	Pearson's r	0.216 *	0.240 ***	0.199 *	—							
	p-value	0.012	< .001	0.020	—							
5. AL4-z	Pearson's r	0.099	0.313 ***	0.084	0.315 ***	—						
	p-value	0.254	< .001	0.335	< .001	—						
6. DB-z	Pearson's r	0.063	0.227 **	0.045	0.108	0.117	—					
	p-value	0.509	0.003	0.638	0.156	0.123	—					
7. DF-z	Pearson's r	-0.006	0.169 *	-0.073	0.127	0.112	0.486 ***	—				
	p-value	0.953	0.024	0.439	0.091	0.136	< .001	—				
8. CVMT-z	Pearson's r	0.213 *	0.168 *	0.231 *	0.220 **	0.317 ***	0.038	0.048	—			
	p-value	0.019	0.020	0.011	0.002	< .001	0.617	0.528	—			
9. Flanker-z	Pearson's r	0.032	0.024	0.121	-0.039	-0.013	0.108	-0.016	-0.027	—		
	p-value	0.713	0.733	0.161	0.578	0.855	0.155	0.830	0.710	—		
10. Hanoi4-z	Pearson's r	0.106	-0.103	0.097	0.043	-0.044	-0.044	-0.064	0.026	-0.023	—	
	p-value	0.244	0.159	0.287	0.558	0.549	0.572	0.405	0.729	0.751	—	
11. Hanoi5-z	Pearson's r	0.108	-0.112	0.121	-0.074	-0.032	-0.030	0.078	-0.099	-0.028	0.760 ***	—
	p-value	0.285	0.202	0.234	0.399	0.717	0.754	0.402	0.265	0.754	< .001	—

19

Model Summary - aptitude_overall-z

Model	R	R^2	Adjusted R^2	RMSE
1	0.000	0.000	0.000	1.045
2	0.317	0.100	0.089	0.998

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
2	Regression	8.877	1	8.877	8.920	0.004
	Residual	79.614	80	0.995		
	Total	88.491	81			

Note. The intercept model is omitted, as no meaningful information can be shown.

20

Swansea University
Prifysgol Abertawe
School of Culture and Communication
Ysgol Diwylliant a Chyfathrebu

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	0.038	0.115		0.331	0.741
2	(Intercept)	8.742 e-4	0.111		0.008	0.994
	CVMT-z	0.357	0.119	0.317	2.987	0.004

Note. The following covariates were considered but not included: age, DB-z, Flanker-z, Hanoi4-z, Hanoi5-z, DF-z.

21

Rogers, Bokander, Meara & Rogers (in prep)

Swansea University
Prifysgol Abertawe
School of Culture and Communication
Ysgol Diwylliant a Chyfathrebu

Subtest	Cronbach's alpha	Lower CI	Upper CI
LLAMA D (all)	0.702	0.668	0.734
LLAMA D (yes)	0.875	0.860	0.888
LLAMA B	0.897	0.844	0.908
LLAMA E	0.903	0.892	0.913
LLAMA F	0.864	0.849	0.879

22