



Swansea University  
Prifysgol Abertawe

# Turning a LLAMA into an ALPACAA and back again: An initial revised attempt at assessing aptitude

Vivienne Rogers & Paul Meara



## Outline

- Background on LLAMA tests
- Methodology: how we've revised the tests
- Results & Discussion
- And back to LLAMA



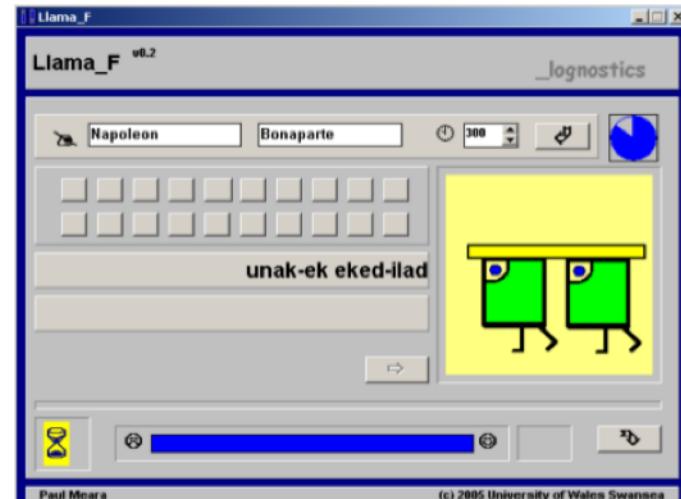
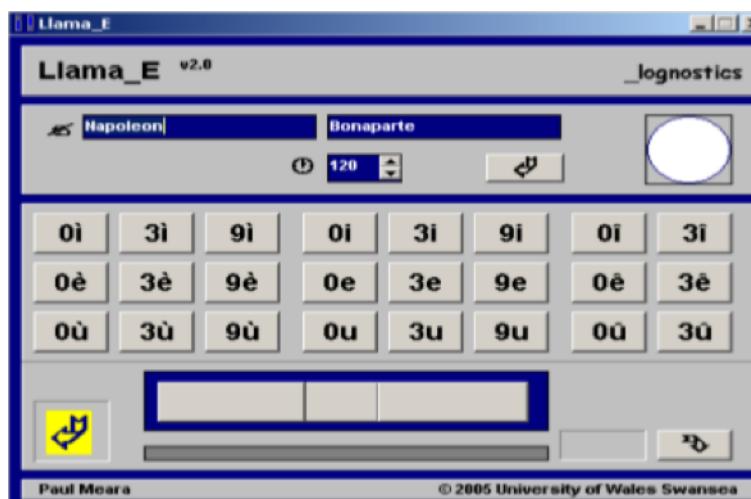
# Swansea LLAMA tests (Meara, 2005)

[www.lognistics.co.uk/tools/llama](http://www.lognistics.co.uk/tools/llama)



Swansea University  
Prifysgol Abertawe

- Free, loosely based on MLAT
- LLAMA B = vocabulary measure
- LLAMA D = sound recognition (implicit learning)
- LLAMA E = sound-symbol correspondence
- LLAMA F = grammatical inferencing
- Has not been fully validated.





## Previous validation work: Grañena

- Grañena (2013):
- Internal consistency, Gender and Language neutrality
- n=187 aged 18-39
- L1s: Spanish, Chinese and English
- internal consistency but two forms of aptitude
- LLAMA D measures implicit and others explicit?
- Grañena (2018):
- Compared 4 LLAMA tests with 4 Hi-LAB (n=135)
- Found 3 underlying constructs across the tests.
- Only the factor with LLAMA D and ALTM Synonym
- (Hi-LAB) significantly predicted L2 fluency (pruned speech rate per min).

Rogers, V., Meara, P., Barnett-Legh, T., Curry, C., & Davie, E. (2017). Examining the LLAMA aptitude tests.. *Journal of the European Second Language Association*, 1(1), 49–60.  
DOI: <http://doi.org/10.22599/jesla.24>



Swansea University  
Prifysgol Abertawe

- **How much of the LLAMA test score variance do the individual factors measures account for?**
- Factors included age, L1, L2 status, education level, gender, playing of logic puzzles.
- 404 participants in total.
- 346 took all 4 parts of the LLAMA tests and background questionnaires.
- Multiple regression analysis for 6 factors.  
Overall variance for:
  - LLAMA B:  $R^2 = 9.1\%$
  - LLAMA D:  $R^2 = 4.8\%$
  - LLAMA E:  $R^2 = 3.4\%$
  - LLAMA F:  $R^2 = 6.6\%$
- Only L2 status consistently was significant  $p < .05$  (not for E).
  - LLAMA B:  $\beta = -.250$ , contribution to variance = 6.0
  - LLAMA D:  $\beta = .136$ , contribution to variance = 1.8
  - LLAMA F:  $\beta = -.165$ , contribution to variance = 2.6



## Previous validation work: Bokander & Bylund (2019)

- Scoring
  - LLAMA B performed well.
  - Others did not (particularly D)
- Generalization
  - Internal consistency
  - LLAMA B & E met .70 criterion
  - LLAMA E: analytic/ strategic use of vowels only rather than sound/symbol
- Explanation
  - Construct & content validity
  - Possibly doesn't reflect Skehan's (1998) three components of aptitude
  - Two component: LLAMA D is different to the others.





Swansea University  
Prifysgol Abertawe

# Purpose/ Research questions

# Purpose/ Research questions

- This study has three purposes:
- to remedy some of the test flaws.
- to revise the scoring method of the LLAMA test
- to examine if the revised tests overlap with working memory measures



Swansea University  
Prifysgol Abertawe

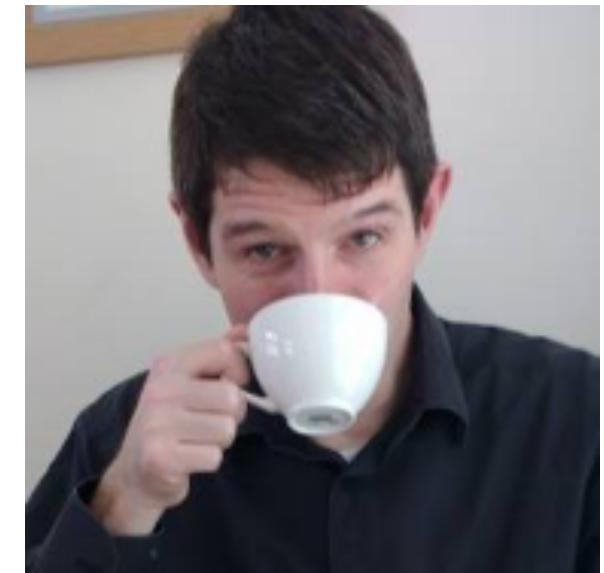
## Research Questions:

- What is the impact of different scoring mechanisms on the distribution of ALPACAA scores?
- Do all the items discriminate between participants?
- What is the relationship between the new scoring method and WM?



# Methodology

- Re-programmed the LLAMA tests into OpenSesame – called ALPACAA
- Changed order of administration:
  - D then B, E, F
  - Kept: 2 mins learning B & E, 5 mins learning F
- Fixed errors in original.
- No feedback to participant during test.
- End: given average RT and total correct.
- Clearer instructions (English)
- Can start test early



# Participants



- Administered to 123 participants
- Age 17-55, ( $M=23.5$ ,  $S.D.=5.576$ )
- Male = 56, Female = 67
- L1 English speakers = 77
  - (63 with L2, 14 L1 English only)
- Bilingual L1 English speakers = 7
- L2 English speakers = 39
- Also administered Stroop, Flanker and auditory Digits backwards.
- Collected by BA dissertation students (L-R, Dafydd, Megan, Amy)



Swansea University  
Prifysgol Abertawe



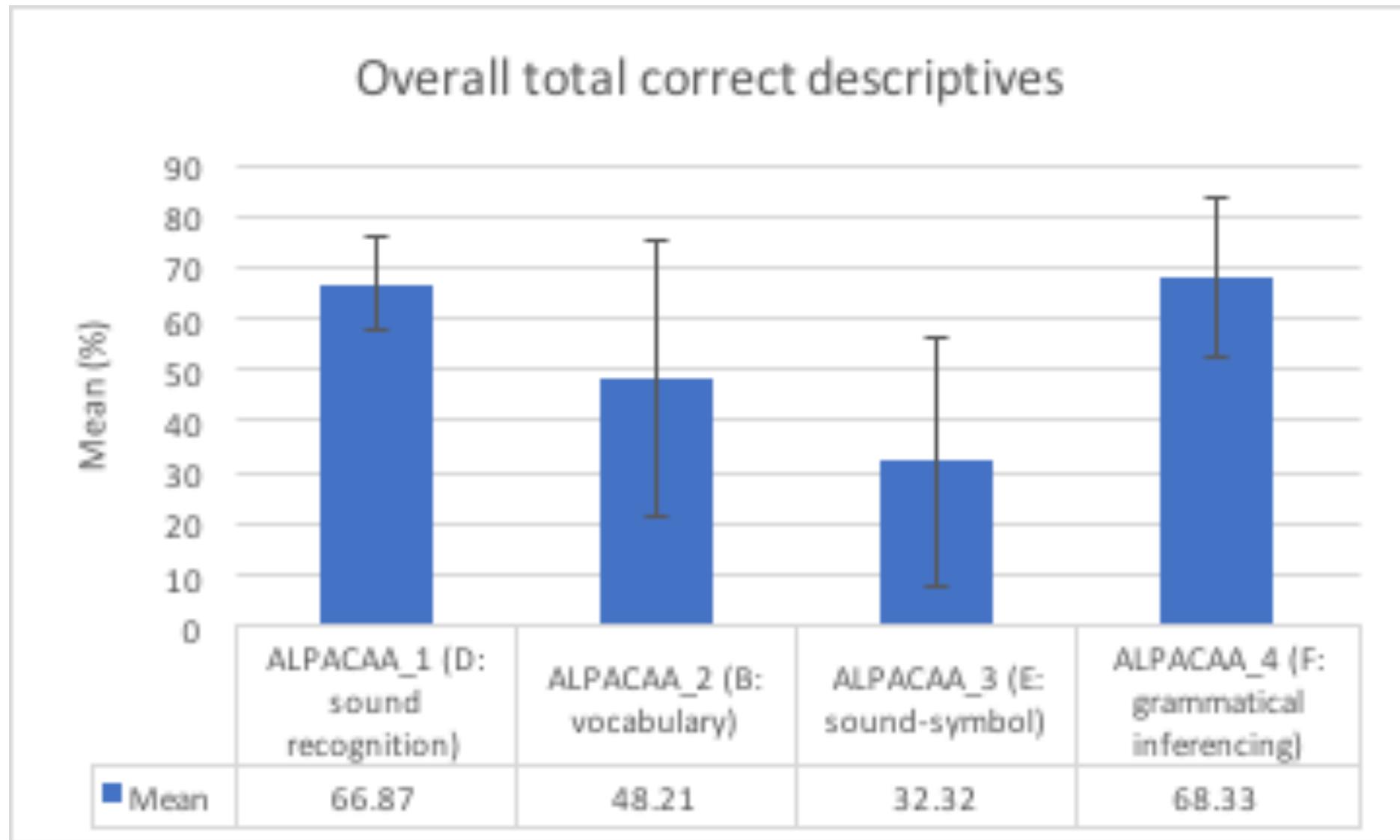
RQ1:

What is the impact of different scoring mechanisms  
on the distribution of ALPACAA scores?

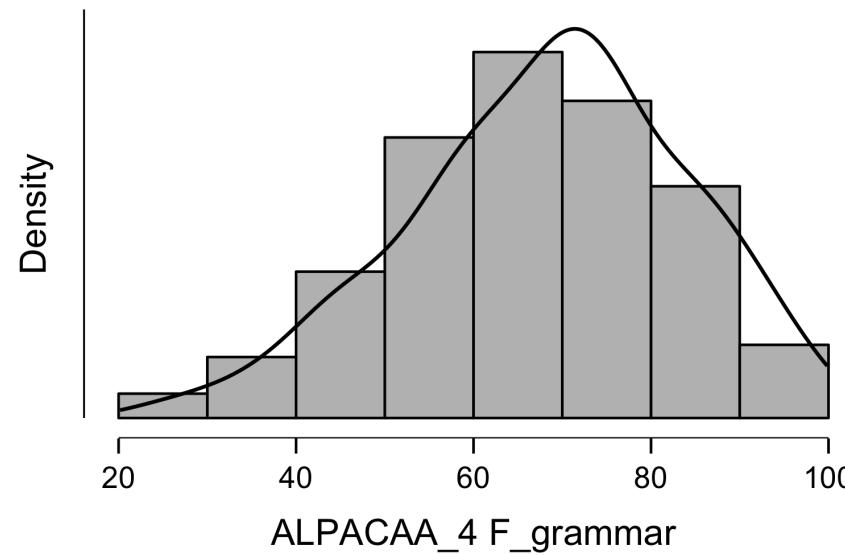
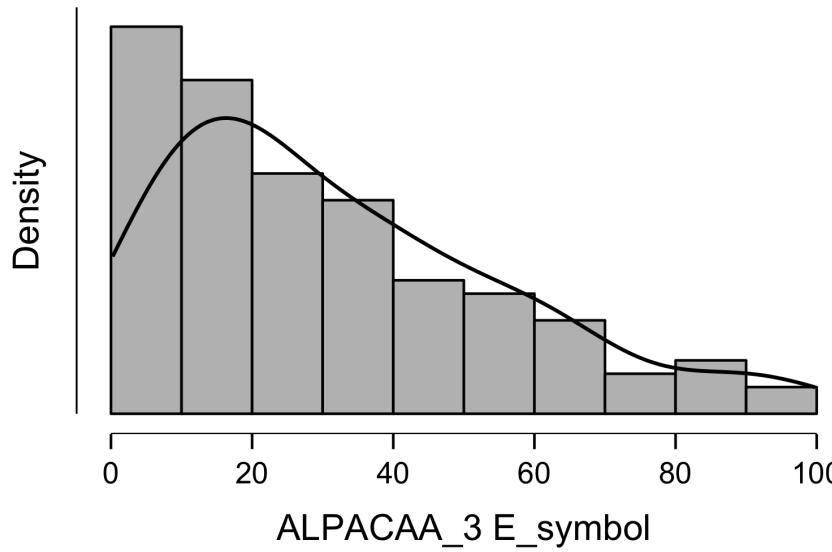
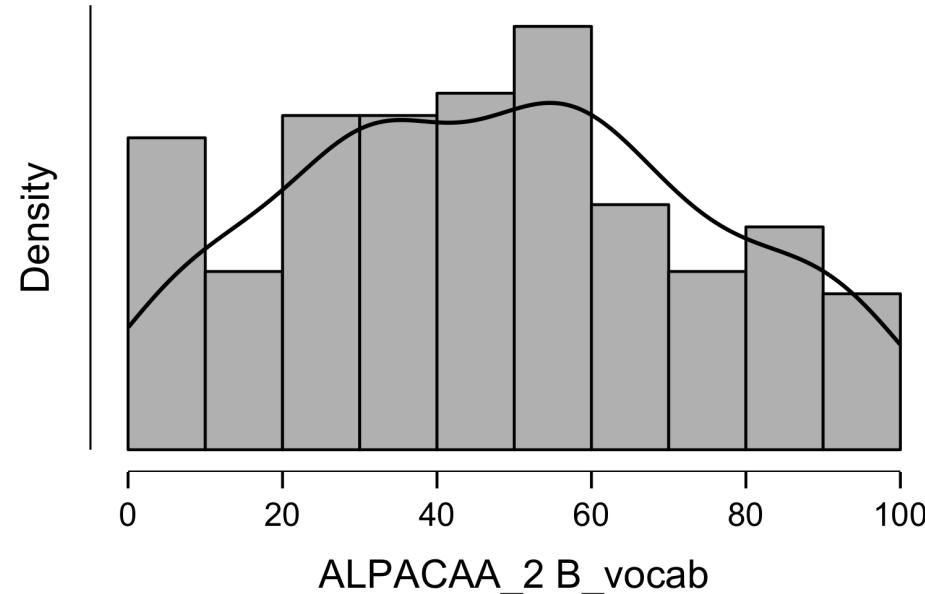
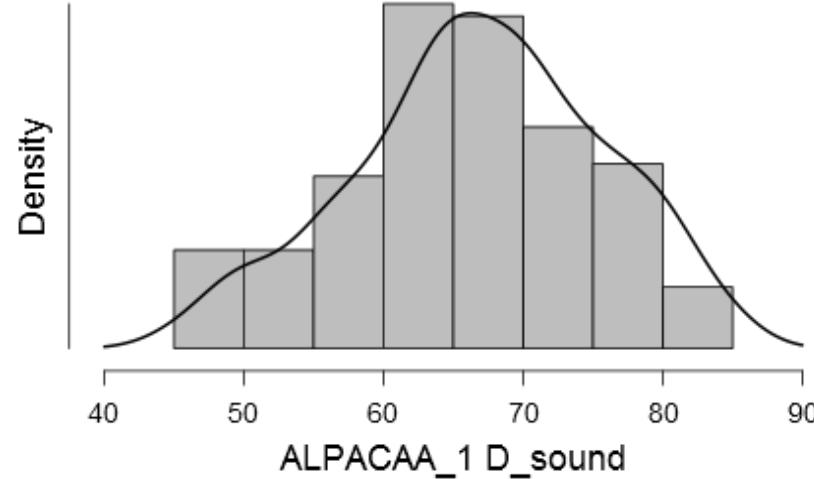
# Overall descriptive: total correct – n=123



Swansea University  
Prifysgol Abertawe



- Distribution of scores in tests (total correct)



Swansea University  
Prifysgol Abertawe

# Adjusting for guessing



Swansea University  
Prifysgol Abertawe

- Step 1
- Adjusted for not doing learning phase (20 items)
- Criteria – must click on each item at least once.
- ALPACAA\_2 (vocab): 6 removed n=117
- ALPACAA\_3 (sound/symbol): 3 removed, n=120
- ALPACAA\_4 (grammatical inferencing): 3 removed, n=120

- Step 2: Applied LLAMA penalties
- LLAMA D, E, F – lose 1 mark (5%) for incorrect answer (binary choice)
- ALPACAA\_1 (D)
  - M=33.74, S.D=17.86
  - Mean was 68.67
  - Range: -10 – 70
- ALPACAA\_4 (F)
  - M=36.50, S.D=31.35
  - Mean was 68.33
  - Range: -50 - 100



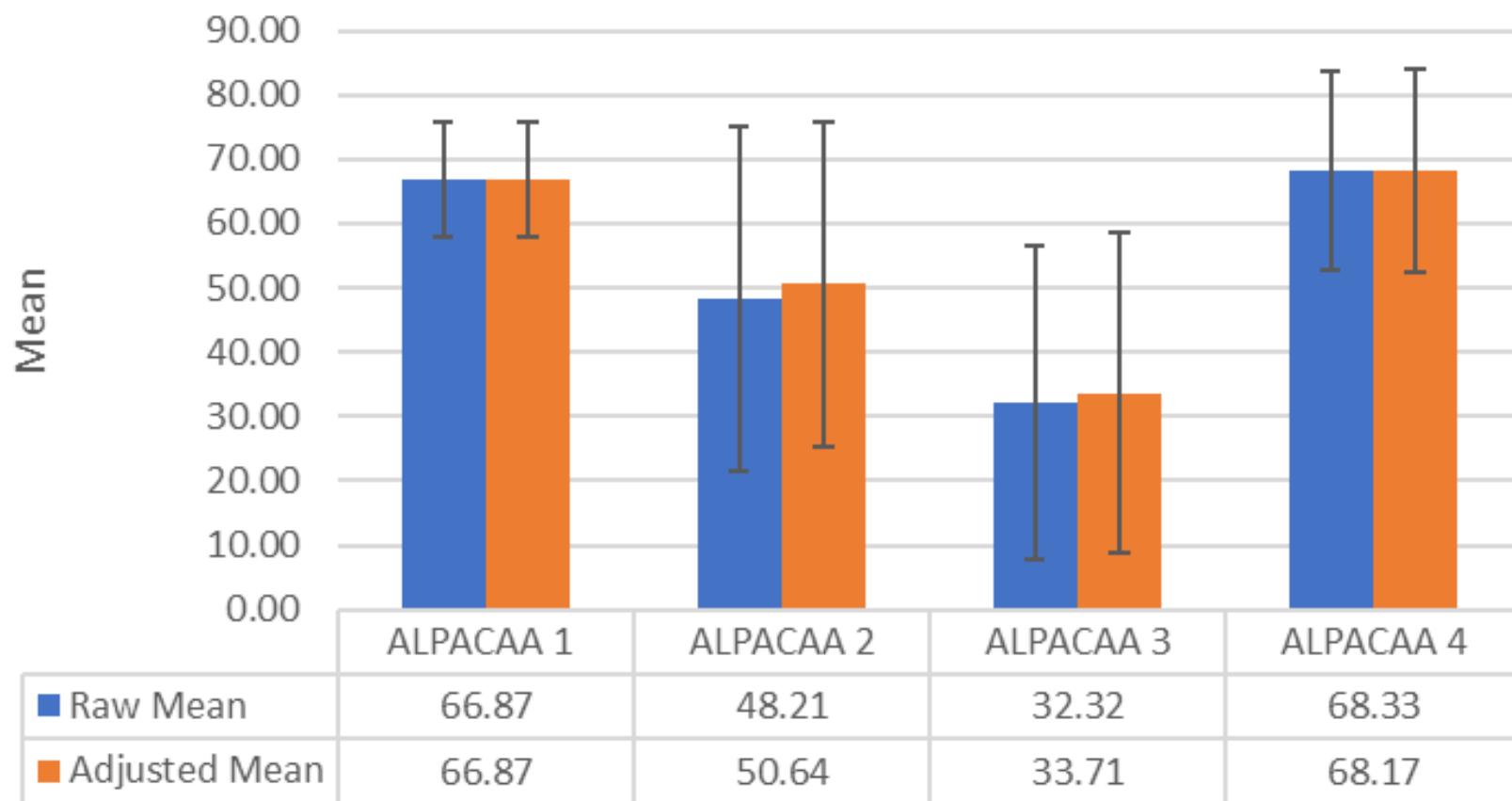
# What about adjusting for guessing? Using RTs

- Have RTs for all test items for all participants..
- Excluded any RTs faster than 200ms.
- ALPACAA\_1\_D: In test phase, heard sound then question prompt then click.
  - More than 200ms after sound so no exclusions.
- ALPACAA\_2\_B: Three items identified (out of  $117 \times 20 = 2340$ )
  - Two were correct: removed.
- ALPACAA\_3\_E: No items
- ALPACAA\_4\_F: Four items identified (out of  $120 \times 20 = 2400$ )
  - Two were correct: removed.

As they have to navigate with mouse then 200ms not an appropriate cut off?



## Total correct and adjusted scores





# Spearman's correlations: average RT and total correct (adjusted for learning phase)

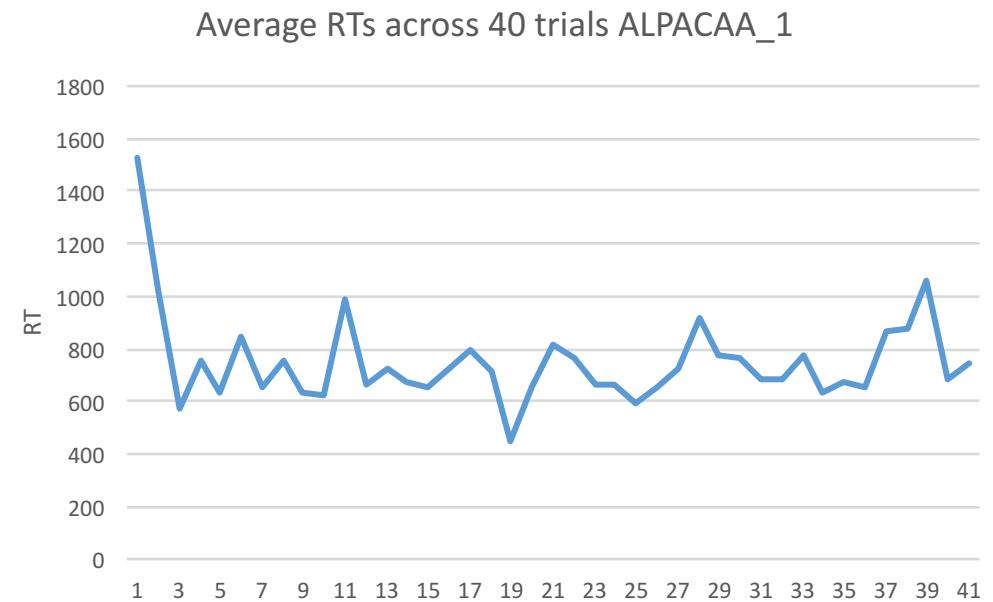
	<i>rho</i>	<i>p</i>	Lower 95% CI	Upper 95% CI
ALPACAA 1 (sound recognition)	-0.009	0.918	-0.186	0.168
ALPACAA 2 (vocabulary)	-0.078	0.406	-0.255	0.106
ALPACAA 3 (sound-symbol)	0.550***	<.001	0.664	0.411
ALPACAA 4 (grammatical inferencing)	0.401***	<0.001	0.239	0.541

# Discussion



Prifysgol Abertawe  
Swansea University

- Participants use the learning phase (12/369)
- Few react within 200ms (7/7140)
- Penalising doesn't change distribution but lowers mean (a lot).
- Lots more could be done with RT data.
- Very slow on first question but then flatten.
- Exclude items based on +/- 2 S.D.'s.





RQ2:

Do all the items discriminate between participants?



# Internal reliability (Cronbach's alpha)

	n	Cronbach's $\alpha$	Average inter item correlation	95% CI Lower	95% CI Higher
ALPACAA_1 (sound recognition) all	123	0.385	0.017	0.329	0.438
ALPACAA_1 (sound recognition) correct	123	0.544	0.502	0.502	0.584
ALPACAA_2 (vocabulary)	117	0.850	0.227	0.836	0.863
ALPACAA_3 (sound/symbol)	120	0.883	0.272	0.872	0.893
ALPACAA_4 (grammatical inferencing)	120	0.617	0.079	0.581	0.650

# ALPACAA\_1 Item Reliability Statistics (all items)



Swansea University  
Prifysgol Abertawe

	mean	sd	item-rest correlation	If item dropped Cronbach's $\alpha$		mean	sd	item-rest correlation	If item dropped Cronbach's $\alpha$
latd11-n	0.463	0.501	-0.137	0.417	latd04-y2	0.756	0.431	0.124	0.371
latd12-n	0.545	0.500	-0.076	0.406	latd06-y2	0.732	0.445	0.195	0.359
latd03-y1	0.894	0.309	0.183	0.368	latd07-y2	0.748	0.436	0.129	0.370
latd13-n	0.236	0.426	-0.048	0.398	latd23-n	0.829	0.378	0.189	0.363
latd08-y1	0.301	0.460	0.161	0.365	latd08-y2	0.382	0.488	0.042	0.385
latd14-n	0.683	0.467	0.160	0.365	latd10-y2	0.659	0.476	0.058	0.382
latd15-n	0.642	0.481	0.003	0.392	latd24-n	0.667	0.473	0.177	0.361
latd05-y1	0.813	0.391	0.015	0.388	latd25-n	0.740	0.441	0.206	0.358
latd04-y1	0.691	0.464	0.243	0.350	latd26-n	0.699	0.460	0.163	0.364
latd06-y1	0.780	0.416	0.040	0.385	latd03-y2	0.764	0.426	0.057	0.382
latd16v-n	0.740	0.441	0.092	0.377	latd27-n	0.780	0.416	-0.010	0.392
latd09-y1	0.585	0.495	-0.168	0.422	latd05-y2	0.675	0.470	0.099	0.375
latd17-n	0.740	0.441	-0.030	0.396	latd02-y2	0.561	0.498	0.128	0.370
latd10-y1	0.602	0.492	0.048	0.384	latd01-y2	0.846	0.363	0.222	0.360
latd07-y1	0.732	0.445	0.146	0.368	latd28-n	0.675	0.470	0.054	0.383
latd18-n	0.496	0.502	0.012	0.391	latd09-y2	0.618	0.488	-0.033	0.398
latd19-n	0.732	0.445	0.082	0.378	latd29-n	0.748	0.436	0.179	0.362
latd20-n	0.366	0.484	-0.006	0.393	latd30-n	0.764	0.426	0.096	0.376
latd01-y1	0.951	0.216	0.115	0.378					
latd02-y1	0.553	0.499	0.160	0.364					
latd21-n	0.821	0.385	0.051	0.383					
latd22-n	0.740	0.441	0.157	0.366					



# ALPACAA\_1 Inter-Item Reliability (all items)

	n	Cronbach's $\alpha$	Average inter item correlation	95% CI Lower	95% CI Higher
ALPACAA_1 (sound recognition) all	123	0.385	0.017	0.329	0.438
ALPACAA_1 (sound recognition) revised	123	0.535	0.036	0.492	0.575

# ALPACAA\_1 Item Reliability Statistics (yes only)



Swansea University  
Prifysgol Abertawe

	mean	sd	item-rest correlation	If item dropped Cronbach's $\alpha$
latd03-y1	0.894	0.309	0.180	0.532
latd08-y1	0.301	0.460	0.106	0.543
latd05-y1	0.813	0.391	0.091	0.544
latd04-y1	0.691	0.464	0.258	0.517
latd06-y1	0.780	0.416	-0.129	0.576
latd09-y1	0.585	0.495	-0.007	0.564
latd10-y1	0.602	0.492	0.217	0.524
latd07-y1	0.732	0.445	0.162	0.533
latd01-y1	0.951	0.216	0.124	0.539
latd02-y1	0.553	0.499	0.284	0.511
latd04-y2	0.756	0.431	0.185	0.530
latd06-y2	0.732	0.445	0.169	0.532
latd07-y2	0.748	0.436	0.300	0.510
latd08-y2	0.382	0.488	0.103	0.544
latd10-y2	0.659	0.476	0.219	0.523
latd03-y2	0.764	0.426	0.216	0.525
latd05-y2	0.675	0.470	0.178	0.531
latd02-y2	0.561	0.498	0.268	0.514
latd01-y2	0.846	0.363	0.336	0.509
latd09-y2	0.618	0.488	0.252	0.517



# Inter-item reliability (Cronbach's alpha)

	n	Cronbach's $\alpha$	Average inter item correlation	95% CI Lower	95% CI Higher
ALPACAA_1 (sound recognition) all	123	0.385	0.017	0.329	0.438
ALPACAA_1 (sound recognition) revised	123	0.535	0.036	0.492	0.575
ALPACAA_1 (sound recognition) correct	123	0.544	0.502	0.502	0.584
ALPACAA_1 (sound recognition) correct revised	123	0.593	0.075	0.555	0.629

# ALPACAA\_4 Item Reliability Statistics



Swansea University  
Prifysgol Abertawe

	mean	sd	item-rest correlation	If item dropped Cronbach's α
eket-arap-sa	0.825	0.382	0.173	0.609
ipod-ilad-za	0.850	0.359	0.247	0.601
eket-arap	0.733	0.444	0.316	0.591
atak-arap-sa	0.767	0.425	0.299	0.594
ipot-arap	0.592	0.494	0.095	0.621
atag-ilad	0.583	0.495	0.349	0.584
unak atak-arap-sa	0.875	0.332	0.321	0.595
umush-ek ipot-arap	0.783	0.414	0.306	0.593
unak-ek ipot-arap	0.642	0.482	0.255	0.599
inut-ek eket-arap	0.708	0.456	0.222	0.603
unak-em eked-ilad	0.592	0.494	0.245	0.600
umush-em ipod-ilad	0.675	0.470	0.253	0.599
unak ipot-arap-sa	0.692	0.464	0.253	0.599
umush ipot-arap-sa	0.633	0.484	0.137	0.615
ipod-orad-za	0.817	0.389	0.220	0.604
atag-orad-za	0.508	0.502	-0.003	0.635
eked-orad-za	0.650	0.479	0.315	0.590
umush-ek atag-orad	0.658	0.476	0.334	0.587
unak-em atag-orad	0.650	0.479	0.291	0.593
ipod-orad	0.400	0.492	-0.219	0.662



# Internal reliability (Cronbach's alpha)

	n	Cronbach's $\alpha$	Average inter item correlation	95% CI Lower	95% CI Higher
ALPACAA_4 (grammatical inferencing)	120	0.617	0.079	0.581	0.650
ALPACAA_4 (grammatical inferencing) revised	120	0.682	0.108	0.652	0.710



# Discussion

- ALPACAA\_2 & 3 (vocab and sound/symbol) discriminate well.
  - Participants chose from 20 pictures.
- ALPACAA\_1 & 4 (sound recognition and grammatical inferencing) do not discriminate well.
  - Participants given binary choice.
- Need more participants.
- More detailed analysis of items.
- Follow Bokander & Bylund (2019)

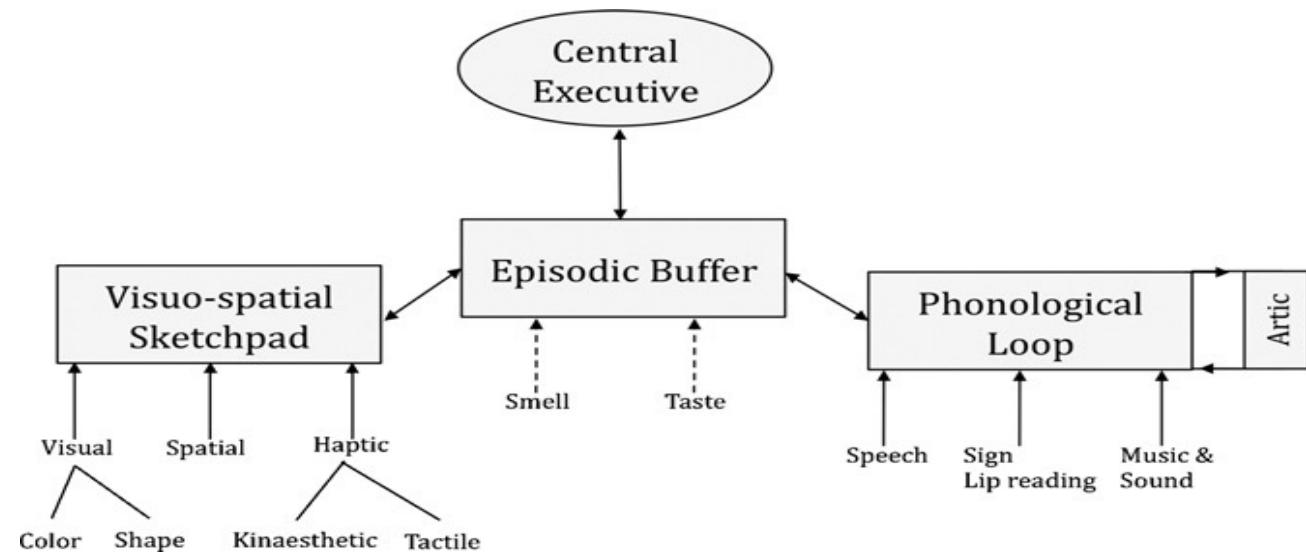


RQ3: What is the relationship  
between the new scoring method  
and WM



# What is working memory?

“Working memory refers to the system or systems that are assumed to be necessary in order to keep things in mind while performing complex tasks such as reasoning, comprehension and learning.”  
Baddeley (2010, p. 136)



STM: maintenance of information

WM: maintenance and manipulation



# Previous work with LLAMA

## (presented at EUROS LA 2017)

- Data collected by BA dissertation students:
  - Tesni Galvin, Amelia Cobner, Martha Chisholm, Jake Clothier & Issy Greenfield
- 127 participants
  - predominantly students
- Typically L1 English speakers

**Table I – Participant Data**

No. Females	60
No. Males	67
Age Range	16-78
Average Age	33.5



# Results: PCA

- No LLAMA test loads on the same factor as any of the working memory and attention tests.

	Pattern Matrix <sup>a</sup>	
	Component 1	2
LLAMA E	.807	
LLAMA F	.799	
LLAMA B	.670	
LLAMA D	.546	
WM3 (A)		.906
WM3 (B)		.877
WM1 (Visual)		-.498
WM2 (Digits)		-.392

Extraction Method: Principal Component Analysis.

Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 6 iterations.

# Results: PCA part 2

- Even if forced to four factors, LLAMA tests load differently to the WM/attention tests.
- LLAMA B, E & F measure something different to LLAMA D (similar to Grañena 2013).
- TMT parts A & B measure different aspect of WM to the digits backwards (PSTM) and visuo-spatial/ storage measures.

	Pattern Matrix <sup>a</sup>			
	Component			
	1	2	3	4
LLAMA F	.831			
LLAMA E	.828			
LLAMA B	.672			
WM3 (A)		.914		
WM3 (B)		.867		
WM2 (Digits)			.897	
WM1 (Visual)			.586	
LLAMA D				.947

Extraction Method: Principal Component Analysis.  
Rotation Method: Oblimin with Kaiser Normalization.  
a. Rotation converged in 6 iterations.



# Working memory results (n=123)

	flanker conflict cost	stroop conflict cost	DB_span
Mean	45.440	127.297	5.537
Std. Deviation	20.379	114.391	1.317
Minimum	0.025	-43.880	3.000
Maximum	110.550	1123.434	9.000

Only significant correlation (Spearman's) between  
Stroop and Digits Backwards ( $r = -0.252, p = .005$ )

## Spearman Correlations

		flanker cost	stroop cost	DB_span	A1	A2	A3	A4
flanker conflict cost	Spearman's rho	—						
conflict cost	p-value	—						
stroop conflict cost	Spearman's rho	0.065	—					
conflict cost	p-value	0.478	—					
DB_span	Spearman's rho	0.049	-0.252**	—				
	p-value	0.590	0.005	—				
A1_total_corr ect	Spearman's rho	0.006	-0.046	0.073	—			
	p-value	0.948	0.616	0.428	—			
A2_total_corr ect	Spearman's rho	-0.012	-0.186*	0.432***	0.200*	—		
	p-value	0.901	0.045	1.161e -6	0.031	—		
A3_total_corr ect	Spearman's rho	0.019	-0.178	0.252**	0.178	0.467***	—	
	p-value	0.834	0.052	0.005	0.052	1.253e -7	—	
A4_total_corr ect	Spearman's rho	-0.021	-0.193	0.200*	0.191*	0.523***	0.455***	—
	p-value	0.821	0.036	0.029	0.037	1.988e -9	2.254e -7	—

\* p < .05, \*\* p < .01, \*\*\* p < .001

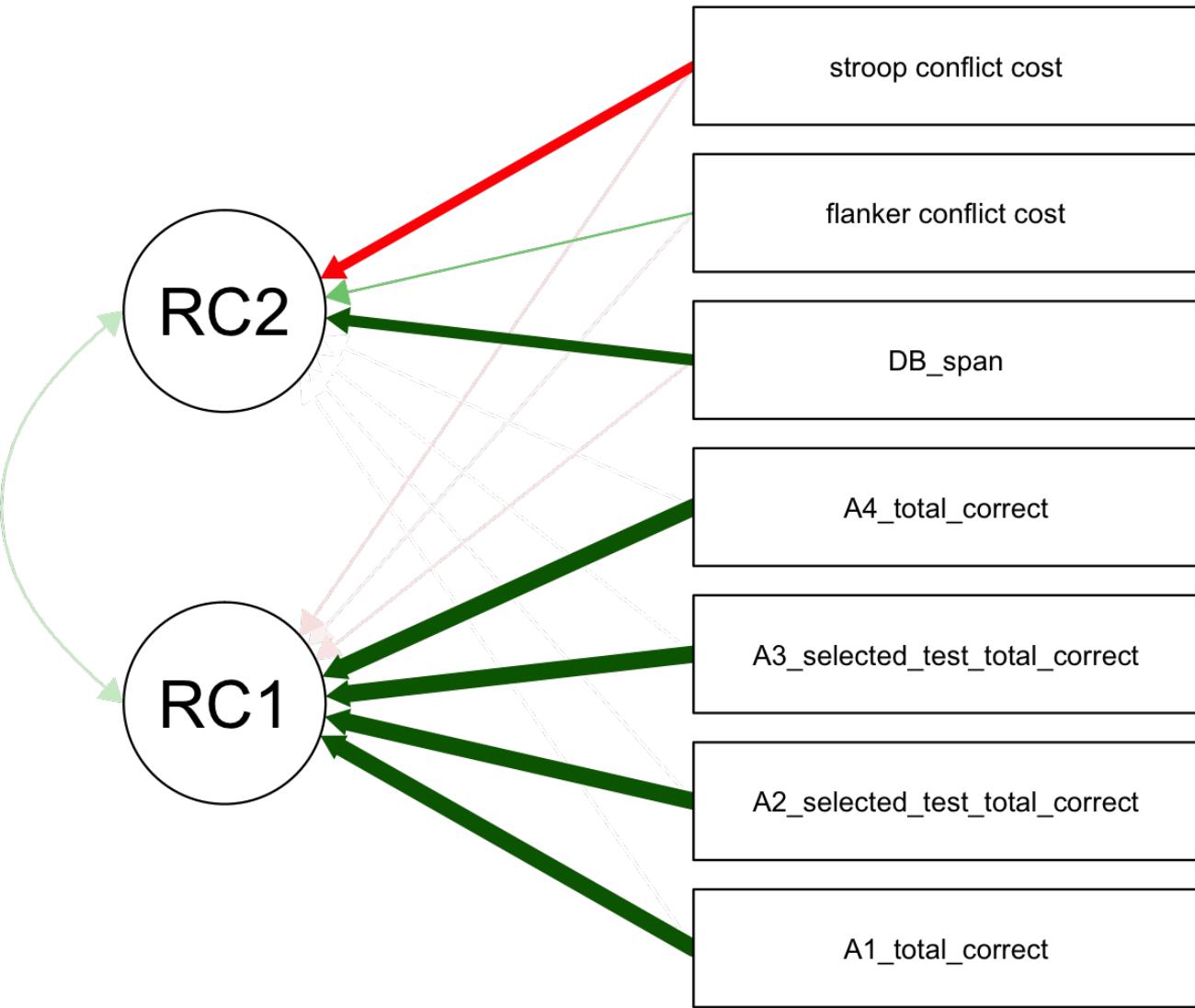


# PCA analysis: WM and ALPACAA component total correct scores

Component Loadings

	RC 1	RC 2	Uniqueness
A1_total_correct	1.000	.	1.821e -4
A2_selected_test_total_correct	1.000	.	1.895e -4
A3_selected_test_total_correct	1.000	.	2.005e -4
A4_total_correct	1.000	.	1.792e -4
DB_span	.	0.765	0.417
flanker conflict cost	.	.	0.958
stroop conflict cost	.	-0.714	0.484

Note. Applied rotation method is promax.





# Discussion

- WM tests and ALPACAA aptitude tests (total correct) are measuring different things.
- WM may be part of aptitude but doesn't replace it
  - (cf Wen, 2016)
- Comparable to previous findings on LLAMA and WM.
  - Different WM tests (Corsi block, TMT A&B & Digits backwards)
- Didn't find difference with sound recognition and other tests.
  - Scores to 100 and no penalties?



# Overall conclusion

- ALPACAA are an (initial) attempt to refine the LLAMA tests.
- Further work needed on:
  - ALPACAA\_4 (grammatical inferencing) and ALPACAA\_1 (sound recognition) in terms of reliability.
  - ALPACAA\_3 (sound/symbol) in terms of negative skew.
  - Are layout revisions enough?
- More detailed analysis of RT and items needed.
- More detailed analysis of predictor variables.

# Back to LLAMA

- Updated test versions on Paul Meara's website ([www.lognistics.co.uk](http://www.lognistics.co.uk))
  - Web-based, work on different browsers.
  - LLAMA B (vocabulary) = same
  - LLAMA D (sound recognition) = now out of 100
  - LLAMA E (sound/symbol) = revised test layout
  - LLAMA F (grammatical inferencing) = due Dec 2019
- Download versions will not be supported after Jan 2020



Thank you!

Diolch yn fawr!

謝謝

Vivienne Rogers: [v.e.rogers@swansea.ac.uk](mailto:v.e.rogers@swansea.ac.uk)

twitter: @RogersVivienne

Paul Meara: [p.m.meara@gmail.com](mailto:p.m.meara@gmail.com)

