Further Investigation of the LLAMA aptitude tests

Vivienne Rogers

University of Essex
28th July 2017

Swansea University
Prifysgol Abertawe

Outline

Carroll's definition of aptitude

> *the amount of time a student needs to learn a given task, unit of instruction, or curriculum to an acceptable criterion of mastery under optimal conditions of instruction and student motivation. (Carroll 1990 p. 26)*

Components of aptitude:

- phonemic coding ability: capacity to retain unfamiliar sounds
- grammatical sensitivity: ability to identify functions of words in a sentence
- inductive language learning ability: talent to find generalisations based on input
- associative learning ability: make links between L1 and L2 words.

MLAT: Modern Languages Aptitude Test

Carroll & Sapon (1959)

- Most widely used aptitude test in the USA.
- Predictive test for learning rate in instructed learners.
- Three components:
  - grammatical sensitivity
    - Words in sentence
  - phonetic coding ability
    - number learning (aural)
    - phonetic script (aural)
    - spelling cues
  - memory capacity
    - Paired associates.

Criticisms of Carroll's approach

- Main teaching approach at the time = audiolingualism
- Krashen's (1981) acquisition vs learning - aptitude not relevant.
- Skehan (2002)
  - Outdated - particularly in terms of memory capacity
- Robinson (2005)
  - not so interested in rate of learning any more.
  - more interested in ultimate attainment.
  - relevance of aptitude in various conditions.

Other work on aptitude

- L1 (2015): meta-analysis
- aptitude = independent of other individual differences. (contra Pimsleur 1966)
- strong predictor of general proficiency but not vocabulary learning or L2 writing.
- YET different test sub-components predicted different aspects of learning.
- Supports multi-component approach to aptitude.
- more associated with Executive WM then PSTM.
- PSTM may be more relevant at earlier stages (Linck et al 2013)

LLAMA tests: background

- Developed by Paul Meara (2005)
- free, loosely based on MLAT
- increasingly used in research projects (over 1000 papers reference the LLAMA tests).
- Has not been fully validated.

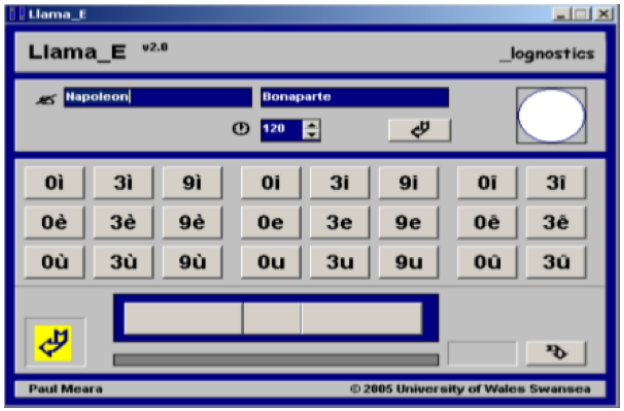
LLAMA tests

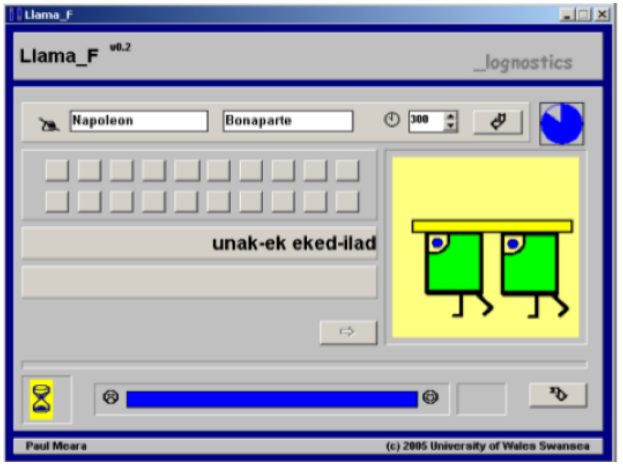
- Not only designed for English - language neutral
- Four components
- LLAMA B = vocabulary measure
- LLAMA D = sound recognition (implicit learning)
- LLAMA E = sound-symbol correspondence
- LLAMA F = grammatical inferencing

LLAMA tests


Llama_B v1.0
_lognostics
Napoleon
Bonaparte
Paul Meara
©2005 University of Wales Swansea


Llama_D
Llama_D v1.0
_lognostics
Napoleon
Bonaparte
Paul Meara
(c) 2005 University of Wales Swansea


Llama_E
Llama_E v2.0
_lognostics
Napoleon
Bonaparte
Paul Meara
© 2005 University of Wales Swansea


Llama_F
Llama_F v0.2
_lognostics
Napoleon
Bonaparte
unak-ek eked-ilad
Paul Meara
(c) 2005 University of Wales Swansea

Previous validation work: Grañena (2013)

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 measuring something different to the others
- LLAMA D measures implicit and others explicit?

- More and more researchers are using the LLAMA tests.
- Difficulty in obtaining other tests.
- Paul Meara is concerned as they have not been validated.
- Initial cross-sectional validation attempts.

Individual differences in LLAMA scores

- With 2013-14 BA dissertation students.
- Rachel Aspinall, Louise Fallon, Tom Goss, Emily Keey, Rosa Thomas.
- Published in EUROSLA Yearbook 2016
- Looked at a range of factors that might influence test performance, including age, L1, L2 status, education level, gender, playing of logic puzzles and timings of the test.

Louise, Rosa, Emily, Tom & Rachel

Research Questions

1. What is the role of gender?
2. Are the LLAMA tests language neutral?
3. What is the role of age?
4. What is the role of formal education qualifications?
5. Does playing logic puzzles affect LLAMA scores?
6. What difference would changing the test timings make to scores?

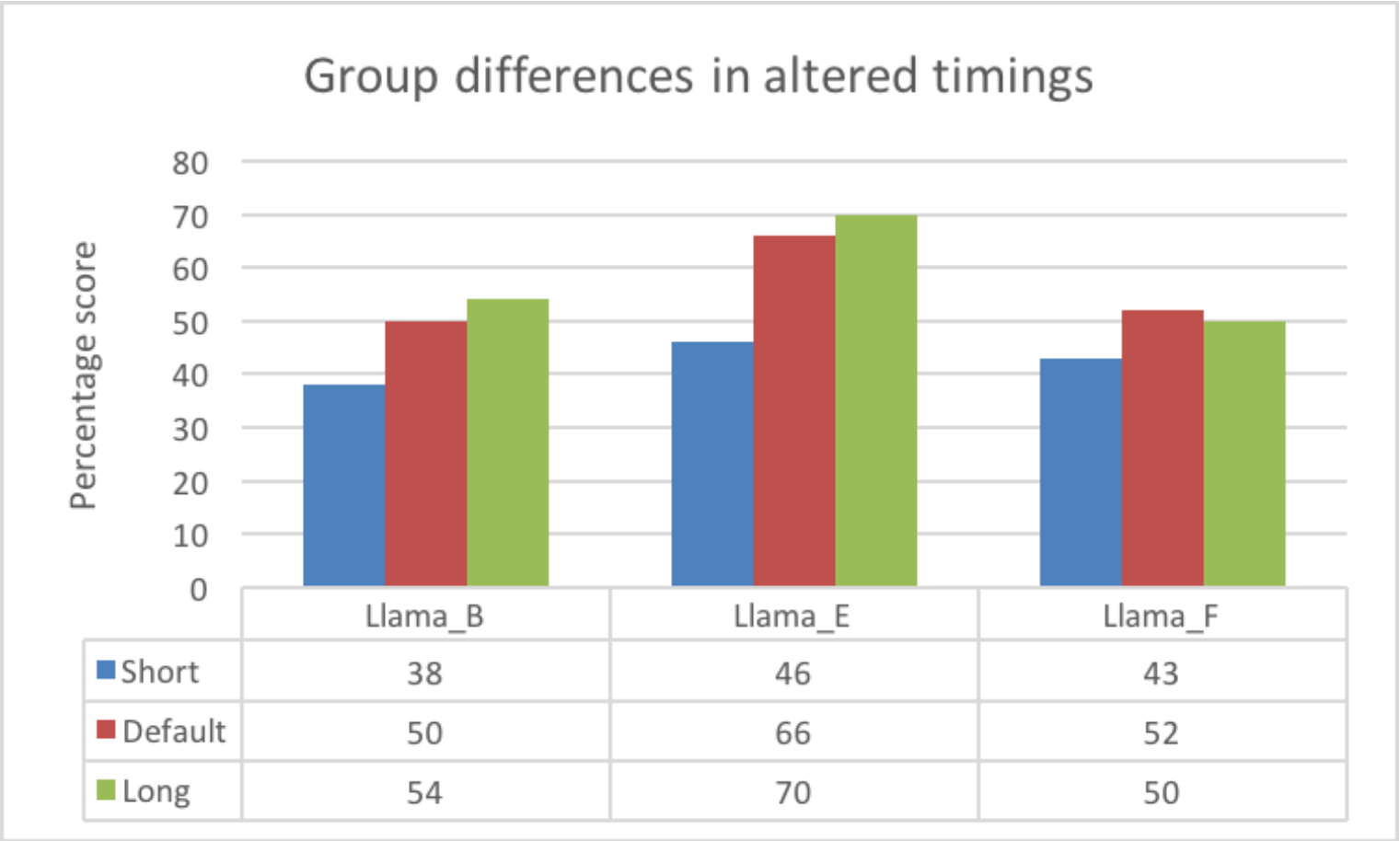
Methodology

- 164 participants at standard length
- 65 participants at altered lengths
- aged 10-75
- Limesurvey background questionnaire
- Data collected via individual and drop-in sessions.

RQ6: Timings

- Default timings:
  - LLAMA B, E = 2 mins
  - LLAMA F = 5 mins
- LLAMA D not included (recording)
- Shorter condition = minus 1 minute
- Longer condition = plus 1 minute
- Participants (n=98)
  - 32 shorter timing
  - 33 default timing
  - 33 longer timing

Results for RQ6



	Llama_B	Llama_E	Llama_F
Short	38	46	43
Default	50	66	52
Long	54	70	50

Stats for Timings

- Not normally distributed (non-parametric)
- Overall timing effects for:
  - LLAMA B (vocab) p=.011.
  - LLAMA E (sound-symbol) p=.004.
- Within groups:
  - Significant difference between default time and shorter time (LLAMA B & E).
  - Significant difference between shorter time and longer time (LLAMA E).
- No effect of timing on LLAMA F
  - Even 4 mins may be too long.
  - Students seem to have finished early (no notes).

Co-variates?

- Participants were matched gender, age, education and L2 status.
- Effect of L2 status on changed times with LLAMA B (vocab) and LLAMA E (sound-symbol).
  - Monolingual scores more affected in B & E.
- Males more affected than females by changes for LLAMA B and LLAMA E.

Overall 2013-14 results

- Results
  - Comparable results to Grañena (2013) for age but inconclusive for language neutrality (LLAMA E).
  - Significant effect of formal education and playing logic puzzles on LLAMA E (sound-symbol).
  - Default timings for LLAMA B (vocab) and E (sound-symbol) appear optimal.
  - LLAMA F could be shortened if participants do not take notes.
- Limitations
  - Over-dominance of UG monolingual participants.
  - Some groups were very small (age, language neutrality)

Follow up study

- Follow up study to 2013-14 work.
- With 2014-15 BA dissertation students.
- Tom Barnett-Legh, Clare Curry & Emma Davie.
- In press, JESLA2017.
- Looked at L1 (language neutrality), L2 status, age.

Tom, Clare & Emma

Research Questions

1. Are the LLAMA tests language neutral?
2. What effect does instructed L2/ bilingual status have on LLAMA scores?
3. Does age affect aptitude as measured by the LLAMA tests?

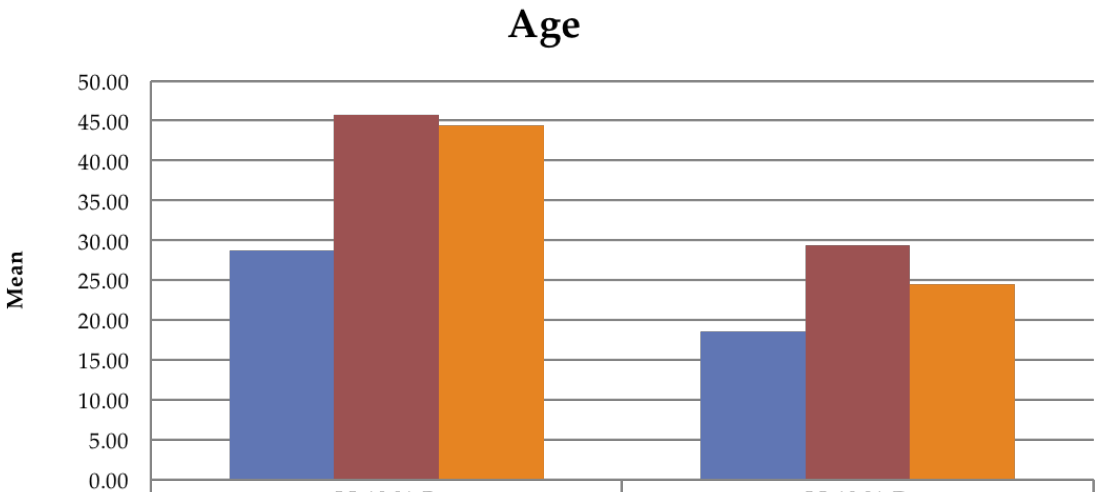
Methodology

- Most data collected by BA dissertation students.
- Data also collected from pre-sessional course and by Khaled Alamri (PhD student).
- Data collected individually or in large computer sessions.
- Background questionnaire on Limesurvey.
- Total number of participants = 240.

RQ3: age

- Only 2 subcomponents tested.
- Three age groups (10-11, 20-21, 30-70)
- Hypothesis 1: no difference on LLAMA B (vocab) as vocabulary learning is lifelong.
- Hypothesis 2: older learners will outperform younger learners due to increased cognitive capacity and maturity on LLAMA B (Miralpeix 2006, 2009)
- Hypothesis 3: younger participants will outperform older participants in LLAMA D (implicit learning) due to critical period effect for implicit learning.
- Subset of participants (n=104) matched for age and gender.

RQ3: Results - graph



	LLAMA B	LLAMA D
10-11 n=30	28.67	18.50
20-21 n=44	45.68	29.32
30-70 n=30	44.33	24.50

Results - stats

- LLAMA B (vocabulary)
  - 10-11 year olds performed significantly worse than both older groups $p<.05$.
  - No significant differences between 20-21s and 30-70s.
  - Hypothesis 1 disconfirmed as younger participants were worse.
  - Hypothesis 2 confirmed.
- LLAMA D (implicit)
  - 10-11 year olds performed significantly worse than 20-21a $p<.05$ but not 30-70s.
  - No significant differences between 20-21s and 30-70s.
  - Hypothesis 3 disconfirmed as younger participants did not perform better than either older group.
- However, 10-11 year olds were able to do the tests.
- No conceptual or interface problems.
- May need different norms.

Bringing year 1 & year 2 together: Rogers et al 2017 to appear JESLA 2017

- RQ1: Are the LLAMA tests language neutral?
- RQ2:What is the effect of monolingualism on LLAMA scores?
- RQ3: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.

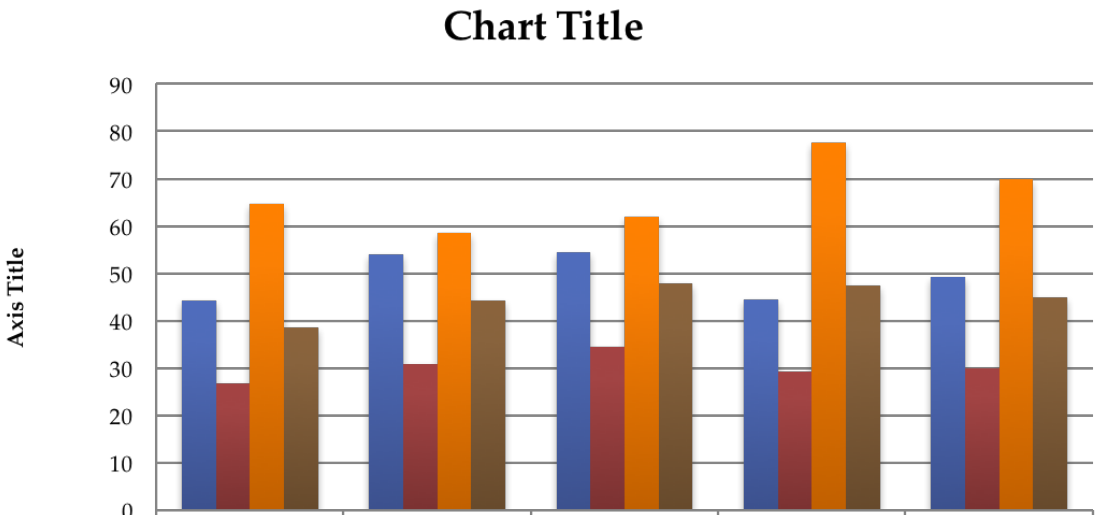
Language neutrality

- Several studies have suggested the distance between L1 and L2 plays a role in word processing and retention of the L2.
- If the language script of the L1 can influence the L2, then does the L1 script influence LLAMA aptitude scores?
- LLAMA B & LLAMA F both contain roman alphabet letters.
- Chinese = morphosyllabic (Tolchinsky et al 2011).
- Arabic = consonant alphabetic script (common ancestor with Roman alphabet).

Hypotheses

- Hypothesis 1: English native speakers will outperform Chinese and Arabic native speakers on LLAMA B & LLAMA F as the script will not require such a strong processing load.
- Hypothesis 2: Arabic speakers will outperform Chinese speakers as it is an alphabetic script.
- Participants: English (n=206), Chinese (n=69), Arabic (n=34), other Roman script (n=24), other (n=6). Total n=339.

Results



	English (n=206)	Chinese (n=69)	Arabic (n=34)	Roman alphabet (n=26)	other (n=6)
LLAMA B	44.22	53.91	54.56	44.38	49.17
LLAMA D	26.82	30.87	34.41	29.17	30
LLAMA E	64.66	58.48	62.06	77.5	70
LLAMA F	38.62	44.35	47.94	47.5	45

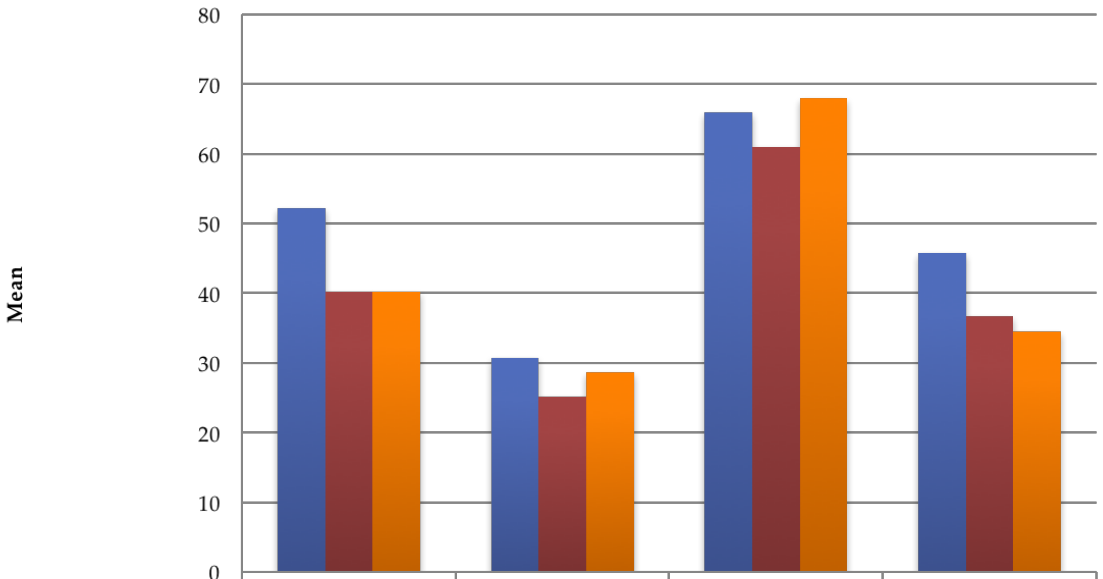
Discussion

- No significant differences for any test.
- possibly due to large differences in group size and large standard deviations.
- However, English native speakers were outperformed in LLAMA B, LLAMA D & LLAMA F.
- Could this be due to high number of monolinguals? >RQ2

RQ2: Effect of monolingualism/ L2 status?

- Compare monolinguals, instructed L2 and bilinguals (defined as two L1s before age 5).
- Hypothesis 1: L2 learners will outperform the other groups as they will have developed conscious strategies.
- Hypothesis 2: Bilinguals will outperform monolinguals as they are more aware of language.
- Participants: L2ers (n=198), monolinguals (n=118), bilinguals (n=29)

RQ2: L2 status results graph



	LLAMA B	LLAMA D	LLAMA E	LLAMA F
L2er (n=198)	52.17	30.71	66.01	45.78
monolingual (n=118)	40.17	25.21	60.97	36.65
bilingual (n=29)	40.17	28.62	67.93	34.48

RQ2: L2 status statistics

- LLAMA B (vocabulary)
  - L2ers significantly outperformed monolinguals and bilinguals.
  - No difference between monolinguals and bilinguals.
- LLAMA D (implicit)
  - L2ers significantly outperformed monolinguals but not the bilinguals.
- LLAMA E (sound-symbol)
  - No difference between any groups.
- LLAMA F (grammatical inferencing)
  - L2ers significantly outperformed monolinguals but not bilinguals.
  - No difference between monolinguals and bilinguals.
- Hypothesis 1: confirmed for LLAMA B & F.
- Hypothesis 2: not confirmed (possibly due to small sample size (n=29))

RQ3 Results

- 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: β = -.250, contribution to variance = 6.0
  - LLAMA D: β = .136, contribution to variance = 1.8
  - LLAMA F: β = -.165, contribution to variance = 2.6

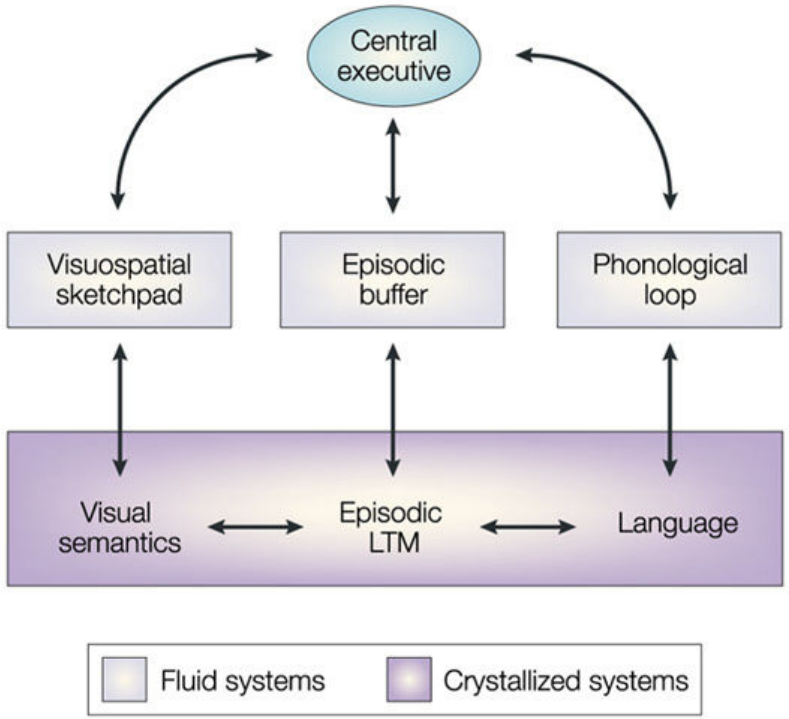
Some other interesting findings

- Learners consistently perform best on LLAMA E (ceiling?).
- Possible pattern element due to layout.
- LLAMA D is out of 75 due to error in program.
- LLAMA F: manual says to take notes but...
  - Two versions of test 2016 paper = no notes (n=135), 2017 paper = notes (n=211).
  - A t-test did not show any difference ($t(344)=0.268$, $p=0.789$).
  - Participants allowed to take notes (M=41.42, s.d.=26.28) and not allowed (M=42.22, s.d. 28.35).
  - No notes completed quicker.
  - Notes group drew pictures and wrote sentences not work out rules.
- Only LLAMA B does not penalise for marking. Need over 50% correct to score above 0% for D, E & F.

Current experiment: LLAMA and Working memory

- BA dissertation group: Tesni Galvin, Izzy Greenfield, Martha Chisholm, Jake Clothier, & Amelia Cobner (not pictured), .

Baddeley's (2003) model of working memory



Background: Wen (2016)

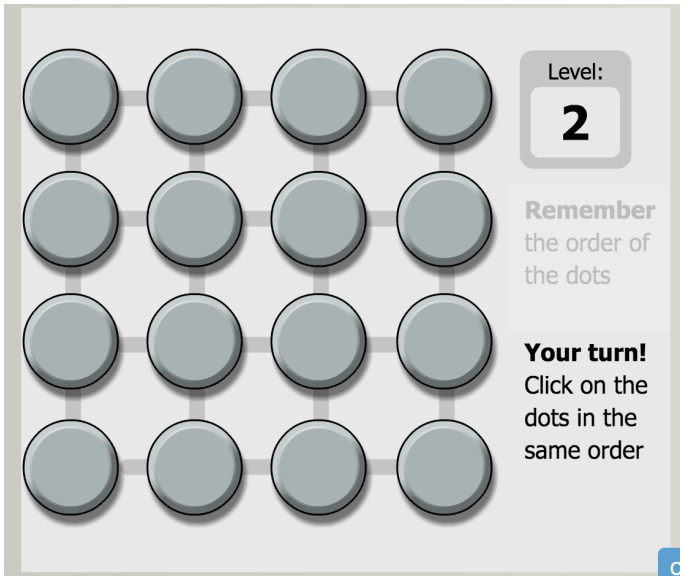
- different components of aptitude relate to different components of working memory:
- PSTM = 'language learning device'
- Central executive = 'language processes'

Methodology

- All four LLAMA sub-tests.
- Working memory tests:
  - Visuo-spatial task (reading)
  - Auditory digits backwards task (PSTM)
  - TMT part 1 & 2: attentional control (Central executive)
- Background questionnaire

Working memory tasks

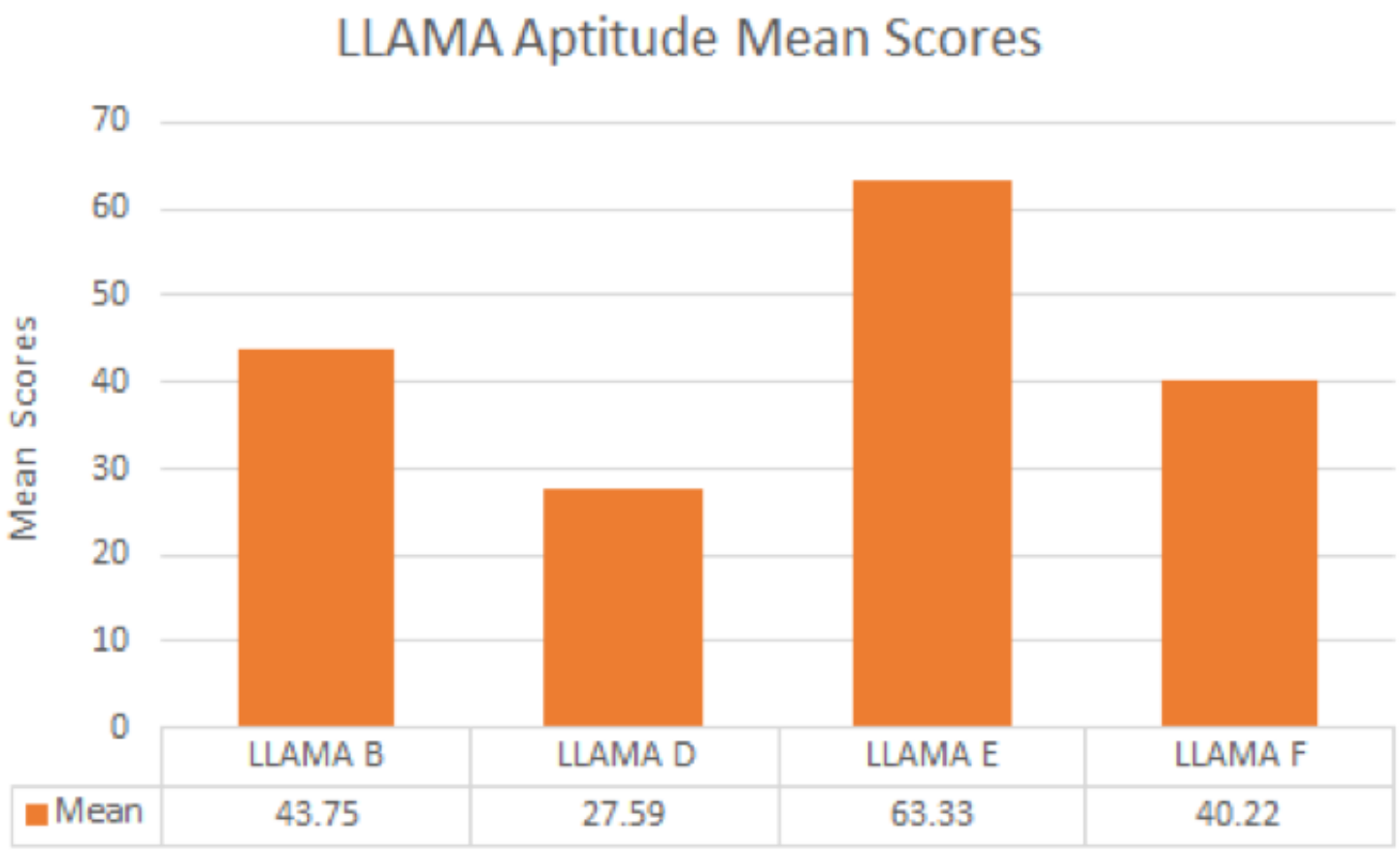
storage task: http://www.cogmed.com/working-memory-challenge
digits backwards: http://cognitivefun.net/test/11



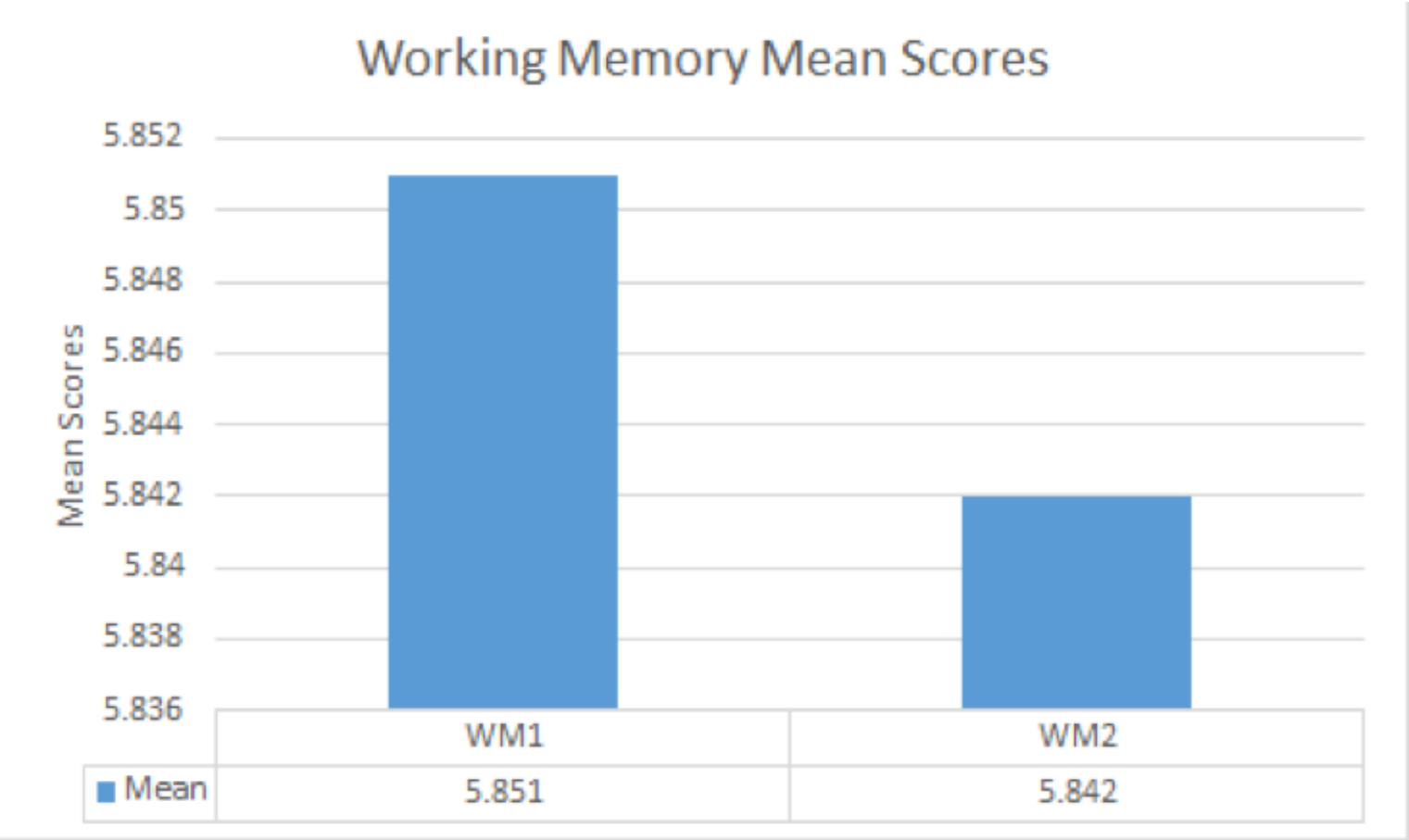
Participants

- 123 participants
- aged 16-78 (mean = 34.29, SD=19.01, Median=22)
- 46% female, 54% male
- predominantly students

Results: LLAMA tests

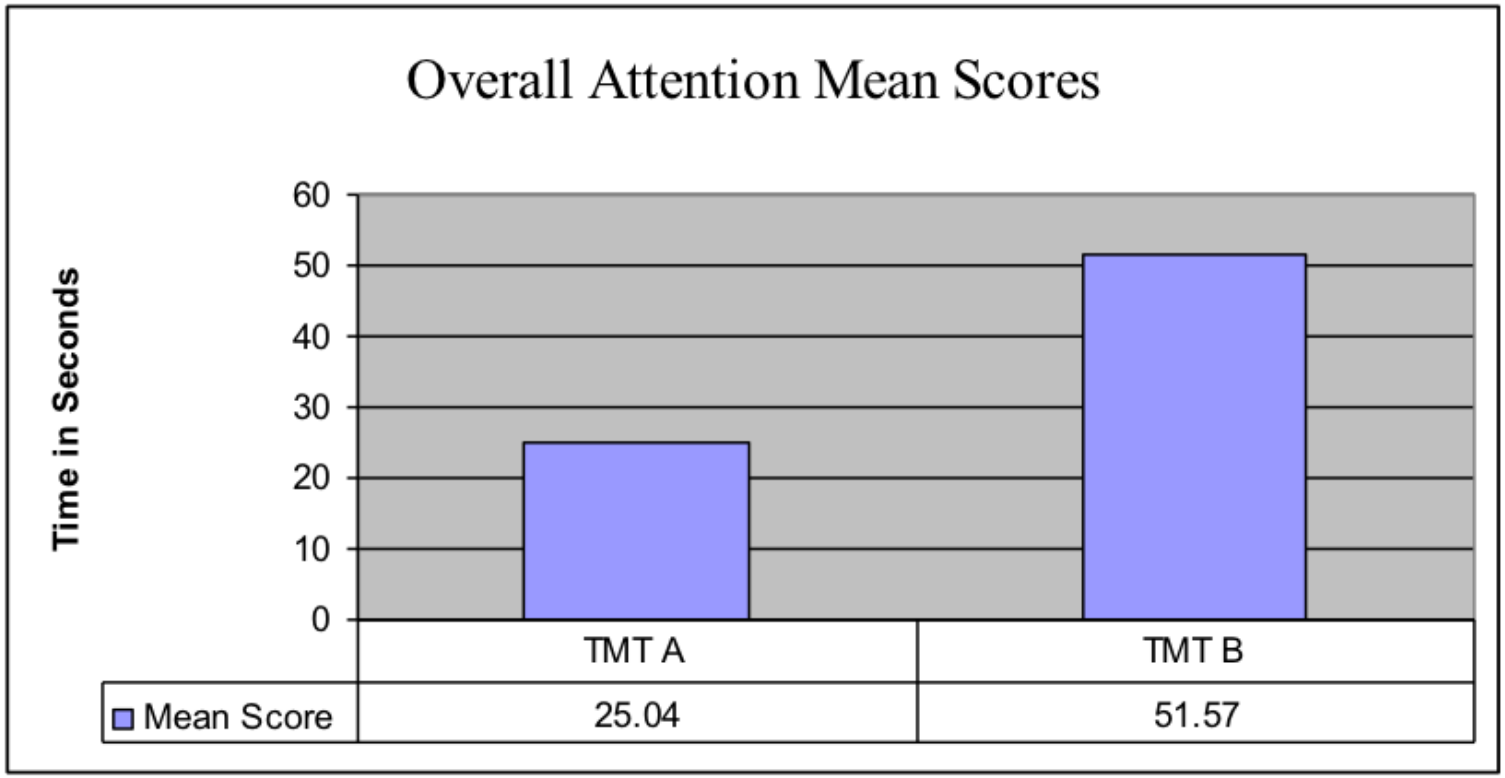


Results: Working memory tests



WM1 = visuo-spatial, WM2 = digits backwards

Results: Working memory (attention) tests



Correlations

- Significant weak correlations found with LLAMA B, E & F with WM scores.
- Significant weak-moderate correlations between WM scores and attention scores.
- Significant weak correlations between TMT2 and LLAMA B, E & F.
- Significant weak correlations between TMT1 and LLAMA B & F

Principal Components Analysis

- oblique, rotated PCA.
- Four components identified.

Component	1	2	3	4
F	.819	-.125	.174	
E	.795		.231	
B	.698	-.276		.270
WM3		.890	-.127	-.115
WM4	-.185	.848	-.169	
WM2			.878	.156
WM1	.321	-.332	.638	-.110
D	.189			.945

Discussion

- LLAMA aptitude tests measure two contructs (similar to Grañena 2013).
- LLAMA B, E & F measure something different to LLAMA D.
- No LLAMA test loads on the same factor as any of the working memory and attention tests.
- TMT parts 1 & 2 measure different aspect of WM to the digits backwards (PSTM) and visio-spatial/ storage measures.
- Even if forced to two or three factors, LLAMA tests load differently to the WM/attention tests.
- Possible evidence against Wen's integrated Model.

Some overall conclusions

- LLAMA tests are robust and not unduly influenced by individual factors.
- Caution is advised if used with younger learners or L3 learners as different norms may be needed.
- Timings seem optimal.
- LLAMA E is negatively skewed and too easy. No notes should be allowed.
- No differences with presence of absence of note-taking with LLAMA F.
- LLAMA measures something different to WM and attention measures.

Next steps

- Need to further analyse current data set.
- Interesting findings on over 55s bilingual vs monolingual.
- Improve WM measures.
- Developing online versions and also reaction time versions.
- Revise layout of LLAMA E.
- Consider revising the scoring.
- Need macro validation to put LLAMA on par with MLAT etc.

Thank you
Any questions?

Contact details

v.e.rogers@swansea.ac.uk
www.viviennerogers.info
@RogersVivienne
ResearchGate