INSTITUT FÜR MEHRSPRACHIGKEIT
INSTITUT DE PLURILINGUISME
ISTITUTO DI PLURILINGUISMO
INSTITUT DA PLURILINGUITAD
INSTITUTE OF MULTILINGUALISM

CSP Center scientific da cumpetenza per la plurilinguitad Cogniziun Società Formation Bildung Migration Furmaziun Gesellschaft
CSP Centro scientifico di competenza per il plurilinguismo Scuola Arbeit Politique Communitad School Travail Ecole Community
CSP Centre scientifique de compétence sur le plurilinguisme Migrazione Furmaziun Societad Cognition Society Scola Migration
KFM Wissenschaftliches Kompetenzzentrum für Mehrsprachigkeit Societé Cognizione Migraziun Schule Communauté Kognition
RCM Research Centre on Multilingualism Formazione Lavoro Politics Comunità Work Politik Lavur Politica Formation Gemeinschaft

Investigating test method effects in French L2 reading items for young learners

Peter LENZ
Katharina KARGES
Malgorzata BARRAS

Institute of Multilingualism,
University of Fribourg & HEP Fribourg (CH)

31 May 2019, 16th EALTA Conference, Dublin

Content

Introduction

- Findings on multiple-choice vs. open-ended items
- Research questions
- The Task Lab Project

Psychometric item analyses

Regression analyses on the construct

Summary and discussion

Terminology

- item type $\subset$ test method
- multiple-choice (MC) items $\subset$ selected-response items
- short-answer (SA) items $\subset$ constructed response (CR) or open-ended (OE) items

Prior findings on test method effects

- In practice, MC and CR items seldom tap into the same construct. If they do, the correlations between scores are high. Rodriguez (2003)
- When items are stem-equivalent, correlations between MC and CR are particularly high. Rodriguez (2003)
- In EFL reading, MC items are easier on average than OE items. Less proficient students are more affected by harder conditions. Shohamy (1984)
- Reading scores on MC and OE items are more highly correlated when the text prompt is unavailable while answering. Ozuru et al. (2007)
- OE items measure more sensitively the quality of active generative processing during reading comprehension. MC items tap into more passive recognition. Ozuru et al. (2013)

Research questions

A. Are there any systematic differences in the psychometric functioning of stem-equivalent SA and MC items?

If there are:

B. How dramatic are they for a measurement instrument consisting of these two item types?

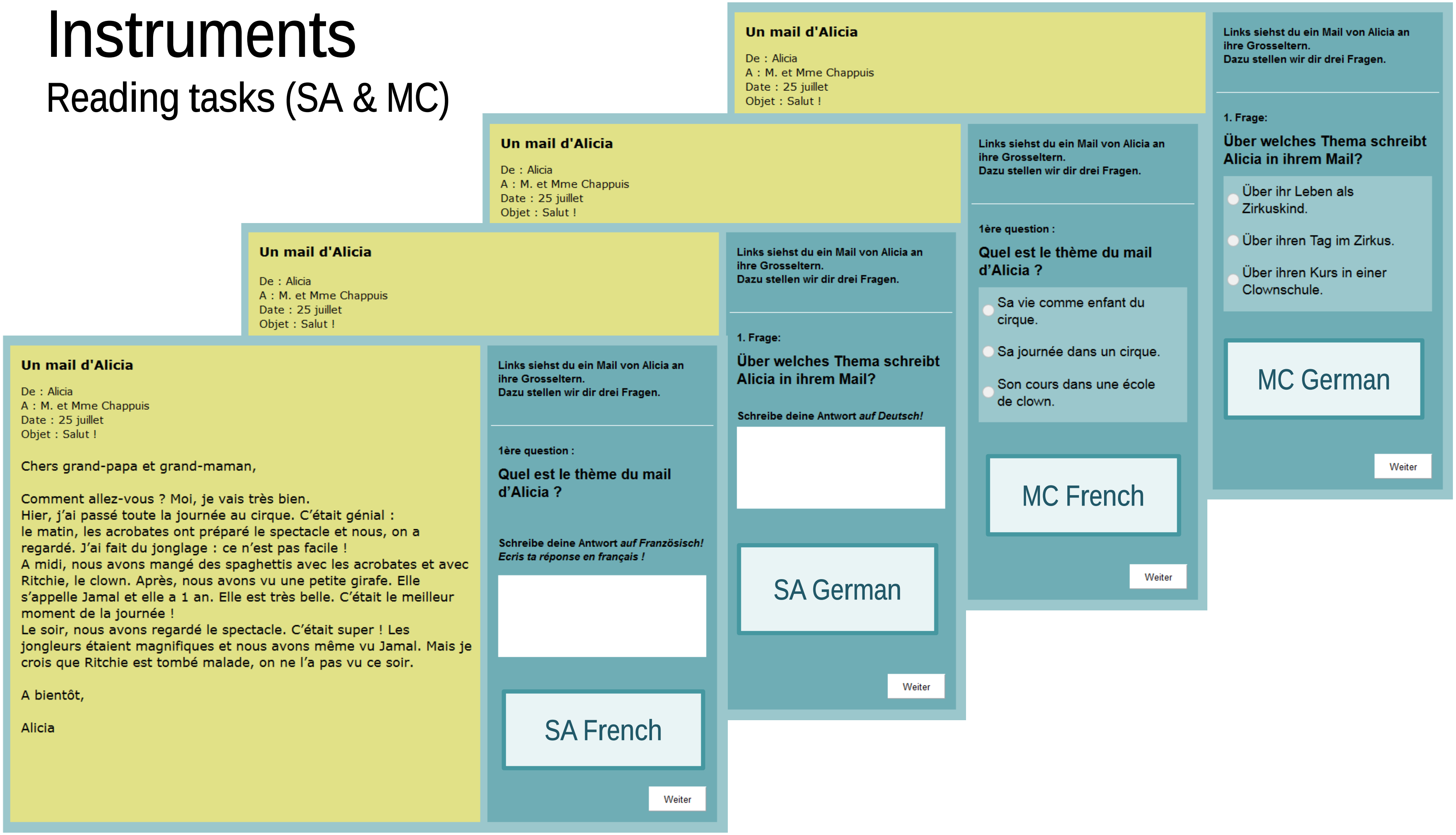
C. In what way do the constructs represented by either of the two item types differ?

The Task Lab project

- **Practical interest:** Inform upcoming test development for large-scale assessments in Switzerland
- Objectives for research
  - – Understand computer-based reading assessment
  - – Investigate test method effects
    - item types: SA – MC – Matching
    - language of questions and responses
  - – Investigate covariates of reading proficiency, e.g. vocab knowledge
- Participants
  - – Pupils age 12, grade 6, German = language of schooling
  - – French = first foreign language, 4 years of instruction (≈ 400 lessons)
  - – Main study: 35 classes ≈ 600 learners

Instruments

Reading tasks (SA & MC)



Un mail d'Alicia

De : Alicia
A : M. et Mme Chappuis
Date : 25 juillet
Objet : Salut !

Chers grand-papa et grand-maman,

Comment allez-vous ? Moi, je vais très bien.
Hier, j'ai passé toute la journée au cirque. C'était génial : le matin, les acrobates ont préparé le spectacle et nous, on a regardé. J'ai fait du jonglage : ce n'est pas facile ! A midi, nous avons mangé des spaghettis avec les acrobates et avec Ritchie, le clown. Après, nous avons vu une petite girafe. Elle s'appelle Jamal et elle a 1 an. Elle est très belle. C'était le meilleur moment de la journée !
Le soir, nous avons regardé le spectacle. C'était super ! Les jongleurs étaient magnifiques et nous avons même vu Jamal. Mais je crois que Ritchie est tombé malade, on ne l'a pas vu ce soir.

A bientôt,

Alicia

**Links siehst du ein Mail von Alicia an ihre Grosseltern.
Dazu stellen wir dir drei Fragen.**

1ère question :

Quel est le thème du mail d'Alicia ?

Schreibe deine Antwort *auf Französisch!*
Ecris ta réponse en français !

SA French

Weiter

**Links siehst du ein Mail von Alicia an ihre Grosseltern.
Dazu stellen wir dir drei Fragen.**

1. Frage:

Über welches Thema schreibt Alicia in ihrem Mail?

Schreibe deine Antwort *auf Deutsch!*

SA German

Weiter

**Links siehst du ein Mail von Alicia an ihre Grosseltern.
Dazu stellen wir dir drei Fragen.**

1ère question :

Quel est le thème du mail d'Alicia ?

○ Sa vie comme enfant du cirque.
○ Sa journée dans un cirque.
○ Son cours dans une école de clown.

MC French

Weiter

**Links siehst du ein Mail von Alicia an ihre Grosseltern.
Dazu stellen wir dir drei Fragen.**

1. Frage:

Über welches Thema schreibt Alicia in ihrem Mail?

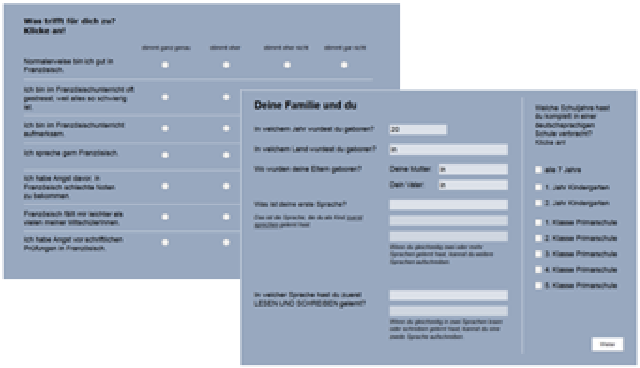
○ Über ihr Leben als Zirkuskind.
○ Über ihren Tag im Zirkus.
○ Über ihren Kurs in einer Clownschule.

MC German

Weiter

Independent variables

Social and conative variables



**Was trifft für dich zu?
Klicke an!**

	stimmt ganz genau	stimmt eher	stimmt eher nicht	stimmt gar nicht
Normalerweise bin ich gut in Französisch.	○	○	○	○
Ich bin im Französischunterricht oft gestresst, weil alles so schwierig ist.	○	○		
Ich bin im Französischunterricht aufmerksam.	○	○		
Ich spreche gern Französisch.	○	○		
Ich habe Angst davor, in Französisch schlechte Noten zu bekommen.	○	○		
Französisch fällt mir leichter als vielen meiner Mitschüler/innen.	○	○		
Ich habe Angst vor schriftlichen Prüfungen in Französisch.	○	○		

Deine Familie und du

In welchem Jahr wurdest du geboren? 20

In welchem Land wurdest du geboren? in

Wo wurden deine Eltern geboren? Deine Mutter in
Dein Vater in

Was ist deine erste Sprache?
Das ist die Sprache, die du als Kind zuerst sprechen gelernt hast.

Wenn du gleichzeitig zwei oder mehr Sprachen gelernt hast, kannst du weitere Sprachen aufschreiben.

In welcher Sprache hast du zuerst LESEN UND SCHREIBEN gelernt?

Wenn du gleichzeitig in zwei Sprachen lesen oder schreiben gelernt hast, kannst du eine zweite Sprache aufschreiben.

Welche Schuljahre hast du komplett in einer deutschsprachigen Schule verbracht? Klicke an!

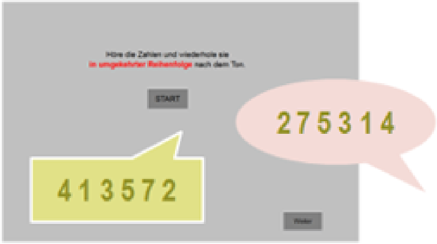
- □ alle 7 Jahre
- □ 1. Jahr Kindergarten
- □ 2. Jahr Kindergarten
- □ 1. Klasse Primarschule
- □ 2. Klasse Primarschule
- □ 3. Klasse Primarschule
- □ 4. Klasse Primarschule
- □ 5. Klasse Primarschule

Student Questionnaire

- Gender
- Language background
- Motivation (enjoyment)
- Motivation (ought)

Independent variables

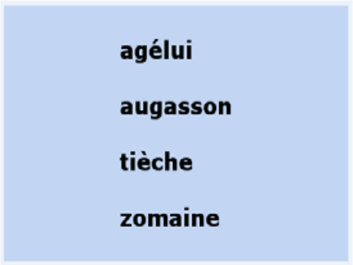
Component tests (I)



Höre die Zahlen und wiederhole sie
in umgekehrter Reihenfolge nach dem Ton.

Backward Digit Span Task

Working memory/ processing



agélui

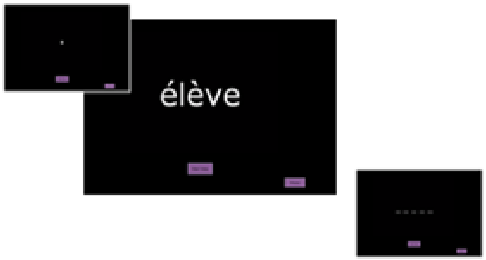
augasson

tièche

zomaine

Phonological awareness

Pronounce French graphemes

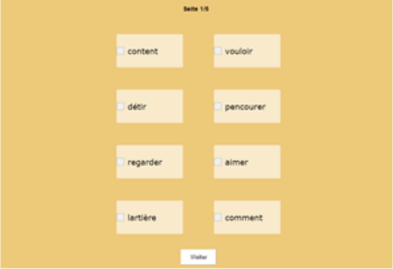


Sight-word recognition

Word decoding (gestalt)

Independent variables

Component tests and integrative measures



Seite 1/5

□ content □ vouloir

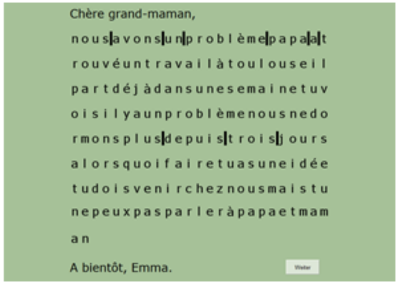
□ détir □ pencourer

□ regarder □ aimer

□ lartière □ comment

Yes-No Task

Vocabulary breadth (receptive)



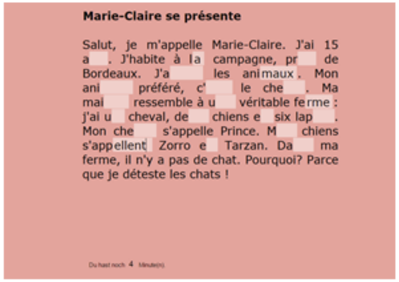
Chère grand-maman,

nous|avons|un|problème|papa|a|t
rouvéuntravailàtoulouseil
partdéjàdansunesemainetuv
oisilyaunproblèmenousnedo
rmonsplus|depuis|trois|jours
alorsquoifairetuasuneidée
tudoisvenircheznousmaistu
nepeuxpasparleràpapaetmam
an

A bientôt, Emma.

Text segmentation

Lexico-syntax / integrative measure



Marie-Claire se présente

Salut, je m'appelle Marie-Claire. J'ai 15 a___. J'habite à la campagne, pr___ de Bordeaux. J'a___ les animaux. Mon ani___ préféré, c'___ le che___. Ma mai___ ressemble à u___ véritable ferme : j'ai u___ cheval, de___ chiens e___ six lap___. Mon che___ s'appelle Prince. M___ chiens s'appellent Zorro e___ Tarzan. Da___ ma ferme, il n'y a pas de chat. Pourquoi? Parce que je déteste les chats !

Du hast noch 4 Minute(n)

C-Test

Integrative measure / written text reconstruction

Instrument development and data collection

- **Pre-piloting** (cog lab): instrument usability; construct validity
  - Retrospective interviews/stimulated recall for all instruments (34 students)

- **Piloting** (field study): data collection process; data samples
  - Piloting of the data collection and revised instruments (97 students)

- Main data collection
  - 35 classes, ≈ 600 learners of French in 6th grade

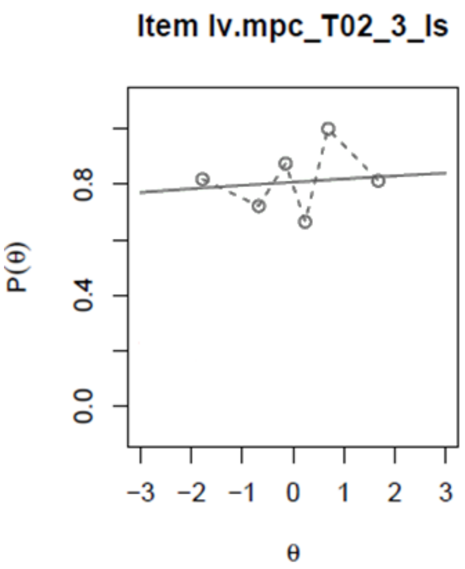
PSYCHOMETRIC ITEM ANALYSES:
multiple choice vs. short answer

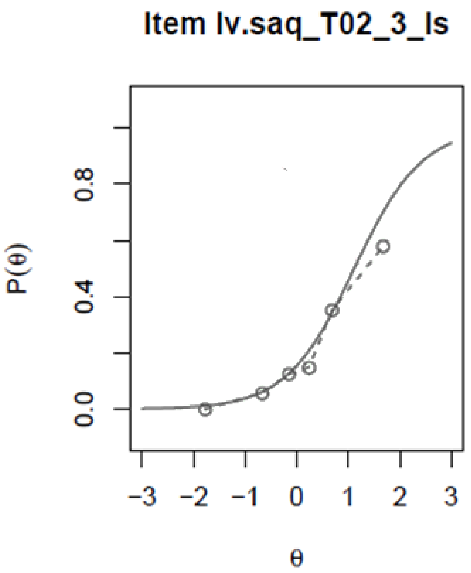
Analysis of reading items

Creating a sound item basis: **item selection**

- Fundamental considerations, e.g. exclude the rare comics items
- Bad fit to 2PL model or/and low discrimination (< .2)

Main source for exclusions: graphic inspection of ICCs





Always MC/SA **pairwise exclusions**
Remaining: **98 item variants**
on 10 text passages
83-154 (mean = 117.9) responses
per item variant
588 students (f = 290, m = 298)

Analysis of reading items

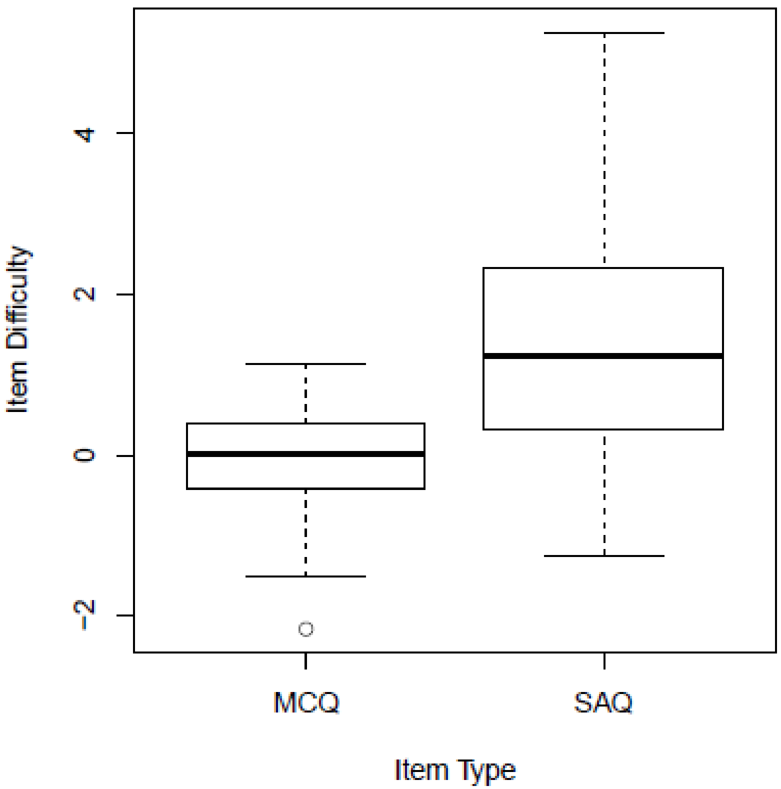
Item difficulty (2 PL model)

Means
MC: -0.126 (0.106)
SA: 1.349 (0.218)

Significance: paired t-test
$t_{(48)} = 7.67$, $p < .001$

Effect size
$d = 1.10$ (large)

Item Difficulties per Item Type



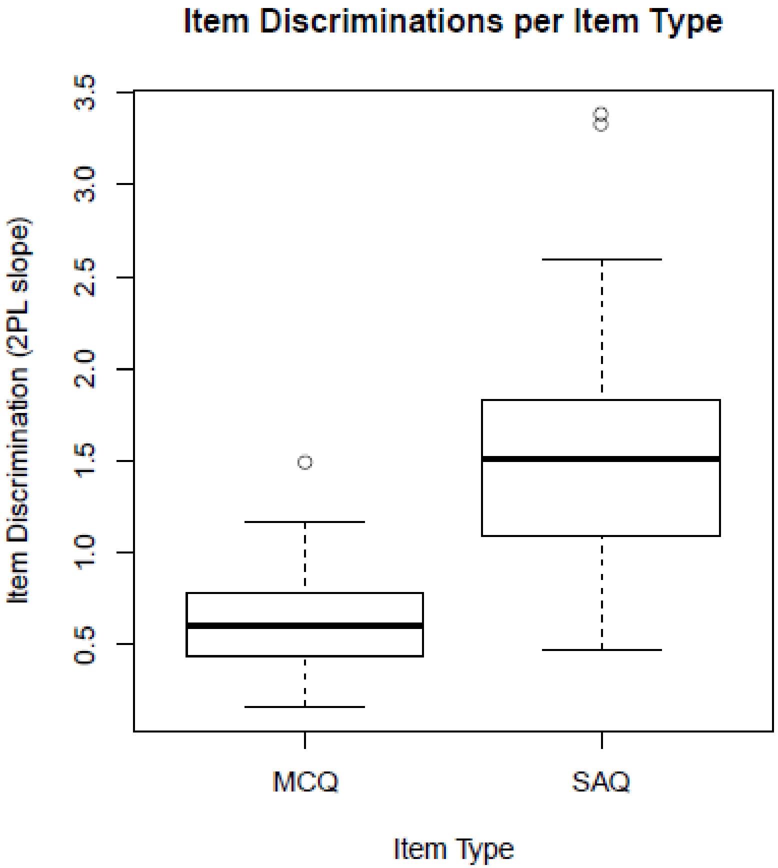
Analysis of reading items

Item discrimination (slopes in 2 PL model)

Means
MC: 0.657 (0.041)
SA: 1.535 (0.091)

Significance: paired t-test
$t_{(48)} = 9.26$, $p < .001$

Effect size
$d = 1.32$ (large)



Analysis of reading items

Effects of combining MC & SA items in a Rasch framework

Rasch premise: **specific objectivity**

Any subsample of items taken from a test would classify test-takers in the same order (cf. Rasch, 1977).

Q: Is the principle of **specific objectivity met** by our collection of items?

→ Calculate the **Mean deviation profile** from ***Profile Analysis*** (Verhelst, 2011)

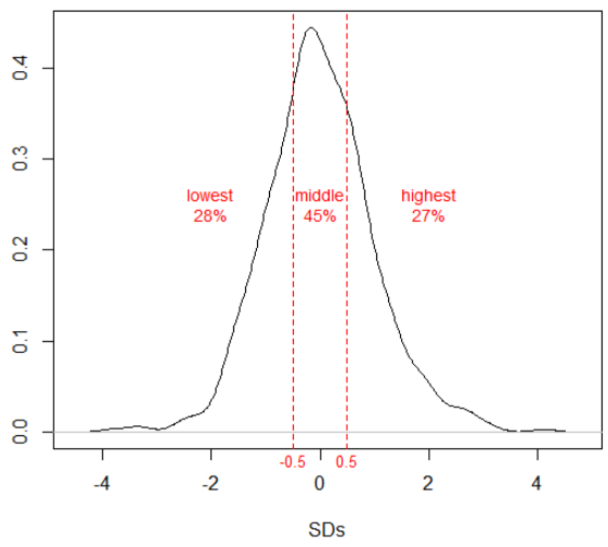
Analysis of reading items

Effects of combining MC & SA items in a Rasch framework

Establish *mean deviation profile* for 2 item and 3 ability groups

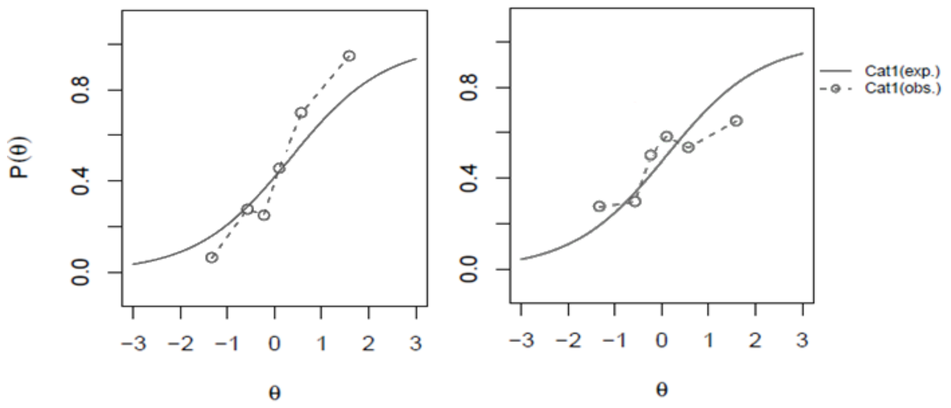
- Individual deviation profiles:
  Add differences between observed scores (0 or 1) and expected scores for each MC or SA item
- Calculate 3 group means from the individual profiles

WLE Plot - Ability Groups



Item lv.saq_T04_1_tl

Item lv.mpc_T05_3_tl



Analysis of reading items

Effects of combining MC & SA items in a Rasch framework

Profile analysis: Mean deviation profile

Ability group	SA items	MC items	SE	z	p
lowest	-0.394	0.394	0.062	-6.352	< 0.001
middle	-0.004	0.004	0.056	-0.064	0.475
highest	0.376	-0.376	0.073	5.159	< 0.001
lowest - highest	-0.77	0.77	0.096	-8.056	< 0.001

Table 5 Mean deviation profile for three ability and two item groups

The subsample of items submitted *does* matter → DIGF.
The least able students according to the model will score higher more easily on MC items than on SA items. The opposite is true for the most able group.
Raw score is not a sufficient statistic for ability → choose 2PL or OPLM model.

REGRESSION ANALYSES ON THE SA AND MC CONSTRUCTS

Exploring the MC & SA reading constructs

through (mixed) multiple regression

1) Separate hierarchical regressions of the MC and the SA scales
2) Concurrent estimation of a LM model for the MC and SA scales

- **Dependent variables**: 2 ability scales based on a) the MC and b) the SA items (WLEs from two-dimensional Rasch analysis) (Latent – ‘error free’ – correlation MC/SA reading: 0.91)
- **Independent variables**: questionnaire and test data (as introduced above)

- **Data** for regression: 40 complete **imputed datasets** reflecting measurement error of missing data and the test scales

Mean correlations between test variables

	De-	S-w	Y/N	Text	C-	Read.	Read.
	coding	recog.	diff.	segm.	Test	SA	MC
Backward digit span (z)	0.18	0.28	0.18	0.13	0.13	0.23	0.18
Decoding (z)		**0.77**	**0.72**	0.66	0.67	0.51	0.48
Sight-word recognition (z)			**0.74**	0.61	0.66	0.56	0.52
Y/N Test, difference (z)				**0.75**	**0.78**	0.62	**0.70**
Text segmentation (z)					**0.84**	0.56	0.52
C-Test (z)						0.58	0.51
Reading SA items							0.63

	Y/N Test, words (z)
Y/N Test, pseudowords (z)	0.58

Hierarchical regression of MC & SA-based reading

Romance L1
Motivation/Enjoy

		SA reading items		MC reading items	
		R^2	R^2 Change	R^2	R^2 Change
	Background variables	0.157	-	0.107	-
French	Backward digit span (z)*	0.196	0.039	0.13	0.023
	Decoding (z)*	0.335	**0.139**	0.262	**0.132**
Voc	Sight-word recognition (z)*	0.389	0.054	0.309	0.047
?	Y/N Test, words (z)*	0.417	0.028	0.337	0.028
	Y/N Test, pseudowords (z)*	0.486	**0.069**	0.574	**0.237**
Txt	Text segmentation (z)	0.504	0.018	0.577	0.003
Wri	C-Test (z)	0.516	0.012	0.584	0.007

9.7%

26.5%

3%

* = sign. when introduced | R^2 = mean pseudo R^2 (Nakagawa & Schielzeth, 2013)

Stepwise regression of MC & SA-based reading

	SA items				MC items			
	R^2	R^2 Change	AIC	AIC change	R^2	R^2 Change	AIC	AIC change
Sight-word recognition (z)	0.389	0.054	6780.0	-33.4	0.309	0.047	6704.0	-47.3
Y/N Test, pseudowords (z)	0.428	0.039	6780.9	0.9	0.332	0.023	6686.0	-18.0
Y/N Test, words (z)	0.486	0.058	6734.0	-46.9	0.574	**0.242**	6530.2	**-155.8**

	SA items				MC items			
	R^2	R^2 Change	AIC	AIC change	R^2	R^2 Change	AIC	AIC change
Sight-word recognition (z)	0.389	0.054	6780.0	-33.4	0.309	0.047	6704.0	-47.3
Text segmentation (z)	0.448	0.059	6706.0	-74.0	0.361	0.052	6645.5	-58.5
C-Test (z)	0.474	0.026	6691.9	-14.1	0.371	0.010	6643.2	-2.3

Sight-word recognition (z)	0.389	0.054	6780.0	-33.4	0.309	0.047	6704.0	-47.3
↑ C-Test (z)	0.465	0.076	6703.3	-76.7	0.355	0.046	6658.8	-45.2
↓ Text segmentation (z)	0.474	0.009	6691.9	-11.4	0.371	0.016	6643.2	-15.6

Differential prediction per item type

Statistical significance

	SA reading measure				
	coeff.	SE	t	df	p
Main effects (extract)					
Backward digit span (z)	8.62	4.22	2.04	113.1	**0.042**
Decoding (z)	-10.26	14.38	-0.71	37.2	0.476
Sight-word recognition (z)	4.32	17.65	0.24	34.2	0.807
Y/N Test, words (z)	63.08	28.58	2.21	29.0	**0.028**
Y/N Test, pseudowords (z)	-46.84	27.68	-1.69	27.3	**0.092**
Text segmentation (z)	11.55	12.35	0.93	52.2	0.351
C-Test (z)	14.12	15.81	0.89	44.9	0.373
Interactions: item type x predictors (extract from output)					
	'Correction' for MC measures				
Backward digit span (z)	-4.20	5.95	-0.71	75.56	0.48
Y/N Test, words (z)	52.80	25.98	2.03	33.8	**0.043**
Y/N Test, pseudowords (z)	-56.39	25.04	-2.25	31.8	**0.025**
Text segmentation (z)	-7.14	13.31	-0.54	56.2	0.592
C-Test (z)	-26.07	16.70	-1.56	46.7	0.120

Association with MC sign. different

Summary and discussion

Psychometric analyses show differences in the way stem-equivalent SA and MC items function (similar to Shohamy, 1984).

- The average MC **item is considerably easier** than the average SA item. Reasons may be: possibility of guessing with MC and a productive element in SA items.
- **SA items discriminate considerably better** than MC items, i.e. they have a stronger relationship to the common latent dimension. MC items may allow for a variety of compensatory test-taking strategies while SA items may engage mainly (and more) linguistic knowledge and skills. However, providing SAs goes beyond reception.
- *Profile Analysis* provides evidence that our MC and SA items show **non-uniform DIGF**. Different samples of MC and SA items would *not* rank test-takers invariably – thus violating a principle of Rasch measurement.

Summary and discussion

- The MC-based and SA-based **reading constructs seem closely related** (latent correlation of Rasch dimensions = 0.91). (cf. Rodriguez' meta-analysis)
- The *words* and *pseudo-words* dimensions of the **Y/N Test** together are the **best predictor** of MC and SA-based reading → No surprise: it is **vocabulary!**
  - Y/N Test predicts **MC-based reading** significantly better than SA-based reading.
  - **What is in the Y/N Test?** – Strictly receptive vocabulary breadth; a penalty for adventurous guessing, …?
    → Y/N Test may mirror **selection and deselection of options** in MC-based reading.
- **Text segmentation and the C-Test** predict SA-based reading almost equally well as the Y/N Test. Also, **working memory** capacity is associated with SA-based reading.

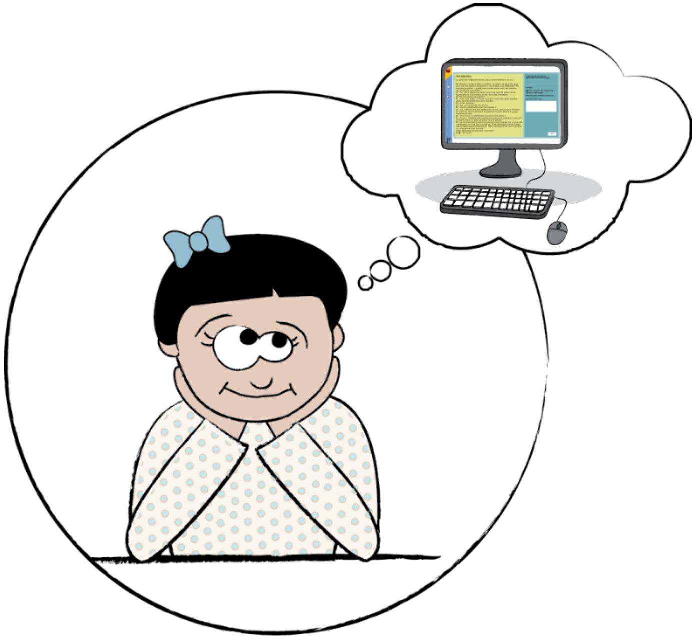
  Reasons: The productive element? More active text processing in the case of SA (Ozuru, 2013)?

Limitations and outlook

- **Reliability** of test scales: each test taker should complete a larger number of items. More items should be involved.
- A better targeted and more **complete set of measures of component/ precursor skills** of reading is desirable.
- Other population samples (age, level of reading proficiency) need to be studied.
- Test method is a superficial characteristic of an item. More **fine-grained item studies** are necessary to help item writing and interpretation.
- Statistical associations between reading measures and predictor variables cannot substitute **introspection and eye-tracking.**
- …

Contact

Research Centre on Multilingualism
Institute of Multilingualism
Rue de Morat 24
CH-1700 Fribourg
Switzerland



Mail: peter.lenz@unifr.ch
katharina.karges@unifr.ch
malgorzata.barras@unifr.ch

Web: www.centre-multilingualism.ch

Lenz, P., Karges, K., & Barras, M. (2019). Investigating test method effects in French L2 reading items for young learners. In A. Huhta, G. Erickson, & N. Figueras, Developments in Language Education: A Memorial Volume in Honour of Sauli Takala (S. 182–202). University of Jyväskylä & EALTA.

Mean correlations between test variables

	De-coding	S-w recog.	Y/N diff.	Text segm.	C-Test	Read. SA	Read. MC
Backward digit span (z)	0.18	0.28	0.18	0.13	0.13	0.23	0.18
Decoding (z)		**0.77**	**0.72**	0.66	0.67	0.51	0.48
Sight-word recognition (z)			**0.74**	0.61	0.66	0.56	0.52
Y/N Test, difference (z)				**0.75**	**0.78**	0.62	**0.70**
Text segmentation (z)					**0.84**	0.56	0.52
C-Test (z)						0.58	0.51
Reading SA items							0.63

	Y/N Test, words (z)
Y/N Test, pseudowords (z)	0.58

Literature

Ozuru, Y., Best, R., Bell, C., Witherspoon, A., & McNamara, D. S. (2007). Influence of question format and text availability on the assessment of expository text comprehension. *Cognition and Instruction*, *25*(4), 399–438.

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *67*(3), 215–227.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. In *The Danish Yearbook of Philosophy* (Bd. 14, S. 58–93). Copenhagen: Munksgaard.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: a random effects synthesis of correlations. *Journal of Educational Measurement*, *40*(2), 163–184.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, *1*(2), 147–170.

Verhelst, N. D. (2011). Profile Analysis: a closer look at the PISA 2000 reading data. *Scandinavian Journal of Educational Research*, 1–18.