Quantitative approaches to the validation of language test tasks

Peter LENZ

Research Centre on Multilingualism
University of Fribourg/Switzerland

PhD Workshop, October 13, 2017

Relevance

- *Quantitative item and test analysis* has a long-standing tradition, particularly in Anglo-Saxon professional language testing (and, of course, psychological testing, educational measurement, etc.).
- In *applied linguistics*, actual development and validation of measurement instruments is often ignored or neglected, but this is changing (cf. Purpura, Brown & Schoonen (2015) in *LL*).
- Quantitative (psychometric) approaches to item and test validation can
  - make the properties of measurement instruments and the properties of person measures based on them known,
  - improve measurement instruments,
  - add to the validity evidence from qualitative sources.

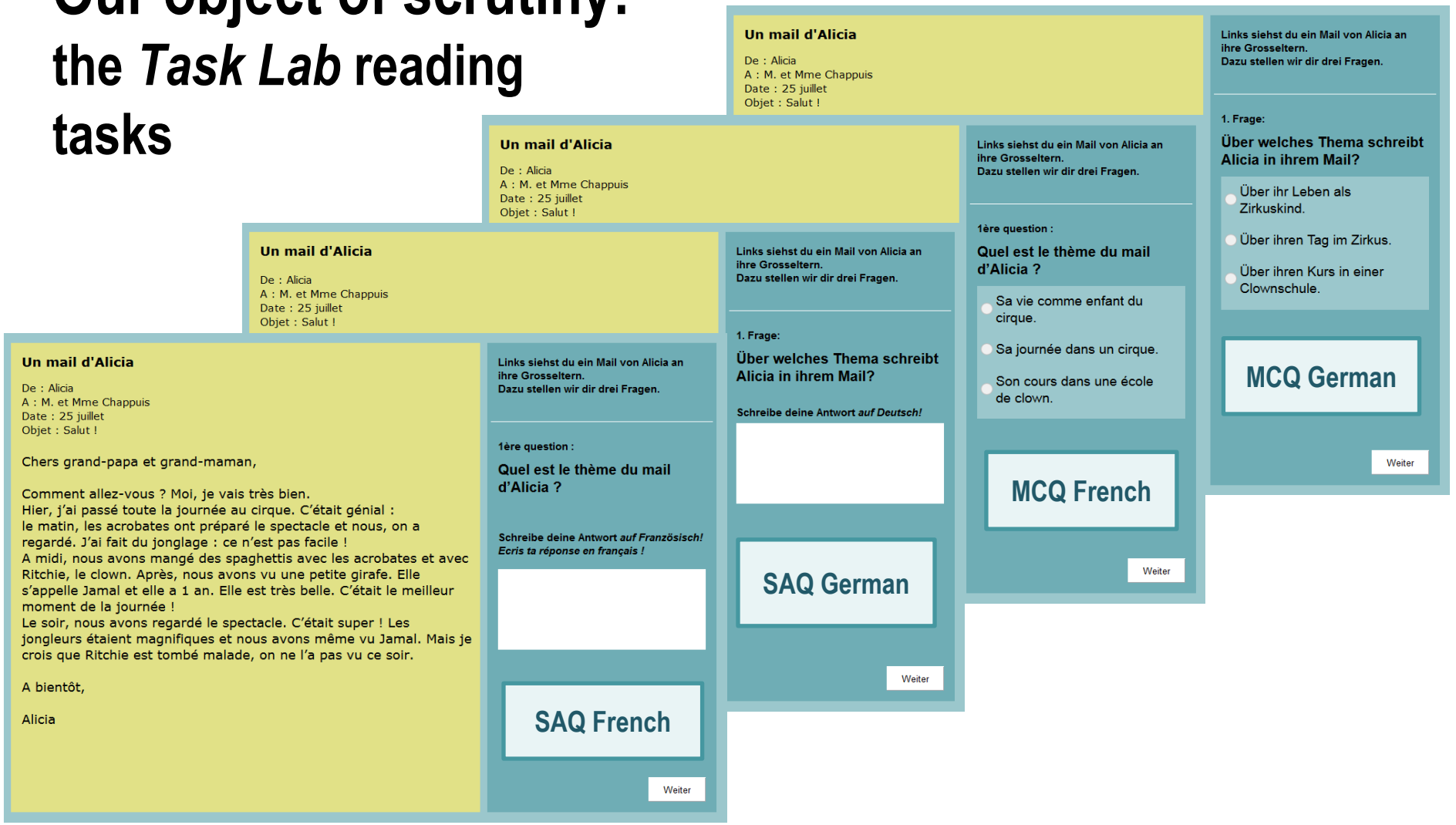
Construct validity from a quantitative perspective

- Measurement instrument is unidimensional or controlled multidimensional
- Instrument has specific objectivity (measures persons with 'more' or 'less' of the construct equally along the scale)
- Test-takers are classified similarly when tested on a recognized instrument (criterion) measuring the same construct.

Questions for analysis:

- Are the items that are assumed to be equal – equal?
- Do all of the items and item groups in the instrument 'fit', i.e. function according to the model used across all items?
- Do the items perform equally for relevant groups of users? (DIF)

Our object of scrutiny: the *Task Lab* reading tasks



Some sample analyses of the *Task Lab* items

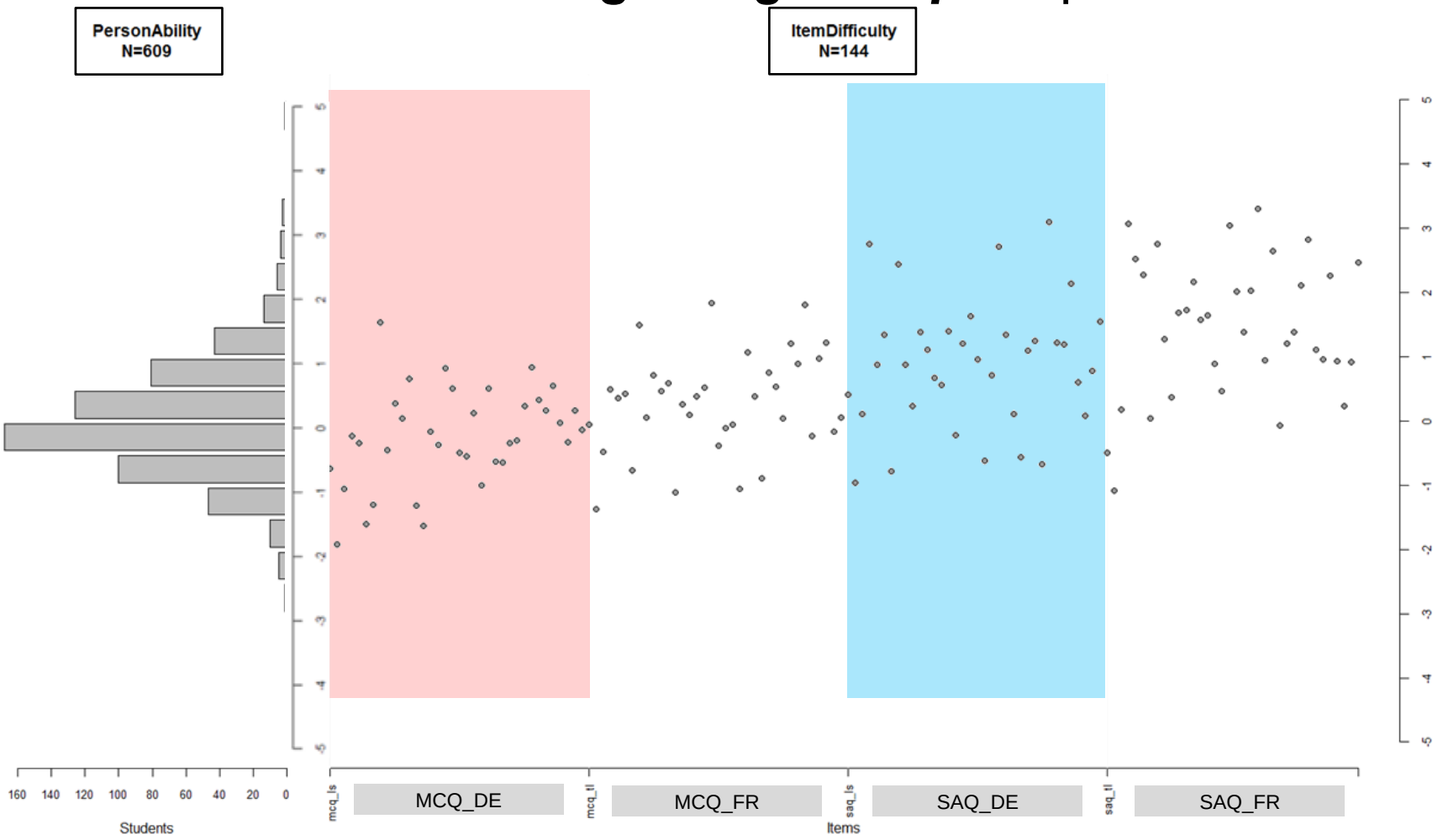
…focusing on the two format aspects *language* and *item type* (*test method*)

		Language of Qu.&Re	
		DE	FR
Item Type	MCQ	MCQ_DE	MCQ_FR
	SAQ	SAQ_DE	SAQ_FR

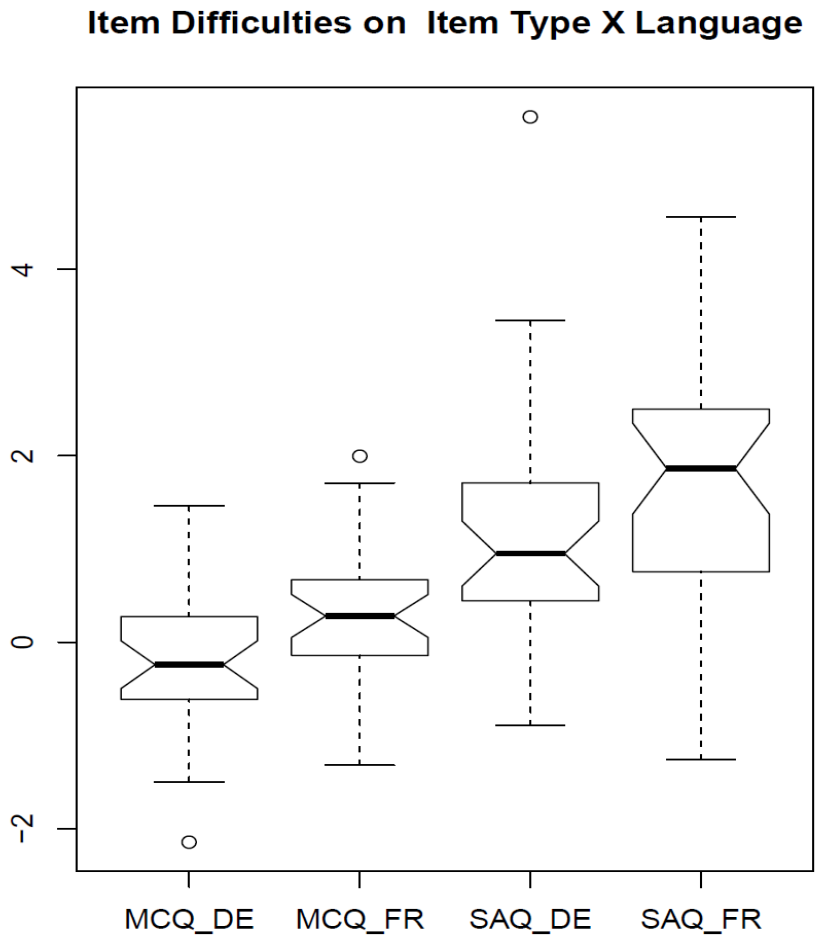
N items: 4 x 36 such items (on 12 different reading passages)

N pers.: 35 classes, 609 students (Ø 120 students per reading task)

Result of Rasch scaling: *Wright Map* for persons & items



Item difficulty per item group

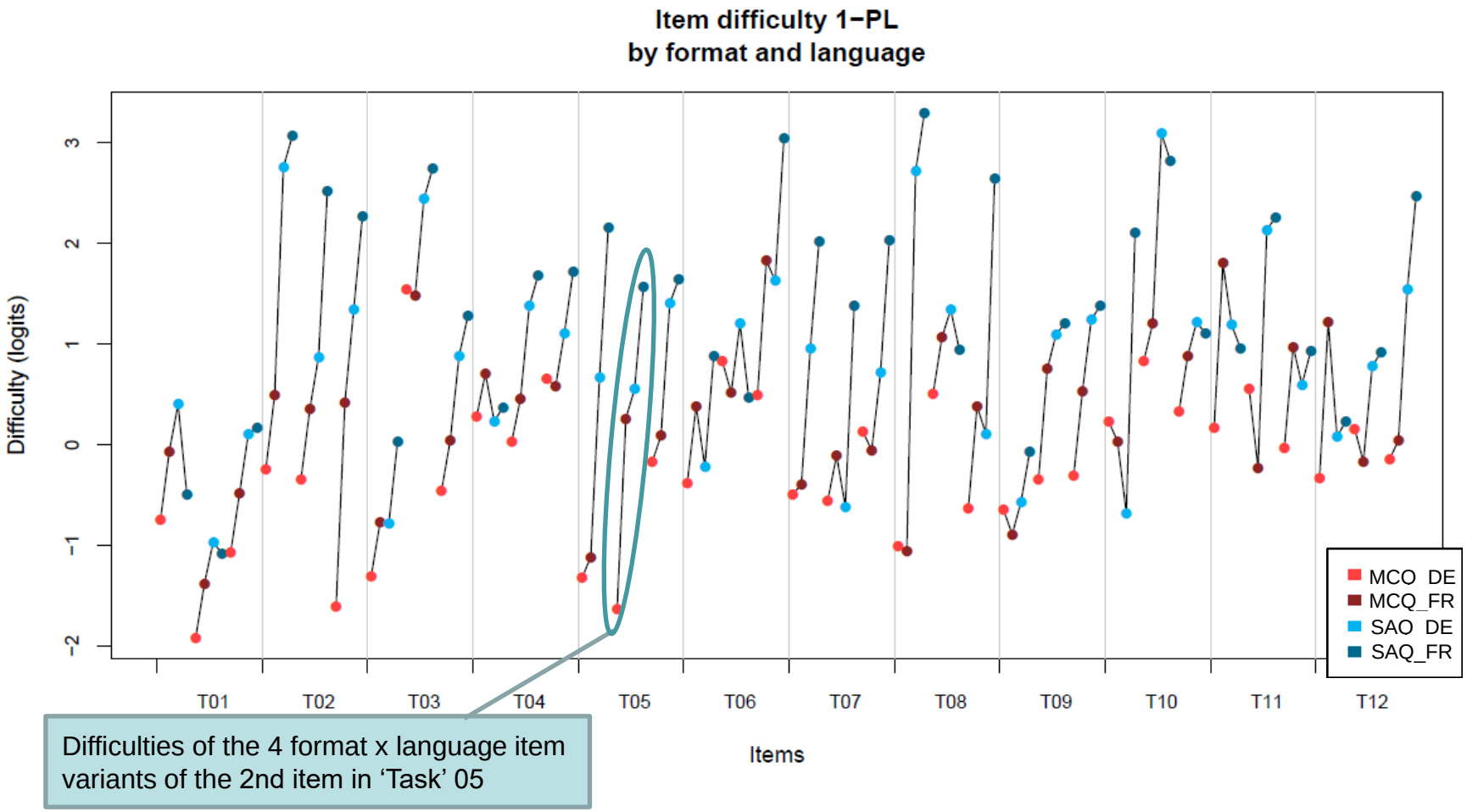


Should this bother us?

- Possibly: "same items" are not equal in difficulty
- There are factors around we don't know, i.e. which are even undesirable

=> A detailed look may help understand and improve item construction

Difficulty of individual items per item group (I)

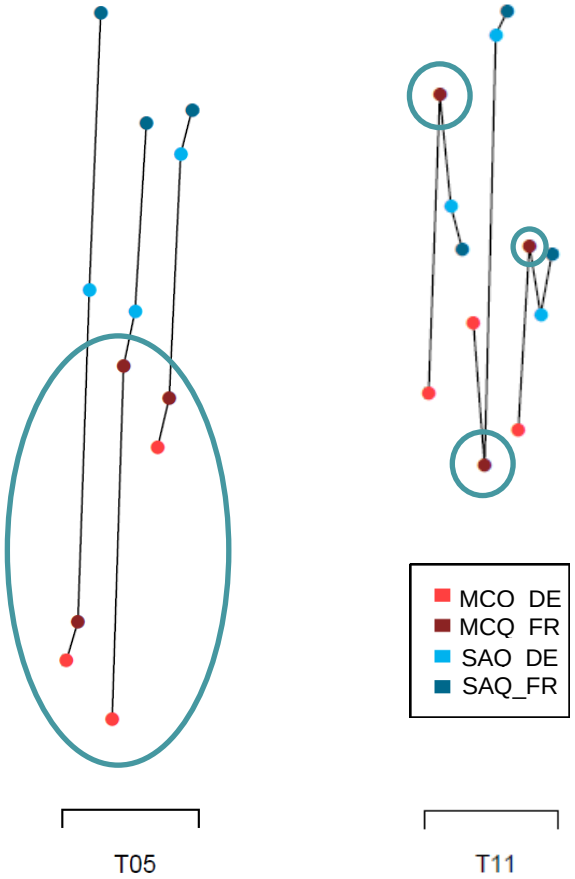


Difficulties of the 4 format x language item variants of the 2nd item in 'Task' 05

Difficulty of individual items per item group (II)

'Normal' patterns:

- MCQ-FR: less difficult compared to MCQ-DE when answer options contain more **words from text**
- SAQ more difficult than MCQ when answer cannot be **copied from text**

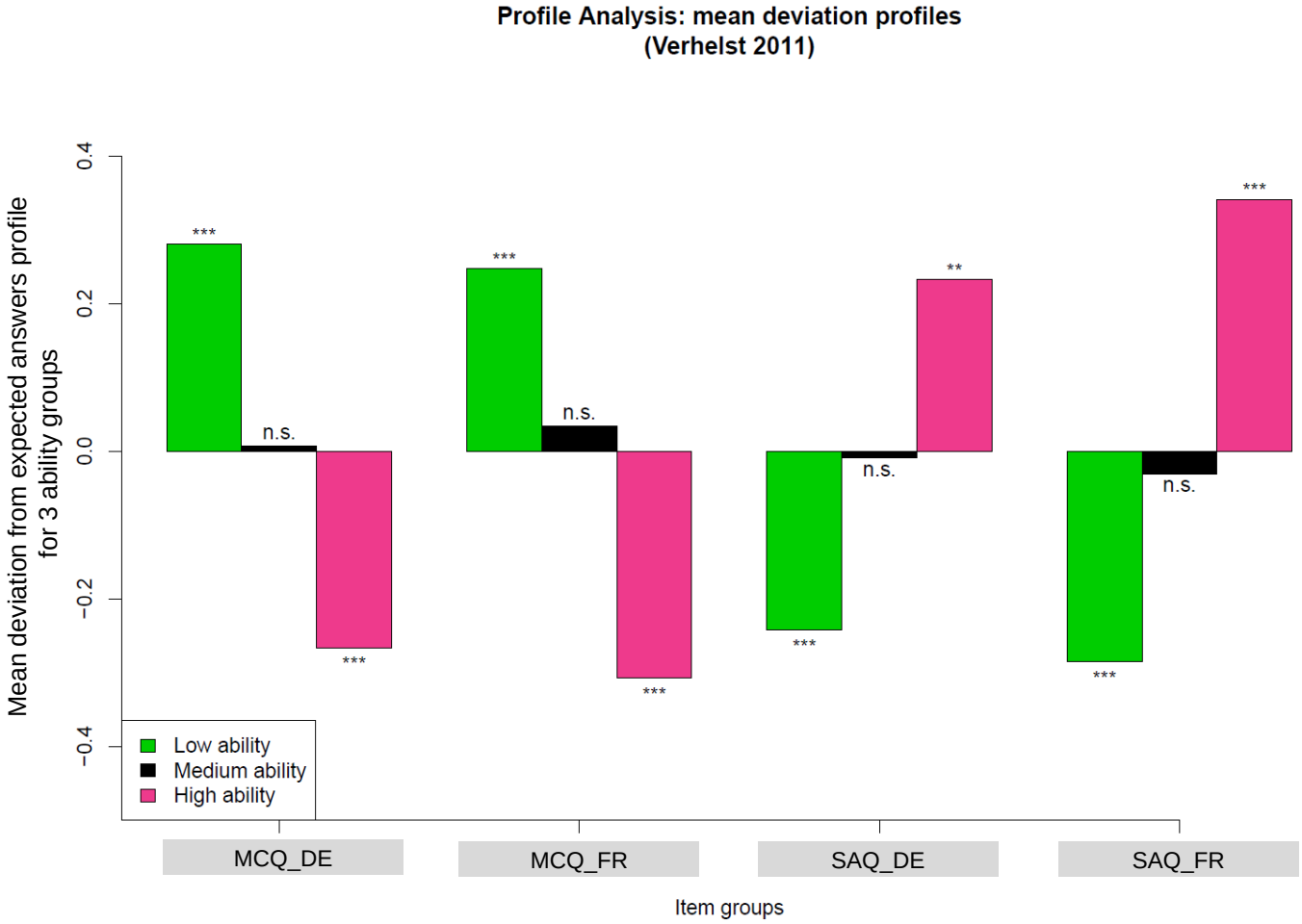


MCQ-FR

- harder when **easy keywords** appear **in distractors** (i1).
- **easier** when **known keywords in question appear in the text**, or when **easy keywords in correct option appear in the text** (i2).
- **Short and simple options** help (i2), less accessible options make things more difficult (i3).

SAQ_FR easier when **answer can be copied from text** (i1, i3)

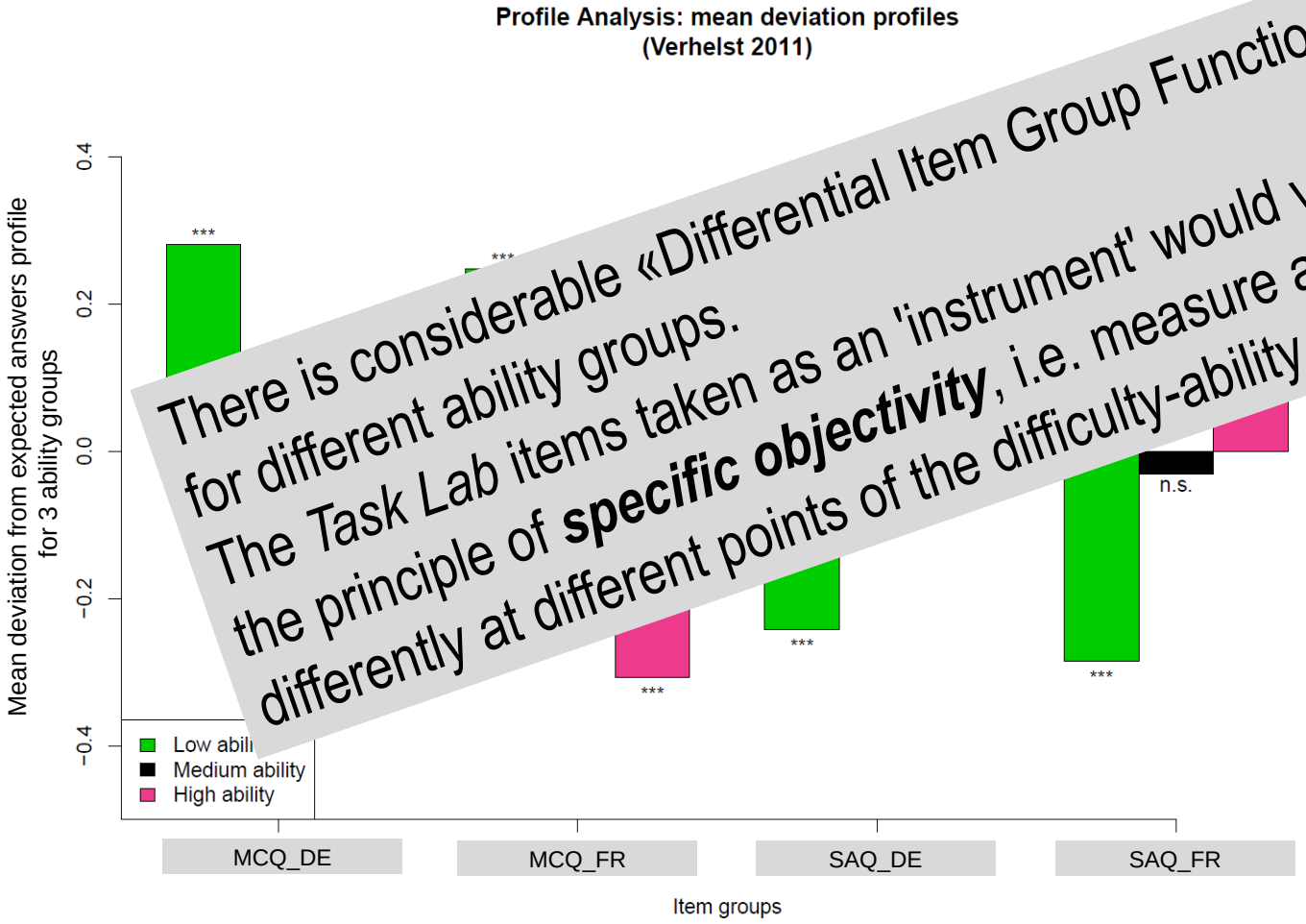
Item difficulty per item group: *Profile Analysis*



- Low-ability students score higher on MCQ items than the overall Rasch model predicts.
- High-ability students score higher on SAQ items than the overall Rasch model predicts. This effect is more pronounced when the language of rubrics and responses is the target language.

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

Item difficulty per item group: *Profile Analysis*



There is considerable «Differential Item Group Functioning» for different ability groups.
The *Task Lab* items taken as an 'instrument' would violate the principle of **specific objectivity**, i.e. measure ability differently at different points of the difficulty-ability scale.

…y students score …MCQ items than …Rasch model

…udents score …on SAQ items than the overall Rasch model predicts. This effect is more pronounced when the language of rubrics and responses is the target language.

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

Item fit analysis – a heuristic tool for quality assess.

Two well-known (but not uncontroversial) statistics:

Outfit and **Infit Mean Square**.

Outfit mean square summarizes the difference between the observed person scores on an item and the person scores predicted by the model on that same item.

The **infit mean square** weights extreme differences (e.g. very weak student succeeds on a very hard item) less than differences in the central region. It is often preferred. **Infit > 1.2** may indicate a problem (> 15% of unexpected variablity relative to the model). For more precise information check **ICC**!

	parameter	Infit	Infit_t	Infit_p	Infit_phol
1	lv.saq_T01_1_ls	0.9091704	-1.6102363454	0.1073462697	1.000000
2	lv.saq_T01_2_ls	0.9982843	-0.0011942772	0.9990471049	1.000000
3	lv.saq_T01_3_ls	0.9009748	-1.9418915319	0.0521502328	1.000000
4	lv.saq_T02_1_ls	0.8199686	-0.5014013947	0.6160886584	1.000000
5	lv.saq_T02_2_ls	1.0371765	0.3794236195	0.7043733137	1.000000
6	lv.saq_T02_3_ls	0.8867375	-0.8128006135	0.4163323837	1.000000

Visualizing item fit: the Item Characterstic Curve

Low discrimination, underfit & misfit

High discrimination: overfit

Un mail de Samuel

Links siehst du ein Mail, das Samuel an seinen Freund Benjamin schreibt.

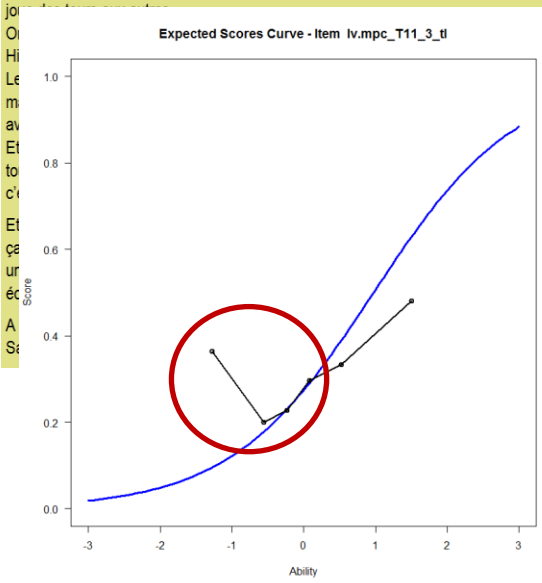
Kreuze die richtige Antwort an.

De : Samuel
A : Benjamin
Date : 2 avril, 8h15
Objet : Hello from Lausanne !

Salut Benjamin !

Ça va ? Moi je vais très bien : mon test d'anglais d'hier a été annulé !
Je te raconte ce qui s'est passé hier. C'était le 1er avril. En Suisse, le 1er avril, on

1. **Quel est le thème du mail de Samuel ?**
   - ☐ L'école en Suisse.
   - ☐ Le test d'anglais du jour d'avant.
   - ☐ Les blagues qu'on lui a faites.
2. **Quelle était la blague préférée de Samuel ?**
   - ☐ Le test d'anglais.
   - ☐ Le sel dans le thé.
   - ☐ Le papier toilette dans les chaussures.
3. **Qu'est-ce que Samuel fera le 1er avril prochain ?**
   - ☐ Il ne sait pas encore. Il demande à Benjamin quoi faire.
   - ☐ Il va mettre le sel dans la boîte à sucre.
   - ☐ Il va faire attention aux blagues d'autres.



Un chat entre Timo et Martin

Martin *:* Salut Timo ! Ça va les vacances ? Toujours dans la grande capitale japonaise ? A bientôt, Martin.

Timo: Salut ! Oui, on est toujours là, c'est super !!! C'est vraiment incroyable : plus de 13 millions de personnes habitent dans cette ville ! J'adore ça ! Hier on est allé sur la tour « Tokyo Skytree ». Là, on a une vue magnifique sur toute la ville. C'était trop cool !

On mange souvent dans les restaurants : les sushi, les « ramen » (des pâtes), et les glaces sont trop bons. Et il y a des automates avec des boissons ou même « l'udon » (une soupe) partout. Et puis, aujourd'hui, à midi, on a mangé du « takoyaki » dans un restaurant. Ça n'avait pas l'air bon, mais
c'était délicieux ! Maintenant,

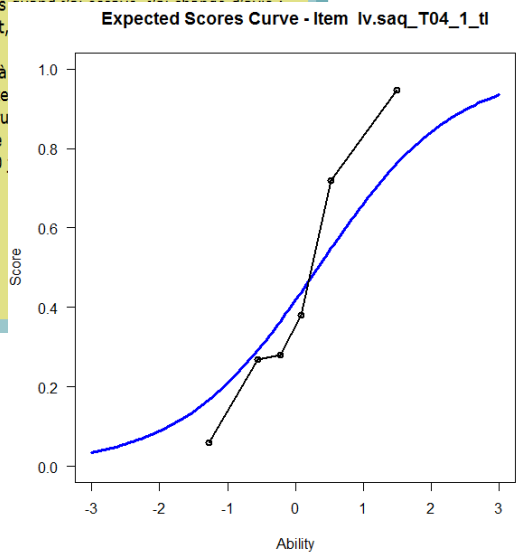
On reste encore deux jours à
une statue en bronze gigante
Là-bas, on a beaucoup de tru
du karaté ! Je t'enverrai une
On rentre en Suisse dans 10

Links siehst du einen Chat zwischen Timo und Martin.
Dazu stellen wir dir drei Fragen.

1. Frage:

Was ist das Thema von Timos Nachricht?

Schreibe deine Antwort *auf Deutsch!*



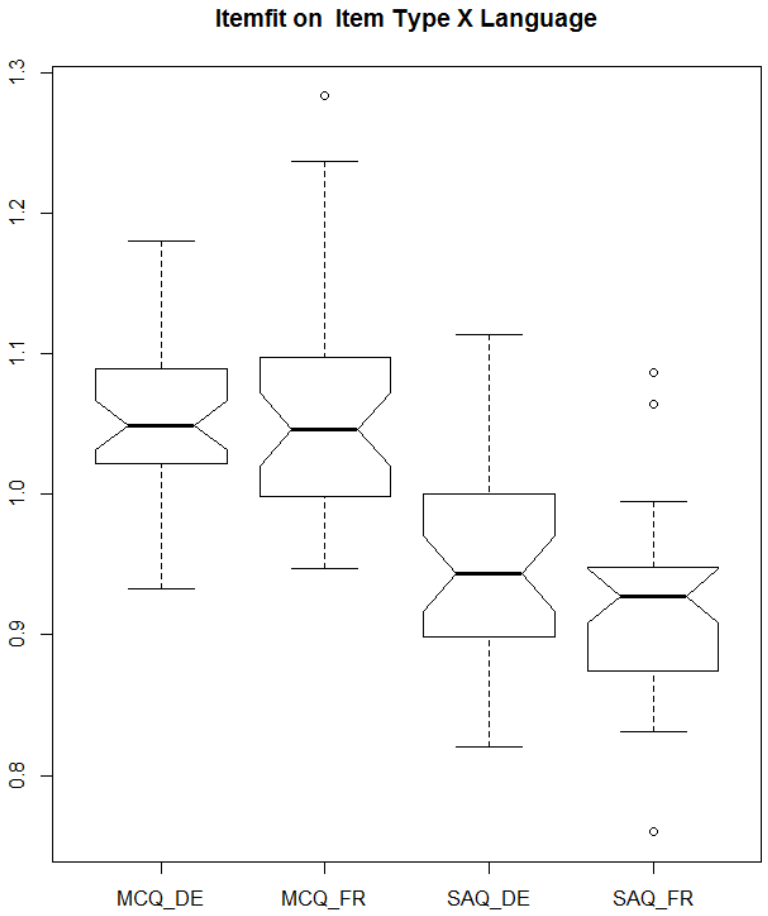
Misfit analysis if different item groups are present as in our case

General tendencies:

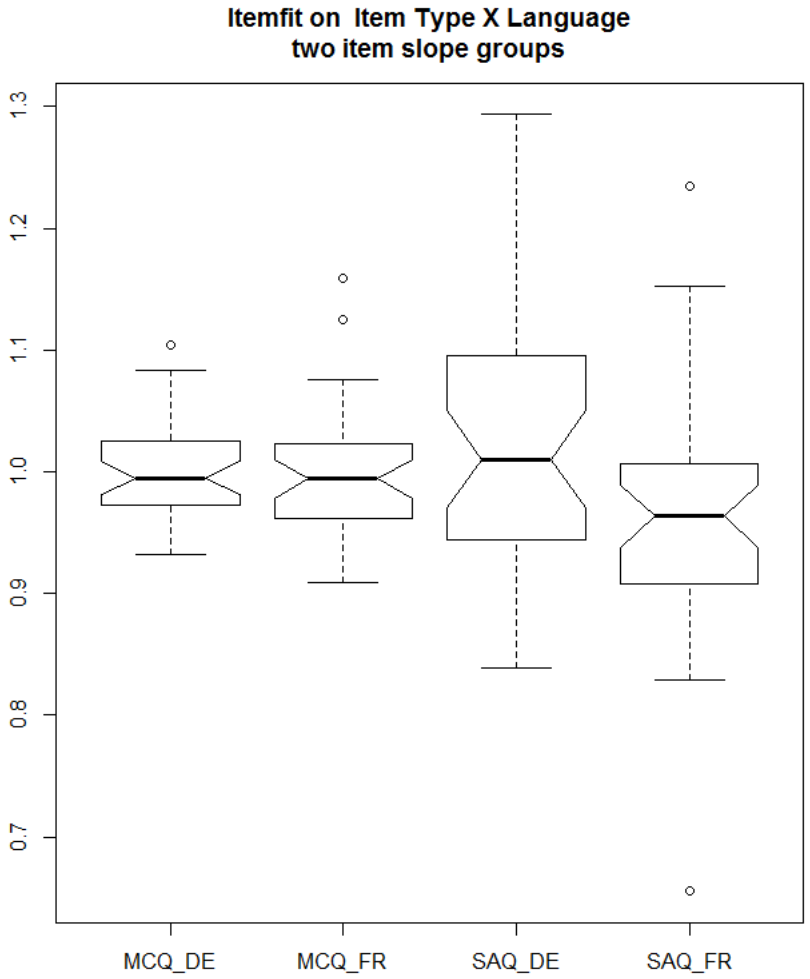
MCQ items generally have infit values > 1 (*underfit*) because they discriminate less.

SAQ items generally have infit values < 1 *overfit*) because they discriminate more.

=> **Be careful**: Infit values are misleading!
Carry out **separate** fit analyses for 2 item groups or **define slope groups**



Misfit analysis based on two different slope groups



When two slope groups are defined, the infit statistic can be used to detect problems causing misfit.

Check items with high infit. It is usually a sign that *construct-irrelevant variance* has an influence.

On the low infit side, only very low fit values are worth inquiring.

Detecting misfit in items

	parameter	Infit	Infit_t	Infit_p	Infit_pholm
2	lv.saq_T01_2_ls	1.1034843	1.1016794	0.2706010881	1.00000000
5	lv.saq_T02_2_ls	1.1767505	1.4471564	0.1478531316	1.00000000
7	lv.saq_T03_1_ls	1.1209062	1.2526753	0.2103238886	1.00000000
16	lv.saq_T06_1_ls	1.1538618	1.9649232	0.0494231275	1.00000000
17	lv.saq_T06_2_ls	1.1642487	1.2992073	0.1938728133	1.00000000
24	lv.saq_T08_3_ls	1.2933722	3.4302113	0.0006031114	0.08684804
27	lv.saq_T09_3_ls	1.1859465	1.2954973	0.1951487337	1.00000000
31	lv.saq_T11_1_ls	1.1957725	1.5111295	0.1307554677	1.00000000
58	lv.saq_T08_1_tl	1.1519310	0.4698663	0.6384505627	1.00000000
67	lv.saq_T11_1_tl	1.2343125	2.0947628	0.0361920814	1.00000000
106	lv.mpc_T06_2_tl	1.1588436	2.1831165	0.0290272283	1.00000000
126	lv.mpc_T09_3_tl	1.1245595	2.3171081	0.0204978457	1.00000000
139	lv.mpc_T12_1_ls	1.1045118	1.9067241	0.0565563227	1.00000000
54	lv.saq_T06_3_tl	0.6553845	-1.3290684	0.1838254179	1.00000000

>1 reasonable answer?

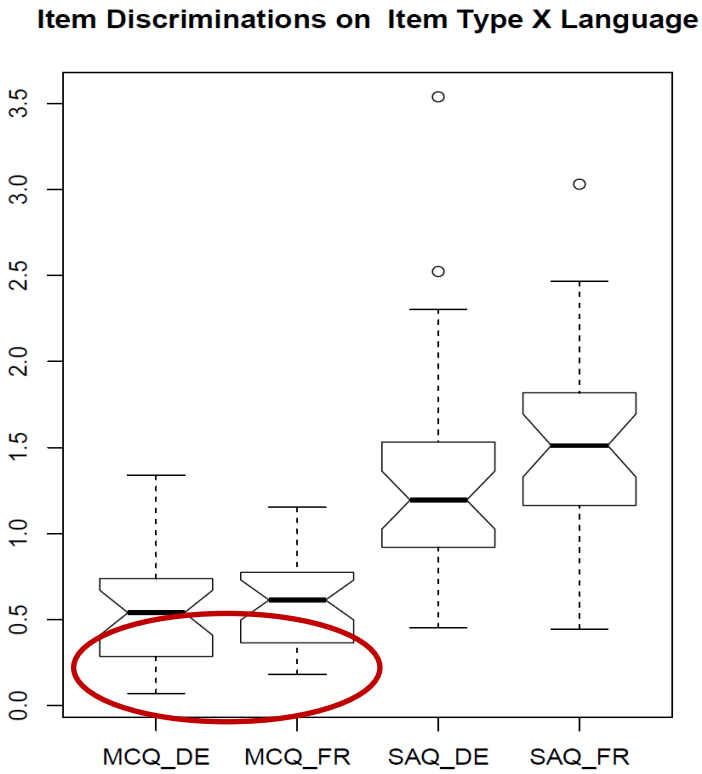
A technical solution for DIGF: a 2PL IRT model

The items of the four groups don't contribute equally to the measurement of the construct

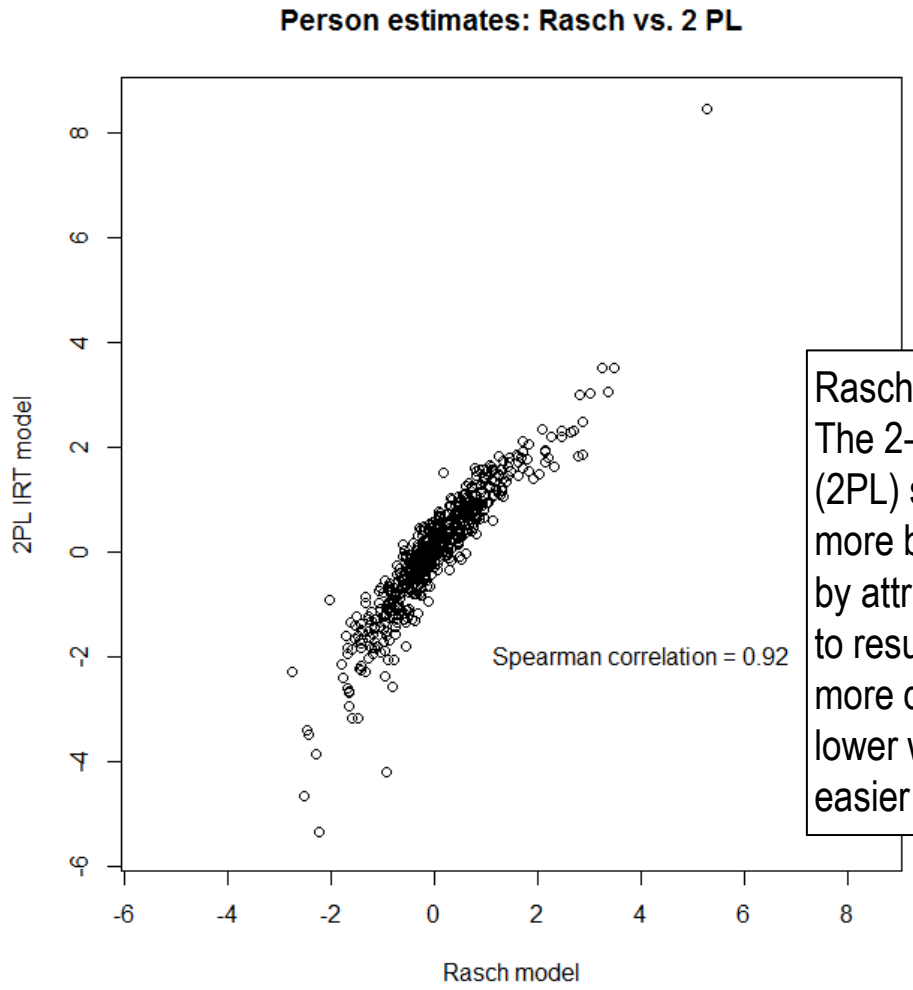
A 2PL IRT model takes this into consideration by **estimating a slope parameter (discrimination) for the individual items.**

=> Discrimination of MCQ items is generally much lower.
Some items hardly separate generally strong from generally weak students

=> 2PL model weights the scores on the items according to discrimination.



Person (WLE) measures: Rasch vs. 2 PL models



Rasch scale is shorter.
The 2-parameter logistic (2PL) scale differentiates more between the students by attributing higher weights to results on the generally more difficult SAQ items and lower weights to the generally easier MCQ items.

Person scores based on

Rasch	2 PL
8	4.84752067
5	3.19724245
8	6.12784231
21	23.36547804
16	17.06364283
9	6.82176131
1	0.92541771
7	3.31142503
11	9.50736777
9	6.89213626
6	3.55231858
13	10.81206406
17	11.64165249
9	6.67295301
5	2.44583516
5	3.68716835
9	5.04079237
10	6.98628406
8	8.54370435
9	4.72816822

DOES SUCCESS ON MCQ AND SAQ HAVE THE SAME PREDICTORS? – LATENT REGRESSION ON 2 DIMENSIONS

A 2-dimensional (per item type) model fits better than the 1-dimensional model

Mod.2PL.1Dim vs. Mod.2PL.2Dim (dev. Diff.): $\chi^2(10.85, 1)$, $p < 0.001$

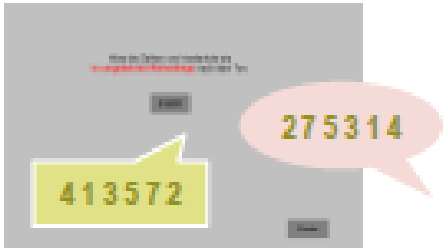
Results of latent ('error-free') regression on 2 dimensions:

Predictors	Predictor type	Dim 1 SAQ	Dim 2 MCQ
Gender: male	dummy	0.095	-0.115
Romance lang. background	dummy	0.512	0.112
Motivation: enjoyment	z-std.	0.181	0.109
Motivation: ought	z-std.	0.038	-0.013
Backward digit span	z-std.	0.150	0.145
Sight-word recognition	z-std.	0.159	0.145
Yes-No Test (recognise word)	z-std.	0.142	0.292
Segmentation task	z-std.	0.406	0.293
C-Test	z-std.	0.221	0.156

PREDICTORS USED

Student Questionnaire

- Gender
- (Rom, lang, background
- Motivation (enjoyment)
- Motivation (ought)

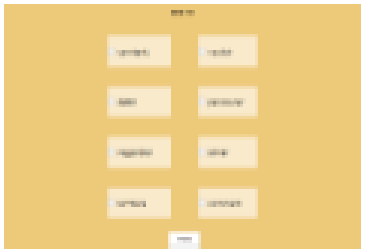


Backward Digit Span Task

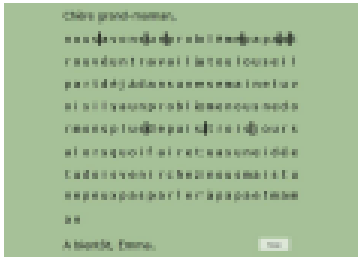
Working memory/ processing

Sight-word recognition

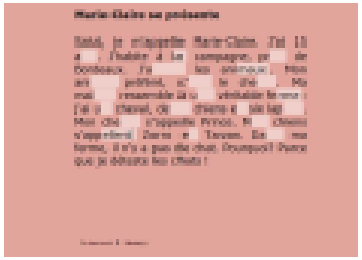
Word decoding (gestalt)



Yes-No Task
Vocabulary breadth (receptive)



Text segmentation
Morpho-syntax & integrative measure



C-Test
Integrative measure / written text reconstruction

Predictor matrix was completed through **imputation** using the Amelia II R package (max. 10.7 % missings)

Predictors	Predictor type	Dim 1 SAQ	Dim 2 MCQ
Gender: male	dummy	0.095	-0.115
Romance lang. background	dummy	0.512	0.112
Motivation: enjoyment	z-std.	0.181	0.109
Motivation: ought	z-std.	0.038	-0.013
Backward digit span	z-std.	0.150	0.145
Sight-word recognition	z-std.	0.159	0.145
Yes-No Test (recognise word)	z-std.	0.142	0.292
Segmentation task	z-std.	0.406	0.293
C-Test	z-std.	0.221	0.156

Some Observations

- **Known correlates** of better language knowledge predict success on **SAQ** items *particularly well*: Integrative measures, Motivation (enjoyment ≈ intrinsic), and a romance family language background (13.6% of sample).
- The strictly **receptive Yes-No** word recognition test is a better predictor for success on **MCQ**. The possibility of success through **guessing** may be a commonality (despite a correction for guessing made on YNT).

Discussion

- It seems desirable to be able to pinpoint more **specific component knowledge & skills – which ones?**.
- What **item characteristics** should be taken into account for a rigorous **person-item explanatory model?**

Contact

Peter Lenz

Institute of Multilingualism

e-mail: Peter.Lenz@unifr.ch

www.centre-multilingualism.ch