Appeared in: Huhta, Erickson, Figueras (2019):
Developments in Language Education,
University of Jyväskylä + EALTA. 182-202.

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

Peter Lenz, Katharina Karges and Malgorzata Barras
Institute of Multilingualism, University of Fribourg &
University of Teacher Education Fribourg

1. Introduction

1.1 Background

From 2014-16 the Competence Centre on Multilingualism carried out the 'Task Lab' study to have more firm ground for the upcoming item development for a country-wide computer-based survey (system monitoring) of sixth graders' receptive skills in their first foreign language learned at school.

The main emphasis of the 'Task Lab' project was on the exploration of specific design options for the French reading test. These options included, first, item type (test method) – short open answers (SA), multiple choice (MC) and matching (MTC); second, the language of the items – French, the target language, or German, the language of schooling. Although seemingly formal features, we suspected the choice of item type and language to have an influence on what is actually tested and, therefore, what the test scale stands for. The present paper focuses on the comparability of SA and MC items testing French reading skills in the CEFR A1 to A2+ range of levels.

1.2 Literature

There is a longstanding tradition of investigating test method effects in the field of education. Due to the wide use of MC items, they are often under scrutiny. Rodriguez (2003) performed a meta-analysis on the construct equivalence of MC and constructed-response[24] (CR) items. For this purpose, he formally summarised 56 correlations between MC and CR-based results. Almost 60 percent of these correlations stemmed from studies in language arts, the rest from various other fields. The main finding was that whenever item writers intended to tap the same construct using both item types, the test results on items of the two types were highly correlated. The average correlation

[24] 'Constructed response' stands in opposition to 'selected response'. Constructed responses may be short or extended. Multiple choice is one of several selected-response formats. In the following, when referring to studies, we use the terms for item formats that are used in these studies, e.g. 'open-ended items' for (a type of) constructed-response items.

turned out highest in studies that used stem-equivalent items, i.e. items using the same question, instruction or beginning of sentence to initiate the response process (MC or CR). The disattenuated correlation across studies amounted to 0.95 in this case.

In reading assessment, there is a tendency to use MC items to test lower-level skills and CR items to give test takers the opportunity to demonstrate higher-level reading skills such as global inferencing or reflecting on content. Obviously, in such circumstances construct equivalence between MC and CR items cannot be expected. Rauch and Hartig (2010) applies a two-dimensional latent regression model to investigate construct-differences between a general (L1) reading dimension, based on all MC and open-ended (OE) items, and a specific reading dimension, based on unaccounted variance from the OE items. The regression analyses showed several differential associations of social, cognitive and linguistic predictor variables with the two reading dimensions. However, it was impossible to attribute these findings with any certainty to item type because test method and construct(s) were confounded due to test design.

Ozuru and colleagues (Ozuru, Best, Bell, Witherspoon, & McNamara, 2007; Ozuru, Briner, Kurby, & McNamara, 2013) investigated construct equivalence of MC and OE items by using the same questions for both formats on the same tests. In all experiments described, the participants (U.S. college undergraduates) started by answering the OE items and then proceeded to the set of corresponding MC items. They were not allowed to go back and forth between OE and MC items. In the 2007 study (two experiments), half of the participants answered the OE and MC questions without having the opportunity to go back to the text passage, the other half could use the passage in the answering process. The results showed different test method effects depending on the availability of the passage. When the passage was unavailable, the effect size of the correlation between the scores based on the OE and the MC items respectively was large while it was only modest (and the correlation statistically nonsignificant) when the text was available during the response process. In the latter case, construct equivalence is doubtful. Ozuru et al. (2013) focuses on reading processes that might explain differential success on OE and MC items. While reading the text passage, the participants had to explain the meaning of some highlighted sentences in the text, thereby integrating information from different locations. After reading the passage, they first answered a series of OE items, then the corresponding series of MC items. When compared with the scores on both item types, the quality of the sentence explanations was moderately correlated with success on the OE items but not the MC items. The authors conclude that OE items measure more sensitively the quality of active generative processing during comprehension, while MC items tap in more passive recognition processes.

In L2-related research, Shohamy (1984) undertook an early systematic investigation of the effects of item design features on the measurement of the construct. She produced a total of eight English reading test versions by varying text prompt (two topics), test method (MC or OE items) and the language of the questions and options/answers (L1 Hebrew, the participants' stronger language, or L2 English). The first part of the test was the same for all participants: Eight identical questions relating to the same eight passages served as a link between the test versions. Two main findings were that, on average, MC and L1 items were easier than OE and L2 items and that the effect of the harder conditions was stronger among less English-proficient students in the wide proficiency range represented in the sample.

1.3 Approach of the present study

Somewhat similarly to the Shohamy study, we also investigated test method effects by systematically varying the language of the items and the type of response. In addition, we collected information on precursor skills of reading as well as data from integrative tests that are usually strong correlates of reading comprehension.

The purpose of the present paper is to explore the equivalence of SA and MC items as test-methods for measuring the L2 French reading proficiency of young learners in the A1-A2+ level range. We refrain from the language-of-the-items issue and focus on item format effects using quantitative data from the main survey.

We investigate the following research questions:

a) Are there any systematic differences in the psychometric functioning of the SA and MC items used?

If there are differences –

b) how dramatic are they for the quality of a measurement instrument consisting of these item types?
c) in what way do the constructs represented by either of the two item types differ?

2. Method

2.1 Reading task development

We created the reading tasks for the study around 18 different text inputs. Twelve text inputs served as a basis for 36 (12 x 3) short-answer (SA) and 36 stem-equivalent multiple-choice (MC) items. The six remaining text inputs were used as a basis for 18 matching (MTC) items. Each of the SA, MC and MTC items came in two language versions, one with items in German[25] (the students' language of schooling), the other with items in French (the target language to be assessed). So, the complete test consisted of 144 SA or MC and 36 MTC (i.e. a total of 180) item versions[26].

[25] This means that all components of the items were in German, except for the text passages: in the case of SA, the question and the expected open answer; in the case of MC, the question and the three options; in the case of MTC, the question.

[26] As the matching items are quite different from the other items (no stem or content equivalence intended), we did not include them in this study.

Table 1. The task and item versions on the French reading test.

	short-answer	multiple-choice	matching
German items	French text inputs numbers 1-12 (12 x 3 items)	French text inputs numbers 1-12 (12 x 3 items)	French text inputs numbers 13-18 (6 x 3 items)
French items	French text inputs numbers 1-12 (12 x 3 items)	French text inputs numbers 1-12 (12 x 3 items)	French text inputs numbers 13-18 (6 x 3 items)

The tasks were designed as transfer-of-learning tasks for students who are all learning French in the same curricular region and with the same core of textbook materials. Task development followed a set of guidelines concerning types of reading (Urquhart & Weir, 1998) range of topics and the number of items per text input.

The writing of the SA and the MC items was marked by the decision to have all items in four versions by varying item language and item format. For example, the text input had to contain text references for the MC distracters and correct choices, even when the items were of the SA type. Conversely, all questions needed to be formulated precisely enough to narrow down the number of correct answers to one to have reliable short-answer items.

From a previous project, we had a corpus of all textbook materials available which the students had (at least potentially) worked with. Based on the corpus, we compiled a word frequency list, which served as a basis for component skills tests (e.g. vocabulary). We also used the corpus to check the familiarity of vocabulary items in text input and items.

2.2 Pre-piloting of reading tasks

The reading tasks were implemented in CBA ItemBuilder (DIPF & Nagarro IT Services, n.d.), a server-based test environment. Quality assurance was a major concern all along the test development and administration process. In a first phase, prototype reading tasks underwent usability testing to set the relevant screen design parameters and to improve functionality. After moderation by native speakers and experts, the reading tasks were pre-piloted by eight sixth grade classes[27]. We collected statistical routine information on the items and also a sample of the short answers we had to expect from the SA items (the “outcome space” according to Wilson, 2005) in order to prepare a coding key. In addition, we did one-on-one stimulated recall interviews (Gass & Mackey, 2000) with 34 students to collect evidence on the cognitive validity (Field, 2012) of our items.

[27] In Switzerland, sixth grade is the final grade of primary school. The great majority of students are between 11 and 12 years old. Average class size is slightly below 20. In primary school, the students of a class are normally taught together in all academic subjects. In the region we did our study, French teaching starts in third grade. From third to sixth grade, the average number of weekly French lessons amounts to 2.5.

2.3 Component skills assessments and integrative language tests

In addition to the reading tasks, we selected and developed a series of relatively short assessments of known correlates of reading comprehension, expecting that we could, among other things, use these additional measures to explore the construct or constructs embodied by the different types of items.

We settled for the following assessment instruments:

Table 2. Measurement instruments for component and reading task-related skills.

Test instrument		Cognitive component(s) targeted
1	Backward digit span task: repeat orally, in reverse order, a series of digits of increasing length	Working memory capacity (processing)
2	Read aloud French pseudowords	Phonemic awareness, French decoding/grapheme-to-phoneme conversion
3	Sight-word recognition	French sight-word reading; automatised receptive knowledge of whole written word forms
4	Yes/No Test	Breadth of French receptive vocabulary
5	Text segmentation (identifying word boundaries in text)	Receptive knowledge of French vocabulary and syntax; text segmentation accuracy
6	C-Test (integrative written gap-filling task)	French word/sentence/text comprehension in conditions of reduced redundancy; lexically and grammatically accurate word writing

Some brief comments on these assessment instruments:

1) Success on the backward digit span (BDS) task is a well-known predictor of success in reading, which, however, does not necessarily imply a substantive causality between working memory capacity and success in reading (Alderson et al., 2015). The BDS task is a simple and widely known working memory capacity (WMC) test that includes a secondary processing task (repeating the input backwards). Secondary processing also takes place when readers manipulate verbal information in their working memory. WMC accounts for a significant portion of variance in general intellectual ability (Conway et al., 2005). Our final BDS test included ten items, each two to six digits long, two of each length. Our students heard the ten series of digits in German on the computer headphones and repeated them orally.

2) The pseudoword reading aloud task assesses a learners' phonemic awareness and decoding skills in a language. Beginning readers rely a great deal on decoding. According to Geva and Siegel (2000) phonological and orthographic processing are

involved in decoding. The test uses pseudowords to make sure that grapheme-to-phoneme conversion actually needs to take place. We created the 20 items we needed to suit our student population with the help of a corpus-based web tool (New & Pallier, 2001).

3) The sight-word recognition task is a measure of sight-word reading, an advanced, automated form of word recognition that is crucial for fluent reading (cf. Alderson et al., 2015; Sabatini, Bruce, & Steinberg, 2013). We presented the students 20 French words (two to eight letters long). The words were visible on screen for just 80 milliseconds. Then the test takers spoke the words they had seen into a microphone.

4) The Yes/No Test (or Vocabulary Size Placement Test) (Meara & Buxton, 1987) is a well-known measure of receptive vocabulary breadth that is often used as a placement test. A Yes/No Test consists of real words and pseudowords. Test takers declare for every item they encounter whether they know it as a word of that language or not. The score on the pseudowords provides a false-alarm rate that can be used to correct the score on the existing words for guessing.

Our Yes/No Test consisted of 21 French words from the textbook corpus and 19 pseudo-French words that were generated in the same manner as the pseudowords for decoding.

5) The segmentation task is considered a combined (receptive) grammar-vocabulary task. In the DIALUKI study (Alderson et al., 2015), segmentation tasks proved to be strong predictors of reading proficiency. In a text segmentation task, test takers need to mark the word boundaries in one or more texts without blanks between the words.

6) The C-Test (Klein-Braley, 1985) is an integrative language test format whose strong association with language proficiency measures was established in many studies (e.g. Eckes & Grotjahn, 2006; Harsch & Hartig, 2016). A C-Test consists of a series of different texts (often four or five), in which, starting with the second word of the second sentence, the second half of every second (suitable) word is missing while the final sentence remains intact. Unlike the component skills tests, the C-Test involves written production of French, which is also the case for SA items. We used a C-Test from the Lingualevel collection (Lenz & Studer, 2007) with a total of 60 gaps.

We had the students of two classes do the six tasks described. Fourteen students from another class did the oral tasks (1-3), as well. In addition, they talked them through with a researcher in a one-on-one setting. The information gained in this manner helped to improve and customise the instruments.

2.4 Student questionnaire

The assessment instruments for the Task Lab study were accompanied by a short student questionnaire on social and language background, reading habits, language learning motivation and perceived characteristics of the language teaching the students were experiencing. The items used in the questionnaire came from other questionnaires we had used in previous studies and were not pre-piloted again.

2.5 Piloting

For piloting the main survey, all data collection instruments (i.e. a brief questionnaire, the instruments presented in Table 2, and the reading tasks) were deployed on the CBA ItemBuilder system. This software allows access to customised test sets residing on a remote server by means of a current web browser. The goal for the reading test was to confront every student with a balanced sample of the existing task variants while never confronting the same student with the same task in two language or item format variants. Due to a time limit of 90 minutes for all written tasks, including the questionnaire, we confronted each student with a selection of 13 out of 18 available reading tasks (i.e. 39 items). A total of 24 different test sets was used. In these, the reading tasks appeared in different item format and language variants and positions. Overall, 119 sixth graders participated in these trial runs for the main survey.

2.6 The main survey

Overall, 609 sixth graders from 33 self-selected classes in 13 different schools located in German-speaking Switzerland were involved in the main data collection. All students were to do all tasks in the manner described for the piloting. Integral classes worked in the school's computer lab for 90 minutes, then went back to normal schoolwork. During the following lessons, small groups of students came to a separate room where they did the tasks with an oral component.

3. Results

We used the data obtained in the main study in various ways to identify differences, if they exist, with regard to a) the quality of the two item types as measurement instruments (3.1), and b) the constructs embodied by the two item types (3.2). While section 3.1 performs item analyses on the reading data, section 3.2 uses the results on the component and integrated measures tests as predictors of reading proficiency, measured separately by MC or SA items.

3.1 Format effects among the reading items

3.1.1 Data preparation and item selection

In a preliminary step, the answers to the SA items had to be coded. The provisional instrument from the pilot needed further refinement. Initial efforts to use partial-credit scoring were finally abandoned in favour of quasi-objectively applicable guidelines for dichotomous scoring. Interrater reliability was not evaluated statistically as all answers were double-rated and all issues discussed in short intervals until mutual agreement between the two raters and a third person was reached.

Before investigating differences in the functioning of the SA and MC items for the present study, we first selected a set of quality items. For this purpose, all items were scaled[28] using the Rasch model and the 2PL (2-parameter logistic) IRT model[29]. There were between 83 and 154 (mean = 117.9) responses available per item variant. These relatively modest numbers are owed to the fact that each of the 609 students only solved a subset of 30 of the available 144 SA or MC item variants[30]. A total of 46 item variants was removed from the present analysis for various reasons (e.g. low discrimination, misfit). In order to diagnose misfit, mainly visual inspection of the empirical versus model item characteristic curve under the Rasch and the 2PL model was used, complemented by an inspection of the actual items and the answers provided. Whenever an item variant was excluded, its counterpart in terms of format (and language) was also excluded so that, now, for every MC item the corresponding SA item is also in the final set of items (and *vice versa*). The final set contains 98 item variants relating to 10 different passages; 588 students (290 females, 298 males) contributed usable responses. The Expected A Posteriori (EAP) reliabilities (Adams, 2005) amounted to 0.74 for the Rasch scale and 0.78 for the 2PL scale.

3.2 Analyses

A comparison of the difficulties of the items in both formats in the 2PL model reveals considerable differences.

Table 3. Mean difficulties of SA and MC items.

	short answer (SA)	multiple-choice (MC)
mean difficulty (logits)	1.349	-0.1256
SE (logits)	0.218	0.106

The difference is statistically significant on a paired t-test ($t = 7.67$, $df = 48$, $p < 0.001$), the standardised effect size ($d = 1.10$) large according to Cohen's rule-of-thumb interpretation. The findings for the item slopes (item discriminations in the 2PL model) are similarly clear:

[28] All statistical analyses were carried out using R software, for IRT the 'TAM' package (Kiefer, Robitzsch, & Wu, 2015).

[29] The Rasch model assumes that all items discriminate equally between weaker and stronger students. The 2PL model, however, estimates an individual discrimination parameter (the slope) for each item (Embretson & Reise, 2000). In order for the 2PL slope estimates to be stable over time, much larger numbers of test takers would be needed. However, generalisation of these parameters is not an issue here.

[30] Since four item variants were always based on the same question, the maximum workload would have been 36 items. Time limits made further reduction necessary.

Table 4. Mean slopes of SA and MC items under the 2PL model.

	short answer (SA)	multiple-choice (MC)
mean 2PL slope	1.535	0.657
SE	0.091	0.041

A difference of 0.878 is statistically significant on a paired t-test (t = 9.26, df = 48, p < 0.001), and the effect size (d = 1.32), again, is large. A difference in slope (i.e. discrimination) between SA and MC items is not unexpected considering what it generally takes to answer items of either type. In the case of SA items, it is not enough to understand and answer a question – the answer also needs to be formulated and written down (in German or French, depending on the item variant), otherwise comprehension remains unnoted. In the case of the purely receptive MC items, better students have fewer opportunities to prove that they actually are better. The theoretical 33% chance of guessing the right answer mitigates the power of an item to discriminate between weaker and stronger test takers and so does the fact that comprehension can be documented by simply ticking a box.

If person measures are produced based on a 2PL model, the slope or discrimination parameter is used to weight the scores on the individual items. So, getting an item right or wrong, counts more if the slope of an item is steeper. In the present case, the average SA item would contribute more than twice as much as the average MC item to the weighted person scores.

The frequently used Rasch model assumes equal slopes and therefore weights every item equally. Consequently, the raw score (number of correctly solved items) is considered a sufficient statistic. Test takers who have a higher total score on the same test, no matter which of the items they solved correctly, have higher ability according to the model. In addition to this principle of sufficiency of the raw score, Rasch (Rasch, 1977) postulates the related principle of specific objectivity as a fundamental property of the Rasch model: Any sub-sample of items from this test would classify any sub-group of test takers in the same order. From a Rasch measurement perspective, our findings regarding the two item types indicate a (undesirable) case of differential item group functioning. In practice, differential item or item group functioning is commonly observed due to person or item groups that have something in common others do not have. Profile Analysis (Verhelst, 2011; Yildirim, Yildirim, & Verhelst, 2014) provides the statistical means to evaluate the strength and significance of such effects.

Our statistic of interest was the mean deviation profile for several ability groups that we formed along the common Rasch scale constructed from our SA and MC reading items. An individual deviation profile is calculated as follows: The expected score, based on the Rasch model, on the completed items of each item group (SA or MC) is subtracted from the observed score on the items of each group. The differences on all item groups (here two) form the deviation profile. The mean deviation profile is an aggregation of the individual deviation profiles of the test takers per ability group. For our analysis we defined three ability groups based on the Rasch scale: a middle group including person scores +/- 0.5 SDs around the mean, and the two groups left and right of this band. The actual group sizes were 161 (weakest group), 257 and 155 students (28%, 45%, 27%). 15 students were excluded from the analysis because they had either extreme scores or no data on one of the two item types.

The resulting mean deviation profiles show highly significant deviations from the Rasch model-based score predictions for the lowest and the highest-scoring groups (cf. Table 5).

The results for the three ability groups show how much the average of the observed scores differs from the average of the expected scores in each item group. The results for the two item types add up to zero in each ability group. The least-ability group scored significantly higher than predicted by the model on the MC items while the highest-ability group scored significantly higher than expected on the SA items. Evidently, the contrast between these two ability groups on the two item types is even larger than the difference between the observed and the expected mean for each group.

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

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	-0.770	0.770	0.096	-8.056	< 0.001

The above findings show that the assumption of specific objectivity is not appropriate for our set of items, nor is the test score a sufficient indicator for a person's ability. Depending on the sub-sample of items (esp. types of items) they are confronted with, the Rasch model may classify test takers in different ability groups either too low or too high on the latent ability scale.

3.2.1 Exploring construct-equivalence of SA and MC items

In order to explore potential systematic differences in the demands the items in both formats make, we used the results of the component skills and integrative tests to find associations between the constructs they embody and the constructs underlying the SA and MC-based reading tests (similarly: Rauch & Hartig, 2010). For this purpose, we first scaled the reading data using yet another IRT model, and prepared the scores from the different component skills and integrative measurements for further analysis. Then, we combined them with other (i.e. structural and questionnaire) variables in a single dataset, and performed multiple imputation on this dataset. Multiple imputation produced a series of complete datasets that could easily be used in multiple regression analysis to explore associations between the component skills or integrative tests (independent variables) and reading comprehension through SA or MC items (dependent variables).

3.3 Data preparation

In order to suit the purpose of this part of the study, the SA-based and the MC-based reading scores were additionally scaled using two-dimensional Rasch analysis (Reckase, 2009). Dimension 1 (EAP reliability 0.74) was based on the SA items, dimension 2 (EAP reliability 0.69) on the MC items[31]. The latent correlation between both dimensions amounted to 0.91, suggesting closely related constructs. WLEs (Warm's weighted likelihood estimates, Warm, 1989) were output as person estimates for subsequent use.

For the backward digit span task, we defined the score as the length of the longest string of numbers the students correctly repeated backwards. The maximum string-length metric (ML) is one of two metrics Woods et al. (2011) recommend based on their comparative study.

Coming up with a coding scheme for the decoding task proved challenging. We finally settled on 37 different syllables as our items. A single rater coded them once. The Rasch scale produced had an EAP reliability of 0.86. Again, WLEs were estimated as person measures.

From the 20 sight-word recognition items we selected 14 for coding, the six remaining being too easy. We managed to apply partial credit scoring quasi-objectively. The items were Rasch scaled (EAP reliability = 0.78), and, again, WLE person estimates were produced.

When Yes/No Test scores are used in practice, the number of 'yes' on the existing-word items is usually corrected by the number of 'yes' on pseudoword items (false-alarm score or rate) (cf. Huibregtse, Admiraal, & Meara, 2002). Without a correction, a test taker could attain the maximum score by simply choosing 'yes' for every item. Following a recommendation by Harsch & Hartig (2016), we constructed separate measures for the 21 words and the 19 pseudowords using a two-dimensional Rasch model (EAP reliabilities: words 0.77; pseudowords 0.70). For subsequent analyses, two WLE scales were output.

The text segmentation scale was produced by counting the correct segmentations in each text (after some exclusions). We standardised both score scales, added the resulting values and standardised the sums again to get the final standardised score.

In the case of the C-Test, each of the three texts was treated like a polytomous item with up to 17 categories after collapsing a few categories with too sparse data. The items were Rasch scaled (EAP reliability = 0.70), and WLEs were produced as person measures.

To prepare for data imputation, all test scores and scales were merged into a single dataset and complemented with indicator variables (e.g. students' class membership) and variables from the student questionnaire covering topics such as reading habits and motivations for learning French. On this dataset, we performed

[31] Thanks to this split, the problems with the Rasch scale detected in the previous section are not an issue here.

multiple imputation using the R package 'mice' (multivariate imputation by chained equations, Buuren & Groothuis-Oudshoorn, 2011). Our data imputation had two objectives: first, replacing missing data with plausible data, and second, accurately representing person measures containing measurement error (here the WLEs). For the purpose of the present study, 240 imputed datasets were produced, from which we used 40, i.e. every sixth set, in the data analysis[32]. All statistical analyses based on imputed datasets need to be carried out 40 times[33] independently. The results are combined according to Rubin's rules (Rubin, 1987).

3.4 Multiple regression analyses

The intercorrelations of the cognitive (backward digit span) and the various language-related predictor and criterion variables afford an overview of existing associations between variables.

Table 6. Correlations between cognitive and ling. variables (mean correlations from 40 imp. sets).

	De--	S-w	Y/N	Y/N	Y/N	Text	C-	Read.	Read.
	coding	recog.	words	pseud.	diff.	segm.	Test	SA	MC
Backward digit span (z)	0.18	0.28	0.05	-0.11	0.18	0.13	0.13	0.23	0.18
Decoding (z)		**0.77**	0.44	-0.23	**0.72**	0.66	0.67	0.51	0.48
Sight-word recognition (z)			0.42	-0.26	**0.74**	0.61	0.66	0.56	0.52
Y/N Test, words (z)				0.58	0.46	0.39	0.41	0.43	0.40
Y/N Test, pseudowords (z)					-0.46	-0.29	-0.30	-0.14	-0.25
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

The Pearson product-moment correlations presented in Table 6 are averages from the 40 imputation sets. The highest correlations (> 0.7) are highlighted in bold type. In most of these high correlations, the ad-hoc variable 'Y/N Test, difference (z)' is involved. This is the standardised difference between the standardised 'words' score and the standardised 'pseudowords' score of the Yes/No Test. Neither of these two shows strong associations with any of the other variables, but the difference does. This difference also shows the strongest association with either of the item type-specific reading scores. With the MC reading subtest it shares nearly 50% of the variance

[32] In each of these datasets, the former WLE measures differ slightly as they were drawn from the error distribution of the original WLE measures during imputation.

[33] We chose to work with such a high number of imputed datasets because of the partly only moderate scale reliabilities. More datasets can better reflect a wider error distribution (uncertainty) of the person measures.

(squared correlation $R^2 = 0.49$). Another noteworthy observation is the fact that the short and simple phonemic awareness/decoding test and the sight-word recognition test show similarly strong associations with both reading subscores as the more integrative text segmentation task and the C-Test.

In order to explore to what extent each of the component skills and integrative tests share variance with the SA-based and the MC-based reading test, we used stepwise multiple regression (cf. Sabatini, O'Reilly, Halderman, & Bruce, 2014). Building on a background model (containing some social and attitudinal variables), we first added the scores from the cognitive backward digit span task, then the fundamental, language and reading-related decoding and sight-word recognition tasks, and so on, until we finally arrived at the measure from the integrative C-Test. With every additional variable, we recorded the variance shared (R^2) between the updated model and the reading measures as well as the AIC. Since we chose a linear mixed-effects model (LMM)[34], we actually used a (marginal[35]) pseudo-R^2 based on Nakagawa & Schielzeth (2013), implemented in the R package 'piecewiseSEM' (Lefcheck, 2016). The numbers reported in Table 7 represent the mean of the results obtained from our 40 datasets. The AIC (Akaike Information Criterion) is an indicator of the fit of the model to the data (lower numbers mean better fit). The AIC statistic also penalizes higher numbers of predictor variables, i.e. it favours leaner models to some degree.

The predictors marked with an asterisk (*) in Table 7 all significantly improved the multiple regression models for both reading scales when they were first introduced in the given order. With the introduction of (correlated) predictors that capture similar but more comprehensive reading-related skills, the earlier predictors essentially lost this function and turned insignificant except for the best predictors (for details see Table 11).

[34] The LMMs were estimated with the 'lmer' function from the 'lme4' R package (Bates, Mächler, Bolker, & Walker, 2015). The pooling was done with the 'pool' function from the 'mice' package (Buuren & Groothuis-Oudshoorn, 2011).

[35] The marginal R^2 takes into account the variance shared between the fixed effects (background and predictor variables) and the reading scores but not the variance explained by the random effect (school classes).

Table 7. Results of stepwise, hierarchical multiple regression for SA and MC reading items.

	SA reading items				MC reading items			
	R^2	R^2 Change	AIC	AIC change	R^2	R^2 Change	AIC	AIC change
Background variables	0.157	-	6960.6	-	0.107	-	6864.4	-
Backward digit span (z)*	0.196	0.039	6933.8	-26.8	0.13	0.023	6834.2	-30.2
Decoding (z)*	0.335	**0.139**	6813.4	-120.4	0.262	**0.132**	6751.3	-82.9
Sight-word recognition (z)*	0.389	0.054	6780.0	-33.4	0.309	0.047	6704.0	-47.3
Y/N Test, words (z)*	0.417	0.028	6761.0	-19.0	0.337	0.028	6701.4	-2.6
Y/N Test, pseudowords (z)*	0.486	**0.069**	6734.0	-27.0	0.574	**0.237**	6530.2	-171.2
Text segmentation (z)	0.504	0.018	6690.7	-43.3	0.577	0.003	6526.8	-3.4
C-Test (z)	0.516	0.012	6679.5	-11.3	0.584	0.007	6528.5	1.7

Overall, we find that our predictors share more variance (R^2) with the MC reading measure than with the SA reading measure (58.4% vs. 51.6%). This is mainly due to the Y/N Test. When it is added to the model, it contributes an additional 26.5% of shared variances in the case of the MC-based test but 'only' an additional 9.7% in the case of the SA-based test. Text segmentation and the C-Test seem irrelevant as further predictors of the MC test result but keep improving the fit of the model (AIC) that predicts the SA reading score, even though the additional 3% of shared variance seems modest. A closer look (Table 8) at the two scores derived from the Y/N Test reveals that neither one of them is an extraordinary predictor by itself (as the moderate correlations already suggested) but that together they make a great and differential impact on our models.

Table 8. The two Y/N Test dimensions as predictors in reverse order (cf. Table 7).

	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**

If the Y/N Test is excluded from the set of predictor variables, a big difference between the models for the SA and the MC items becomes visible (Table 9, Table 10).

Table 9. Hierarchical multiple regression without the Y/N Test; text segmentation added, then C-Test.

	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

Table 10. Hierarchical multiple regression without the Y/N Test; C-Test added, then text segmentation.

	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
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

Concerning the SA-based reading test, the highly correlated (r = 0.84) text segmentation and C-Test measures together add an amount of shared variance to the model that is comparable to the contribution the Y/N test makes. With regard to the MC-based test, however, their explanatory power remains modest. Text segmentation and the C-Test together add a mere 6.2% of shared variance while the two Y/N Test measures add 26.5%. In the SA model, text segmentation appears almost redundant as a predictor when added second (Table10), in the MC model the same is true for the C-Test (Table 9). In the regression model for SA-based reading that comprises text segmentation and the C-Test but excludes the Y/N test (output not shown here), the C-Test measure improves the model significantly at the 95% confidence level (t = 2.30, p = 0.023) while the text segmentation score only just reaches borderline significance (t = 1.85, p = 0.065). In the corresponding model for MC-based reading, it is the reverse situation, just clearer: text segmentation is a significant predictor (t = 2.25, p = 0.026) while the C-Test is not (t = 1.16, p = 0.248).

In order to evaluate the differential effects the predictors have on SA and MC-based reading by means of inferential statistics, we estimated a joint LMM model in which the SA reading score and the MC reading score are implemented as repeated measures while all predictors interact with item type.

The left and right-hand panels in Table 11 provide extracts from the model output (i.e. fixed effects parameters of interest) of two equivalent multiple regression models. The upper left panel shows the cognitive and language-related fixed effects predictors for the SA reading score. The lower left panel adds the interaction effects representing the 'corrections' that have to be made to the predictors for SA-based reading in order to optimally predict MC-based reading. The p-values in bold type in the left panel indicate that the backward digit span score and both Y/N Test scores are significant predictors of SA-based reading. In addition, the significant interaction effects for the Y/N Test scores statistically endorse the observation that the Y/N measures are differentially associated with the two reading scores. Such a differential effect is not confirmed for the C-Test as a predictor. The right-hand panel displays the same results as the left-hand panel but takes the main effects for the prediction of MC-based reading as a point of departure. It shows that the Y/N Test measures are the only statistically significant predictors for MC-based reading in our set.

The effect size of the predictors can be directly inferred from these tables because we entered the reading measures on scales with a standard deviation (SD) of 100 while all predictors were coerced to a standardised scale (mean = 0, SD = 1). So, for example, a score of 0.5 instead of -0.5 on the 'words' dimension of the Y/N Test predicts an MC-based reading measure that is more than 1 SD (115.88 units) higher.

Table 11. Cognitive and language-related predictors of reading (SA-based vs. MC-based).

	SA reading measure					MC reading measure				
	coeff.	SE	t	df	p	coeff.	SE	t	df	p
Main effects (extract)										
Backward digit span (z)	8.62	4.22	2.04	113.1	**0.042**	4.42	5.56	0.79	60.0	0.428
Decoding (z)	-10.26	14.38	-0.71	37.2	0.476	-13.63	18.45	-0.74	29.6	0.461
Sight-word recognition (z)	4.32	17.65	0.24	34.2	0.807	-9.58	21.13	-0.45	29.0	0.651
Y/N Test, words (z)	63.08	28.58	2.21	29.0	**0.028**	115.88	33.80	3.43	24.8	**0.001**
Y/N Test, pseudowords (z)	-46.84	27.68	-1.69	27.3	**0.092**	-103.23	31.19	-3.31	24.4	**0.001**
Text segmentation (z)	11.55	12.35	0.93	52.2	0.351	4.41	13.43	0.33	47.0	0.743
C-Test (z)	14.12	15.81	0.89	44.9	0.373	-11.95	16.73	-0.71	42.3	0.476
Interactions: item type x predictors (extract from output)										
	'Correction' for MC measures					'Correction' for SA measures				
Backward digit span (z)	-4.20	5.95	-0.71	75.56	0.48	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**	-52.80	25.98	-2.03	33.8	**0.044**
Y/N Test, pseudowords (z)	-56.39	25.04	-2.25	31.8	**0.025**	56.39	25.04	2.25	31.8	**0.026**
Text segmentation (z)	-7.14	13.31	-0.54	56.2	0.592	7.14	13.31	0.54	56.2	0.593
C-Test (z)	-26.07	16.70	-1.56	46.7	0.120	26.07	16.70	1.56	46.7	0.121

4. Discussion

Psychometric item analysis of our stem-equivalent MC and SA reading items revealed large and significant differences in the functioning of MC and SA items with regard to difficulty and discrimination. Our study confirms Shohamy's (1984) findings in a similarly designed study that MC items are easier than SA items. We assume that the relatively high probability of 0.33 of guessing the correct MC option as well as the fact that the answering process involves fewer (or no) productive elements can serve as a general explanation.

The average discrimination of the SA items is more than double the discrimination of the MC items. Generally, if an item has low discrimination in relation to a scale, it has a weak relationship with the specific dimension defined by all the other items (Wilson & Hoskens, 2005). In our case, the difference is particularly remarkable because the complete test consists of an equal number of these two types of item so that, in principle, both could equally contribute to the common dimension (construct). Apparently, the contribution the MC items make, is diluted while the opposite is true for the SA items. It seems likely that a range of different test taking strategies can be applied in the case of the MC items, which tap less intensively and less uniformly into language-related resources than successful test-taking strategies for SA items do as they involve active understanding (no answers suggested) as well as active (productive) answering, both involving language resources.

The latent ('error free') correlation (Wu, Adams, Wilson, & Haldane, 2007) of 0.91 between the MC and SA reading dimensions estimated by the two-dimensional Rasch model is roughly equivalent to the average 0.95 disattenuated correlation of stem-equivalent SA and MC items in Rodriguez' (2003) meta-study. The magnitude of this correlation suggests that a test consisting of both items types is essentially uni-dimensional, i.e. it is appropriate to measure a common construct. We could not necessarily expect high correlation because in our test, text and items were concurrently present, a constellation that did not result in a significant correlation between SA-based and MC-based scores in the study by Ozuru et al. (2007).

However, Profile Analysis reveals that care needs to be taken when MC and SA items are used on the same Rasch scale because they are the source of significant non-uniform differential item group functioning[36]. Concretely, weaker students get a relative advantage from MC items while SA items benefit stronger students (and vice versa) when all items have the same weight, which is the case in Rasch measurement. This issue can be resolved most notably by applying the 2-parameter logistic IRT model instead of the Rasch model. The 2PL model uses item-specific weights and thus takes into account the strength of the relationship an item has with the latent measurement dimension. In the present case, applying the 2PL model instead of the Rasch model increases the variance of the person (WLE) scale by roughly 20%.

[36] The group of SA items and the group of MC items distort the measures to changing degrees (i.e. non-uniformly) along the ability scale.

Our attempt to shed light on differences between the SA and MC reading constructs by means of regression modeling has been somewhat successful. Vocabulary breadth is the best predictor of reading success on both reading scales, but it is even a significantly better predictor with regard to MC-based reading (Table 11). In the complete model for MC-based reading, no other predictor reaches statistical significance. In the model for SA-based reading, however, the backward digit span score also reaches significance. Also, when vocabulary breadth is replaced by the C-Test score, the total variance which the model shares with SA-based reading is only 2.1 percentage points lower (46.5% vs. 48.6%).

The explanatory power of a vocabulary test as such serves as no surprise because in the A1-A2 range of levels reading is usually found to be more of a language than a reading problem (Alderson, Nieminen, & Huhta, 2016; Alderson & Urquhart, 1984). It is tempting to speculate about commonalities between MC-based reading and the Y/N Test. Being successful on our Y/N Test implies the ability to recognise words already encountered before with some certainty. Existing words should not be missed while pseudowords should be discarded. Success on MC items similarly depends on an interplay between selection and deselection based on recognition. SA-based reading on the other hand comprises a productive element – formulating and writing an answer, be it in German or French. This may explain the observed association with the C-Test score. The fact that working memory capacity is a significant predictor of SA-based reading recalls Ozuru et al.'s (2013) finding that success on OE items depends on active generative processing of the input text.

4.1 Limitations and outlook

Our study concerns quite a specific population: German-speaking sixth graders learning basic French in a school context. Research involving learners with more advanced literacy skills, more elaborate test-taking strategies and higher L2 language ability might come to partly different conclusions. Also, the kind of statistical evidence we collected should be complemented by data from introspective research methods and particularly eye-tracking (Brunfaut, 2016; Brunfaut & McCray, 2015) to attain a richer understanding of test takers' actual reading and problem-solving processes when answering SA and MC items.

The set of predictors of reading ability is another point to improve. In order to pinpoint differences on item types and facilitate interpretation, there should be more measures capturing specific component or precursor skills (cf. Alderson et al., 2015).

In addition, the mostly statistical approach we chose takes little notice of individual item characteristics. It would be beneficial if the interplay of text and item characteristics was generally better understood. Item difficulty or discrimination modelling that goes beyond simple test method factors, could greatly enhance the knowledge base item developers can draw on. With respect to our data, so-called retrofitting of task factors will be a logical next step.

5. Conclusion

In our study, we used stem-equivalent SA and MC reading items to explore the equivalence of both item types with respect to scale quality and construct representation in a French-as-an-L2 context with young learners at an elementary ability level. We could show that SA items are, on average, better representatives of the measurement scale embodied by an equal number of SA and MC items. The presence of significant differential item group functioning confirmed through Profile Analysis suggests that simple Rasch scaling is problematic in the presence of SA and MC items because all items are weighted equally.

A latent correlation larger than 0.9 between SA-based and MC-based reading indicate that, overall, both tests methods measure the same construct. As expected for L2 readers at low language ability levels, receptive vocabulary knowledge is the best predictor of reading success, especially when reading is measured through MC items. The fact that working memory capacity is the only other concurrently significant predictor of SA-based reading, may indicate that more active generative processing is involved in answering short-answer items.

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, *31*(2–3), 162–172.

Alderson, J. C., Haapakangas, E.-L., Huhta, A., Nieminen, L. & Ullakonoja, R. (2015). *The diagnosis of reading in a second or foreign language*. New York: Routledge.

Alderson, J. C., Huhta, A. & Nieminen, L. (2016). Characteristics of weak and strong readers in a foreign language. *The Modern Language Journal*, *100*(4), 853–879.

Alderson, J. C. & Urquhart, A. H. (1984). Reading in a foreign language: A reading problem or a language problem? In *Reading in a foreign language* (pp. 1–24). London: Longman.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Brunfaut, T. (2016). *Looking ino reading II: A follow-up study on test-takers' cognitive processes while completing APTIS B1 reading tasks*. British Council. Retrieved from https://www.britishcouncil.org/sites/default/files/brunfaut_final_with_hyperlinks_3.pdf

Brunfaut, T. & McCray, G. (2015). *Looking into test-takers' cognitive processes while completing reading tasks* (ARAGs Research Reports Online No. AR/2015/001). British Council. Retrieved from https://www.britishcouncil.org/sites/default/files/brunfaut-and-mccray-report_final.pdf

Buuren, S. van & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 67.

Conway, A. A., Kane, M., Bunting, M., Hambrick, D. Z., Wilhelm, O. & Engle, R. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786.

DIPF, & Nagarro IT Services. (n.d.). *CBA ItemBuilder*. Frankfurt (Main). Retrieved from https://tba.dipf.de/de/infrastruktur/softwareentwicklung/cba-item-builder/cba-itembuilder

Eckes, T. & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, *23*(3), 290–325.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, N.J: L. Erlbaum Associates.

Field, J. (2012). A cognitive validation of the lecture-listening component of the IELTS listening paper. In L. Taylor & C. J. Weir (Eds.), *IELTS collected papers 2: Research in reading and listening assessment* (pp. 17–65). New York, Cambridge: Cambridge University Press.

Gass, S. M. & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Routledge.

Geva, E. & Siegel, L. S. (2000). Orthographic and cognitive factors in the concurrent development of basic reading skills in two languages. *Reading and Writing*, *12*(1), 1–30.

Harsch, C. & Hartig, J. (2016). Comparing C-tests and Yes/No vocabulary size tests as predictors of receptive language skills. *Language Testing*, *33*(4), 555–575.

Huibregtse, I., Admiraal, W. & Meara, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language Testing*, *19*(3), 227–245.

Kiefer, T., Robitzsch, A. & Wu, M. (2015). *TAM: Test Analysis Modules*. Retrieved from http://CRAN.R-project.org/package=TAM

Klein-Braley, C. (1985). A cloze-up on the C-Test: a study in the construct validation of authentic tests. *Language Testing*, *2*(1), 76–104.

Lefcheck, J. S. (2016). piecewiseSEM: Piecewise structural equation modelling in R for ecology, evolution, and systematics. *Methods in Ecology and Evolution*, *7*(5), 573–579.

Lenz, P. & Studer, T. (2007). *lingualevel: Französisch und Englisch. Instrumente zur Evaluation von Fremdsprachenkompetenzen* (1.). Schulverlag.

Meara, P. & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*(2), 142–154.

Nakagawa, S. & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142.

New, B. & Pallier, C. (2001). Lexique Toolbox - Nonmots, pseudomots, voisins - des outils pour la psycholinguistique. Retrieved January 4, 2016, from http://www.lexique.org/toolbox/toolbox.pub/

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* (Vol. 14, pp. 58–93). Copenhagen: Munksgaard. Retrieved from https://www.rasch.org/memo18.htm

Rauch, D. & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling*, (4), 354–379.

Reckase, M. (2009). *Multidimensional item response theory*. New York; London: Springer,.

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.

Sabatini, J. P., Bruce, K. & Steinberg, J. (2013). SARA reading components tests, RISE form: test design and technical adequacy. *ETS Research Report Series*, *2013*(1), i–25.

Sabatini, J. P., O'Reilly, T., Halderman, L. K. & Bruce, K. (2014). Integrating scenario-based and component reading skill measures to understand the reading behavior of struggling readers. *Learning Disabilities Research & Practice*, *29*(1), 36–43.

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

Urquhart, A. H. & Weir, C. J. (1998). *Reading in a second language: process, product, and practice*. London, New York: Longman.

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

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450.

Wilson, M. (2005). *Constructing measures: an item response modeling approach*. Mahwah N.J.: Lawrence Erlbaum Associates.

Wilson, M. & Hoskens. (2005). Multidimensional item responses: Multimethod-multitrait perspectives. In S. Alagumalai, D. D. Curtis, & N. Hungi (Eds.), *Applied Rasch measurement: a book of exemplars : papers in honour of John P. Keeves* (pp. 287–307). Dordrecht: Springer.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. (2007). *ACER ConQuest Version 2*. Melbourne: ACER Press.

Yildirim, H. H., Yildirim, S. & Verhelst, N. (2014). Profile Analysis as a generalized differential item functioning analysis method. *Education and Science*, *39*(214), 49–64.