Karges, K., Barras, M., & Lenz, P. (in print). Assessing young language learners' receptive skills: Should we ask the questions in the language of schooling? In S. Frisch, E. Romeik, & J. Rymarczyk (Eds.), Current Research into Young FL and EAL Learners' Literacy Skills. Peter Lang.

# Assessing young language learners' receptive skills: Should we ask the questions in the language of schooling?

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

## Abstract

This article discusses empirical evidence concerning the following questions from assessment research and development: Should items in a foreign language (FL) reading test be presented in the language of schooling, or rather, in the target language? Which language would lead to less construct-irrelevant variance? This issue was explored as part of a research project involving young learners studying French as a compulsory subject. In-depth qualitative information was gathered by means of stimulated recall interviews with individual students. Further evidence was collected through a) a questionnaire completed by over 900 students and b) a qualitative analysis of items used in an assessment comprising parallel items in both, the language of schooling and the target language. The combination of the results from these various sources allows for an evidence-based recommendation in favor of the use of the language of schooling in foreign language assessments for young FL learners.

## 1 Introduction[1]

In many European countries, children are now learning a foreign language in primary school. These young learners differ from older students in various respects, making it necessary to adapt both instruction and assessment. For instance, when assessing language skills, teachers and test developers have to take into account that young learners have less experience in reading instructions and completing tasks, have shorter attention spans and limited world knowledge (Bailey, Heritage & Butler 2013; Hasselgreen 2005). They do, however, usually share the official language of schooling (LS), even in contexts where many students speak different languages at home. This makes it possible to phrase instructions, questions about a text and answer options in this language, which may reduce the cognitive load induced by the assessment itself. Yet, according to some, often practitioners, switching languages during an assessment may confuse the test takers. This leads to the essential question that this article attempts to answer based on empirical evidence: What use of languages can actually help improve the quality of language assessments for young learners?

[1] The authors would like to thank the many students and teachers who participated in our studies for their enthusiasm and patience. We also thank Gabriela Lüthi and Patrick Karges for their valuable input and help during the redaction of this article and the editors of this volume for their support and their helpful comments.

## 2 Literature review

In the context discussed in this article, i.e. foreign language reading assessment for educational monitoring purposes[2], the test scores should allow a credible statement of whether or not a population of students has the ability to read in the foreign language at the level prescribed by the national educational standards (Messick 1990, Kane 2006). Since the educational standards themselves do not provide sufficient basis for an operationalizable construct, we needed to complement them with appropriate additional sources[3]. Therefore, our test construct is a conceptualization of foreign language reading ability based on the standards document (EDK 2011), the CEFR (Council of Europe 2001) and, among others, work by Khalifa & Weir (2009: 40ff.).

According to Messick (1995: 742), there are two major threats to the measurement of this (or any other) test construct: construct underrepresentation and construct-irrelevant variance. Construct underrepresentation exists if the test, although framed as a general reading test, only assesses part of the reading construct, e.g. if it elicits only reading for orientation or careful local reading. Construct-irrelevant variance is present when test scores are due not only to the reading ability of a test taker but at least partially to other sources, such as specialized world knowledge, test wiseness or guessing.

Avoiding these pitfalls as well as establishing sufficient evidence for a valid interpretation of the test results hinges on two things: a) a clear idea of the test construct and b) sufficient knowledge of how the input text, the task(s) and the test taker characteristics interact. The latter is the object of research concerned particularly with factors that influence the difficulty of test items. In receptive language assessments, these factors include the propositional density and the topic of the input text (e.g. Freedle & Kostin 1999), the motivation, strategic knowledge and language ability of the test takers (e.g. Jeon & Yamashita 2014; Shiotsu 2010) as well as the characteristics of the test method or items (e.g. In'nami & Koizumi 2009; Ozuru, Briner, Kurby & McNamara 2013; Rodriguez 2003).

One item characteristic which may influence the functioning of an item is the language in which the questions are asked (and the answer options are given). Is it the target language (TL), which is the

[2] The institution which coordinates the Swiss educational monitoring defines the term as follows: "Educational monitoring is the systematic acquisition and compilation of information about an educational system and its environment. It forms the basis for educational planning and policy decisions, accountability and public debate." (SKBF 2019).

[3] In Switzerland, as in many other European countries, the foreign-language curricula are based on the Common European Framework of Reference for Languages (CEFR; Council of Europe 2001). Thus, the minimal standard for reading in a foreign language at the end of primary school (grade 6, age 12) is defined with respect to CEFR level A1.2. By the end of lower secondary school (grade 9, age 15), students are supposed to reach CEFR level A2.2 (EDK 2011). For the actual item development, the user (not constructor) oriented CEFR descriptors (Alderson et al. 2006; Council of Europe 2001: 37f.) were complemented by aspects of a reading test construct based on Khalifa and Weir's reading model (2009: 43ff.). A major feature of this model are the various types of reading which are initiated by the goal setter depending on the purpose of the reading activity (ibid., based on Urquhart & Weir 1998).

object of the test, or another language that the test takers have in common, usually the LS or the test takers' L1? Over the years, several researchers have looked into this issue (e.g. Cox, Bown & Bell 2019; Filipi 2012; Shohamy 1984), but the topic has always been on the sidelines of assessment research. A reason for this may be the preponderance of international language assessments such as the Cambridge exams, the Goethe exams or the DELF/DALF exams. These are designed for test takers all over the world who do not necessarily share a common language (except for the one being tested). Hence, the tests are delivered entirely in the TL. At lower levels of language proficiency, this is not unproblematic: Test items in the TL always carry the risk of being partly or entirely misunderstood (Godev et al. 2002: 204; Gordon & Hanauer 1995: 302), thus limiting a valid interpretation of the test results. To account for this, international exams for lower proficiency levels tend to use well-known item types, which do not need to be explained to the test takers. Also, if at all possible, the language used in the items is often simpler than the reading text itself (Alderson 2000: 86; Green 2014: 113) and in some cases, pictures are used instead of words. Many test takers also familiarize themselves with the specifics of the test beforehand, e.g. by consulting sample tests available online or by taking classes that prepare candidates for certain exams (as is evidenced by the host of material and information available on the internet).

Tests that are written entirely in the TL are not limited to international assessments. Teachers may favor a monolingual approach to foreign language teaching, arguing that exclusively using the TL is more authentic and exposes students to more input, which promotes learning (Godev et al. 2002: 204). Others may fear that using the LS in a foreign language test might confuse students, disadvantage learners who have difficulty in the LS, or introduce a new level of difficulty through the need to translate (ibid.). Finally, teachers may also choose to use target-language assessments in order to prepare students for international examinations. In our own context, compulsory foreign-language learning in Swiss schools, personal discussions with teachers suggest that assessments with questions and answer options in the TL are common practice, and that, from the point of view of those practitioners, the arguments in favor of using the TL hold significant weight.

Yet, the studies which investigated this issue empirically indicate that the use of a strong common language, if it exists, may be a sensible choice in terms of validity. In those studies, most items in the common language of the learners (often the LS) turned out to be easier than items in the TL. This in itself does not necessarily mean that items in the LS lead to more valid test results, but a closer look at the questions and the test takers' answers often points to this conclusion: For instance, Godev et al. (2002: 210f.) describe several instances where the students' short answers in the LS clearly indicate that they "understood the text well enough to respond correctly to [a certain] question" whereas insufficient command of the TL made a correct answer to the otherwise same question less probable (see also Wolf 1993: 482).

In cases where items in the TL were found to be easier, a closer look usually revealed reasons such as a direct match between the wording in the L2[4] item and the L2 text (Godev et al. 2002: 209, 211; Wolf 1993: 481f.) or the presence of transparent cognate words (Filipi 2012: 519). In both cases, it remains unclear whether students who answered correctly did so because they understood the information in the text or because they successfully matched individual words without being aware of their meaning, which would amount to guessing. Shohamy (1984: 158) observed diverging answer patterns in corresponding L1 and L2 multiple choice items, suggesting that students guessed more often whenever they encountered L2 answer options they did not understand. Godev at al. (2002: 210) found evidence of guessing in a number of L2 short answers that did not answer the question in a sensible way but were copied more or less directly from the input text. These findings suggest that students guess more often whenever they encounter items they cannot understand or when they have to give answers they cannot formulate. Such answers are difficult to interpret with respect to the reading construct because guessing is only very inconsistently related to reading proficiency.

Shohamy (1984: 157) argues that the use of the L1 in foreign language assessment may even be considered more authentic "since many students, while processing L2 texts, tend to utilize known elements from their L1 rather than unknown elements from L2". She also maintains that questions and answer options in the LS may offer clues to understand the text better, which she considers to "make the task more natural" (ibid.).

The studies mentioned up to this point mostly focus on low-proficiency learners of the TLs. For learners at higher levels of language proficiency, there is some evidence that the use of the TL may have a lesser effect on the test results (Brantmeier 2006; Shohamy 1984: 155f.). In a recent study by Cox et al. (2019), however, advanced learners of Russian still performed better on a multiple-choice reading test when the items were in their L1. This finding suggests that there is no clear-cut point in language proficiency development where the language of the items becomes irrelevant for the test results. Instead, according to the authors, whether the LS or the TL should be preferred "is likely dependent on the testing situation and population as well as on practical considerations" (Cox et al. 2019: 134).

# 3 Method

In the following subsections, we will describe how we investigated the "test language issue" based on three sources of empirical evidence: stimulated recall interviews, questionnaire data and qualitative item analysis. Most of our evidence stems from the Task Lab project, which investigated task and test-taker characteristics in a low-proficiency French reading assessment. The Task Lab questionnaire data is complemented by students' answers to a questionnaire used in a subsequent

[4] Several of the studies cited here use the terms L1 and L2 instead of LS and TL.

task development project for large-scale educational monitoring conducted in Switzerland in 2017 (ÜGK[5]).

## 3.1 The research projects

The Task Lab project was conducted at the IoM between 2014 and 2016. Its primary aim was to investigate the impact of selected task factors on test scores and test behavior, most importantly item type, i.e. multiple-choice questions (MCQ), short answer questions (SAQ) or matching, and item language, i.e. the LS or the TL. These two task factors were investigated in a reading comprehension assessment of French as a foreign language. The target group were German-speaking Swiss pupils at the end of primary school (grade 6), who had been learning French in a non-intensive two to three-lesson-per-week course for four years. They are expected to reach level A1.2 (i.e. CEFR level A1) by the end of sixth grade according to the national education standards (EDK 2011). The orientation of the project was predominantly quantitative: Around 600 students participated in the main study. During the piloting sessions, qualitative data was gathered by means of stimulated recall interviews (N=34).

The assessment comprised 18 reading comprehension tasks, each consisting of a written text input in French and three successive items (54 items in total). Six of the 18 tasks were matching tasks with items developed in two language variants, i.e. with the same items in German and in French. In the twelve remaining tasks, both the language of the items and the answer format were varied. 36 items were therefore available as multiple-choice as well as short-answer items in both French and German.

During the main data collection, each pupil worked on the French reading tasks for 45 minutes[6], encountering all formats and both languages in about equal measure. The students also completed several cognitive and linguistic component tasks (e.g. a vocabulary test) and a questionnaire. Overall, each student participated in the data collection for 110 minutes during normal school hours. All survey instruments used in the main study were piloted in a field test by 131 students. The tests were delivered on a computer, with the exception of a short paper questionnaire following the reading comprehension test.

Informed by the findings gathered in the Task Lab project, the IoM was later responsible for the task development in the foreign language test of ÜGK 2017, the first nation-wide survey of students' foreign-language skills for educational monitoring purposes in Switzerland. This large-scale assessment targeted the receptive skills of students in their first foreign language at the end of primary school using multiple-choice items in the language of schooling. Pre-piloting of the reading and listening tasks was conducted in three language regions, German-speaking Switzerland and Italian-speaking Switzerland, where French was assessed, as well as French-

[5] Überprüfung des Erreichens der Grundkompetenzen, or Vérification de l'atteinte des compétences fondamentales; literally '"Verification of the achievement of the core competencies".

[6] Each student encountered 13 out of the 18 tasks because of time restraints.

speaking Switzerland, where the students sat a test of German (more details in Table 1). During these pre-piloting sessions, questionnaire data relating to the “test language issue” was collected.

### 3.2 Stimulated recall interviews

We used stimulated recall interviews (Gass & Mackey 2017) to pre-pilot the reading comprehension tasks developed in the Task Lab project. The interviews primarily aimed at finding out how young learners of French proceed when solving computer-based reading comprehension tasks, which strategies they apply, and how their test-taking behavior is influenced by the language of the items and the answer format.

Four previously trained interviewers conducted interviews with a total of 34 pupils. In a 45-minute, audio-recorded session, each participant was confronted with a selection of tasks from the pool described above. Typically, each student completed two to three tasks with three items each. Immediately after each item, the students were asked to explain their approach, their considerations, the strategies used and difficulties they had encountered, if any.

At the beginning of the session, each pupil was informed about the aims of the study and the procedure. While the pupils were reading, the interviewer stepped back. When the pupils indicated that they had finished processing an item, they were interviewed. The interview lasted approximately one to three minutes per item. The task displayed on screen and the answer given by the pupil (the selected answer option or the written short answer, respectively) served as a stimulus for each interview. The interviews were conducted in standard German and/or a Swiss German dialect and were based on a written guideline that was constructed with the research questions in mind. During the interview sessions, the researchers noted down potentially interesting observations to complement the audio recording.

The interviews focused on the students’ responses to the items and the thoughts that led to them. Whenever the questions (and where applicable the answer options) were in the TL, the pupils were asked whether they had understood them. In the case of multiple-choice and matching items, the interviewer asked why a particular answer was chosen (rather than the others), and in the case of short answer items, whether there had been difficulties in formulating the answer. If required, the interviewers asked the pupils additional in-depth questions, for example, how often and how accurately they had read the text, where in the text they had found the answers, whether the task instructions had been clear. The last five minutes of each interview were dedicated to more general questions, e.g. questions related to the use of the computer.

In general, we observed that most of the students were happy to share the thoughts they had during the test, and we had the impression that they actually expressed what they thought. For instance, they readily admitted to have merely guessed or to using test wiseness strategies such as copying text directly from the input text, possibly because they had been informed in advance that their performance in the test had no influence on their school grades.

The interviews were transcribed and coded using the data analysis software *MAXQDA* (VERBI Software 2015). Data analysis was carried out according to the principles of structuring content analysis (Mayring 2010). The coding categories were derived from the research questions (top down) and and inferred from the collected data (bottom up). The results of the data analysis served as a basis for the revision and adaptation of the test tasks for the main study.

## 3.3 Questionnaires

In the stimulated recall interviews of the Task Lab study, we collected some evidence on what the test takers thought about the "test language issue". To expand on this evidence, we added a question to a short questionnaire given to the participants of the main study: At the end of the reading test, students were asked what they had found easier – the questions and answers in German, their LS, or in French, the TL. They checked a box to indicate their preference and had the opportunity to explain their answer.

Later, in small-scale field tests supporting the development of test tasks for the Swiss educational monitoring survey (ÜGK), we asked sixth-graders in three language regions to speculate what they would find easier (or better[7]) regarding the language of the items: the LS they had just encountered during the field test, or the TL. They, too, were asked to explain their answer. Some of these students solved reading tasks similar to the ones we used in Task Lab, others completed listening tasks.

Overall, we collected the opinions of 936 6th-graders on the test language issue. Of these, 879 also wrote a comment[8]. Table 1 gives more details on the sample.

Table 1: Questionnaire data on the "language issue" (overview)

| **Project** | **Assessed skill during data collection** | **LS** | **TL** | **N** |
|---|---|---|---|---|
| Task Lab | Reading | German | French | 591 |
| ÜGK dF[9] (field test) | Reading | German | French | 45 |
| ÜGK fD (field test) | Reading | French | German | 47 |
| ÜGK dF (field test) | Listening | German | French | 46 |
| ÜGK fD (field test) | Listening | French | German | 78 |
| ÜGK iF (field test) | Listening | Italian | French | 129 |

[7] Due to a translation error, the word "easier" was replaced by "better" in some of the questionnaires.

[8] Not all of those comments are meaningful for this article. For instance, some students gave their opinion on how they found the test in general, or they made observations we cannot fully understand in hindsight (e.g. in a test of reading: "The speed was okay overall").

[9] The acronyms dF, iF and fD stand for the German names of the languages involved in the assessment: d, i and f represent the pupils' LS (German, Italian and French respectively), whereas F and D stand for the target languages of the assessment (French and German).

### 3.4 Qualitative interpretation of differences in item difficulty

As pointed out in the literature review, the language of the items does not affect the difficulty of each item in the same way, nor do other characteristics, such as item type. To account for this, we examined the items we used in Task Lab, making use of the empirical item difficulties we obtained from the main study.

To determine the item difficulties, a Rasch model was estimated using the R package "TAM" (Kiefer, Robitzsch & Wu 2015; R Core Team 2014), in which the four item variants (defined by MCQ or SAQ[10]; TL or LS) were considered as separate items. The model used responses[11] from 577 participants. Each of the 144 items was answered by at least 84 test takers (120 on average). The items fit the Rasch model sufficiently well according to common standards (e.g. OECD 2014: 151) as all infit values except one fall between 1.20 and 0.80[12].

The item difficulties thus obtained were grouped so that all four format-by-language variants of one item could be visualized together. We then examined these groups and identified possible reasons for the various patterns of relative item difficulty by referring back to the items, i.e. the input texts, the questions and the students' responses.

## 4 Results

### 4.1 Stimulated recall interviews

The analysis of the transcripts of the stimulated recall interviews conducted in the Task Lab project revealed three major topics with respect to the "language issue": the students' individual preferences regarding the language of the items, the comprehension problems they encountered when confronted with items in the TL and the difficulty of writing short answers in French. In the following, these three perspectives will be presented and illustrated with original quotes from the interviews[13].

#### 4.1.1 Students' preferences regarding the language of the items

As mentioned above, during the Task Lab study each student was confronted with items in each language version, i.e. in the LS German and in the TL French. In the interviews we asked the students (N=34) if they felt bothered or confused by having to switch between German and French when the items were formulated in the LS. A vast majority of the students did not consider this to

[10] The six matching tasks are not considered in this section.

[11] The MCQ items were scored automatically based on the option chosen by the student. The SAQ items were scored manually by two raters according to detailed scoring guidelines. The raters first scored all answers individually and then discussed diverging scores and reached an agreement in each case. Only dichotomous (0 or 1) coding was used.

[12] In fact, the SAQ items generally show overfit (low infit) while the MCQ items show underfit (high infit). This fact has to be taken into account when tests are constructed which consist of both item types (cf. Lenz, Karges & Barras 2019), but it is not relevant for this analysis.

[13] The quotes are English translations of the original German transcripts.

be a problem. We also asked the students if they preferred the items in the TL or in the LS. Many pupils[14] clearly favored the latter, arguing that this allowed them to understand the questions (better), which enabled them to know what to look for in the reading text.

> I: If you could choose, would you choose the questions in German or French?
>
> S: German.
>
> I: And it doesn't bother you that you have to switch between languages?
>
> S: No, it doesn't bother me. (Je116[15])
>
> I: We had some questions in French, and some in German. Which do you like better?
>
> S: In German. Because in French you're busted when you don't understand the question. (Je115)
>
> S: (...) when I have a French question, it sometimes happens that I don't understand a word and can't answer the question because I don't know what to do. But I actually understand the whole text. (Je110)

Not everybody was partial to the use of German in a French test. Some students insisted that the language of the item did not matter to them. Those students were convinced that their test results would remain the same if the items were in the TL[16].

> I: Which did you like better: the questions in German or in French?
>
> S: Both the same. […] If you don't understand individual parts [= words], you can put them together and guess what the question might be.
>
> I: Then would you say that you might have given better answers to the German questions than to the French ones?
>
> S: No, not necessarily. (Ge1103)

Only one of the 34 students we interviewed actually said that he would have preferred to have the questions presented in the TL and found the German items confusing. His home language, interestingly, is Portuguese.

> I: You had some German and some French questions. What was that like?
>
> S: In French it was a bit better.
>
> I: Why?
>
> S: If they are in German, you have to translate the words on top of it. (Ge198)

[14] Due to the qualitative nature of our pilot study we mostly refrain from quantifying the responses we collected in the interviews.

[15] The codes are unique identifiers of individual students used throughout all projects. Their meaning is not relevant for this paper.

[16] Since the students who participated in these stimulated recall sessions only worked on a small number of tasks, it is impossible to say whether they would have indeed performed similarly with items in either language.

It is likely that the linguistic similarity between French and Portuguese made the French items easier for him to understand. Whether this preference for the TL is typical of students who speak Portuguese or other Romance languages remains to be investigated.

#### 4.1.2 Problems understanding items in the TL

While dealing with the items in the TL, students often reported that they did not fully understand the questions. As a result, they could not always be sure whether they were looking for the right answer.

> I: And the questions?
>
> S: If they were in German, maybe I would have had them right. I don't know now if that's right. (Ge1106)
>
> S: I don't have a complete answer for this one, because I didn't understand the question very well. (Ge1101)

Sometimes the students had difficulty understanding a question because one or several words were unfamiliar to them. They usually had a vague idea of what the question could be about but were insecure about whether their assumption was correct. In other cases, students only understood individual words of the question. This was usually not enough to answer the question correctly, and the students were quite aware of that.

> *Question: What occupation do Tom's grandparents have? [Quel est la profession des grands-parents de Tom ?]*
>
> S: […] I don't know exactly what "profession" means.
>
> I: (...) What do you think the question means?
>
> S: It's somehow about Tom's grandparents, but I don't know exactly what it says. (Ge1107)
>
> *Question: Why do the young kids rarely have accidents on the lake? [Pourquoi est-ce que les jeunes enfants ont peu d'accidents sur le lac ?]*
>
> I: (…) Then what do you understand about the question?
>
> S: Not much. I only understand "the lake". And otherwise I didn't understand the question at all. (Je112)

Interestingly, we observed that many of our students struggled to understand the question words and therefore did not know what they were supposed to answer. Even important question words that occur frequently, such as "why" or "where" posed a problem in some cases.

> *Question: Why did Hans Kaufmann start the project? [Pourquoi Hans Kaufmann a-t-il commencé le projet ?]*
>
> I: Why weren't you quite sure?
>
> S: Because I don't know "pourquoi" [= why] exactly... "What" or something. (Je101)

I: If you look at the question: “Où as-tu besoin d’un dictionnaire?” [Where do you need a dictionary?]. Do you understand “où” [= where]?

S: No. (Ge1107)

In multiple-choice items, some students struggled not only to understand the question itself but also the answer options provided.

I: Do you understand the question?

S: No.

I: Did you rather guess?

S: I don’t understand any answer. (Ge1108)

I: Do you know what the other two options mean?

S: No, not really. (Ge1106)

A common mistake we observed in our data was the misinterpretation of French words as German cognates. Such words were translated incorrectly by some students – for example “prof” [teacher] as “Profi” [professional]. This phenomenon may explain a number of incorrect responses.

*Question: Why is Vidal’s way to school special? [Pourquoi le chemin de l’école de Vidal est-il spécial ?]*

I: Do you know what the question is?

S: Why is Vidal’s chemistry lab so special?

*Question: Who prefers languages over mathematics? [Qui préfère les langues aux maths ?]*

I: Do you have a spontaneous idea?

S: Something about math. What is your longest[17] math lesson? (Je114)

During the interviews, we found evidence that at least in some cases, pupils who did not understand the question had actually understood the input text quite well. In the following example, one student was able to write the correct answer in French after the interviewer had translated the French question into German.

I: So you have no idea what it could mean?

S: No.

I: Ok, then I’ll tell you, and after that you’ll try to answer the question. It means: “In this project the children learn to play floorball. What else do they learn?”

S: Ok.

((S reads, types))

[17] The word “longest” refers to the German word “längste”, superlative of “lang”, which the student may have confused with the French word “langues”.

S: So, I wrote that the children learn to play in a team[18]. (Je101)

Whenever the students encountered the types of problem described above, they used various strategies to try and give a suitable answer. For example, if they did not understand a multiple-choice answer option, some students tended to ignore that particular answer option.

I: And why didn't you choose the other options?

S: Concerning the third, I do not know what that means. (Ge196)

S: I read through again. Then I thought it was the one in the middle because I didn't understand the others. (Ge1105)

In many instances, students simply guessed or used test-wiseness strategies such as trying to locating unknown words in the question or answer options in the input texts. Whenever they found a match, the students would try to figure out an appropriate answer with the help of the context provided by the input text. Of course, this strategy was unsuccessful if the unknown word in the question did not appear in the input text, as in the example below.

*Question: What did Alicia like the most? [Qu'est-ce qu'Alicia a aimé le plus ?]*

I: You read the question and didn't understand it, and then you read the text again. What did you do while reading?

S: I looked what the answer could be.

I: And how did you do that?

S: I just read here, and maybe the word "aime" [to like] appears somewhere in the text.

I: You were looking for the word you didn't understand.

S: Yes.

I: And then you didn't find it.

S: Yes.

I: That's why you had to give up.

S: Yes. (Je115)

### 4.1.3 Writing short answers in the TL

In the Task Lab study, when students encountered a short-answer question in French, they also had to answer them in French. During the interview sessions, we observed that for many students at this low proficiency level, writing even the simplest answers in the TL represented a significant problem. The students found it difficult to focus on content, syntax and spelling at the same time. This bothered them even though they had been told at the beginning of the test that linguistic errors would not be taken into account. Many students also mentioned that they lacked vocabulary and were therefore unable to formulate their answers in French.

We found evidence that some students had something very different in mind than what they actually wrote in French.

[18] The student wrote "les enfants apprennent la jouer dans l'equipe" [children learn playing in the team] which can be considered a correct answer to that question.

I: Can you read your answer to me?

S: So, Karusu loses his dad at the zoo.

I: And you wrote “Karusu devenu papa” [Karusu became a dad]. (Je105)

S: I wrote: “Pierre Dumont en danger.” [Pierre Dumont in danger.] (…)

I: What would you write in German if you could write this answer in German?

S: Pierre Dumont is very dangerous. (Je117)

A small number of students stated that it was easier for them to deal with SAQ items in the TL because this gave them the opportunity to copy words or passages directly from the French text or the French question.

I: Was writing in French a problem?

S: No, I could copy that. (Je116)

I: And here you had to write some of the answers in French. Do you think that’s difficult?

S: Well, if the text is in French, it’s not [difficult] because you can copy a lot of things. (Je104)

Having gained these insights, we decided to further annotate the short answers gathered later in the main study. We identified all answers that were copied directly from the text, i.e. answers which contained three or more words in the same order as they appeared in the input text. Overall, more than a third of the French SAQ answers (37 %) were at least partly copied directly from the French text. In these cases, it depended largely on the item whether the strategy was successful or not: Whereas, for instance, 90 % of the copied answers to item T01-2 led to a correct answer (16 out of 18), because the item elicited a concrete piece of information, this was true for only 11 % (7 out of 65) of the copied answers to item T03-2, which demanded an inference from the content of the text.

We also annotated short answers as “absurd” when they did not in any way answer the question. This was the case for 33 % of all short answers in French. Again depending on the item, up to half of these non-sensical answers had been copied directly from the text. Thus, it appears that some students simply chose a random word or text fragment from the input text or the question when they did not know what was asked and to what they were supposed to provide an answer.

*Q: What does Emilie like in school?*

A1: Zurich. (Br625)

A2: [a] world. (Mu712)

A3: He preferred at school Thursday. (Vi129)

*Q: What do the two want to buy?*

A1: dad and boy (Vi154)

A2: and you don’t like to go by bike (Bi366)

We also encountered “absurd” answers in German (10 % of all German answers were annotated as such), but these were often comments unrelated to the text itself, like “no idea” [keine Ahnung] or, supposedly, answers copied from nearby students who were completing a different task. These findings show that the students employed different strategies depending on what language(s) they were dealing with.

## 4.2 Questionnaire

As described in the methods section, the students on the Task Lab main study and on the ÜGK field-test were given a short questionnaire where, amongst other things, they first indicated which language they found easier (or better) for the questions and answers and then justified their choice or left some other comment.

As Table 2 very clearly shows, only a small minority of the students chose the TL when asked what they found easier (or, in two questionnaires, better[19]) (Task Lab: 10 %, ÜGK: 15 %). In the ÜGK field test, where students had only completed items in their language of schooling, another 25 to 35 % of the students indicated that they had no preference concerning the language of the items. This means that overall, a clear majority of the students in our samples preferred the use of the LS or, at least, was indifferent towards it. Based on these results, we assume that the use of the LS in foreign language assessments (reading and listening) does not confuse most students.

Table 2: Distribution of the students’ answers in the questionnaires

| Project | Skill assessed during data collection | Total | In favor of the LS | In favor of the TL | No preference |
|---|---|---|---|---|---|
| Task Lab | Reading | 608 | 537 | 54 | 17 |
| ÜGK dF | Reading | 45 | 22 | 13 | 10 |
| ÜGK fD | Reading | 47 | 24 | 10 | 13 |
| ÜGK dF | Listening | 56 | 34 | 3 | 19 |
| ÜGK fD | Listening | 78 | 49 | 12 | 17 |
| ÜGK iF | Listening | 129 | 68 | 16 | 45 |
| ÜGK all | Listening or reading | 355 | 197 | 54 | 104 |

The reasons the students gave for their choice provide more insight[20]. The most common type of answer is related to the actual languages (LS and TL) concerned, and may not be generalizable in a straightforward manner to any reading or listening test at the proficiency levels in question: Many

[19] Overall, there is no discernible difference between the students’ reactions to the “easier” question and the “better” question. Their written justifications are very similar in both cases.

[20] All answers cited in the following are literal English translations of the students’ handwritten answers.

students argued in favor of the LS either by stating that they were proficient in this language or that they were *not* proficient in the TL.

> Because [Italian] is my language! (ÜGK iF, Ca538)
>
> Because I'm not good at German. (ÜGK fD, Es139)
>
> Because I never speak French and I understand almost nothing! (ÜGK dF, Be635)

Some students who preferred the TL or indicated that they were indifferent, asserted that they either found the TL easy or knew both languages equally well.

> Because everything was easy, because I have spoken French since birth. (Task Lab, Br671)
>
> Because I understand both languages well enough. (ÜGK fD, Fa347)

Some students gave more precise reasons related to the test itself. For instance, students who opted for the LS stated that it helped them understand the questions and answer options. Some students also pointed out that reading the questions and answer options in their LS gave them some idea of what the text was about.

> Because that way you can understand more of what they're saying because [the questions in the LS] tell me a lot. (ÜGK iF, Br448)
>
> Because then I knew what it was about and what the question was. (Task Lab, Si475)

Many students who indicated that they preferred the use of the TL pointed out that words or phrases in the questions or answer options could help them find the answer. With respect to the multiple-choice items, this was mentioned particularly often by the students who had just taken the listening comprehension test in the ÜGK trials. These students argued that seeing a word in its written form may give them clues about the words that were pronounced.

> That way when I heard the text and then read, I understood better, I think. (ÜGK iF, Br441)
>
> I understand German better when I see it written. (ÜGK fD, Fr358)

With respect to the SAQ items, which were only used in the Task Lab project, many of the 10 % of students who preferred the TL argued that writing short answers in that language allowed them to copy words or phrases from the text. This is the same argument we had already encountered in the stimulated recall interviews (cf. section 4.1.3).

> Because then, most of the time, you could look for the words in the text. (Task Lab, Bi383)
>
> Because in French, I could take the answers directly out of the text. (Task Lab, Vi148)

A few students were aware of the benefits of both language versions.

> In French you know because it's my mother tongue. In German, you can locate the words. (ÜGK dF, Bu398)
>
> In German, when you don't know a word in French. In French: When you cannot translate something. (Task Lab Vi182)

Finally, a small number of students stated that mixing the two languages did not appeal to them because they were not used to it. In our entire sample, however, there are less than ten instances of this, and not all of them necessarily imply that the test results suffered from mixing the languages.

> Because I get confused between Italian and French. (ÜGK iF, Br413)

> I've practised it more the other way. (ÜGK dF, Ta251)

Finally, a rather small group of students considered the testing situation to be a learning opportunity. These learners tended to prefer the TL because it gave them more opportunity to practice the language.

> Because you can learn more that way. (Task Lab, Zu302)

> It's easier in French but it would be funnier and more exciting to put the questions in German. (ÜGK fD, Fa351)

Interestingly, it was most often the students from Ticino who brought forth this argument. One possible reason, which is also reflected by the comparably large number of students who answered "I don't care" to the initial question, is that these Italian-speaking learners of French had just encountered listening tasks that were decidedly easier for them compared to the other groups of test takers. This is most probably due to the fact that there is a close typological relationship between Italian, their LS, and French, the TL (both being Romance languages).

> That way it's a little more difficult and a bit more entertaining. (ÜGK iF, Br429)

> Because that way you can practice French better, also with the questions. (ÜGK iF, Ca486)

## 4.3 Qualitative interpretation of differences in item difficulty

As mentioned in the methods section, the following section builds on the results of a Rasch analysis of the reading test in the Task Lab main study. The results of the analysis show that, in general, multiple-choice items are easier than short-answer items, and items in the LS (German) are easier than items in the TL (French). This is the pattern we would expect based on the literature, and it can be observed like that in 13 out of the 36 items. However, for the remaining 23 items, there seem to be three major deviations from this pattern:

- multiple-choice items in French are easier than multiple-choice items in German (6 items),
- short answer questions in French are easier than short answer questions in German (6 items),
- short answer questions in German are easier than multiple-choice items in French (9 items).

The 12 items which are easier in the TL in at least one item format seem to differ from all other items in one important respect: the overlap between the item (question and/or answer options) and the input text in the TL version. Many of these items contain words in the French questions which can be matched directly to the relevant passage in the text. As a result, short answer questions of

this type can be answered correctly by simply copying words or phrases from the input text. Similarly, the TL answer options of multiple-choice items contain words or phrases that can be found verbatim in the relevant passage. This effect can be illustrated particularly well in item T06-2, which is easier in French both in the multiple-choice and in the short-answer version. In this item's input text, the boy Tom presents himself and his family. In the item, students have to indicate the occupation of Tom's grandparents. The correct answer, "paysan" [farmer] appears in the text, in the same sentence as the word "grands-parents", which makes it possible to choose the correct answer option even if neither the meaning of the question nor the answer were quite understood. If, however, the item is in German, the test takers have to know that "Grosseltern" means "grands-parents", and then either be able to identify the word "paysan" and know that it means "Bauer" in German, or – in the multiple-choice version of this item – identify the words "vétérinaire" and "professeur de physique" and discard them as possible answers. A similar argument can be made for the short-answer version of this item and for most of the other items which follow the first two patterns.

The third pattern, where the short-answer item in German is easier than the multiple-choice item in French, is observed in items which seem to roughly belong to two groups. One group is composed of the first items of five tasks. These items ask about the general topic of the given texts, and it seems that some students, when confronted with the multiple-choice question in the TL, failing to understand what was being asked, tried to match words from the text with the answer options. In these items, however, this strategy was less successful because words from the input text usually appeared in all three answer options. Students who encountered the question in their LS knew what they were being asked and thus had the opportunity to show that they understood the general topic of the text. As a result, more correct answers were given in the short-answer items in the LS. The remaining four items which show this pattern contain multiple-choice answer options which do not reveal much about the content of the question. This seems to make it more difficult to decide on an answer without fully understanding the question. A good example of this phenomenon is item 7-02, which asks "Qui utilise son Natel le plus souvent?" [Who uses their cell phone the most often?]. Since the answer options are merely the names of three people who have a conversation in the input text, they do not give much of a clue as to the content of the answer. Thus, students who do not understand the question cannot even make an educated guess about the answer.

# 5 Discussion

In this article, we have presented empirical evidence relating to the use of the language of schooling (LS) and the target language (TL) for the questions and answers in tests of foreign language reading and listening intended for young learners in a compulsory school context. Our data is based on assessments at the lower levels of language proficiency, especially reading comprehension targeting the region around A1.2.

In the Task Lab project, we used items in both the LS and the TL. We observed and interviewed students during task pre-piloting and later analyzed the test results to get a clearer idea of how students process items in the two languages. We also gathered the students' opinions on the "test language issue" through stimulated recall interviews and a questionnaire. More data was collected with a similar questionnaire in the more recent ÜGK task development project, where students only worked on items phrased in the LS.

As discussed in section 2, empirical evidence from earlier studies indicates that for the test questions and answers a language that is familiar to all test takers is preferable with regard to test validity. Our results corroborate these previous research findings. We found almost no evidence for the assumption that the use of the LS in a foreign language assessment might confuse test takers (see also Filipi 2012: 527). On the contrary, using a language in which test takers are more proficient may actually help to better reflect the reading construct: Without the need to spend a lot of time and cognitive resources on understanding the questions, students may be able to concentrate more on the text itself. Fully understanding the questions may also trigger more authentic types of reading because students can choose more consciously whether, for instance, local or global reading is more appropriate for answering a specific question (see also Cox et al. 2019: 123). After all, reading is hardly ever done in the real world without knowing its purpose (see also Shohamy 1984: 157).

Furthermore, whenever the items were presented in the TL, we found evidence that correct answers less often reflected good reading proficiency, and more often depended on chance knowledge of a specific word, (more or less informed) guessing, or test wiseness strategies. This was confirmed by many students we interviewed and it can also be inferred from the qualitative interpretation of the differences in item difficulty. Admittedly, a reading construct may also include the ability to deal with uncertainty and to infer the meaning of unknown words from the surrounding context, pictures or other elements. In Switzerland, and in many other European countries, developing this kind of strategic competence is part of the foreign language curricula. From this point of view, one may argue that guessing a correct answer from a limited number of clues is a desirable skill that should be assessed. However, guessing what one is supposed to do (i.e. which information is to be found in a reading text) does not seem to be part of a construct worth assessing, because guessing the meaning of a test item is not the same as inferring meaning from context during reading. Furthermore, we doubt that it makes sense to combine the measurement of test-strategic competence and reading comprehension in the same items and, as a result, to obtain test results which can be interpreted as being evidence of *either* good (or bad) reading skills *or* good (or bad) test-strategic skills.

In our Task Lab interviews, we encountered students who did understand the information that an item asked for, but were unable to show this understanding, either because they did not understand the question or the answer options or because they could not write a (sufficiently good) short answer in the TL. The item analysis from the main study corroborates this finding. It shows

that TL items were often harder than the same items in the LS, suggesting that the TL items introduced a source of difficulty unrelated to the understanding of the input text.

Of course, the use of the LS can also be problematic: Students with a very weak command of the LS, but good command of the TL may be at a disadvantage. Indeed, one student we interviewed, whose L1 is typologically related to French, indicated that he preferred the items in the TL, and a small number of students who participated in the main study argued similarly. This would have to be more closely examined, however, because it remains unclear whether these students' performance on the LS items was indeed affected negatively. Otherwise, we found very little evidence that the students' proficiency in the LS was an issue: Even though a number of multilingual students participated in our data collections[21], most of them clearly had more contact with the LS than the TL.

Additionally, the students' answers in the questionnaire used in both projects confirm that most of our participants clearly prefer items in the LS to items in the TL, mostly because they understand the LS better. One may argue that the students' opinions should not be overrated because they could have been motivated simply by the desire for a test that can be solved correctly with minimal effort no matter whether this test accurately reflects a reading or listening proficiency construct. In the results section we presented some answers that do indeed reflect such a tendency. However, this does not automatically exclude the test takers' opinions from being considered for validation purposes: One aspect of validity – face validity – is concerned with whether stakeholders (e.g. test takers) think, based on "subjective judgement rather than […] any objective analysis", that a test is "an acceptable measure of the ability they wish to measure" (ALTE Members 1998: 145). This is often rejected as "not [being] a true form of validity" (ibid.). Nevertheless, Hughes (1989: 27) argues that, if a test which lacks face validity is used, "the candidates' reaction to it may mean that they do not perform on it in a way that truly reflects their ability". Bachman (1990: 288) also concludes that "[t]he 'bottom line' in any language testing situation […] is whether test takers will take the test seriously enough to try their best […]". Thus, if the students in our context find the test unusually difficult (e.g. if the questions about the text are hard to understand), if they feel bothered by certain aspects (e.g. if more than one language is used), or if they think that the test is not giving them the chance to show their true understanding (because they do not understand the questions), then they may not do their best, which reduces the validity of the results (see also Cox et al. 2019: 131f.; Filipi 2012: 513). Therefore, the test takers' opinions, whether subjective or not, should carry some weight in the validity argument.

Finally, we must point out that our results are limited to a large-scale assessment context and cannot be applied directly to other settings. Both in the Task Lab project and during the ÜGK task development, we targeted a population of test takers that is relatively homogeneous with respect to

[21] About 10% of the Task Lab students (N=57), asked in which language(s) they had first learned to speak, named one or more languages that were neither German nor Swiss German. Another 30% indicated that they had first learned to speak both German or Swiss German and one or more other languages.

their language skills: The students are usually reasonably proficient in the LS and lower-level learners of the TL. Thus, we could rightfully assume that our items would be easily understood in the LS whereas the same would not be true for the TL. In our specific context, it also makes more practical sense to use the LS for the test items, because our students do not learn according to the same curriculum: Since there is no core vocabulary and no common textbook, it would be very cumbersome to determine what precisely the test takers would be able to understand in the TL and what not. Also, preparing the learners for the test, as is often the case in the context of international language examinations and also in classroom-based assessment, was not feasible for various reasons.

# 6 Conclusion

Our results point to the conclusion that, at least for our target group, there is very little evidence that speaks against using items in the LS in reading and listening tests, and a lot in favor of it. In fact, a large majority of students prefer questions and answers in the LS and have no problem switching between the languages. If items in the TL are used, there is a risk that pupils cannot even start engaging with the text because they have not understood the question, or that they cannot demonstrate their understanding of the text because they do not understand the response options or are not able to formulate a short answer in the TL. Failing to understand, they may apply test-taking strategies that do not involve knowledge of the TL – or simply guess their answers. Furthermore, in our study, the analysis of differences between four variants of the same items led to unexpected findings that mostly have to do with problems in the TL items. All of these sources of failure have very little to do with the ability to carry out the intended reading activities, and therefore introduce construct-irrelevant variance.

Overall, it seems that with the help of questions and answers in the LS, we can measure more reliably what we actually intend to measure: the ability to understand a variety of TL texts in various relevant ways.

# 7 References


Alderson, J. Charles (2000), *Assessing Reading*. Cambridge: Cambridge University Press.

Alderson, J. Charles; Figueras, Neus; Kuijper, Henk; Nold, Guenter; Takala, Sauli & Tardieu, Claire (2006), Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR construct project. *Language Assessment Quarterly* 3: 1, 3-30.

ALTE Members (1998), *Multilingual Glossary of Language Testing Terms*. Cambridge: Cambridge University Press.

Bachman, Lyle F. (1990), *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bailey, Alison L.; Heritage, Margaret & Butler, Frances A. (2013), Developmental considerations and curricular contexts in the assessment of young language learners. In: Kunnan, Anthony John (Ed.) (2013), *The Companion to Language Assessment*. Chichester: Wiley Blackwell, 423-439.

Brantmeier, Cindy (2006), The effects of language of assessment and L2 reading performance on advanced readers' recall. *The Reading Matrix* 6: 1.

Council of Europe (Ed.) (2001), *Common European Framework of Reference for Languages*. Cambridge: Cambridge University Press.

Cox, Troy L.; Bown, Jennifer & Bell, Teresa R. (2019), In advanced L2 reading proficiency assessments, should the Question Language be in the L1 or the L2? Does it make a difference? In: Winke, Paula & Gass, Susan M. (Eds.), *Foreign Language Proficiency in Higher Education*. Cham: Springer International Publishing, 117-136.

EDK (Ed.) (2011), Grundkompetenzen für die Fremdsprachen. Nationale Bildungsstandards. EDK. [Online: http://edudoc.ch/record/96780/files/grundkomp_fremdsprachen_d.pdf].

Filipi, Anna (2012), Do questions written in the target language make foreign language listening comprehension tests more difficult? *Language Testing* 29: 4, 511-532.

Freedle, Roy O. & Kostin, Irene W. (1999), Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing* 16: 1, 2-32.

Gass, Susan M. & Mackey, Alison (2017), *Stimulated Recall Methodology in Applied Linguistics and L2 Research* (2nd ed.). New York & London: Routledge.

Godev, Concepción B.; Martínez-Gibson, Elizabeth A. & Toris, Carol C. M. (2002), Foreign language reading comprehension test: L1 versus L2 in open-ended questions. *Foreign Language Annals* 35: 2, 202-221.

Gordon, Claire M. & Hanauer, David (1995), The interaction between task and meaning construction in EFL reading comprehension tests. *TESOL Quarterly* 29: 2, 299-324.

Green, Anthony (2014), *Exploring Language Assessment and Testing: Language in action*. London, New York: Routledge.

Hasselgreen, Angela (2005), Assessing the language of young learners. *Language Testing* 22: 3, 337-354.

Hughes, Arthur (1989), *Testing for Language Teachers*. Cambridge: Cambridge University Press.

In'nami, Yo & Koizumi, Rie (2009), A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing* 26: 2, 219-244.

Jeon, Eun Hee & Yamashita, Junko (2014), L2 reading comprehension and its correlates: A meta-analysis. *Language Learning* 64: 1, 160-212.

Kane, Michael (2006), Validation. In: Brennan, Robert L. (Ed.), *Educational measurement* (4th ed.). Phoenix: Greenwood, 17-64.

Khalifa, Hanan & Weir, Cyril J. (2009), *Examining Reading*. Cambridge: Cambridge University Press.

Kiefer, Thomas; Robitzsch, Alexander & Wu, Margaret (2015), *TAM: Test Analysis Modules*. [Online: http://CRAN.R-project.org/package=TAM].

Lenz, Peter; Karges, Katharina & Barras, Malgorzata (2019), Investigating test method effects in French L2 reading items for young learners. In: Huhta, Ari; Erickson, Gudrun & Figueras, Neus (Eds.), *Developments in Language Education: A Memorial Volume in Honour of Sauli Takala*. Jyväskylä: University of Jyväskylä & EALTA, 182-202.

Mayring, Philipp (2010), *Qualitative Inhaltsanalyse: Grundlagen und Techniken*. Weinheim, Basel: Beltz.

Messick, Samuel (1990), *Validity of Test Interpretation and Use*. Princeton, NJ: Educational Testing Service.

Messick, Samuel (1995), Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist* 50: 9, 741-749.

OECD (2014), *PISA 2012. Technical Report*. Paris: OECD Publishing.

Ozuru, Yasuhiro; Briner, Stephen; Kurby, Christopher A. & McNamara, Danielle (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.

R Core Team (2014), R: A language and environment for statistical computing. Wien: R Foundation for Statistical Computing.

Rodriguez, Michael 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.

Shiotsu, Toshihiko (2010), *Components of L2 Reading: Linguistic and Processing Factors in the Reading Test Performances of Japanese EFL Learners*. Cambridge: Cambridge University Press.

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

SKBF (2019), Bildungsmonitoring. [Online: http://www.skbf-csre.ch/bildungsbericht/bildungsmonitoring/].

Urquhart, A. H & Weir, Cyril J. (1998), *Reading in a Second Language: Process, Product, and Practice*. London, New York: Longman.

VERBI Software (2015), *MAXQDA 11*. Berlin: VERBI.

Wolf, Darlene F. (1993), A comparison of assessment tasks used to measure FL reading comprehension. *The Modern Language Journal* 77: 4, 473-489.