DE

Leseverstehen in einer Fremdsprache testen

Task Lab – eine empirische Studie zu computerbasierten Aufgaben in Französisch

—

FR

Tester la compréhension de l'écrit dans une langue étrangère

Task Lab – une étude empirique sur des tâches informatisées en français

—

IT

Test di comprensione scritta in una lingua straniera

Task Lab – uno studio empirico sugli esercizi al computer in francese

—

EN

Assessing reading comprehension in a foreign language

Task Lab – an empirical study of computer-based tasks in French

—

Katharina Karges, Malgorzata Barras, Peter Lenz

2021 Bericht des Wissenschaftlichen Kompetenzzentrums für Mehrsprachigkeit
Rapport du Centre scientifique de compétence sur le plurilinguisme
Rapporto del Centro scientifico di competenza per il plurilinguismo
Report of the Research Centre on Multilingualism

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

Herausgeber | Publié par
Institut für Mehrsprachigkeit
www.institut-mehrsprachigkeit.ch
—
Institut de plurilinguisme
www.institut-plurilinguisme.ch

Autor*innen | Auteur·e·s
Katharina Karges, Malgorzata Barras, Peter Lenz

Das vorliegende Projekt wurde im Rahmen des Forschungsprogramms 2016-2020 des Wissenschaftlichen Kompetenzzentrums für Mehrsprachigkeit durchgeführt. Für den Inhalt dieser Veröffentlichung sind die Autor*innen verantwortlich.

Le projet dont il est question a été réalisé dans le cadre du programme de recherche 2016-2020 du Centre scientifique de compétence sur le plurilinguisme. La responsabilité du contenu de la présente publication incombe à ses auteur·e·s.

Übersetzungen | Traductions
Isabelle Affolter, Mary Carozza, Joël Rey

Freiburg | Fribourg, 2021

Layout
Billy Ben, Graphic Design Studio

Unterstützt von | avec le soutien de

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI
Département fédéral de l'intérieur DFI
Dipartimento federale dell'interno DFI
Departament federal da l'intern DFI
Bundesamt für Kultur BAK
Office fédéral de la culture OFC
Ufficio federale della cultura UFC
Uffizi federal da cultura UFC

2021.001-TaskLab

Leseverstehen in einer Fremdsprache testen

Task Lab – eine empirische Studie zu computerbasierten Aufgaben in Französisch

—

Tester la compréhension de l'écrit dans une langue étrangère

Task Lab – une étude empirique sur des tâches informatisées en français

—

Test di comprensione scritta in una lingua straniera

Task Lab – uno studio empirico sugli esercizi al computer in francese

—

Assessing reading comprehension in a foreign language

Task Lab – an empirical study of computer-based tasks in French

Katharina Karges, Malgorzata Barras, Peter Lenz

2021 Bericht des Wissenschaftlichen Kompetenzzentrums für Mehrsprachigkeit
Rapport du Centre scientifique de compétence sur le plurilinguisme
Rapporto del Centro scientifico di competenza per il plurilinguismo
Report of the Research Centre on Multilingualism

Index

Leseverstehen in einer Fremdsprache testen

Task Lab – eine empirische Studie zu computerbasierten Aufgaben in Französisch

—

Katharina Karges, Malgorzata Barras, Peter Lenz

Wie kam es zu dieser Studie?[1]

Im Frühjahr 2017 liess die Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK) erstmals schweizweit das Erreichen der Grundkompetenzen in der ersten Schulfremdsprache überprüfen, und zwar am Ende der Primarstufe. Auf das Frühjahr 2020 hin wurde die zweite „Überprüfung des Erreichens der Grundkompetenzen" (ÜGK[2]) vorbereitet,[3] diesmal am Ende der obligatorischen Schulzeit in der ersten *und* zweiten Schulfremdsprache.[4] Das Kompetenzzentrum für Mehrsprachigkeit (KFM) war bei beiden Erhebungen im Vorfeld mit der Entwicklung von Testaufgaben betraut.

Unabhängig von der EDK wurden im Forschungsprojekt „Task Lab" computerbasierte Leseverstehensaufgaben entwickelt und ihr Funktionieren näher untersucht. Dadurch wurde es möglich, für die erste ÜGK-Aufgabenentwicklung eine Reihe von Entscheidungen zu treffen, die auf empirischen Erfahrungen mit der Zielgruppe basierten. Aus Sicht der Testforschung leistet Task Lab einen Beitrag dazu, genauer zu beschreiben, welche Rolle verschiedene Kenntnisse und Fähigkeiten beim Lösen von Leseverstehensaufgaben in einer schulischen Fremdsprache spielen.

1 Wir bedanken uns bei den vielen Lehrpersonen und Schuldirektorinnen und -direktoren für die Unterstützung unseres Projekts und bei allen Schülerinnen und Schülern, die uns für diese Studie in ihre Klassenzimmer und ihre Köpfe blicken liessen. Auch ohne die tatkräftige Unterstützung von Thomas Aeppli und unseren studentischen Hilfskräften wäre die Datenerhebung nie über die Planung hinausgekommen.

2 Näheres unter https://uegk-schweiz.ch/.

3 Die Haupterhebung konnte wegen der COVID-19-Massnahmen nicht stattfinden und wurde auf einen späteren Zeitpunkt verschoben.

4 In der Westschweiz ist Deutsch die erste Fremdsprache, in der Deutschschweiz entlang der westlichen Sprachgrenze Französisch. In beiden Regionen ist Englisch zweite Fremdsprache, in der übrigen Deutschschweiz wird zuerst Englisch und dann Französisch unterrichtet. Im Tessin ist Französisch die erste und Deutsch die zweite Fremdsprache. Der Kanton Graubünden beteiligte sich an der ÜGK 2017 nicht, an der geplanten ÜGK 2020 nur teilweise (in Deutsch- und Italienischbünden mit Englisch als Fremdsprache).

Was waren die Ziele der Studie?

Für Task Lab wurden Leseverstehensaufgaben entwickelt, wie sie auch in einer schweizweiten Erhebung von Fremdsprachenkompetenzen wie der ÜGK eingesetzt werden könnten. In dem Forschungsprojekt wurde unter anderem das Funktionieren von drei verschiedenen Aufgabenformaten untersucht (Multiple-Choice, Kurzantwort, Matching). Ausserdem wurden die Aufgaben in zwei Sprachversionen eingesetzt, um der Frage nachzugehen, ob die Fragen und Antworten zu den Lesetexten in der Fremdsprache, d.h. in der Sprache der Lesetexte (Französisch), oder in der Schulsprache (Deutsch) formuliert sein sollten.

Daneben sollte auch näher erkundet werden, was Leseverstehenskompetenz eigentlich ausmacht. Deshalb wurden bei den beteiligten Schülerinnen und Schülern weitere Fertigkeiten und Kompetenzen erhoben, von denen man weiss, dass sie mit der Leseverstehenskompetenz in der Fremdsprache zusammenhängen (z. B. rezeptive Wortschatzkenntnisse, das Vorhandensein eines Sichtwortschatzes in der Fremdsprache, die Kapazität des Arbeitsgedächtnisses oder die Motivation für das Sprachfach; vgl. Alderson et al., 2015; Harsch & Hartig, 2016; Sabatini et al., 2013).

Übergeordnetes Ziel war es, die Erfolgsfaktoren beim Leseverstehen besser zu durchschauen und so letztlich zu gültigeren Interpretationen von Testergebnissen und Testskalen bei grossen Untersuchungen zu kommen.

Wer hat an der Studie teilgenommen?

Die Aufgaben wurden Schülerinnen und Schülern vorgelegt, die gerade das letzte Schuljahr der Primarstufe besuchten. Aus praktischen Gründen wurden in Task Lab nur Aufgaben in der ersten Fremdsprache Französisch eingesetzt. Die Studie fand in den fünf Kantonen Basel-Landschaft, Solothurn, Bern, Freiburg und Wallis statt.

Wie bei Testentwicklungen üblich, wurden zuerst sämtliche Aufgaben erprobt und anschliessend pilotiert (vgl. z. B. Kenyon & MacGregor, 2012). Während der Erprobung wurden die einzelnen Aufgaben und die Lösungswege in sogenannten *Stimulated-Recall*-Interviews mit einzelnen Lernenden genau besprochen. Aufgrund der dabei entstehenden Erkenntnisse konnten die Aufgaben weiter optimiert werden. An diesen Interviews beteiligten sich 34 Schülerinnen und Schüler aus zwei Schulklassen. An der Pilotierung beteiligten sich dann 97 Schülerinnen und Schüler aus 5 Klassen. Sie bearbeiteten die gesamte Testbatterie unter realen Bedingungen im Klassenverband. Nach der Pilotierung wurden letzte Anpassungen vorgenommen.

An der Hauptstudie nahmen im Frühsommer 2015 insgesamt 623 Schülerinnen und Schüler aus 34 Klassen teil (309 Jungen, 314 Mädchen, davon hatten rund 20% einen Migrationshintergrund). Die beteiligten Schulen meldeten sich freiwillig für die Teilnahme an der Studie, es handelt sich also nicht um eine repräsentative Stichprobe. Die Studie wurde von den Projektmitarbeitenden und geschulten Hilfskräften in den jeweiligen Schulen während der regulären Unterrichtszeit durchgeführt.

Welche Art von Aufgaben wurden eingesetzt?

Alle Leseverstehensaufgaben bestanden aus einem oder mehreren Lesetexten in französischer Sprache, zu denen jeweils drei Fragen gestellt wurden. Für die Aufgabenentwicklung wurden wichtige Erkenntnisse zum Leseverstehen und zur Konstruktion von Testaufgaben aus der Fachliteratur (u. a. Alderson, 2000; Alderson et al., 2015; Grabe, 2009; Khalifa & Weir, 2009; Lutjeharms & Schmidt, 2010), relevante Lernzielbeschreibungen (Bersinger et al., 2005; D-EDK, 2013; EDK, 2011; Europarat, 2001; Passepartout, 2013 u. a.) sowie das in der Zielregion verwendete Lehrwerk *Mille feuilles* (Bertschy et al., 2011ff.) herangezogen. Alle Aufgaben waren im Umkreis des Niveaus A1 des Gemeinsamen europäischen Referenzrahmens für Sprachen (Europarat, 2001) angesiedelt. Dieses Niveau entspricht den in der Schweiz geltenden Grundkompetenzen für das Leseverstehen in der ersten Fremdsprache am Ende der Primarschule.

Der gesamte Test wurde mittels der Software CBA ItemBuilder[5] für den Einsatz am Computer entwickelt und in den Schulen über einen Internet-Browser abgerufen.[6]

Zwölf der insgesamt 18 entwickelten Aufgaben lagen sowohl im Multiple-Choice-Format (MCQ) als auch als Kurzantwort-Aufgaben (SAQ) vor. Ausserdem existierten beide Itemformate jeweils in zwei Sprachversionen: eine mit französischen Fragen und Antworten und eine mit deutschen Fragen und Antworten (bei den SAQ mussten die Lernenden ihre Antwort in der entsprechenden Sprache schreiben). Insgesamt wurde jede dieser zwölf Leseverstehensaufgaben also in vier Versionen eingesetzt, von denen die einzelnen Schülerinnen und Schüler je eine lösten. → Abb. 1

5 Aktuelle Informationen zu dieser Software: https://tba.dipf.de/de/infrastruktur/softwareentwicklung/cba-itembuilder?set_language=de

6 Wir danken dem Leibniz-Institut für Bildungsforschung und Bildungsinformation (DIPF), insbesondere Helge Einspanier, für die freundliche Unterstützung und die Zurverfügungstellung des Testservers.

Abbildung 1: Eine Task Lab-Aufgabe in vier Versionen

Sechs weitere Aufgaben wurden im Format Matching entwickelt, bei welchem die Lernenden aus mehreren kurzen Texten einen auswählen sollten, der eine gesuchte Information enthielt. Auch dieses Format wurde in zwei Sprachversionen mit Fragen in deutscher bzw. französischer Sprache eingesetzt. In einem Kurzfragebogen wurden die Lernenden befragt, welches Format bzw. welche Sprache sie bevorzugten.

Neben den Leseverstehensaufgaben wurden Aufgaben zu Kompetenzen mit einem bekannten Bezug zur Leseverstehenskompetenz sowie ein Fragebogen eingesetzt. Mit diesen zusätzlichen Aufgaben wurden v. a. weitere Aspekte von Sprachkompetenz erfasst, darunter die allgemeine lexiko-grammatische Kompetenz der Lernenden (C-Test[7] und Wortsegmentierung[8] auf Deutsch und Französisch) sowie Tests zur Wortschatzbreite, automatisierten Worterkennung und zum phonologischen Bewusstsein in der Fremdsprache (rezeptiver Wortschatztest, Sichtwortschatz, Laut-Schrift-Zuordnung[9]). Ausserdem wurden Tests zur Erfassung des Arbeitsgedächtnisses eingesetzt (sog. Digit-Span-Tasks[10]). Im Fragebogen wurden die Schülerinnen und Schüler zu ihrer Herkunft und Erstsprache(n), v. a. aber zu ihren Lesegewohnheiten und ihrer Motivation für den Französischunterricht befragt.→ Abb. 2

7 In mehreren kurzen Texten wird in jedem zweiten Wort die Hälfte gelöscht (der erste und letzte Satz ausgenommen). Die Lernenden müssen diese Wörter rekonstruieren. Dieses Testformat erlaubt es, in kurzer Zeit eine grobe Einschätzung der Sprachkompetenz von Lernenden zu erhalten, da die Lernenden gleichzeitig Wortschatz-, Grammatik- und Orthographiekenntnisse einsetzen müssen.

8 Die Lernenden markieren in einem Text ohne Leerzeichen die Wortgrenzen.

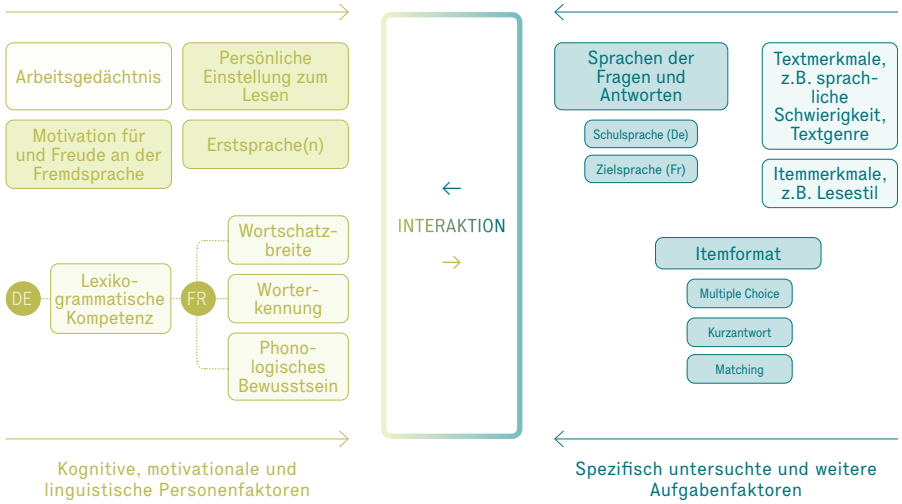


Abbildung 2: Schematische Darstellung der Task-Lab-Testbatterie[11]

9 *Rezeptiver Wortschatztest*: Die Lernenden kreuzen in einer Liste jene Wörter an, die ihnen bekannt sind. Ein Teil der Wörter sind sog. Pseudowörter, also Buchstabenfolgen, die Wörter in der Fremdsprache sein könnten, es aber nicht sind. *Sichtwortschatz*: Die Lernenden sprechen ein Wort laut aus, das nur wenige Millisekunden auf einem Bildschirm erschienen ist. *Laut-Schrift-Zuordnung*: Die Lernenden lesen Pseudowörter laut vor.

10 Die Lernenden geben immer länger werdende Zahlenfolgen wieder, z. T. in umgekehrter Reihenfolge.

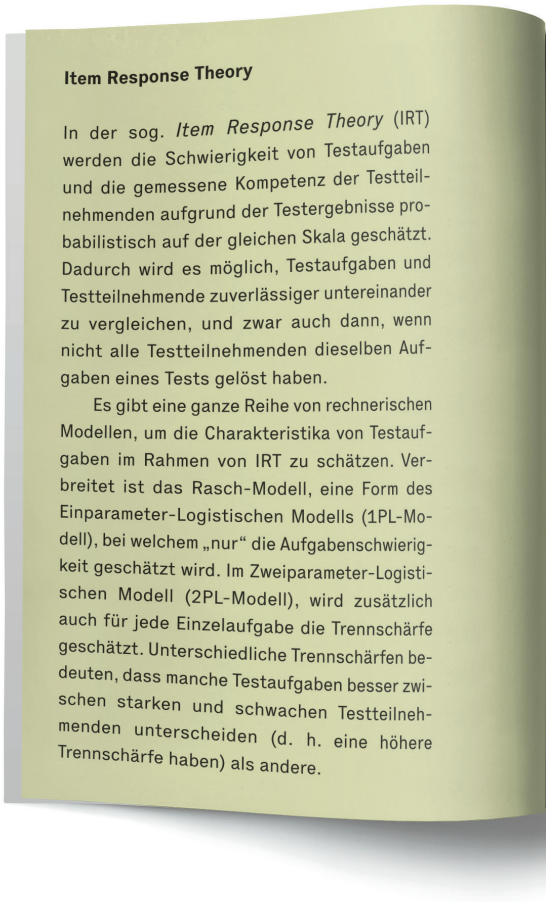
11 In der Grafik sind Aufgabenfaktoren aufgeführt, die im Projekt aus Ressourcengründen nicht analysiert wurden. Da die Daten jedoch im Forschungsdatenarchiv des KFM vorliegen, ist es möglich, auch deren Einfluss auf die Testergebnisse zu analysieren. Interessierte Forschende sind herzlich dazu eingeladen.

Wie wurden die Daten ausgewertet?

Die Interviews aus den Erprobungen sowie die Antworten aus dem Kurzfragebogen wurden transkribiert und mit Blick auf zwei Fragestellungen qualitativ im Sinne der strukturierenden Qualitativen Inhaltsanalyse (Mayring, 2010) ausgewertet: den Lösungsweg der Lernenden und die Sprache der Fragen und Antworten. Ergebnisse dieser Auswertungen können in Barras (2018, auch mit einer Diskussion der Erhebungsmethode) und Karges et al. (i.V., zur Sprache der Fragen und Antworten) nachgelesen werden.

Die Antworten der Testteilnehmenden aus den Tests wurden zum Teil automatisch und zum Teil durch Mitglieder des Projektteams erfasst und bewertet. Die Kurzantworten aus den Leseverstehenstests wurden doppelt beurteilt und dann durch Diskussion vereinheitlicht. Bei den zusätzlichen Tests wurden die Testresultate teilweise doppelt beurteilt, um die Qualität der menschlichen Ratings zu überprüfen.

Die dabei entstandenen Testresultate wurden Qualitätskontrollen unterzogen (z. B. wurden auffällige Antwortmuster identifiziert) und skaliert (u. a. mittels Item Response Theory [IRT], vgl. Kasten „Item Response Theory“). Schliesslich wurden die skalierten Testergebnisse aller Schülerinnen und Schüler über alle Tests hinweg sowie Informationen aus dem Fragebogen in einem einzelnen Datensatz zusammengefasst. Dieser bildete die Grundlage für weiterführende Analysen, insbesondere in Lenz et al. (2019), wo der Effekt des Aufgabenformats näher untersucht wurde.



Item Response Theory

In der sog. *Item Response Theory* (IRT) werden die Schwierigkeit von Testaufgaben und die gemessene Kompetenz der Testteilnehmenden aufgrund der Testergebnisse probabilistisch auf der gleichen Skala geschätzt. Dadurch wird es möglich, Testaufgaben und Testteilnehmende zuverlässiger untereinander zu vergleichen, und zwar auch dann, wenn nicht alle Testteilnehmenden dieselben Aufgaben eines Tests gelöst haben.

Es gibt eine ganze Reihe von rechnerischen Modellen, um die Charakteristika von Testaufgaben im Rahmen von IRT zu schätzen. Verbreitet ist das Rasch-Modell, eine Form des Einparameter-Logistischen Modells (1PL-Modell), bei welchem „nur“ die Aufgabenschwierigkeit geschätzt wird. Im Zweiparameter-Logistischen Modell (2PL-Modell), wird zusätzlich auch für jede Einzelaufgabe die Trennschärfe geschätzt. Unterschiedliche Trennschärfen bedeuten, dass manche Testaufgaben besser zwischen starken und schwachen Testteilnehmenden unterscheiden (d. h. eine höhere Trennschärfe haben) als andere.

Sämtliche Rohdaten, die annotierten qualitativen Interviews sowie die skalierten Testergebnisse sind im Forschungsdatenarchiv des Kompetenzzentrums für Mehrsprachigkeit archiviert und können für weitere Analysen genutzt werden. Auch die eingesetzten Tests sind dort verfügbar, z. B. für Untersuchungen zum Einfluss von Text- und Aufgabenmerkmalen auf Testergebnisse.[12]

Ausgewählte Resultate

Sollten die Fragen und Antworten in einem einfachen Leseverstehenstest in der Fremdsprache verfasst werden?

Eher nicht. Die Resultate unserer Hauptstudie und die Erkenntnisse aus den Erprobungsinterviews weisen klar darauf hin, dass die Nutzung einer gemeinsamen Schulsprache dazu beiträgt, die fremdsprachliche Leseverstehenskompetenz der Schülerinnen und Schüler zuverlässiger zu testen. Zumindest für den von uns untersuchten Kontext (Lernende auf einem niedrigen Sprachniveau und grossflächig anwendbare standardisierte Tests, Existenz einer gemeinsamen Sprache[13]) empfehlen wir daher, die Aufgaben in der Schulsprache zu verfassen.

Im Projekt konnte umfangreiche Evidenz dafür gefunden werden, dass Lernende mit einem Sprachniveau im Bereich A1-A2 Aufgabenstellungen in der Fremdsprache oft zu wenig präzise verstehen und deshalb auf mehr oder weniger erfolgreiche Kompensationsstrategien ausweichen oder einfach raten. Dabei konnten wir verschiedentlich sogar beobachten, dass die Lernenden die gesuchte Information im Lesetext eigentlich gefunden hätten, wenn sie die Frage oder die Antwortoptionen ausreichend verstanden hätten. Ein Misserfolg entstand also nicht wegen eines Mangels an fremdsprachlicher Leseverstehenskompetenz, sondern aus Gründen, die mit dem Verstehen des Textes nichts zu tun hatten.

Auch eine explizite Befragung der Kinder, die an der Studie teilgenommen haben, führt zu diesem Ergebnis: Die meisten Lernenden sind sich einig, dass es für sie einfacher ist, wenn die Fragen zum Text in ihrer Schulsprache Deutsch gehalten sind, da sie diese dann sicher verstehen. Nur rund 10% der befragten Lernenden gaben an, in diesem Fall das Französische zu bevorzugen. Viele dieser Schülerinnen und Schüler begründeten dies damit, dass sie dann einfacher Wörter oder Wendungen im Lesetext wiederfinden könnten, die ihnen Hinweise auf die (vermeintlich) richtige Antwort geben – eine Testlösungsstrategie, die nicht zwingend auf gute Lesekompetenzen schliessen lässt und bei professionell konstruierten Aufgaben oft scheitert.

Auch aus Sicht des Testkonstrukts spricht viel für die Verwendung der Schulsprache: Wer liest, tut dies in der Regel mit einem Ziel (z. B., um eine bestimmte Information zu finden, aber auch zum Vergnügen). Ein Test bzw. eine Testaufgabe gibt dieses Handlungsziel meist vor. Wenn die Testteilnehmenden dieses Handlungsziel aber nicht verstehen, können sie die Lesehandlung nicht so ausführen, wie es der Test vorsieht. So können Testresultate entstehen, die nicht dem Testziel entsprechen

12 https://tinyurl.com/pwcy7p39

13 Bei internationalen Sprachtests, wo keine gemeinsame Sprache vorhanden ist, muss die Verständlichkeit der Fragen auf anderen Wegen hergestellt werden (z. B. durch sehr einfache Formulierungen, Bilder und durch die Möglichkeit eines vorhergehenden Testtrainings). Dies schränkt die Vielfalt der möglichen Fragestellungen aber ein.

und – im schlimmsten Fall – nicht direkt mit der Lesekompetenz der Testteilnehmenden zusammenhängen.

Spielt es eine Rolle, welche Aufgabenformate in einem Leseverstehenstest eingesetzt werden?

Ja und nein. Statistische Analysen der Testergebnisse lassen vermuten, dass die beiden Testformate Multiple-Choice (MCQ) und Kurzantwort (SAQ) zu grossen Teilen das gleiche Konstrukt messen, inhaltlich also kein Format dem anderen vorzuziehen ist.

Allerdings zeigt sich auch, dass MCQs im Durchschnitt a) signifikant einfacher und b) weniger trennscharf sind als die SAQs mit dem *gleichen* Wortlaut. Das bedeutet, dass Lernende „mehr können" müssen, um eine SAQ korrekt zu beantworten als dies bei der gleichen MCQ der Fall wäre. Ausserdem unterscheiden SAQs genauer zwischen schwächeren und stärkeren Lernenden. Beide Effekte lassen sich u. a. durch die unterschiedlichen Antwortprozesse bei SAQ und MCQ erklären: Bei SAQ muss mithilfe des Textes eine Antwort gefunden werden, die dann auch aufgeschrieben werden muss. An dieser Stelle steckt eine zusätzliche Fehlerquelle, was die Wahrscheinlichkeit einer korrekten Antwort senkt und so eine Aufgabe schwieriger und trennschärfer macht. Bei MCQ mit drei Antwortoptionen besteht dagegen bereits eine gewisse Wahrscheinlichkeit, die richtige Lösung zu finden, ohne dass überhaupt etwas gelesen wird. Auch die Möglichkeit, durch geschicktes Raten, statt durch Sprachverstehen eine richtige Antwort zu geben, reduziert die Qualität der MCQ-Items als Messinstrumente für Sprachkompetenzen.

Im Hinblick auf ein Testergebnis bedeutet das, dass eine korrekte Kurzantwort im Durchschnitt mehr darüber aussagt, welches Sprachkompetenzniveau der oder die Getestete hat, als dies eine korrekte Multiple-Choice-Antwort tut. Werden die beiden Formate gemischt, sollte dies berücksichtigt werden: Die Testteilnehmenden sollten eigentlich für korrekte Antworten zu SAQ-Aufgaben mehr Punkte bekommen (in unserer Studie wäre im Durchschnitt eine doppelte Gewichtung der SAQ-Antworten im Vergleich zu den MCQ-Aufgaben angemessen gewesen). Sollen die Testergebnisse psychometrisch mittels Item Response Theory (IRT) skaliert werden, so ist die Verwendung eines 2PL-Modells zu empfehlen, das die Trennschärfe jedes Items direkt schätzt und daraus dessen Gewichtung für die Schätzung der Fähigkeit der Testteilnehmenden ableitet. → Kasten IRT

Wie lösen junge Schülerinnen und Schüler einfache Leseverstehensaufgaben in einer Fremdsprache?

Neben den Leseverstehensaufgaben wurden weitere Tests eingesetzt, die Teilkompetenzen erheben sollten, von denen bekannt ist, dass sie mit Leseverstehen zusammenhängen. Insgesamt widerspiegeln die Ergebnisse der verschiedenen Analysen die in der Literatur postulierten Zusammenhänge. Insbesondere zeigte sich, dass auf dem getesteten elementaren Sprachniveau (A-Niveaus des GER) der Wortschatz eine wichtige Rolle spielt: Die rezeptiven Wortschatzkenntnisse sagen das Ergebnis beim Leseverstehenstest relativ gut voraus.

Auch die beiden kurzen Tests zum phonologischen Bewusstsein und zur (automatisierten) Worterkennung stehen in einem deutlichen Zusammenhang mit den Leseverstehenstests, und zwar ähnlich deutlich wie der komplexere C-Test und die Aufgabe zur Wortsegmentierung (bei diesen beiden Aufgaben spielen auch Kenntnisse auf der Textebene eine Rolle). Dies weist darauf hin, dass auch kurze, wenig kontextualisierte Tests einiges über die Sprachkompetenz von Anfängerinnen und Anfängern aussagen können, deren Leseverstehen auf Textebene noch wenig stabil und daher mit Leseverstehenstests schwer messbar ist.

Welche Folgen hatte Task Lab?

Die ÜGK-Erhebungen 2017 und 2020 → vgl. Fussnote 3 in den Schul- und Fremdsprachen profitierten massgeblich von den Erkenntnissen und Erfahrungen aus Task Lab, insbesondere auch in technischer Hinsicht. Das Design und die zugrundeliegende Struktur der Aufgaben wurden in nur leicht geänderter Form übernommen. Fragen und Antworten zum Text werden in den ÜGK-Fremdsprachentests grundsätzlich in der Schulsprache gestellt. Auch Materialien und Erfahrungen aus den Task-Lab-Erprobungen konnten für die Aufgabenentwicklung im Rahmen der ÜGK genutzt werden.

Für den direkten Nachfolger des Projektes Task Lab („Innovative Formen der Beurteilung", IFB[14]) konnten Teile des Erhebungsdesigns übernommen bzw. weiterentwickelt werden. Auch die technische Grundstruktur der Aufgaben aus Task Lab kam dem neuen Projekt sowie weiteren Projekten zugute.

Aus Sicht der Testforschung ist zu betonen, dass Task Lab eines von sehr wenigen Forschungsprojekten ist, das bei jungen Lernenden (Zwölfjährigen) mit elementaren Sprachkenntnissen (A-Niveaus des GER) die Struktur von fremdsprachlicher Lesekompetenz in einer anderen Sprache als Englisch (nämlich Französisch) untersucht hat. Eher selten ist auch die systematische Herangehensweise bei der Untersuchung der Effekte von Merkmalen von Leseverstehensaufgaben, insbesondere der Sprache von Fragen und Antworten (Schulsprache Deutsch oder Zielsprache Französisch) und der Antwortformate (MCQ oder SAQ).

14 Vgl. auch https://centre-plurilinguisme.ch/de/forschung/innovative-formen-der-beurteilung.

Wo finde ich mehr Informationen?

Die folgenden beiden Artikel enthalten detaillierte Darstellungen und Diskussionen zu den oben beschriebenen Erkenntnissen aus dem Forschungsprojekt (beide in englischer Sprache):

- Lenz, P., Karges, K. & Barras, M. (2019). Investigating test method effects in French L2 reading items for young learners. In A. Huhta, G. Erickson & N. Figueras (eds), *Developments in language education: a memorial volume in honour of Sauli Takala*. Jyväskylä: University of Jyväskylä & EALTA, 182-202.
- Karges, K., Barras, M. & Lenz, P. (forthcoming). 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*. Berlin: Peter Lang.

Im folgenden Artikel wird das qualitativ-methodische Vorgehen anlässlich der Aufgabenerprobung vorgestellt. Der Artikel enthält auch ausgewählte Ergebnisse des Projekts Task Lab:

- Barras, M. (2018). „Stimulated Recall" in der Sprachtestforschung. Ein praktisches Beispiel aus der Erprobung eines computerbasierten Leseverstehenstests. In K. Aguado, C. Finkbeiner & B. Tesch (Hrsg.), *Lautes Denken, „Stimulated Recall" und Dokumentarische Methode. Rekonstruktive Verfahren in der Fremdsprachenlehr- und -lernforschung*. Frankfurt: Peter Lang, 69-86.

Schliesslich wurden die folgenden kurzen Übersichtsartikel veröffentlicht:

- Barras, M., Karges, K. & Lenz, P. (2016). Leseverstehen überprüfen: Welche Sprache für die Fragen und Antworten in den Testitems? *Babylonia, 16*(2), 13-18.
- Karges, K., Barras, M. & Lenz, P. (2016). Task Lab: Untersuchungen zum besseren Verständnis von computerbasierten kommunikativen Testaufgaben zum Leseverstehen in Französisch. *Babylonia, 2016*(3), 56.

Daneben können auf der Webseite des Kompetenzzentrums für Mehrsprachigkeit diverse Poster und PDFs von Vorträgen eingesehen werden, die an verschiedenen Fachkonferenzen präsentiert wurden.[15]

15 https://centre-plurilinguisme.ch/de/forschung/task-lab-untersuchungen-zum-besseren-verstaendnis-und-zur-erhoehung-der-validitaet-von.

Tester la compréhension de l’écrit dans une langue étrangère

Task Lab – une étude empirique sur des tâches informatisées en français

—

Katharina Karges, Malgorzata Barras, Peter Lenz

Les origines de cette étude ?[1]

Au printemps 2017, la Conférence suisse des directeurs cantonaux de l'instruction publique (CDIP) a, pour la première fois, mené une enquête nationale et fait tester les compétences fondamentales des écoliers dans leur première langue étrangère à la fin de l'école primaire. La seconde « Vérification de l'atteinte des compétences fondamentales » (COFO[2]) était prévue pour le printemps 2020.[3] Cette fois, la vérification devait porter sur les compétences dans la première *et* la seconde langue étrangère.[4] Le Centre scientifique de compétence sur le plurilinguisme (CSP) était impliqué dans les phases préparatoires de ces deux enquêtes et chargé du développement des tâches de test.

Indépendamment des enquêtes de la CDIP, le CSP avait développé des tâches informatisées de compréhension de l'écrit dans le cadre du projet « Task Lab » et approfondi l'étude de leur fonctionnement. Les expériences empiriques faites avec le groupe cible dans Task Lab ont ensuite permis de prendre un certain nombre de décisions quant au développement des tâches destinées à la première COFO. En termes de recherche, Task Lab a permis de décrire plus précisément le rôle que jouent les différentes connaissances et compétences dans la résolution de tâches de compréhension de l'écrit dans une langue étrangère.

1 Nos remerciements vont aux nombreux et nombreuses enseignants et enseignantes, directeurs et directrices d'école qui ont soutenu notre projet ainsi qu'aux élèves qui nous ont ouvert leur classe et leurs esprits pour cette étude. Nous remercions également Thomas Aeppli et nos asssistant·e·s scientifiques : sans leur soutien efficace, notre récolte de données n'aurait pas dépassé le stade de la planification.

2 Pour plus d'informations : https://cofo-suisse.ch/.

3 La récolte de données principale n'a pas pu être réalisée en raison des mesures liées à la COVID-19 et aura lieu ultérieurement.

4 La première langue étrangère est l'allemand en Suisse romande et le français en Suisse alémanique, le long de la frontière linguistique occidentale. Dans les deux régions, l'anglais est la deuxième langue étrangère. Dans le reste de la Suisse alémanique, l'anglais est enseigné en 1er lieu, puis le français. Au Tessin, le français est la première langue étrangère et l'allemand la seconde. Le canton des Grisons n'a pas participé à la COFO de 2017, et ne devait participer que partiellement à la COFO prévue pour 2020 (communes germanophones et italophones ayant l'anglais comme langue étrangère).

Quels étaient les buts de l'étude ?

Les tâches de compréhension de l'écrit développées dans le projet ont été conçues de manière à pouvoir être utilisées pour les besoins d'une enquête d'envergure nationale portant sur les compétences en langue étrangère, telle que la COFO. Le projet de recherche s'est notamment penché sur le fonctionnement de trois formats de tâches différents (questions à choix multiple, questions à réponse courte ouverte, questions d'appariement). De plus, les tâches ont été construites en deux langues pour déterminer si les questions et les réponses se rapportant aux textes à lire devaient être formulées dans la langue étrangère, donc la langue des textes (le français), ou dans la langue de scolarisation (l'allemand).

Il s'agissait également de mieux comprendre en quoi consiste réellement la compétence en compréhension de l'écrit. Par conséquent, d'autres aptitudes et compétences, dont on sait qu'elles sont liées à la compréhension de l'écrit dans la langue étrangère (par exemple, le vocabulaire réceptif, la présence d'un vocabulaire visuel dans la langue étrangère, la capacité de la mémoire de travail ou la motivation à étudier la langue en tant que matière ; cf. Alderson et al., 2015 ; Harsch & Hartig, 2016 ; Sabatini et al., 2013), ont été étudiées chez les élèves participant au projet.

L'objectif général était de mieux comprendre les facteurs de réussite de la compréhension de l'écrit et d'ainsi augmenter la validité des résultats et échelles de tests dans les études de grande ampleur.

Qui a participé à l'étude ?

Les tâches ont été proposées à des élèves de dernière année d'école primaire et, pour des raisons pratiques, seulement en français, la première langue étrangère. L'étude a été menée dans cinq cantons : Bâle-Campagne, Soleure, Berne, Fribourg et Valais.

Comme il est d'usage lors de développements de tests, toutes les tâches ont d'abord été testées puis implémentées au cours d'une phase pilote (voir par exemple, Kenyon & MacGregor, 2012). Dans la phase de test, chaque tâche et solution a été précisément discutée avec des apprenant·e·s dans le cadre d'entretiens individuels de rappel stimulé et les informations ainsi obtenues ont permis d'optimiser les tâches. Ces entretiens ont été réalisés avec 34 élèves issus de deux classes. Le projet pilote, lui, a été mené avec 5 classes, soit 97 élèves, qui ont passé toute la batterie des tests en conditions réelles de classe. Cette phase pilote a donné lieu aux derniers ajustements.

L'étude principale a été menée au début de l'été 2015 et a réuni un total de 34 classes, soit 623 élèves (309 garçons, 314 filles) parmi lesquels environ 20 % étaient issus de la migration. Les écoles s'étant portées volontaires pour participer au projet, il ne s'agissait pas d'un échantillon représentatif. L'étude a été menée dans les écoles par les collaborateurs et collaboratrices du projet et des assistant·e·s scientifiques pendant les heures de cours normales.

A quelles tâches a-t-on eu recours ?

Toutes les tâches de compréhension de l'écrit consistaient en un ou plusieurs textes à lire en français, pour chacun desquels trois questions étaient posées. La construction des tâches s'est appuyée sur la littérature spécialisée et notamment d'importantes découvertes en matière de compréhension de l'écrit et d'élaboration de tâches de test (par exemple, Alderson, 2000 ; Alderson et al., 2015 ; Grabe, 2009 ; Khalifa & Weir, 2009 ; Lutjeharms & Schmidt, 2010), des descriptions d'objectifs d'apprentissage (Bersinger et al., 2005 ; D-EDK, 2013 ; EDK, 2011 ; Conseil de l'Europe, 2001 ; Passepartout, 2013 ; etc.) ainsi que *Mille feuilles,* le manuel scolaire utilisé dans la région concernée (Bertschy et al., 2011ff.). Le niveau de langue ciblé par les tâches se situait autour du niveau A1 du Cadre européen commun de référence pour les langues (Conseil de l'Europe, 2001), ce qui correspond en Suisse aux compétences de base à atteindre à la fin de l'école primaire pour la compréhension de l'écrit dans la première langue étrangère.

Développé à l'aide du logiciel CBA ItemBuilder,[5] l'ensemble du test était à réaliser par ordinateur et accessible dans les écoles via un navigateur Internet.[6]

12 tâches sur les 18 que comptait le test étaient de format choix multiple (QCM) et de type question à réponse ouverte courte (QROC). Ces deux formats d'items existaient en deux versions de langue : l'une avec des questions et réponses en français et l'autre avec des questions et réponses en allemand (dans le cas des QROC, les apprenant·e·s devaient écrire leurs réponses dans la langue correspondante). Au total, chacune des 12 tâches de compréhension de l'écrit existait dans 4 versions, et chaque élève devait en résoudre une. → Fig. 1

5 Informations supplémentaires concernant ce logiciel (en anglais ou allemand) : https://tba.dipf.de/de/infrastruktur/softwareentwicklung/cba-itembuilder?set_language=de.

6 Nous tenons à remercier le Leibniz-Institut für Bildungsforschung und Bildungsinformation (DIPF), en particulier Helge Einspanier, pour son aimable soutien et pour avoir fourni le serveur de test.

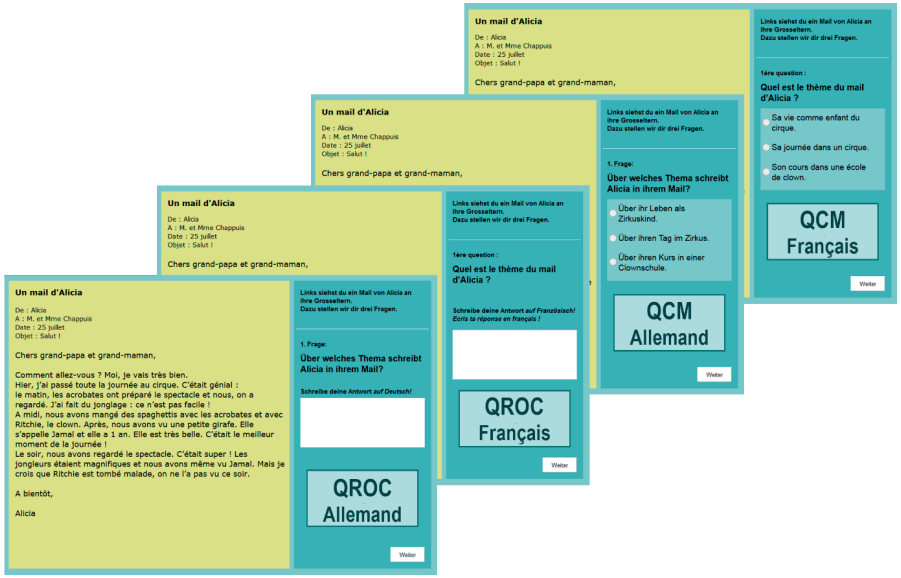


Figure 1 : Les 4 versions d'une tâche développée dans Task Lab

A cela se sont ajoutées six autres tâches de type appariement, un format dans lequel les apprenant·e·s devaient choisir parmi plusieurs textes courts celui qui contenait l'information recherchée. Ce format aussi était disponible dans deux versions de langues (avec des questions en allemand et en français). Les apprenant·e·s ont été invité·e·s à indiquer le format ou la langue qu'ils préféraient en complétant un bref questionnaire.

En plus des tâches de compréhension de l'écrit, le projet a proposé un questionnaire et des tâches portant sur des compétences ayant une relation avérée avec la compréhension de l'écrit. Ces tâches ont principalement permis d'évaluer d'autres aspects de la compétence langagière, notamment la compétence lexico-grammaticale générale des apprenant·e·s (C-Test[7] et segmentation des mots[8] en allemand et en français) ainsi que de procéder à des tests sur la richesse du vocabulaire, la reconnaissance automatique des mots et la conscience phonologique dans la langue étrangère (test de vocabulaire

7 Dans plusieurs textes courts, on supprime la moitié d'un mot sur deux (à l'exception de la première et de la dernière phrase) et les apprenant·e·s doivent les reconstruire. Ce type de test permet d'obtenir en peu de temps une évaluation globale des compétences linguistiques des apprenant·e·s, ces dernier·e·s devant à la fois mobiliser leurs connaissances en matière de vocabulaire, de grammaire et d'orthographe.

8 Dans un texte sans espaces, les apprenant·e·s doivent délimiter les mots.

réceptif, de vocabulaire visuel, d'association son et lettres[9]). L'on a également eu recours à des tests de mémoire de travail (tâches d'empan de chiffres[10]). Quant au questionnaire, il a permis de relever des informations sur l'origine et la(les) langue(s) première(s) des élèves, mais surtout sur leurs habitudes de lecture et leur motivation à apprendre le français. → Fig. 2

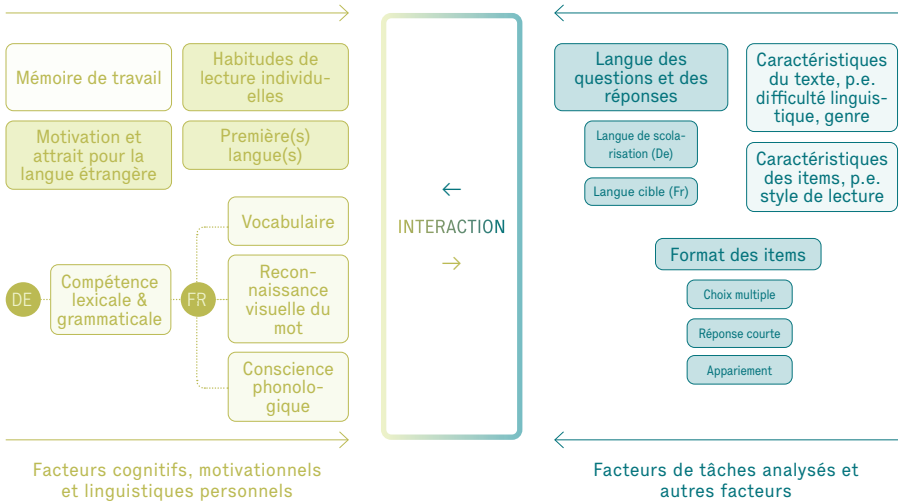


Figure 2 : Représentation schématique de la batterie de tests Task Lab[11]

9 *Test de vocabulaire réceptif* : dans une liste, les étudiant·e·s cochent les mots qu'ils-elles connaissent. Certains de ces mots sont des pseudo-mots, c'est-à-dire des séquences de lettres qui pourraient être des mots dans la langue étrangère mais qui ne le sont pas. *Vocabulaire visuel* : les apprenant·e·s prononcent à haute voix un mot qui n'est apparu sur un écran que pendant quelques millisecondes. *Association son et lettres* : les apprenant·e·s lisent des pseudo-mots à voix haute.

10 Les apprenant·e·s répètent des séquences de nombres de plus en plus longues, parfois à rebours.

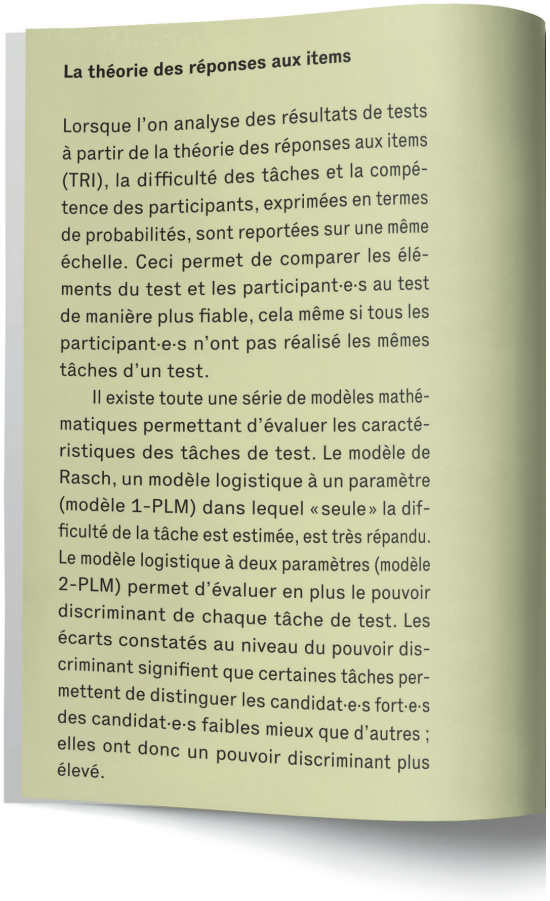
11 Le graphique montre des facteurs de tâches qui n'ont pas été analysés dans le cadre du projet pour des raisons de ressources. Les données étant toutefois disponibles dans les archives de recherche du CSP, il serait possible d'analyser leur influence sur les résultats des tests. Les chercheur·e·s intéressé·e·s sont cordialement invité·e·s à le faire.

Comment les données ont-elles été analysées ?

Les entretiens menés lors des tests préalables ainsi que les réponses au questionnaire court ont été transcrits puis soumis à une analyse qualitative de contenu (Mayring, 2010) visant à évaluer deux éléments : d'une part, l'approche utilisée par les apprenant·e·s pour résoudre la tâche et d'autre part, la langue utilisé dans les questions et les réponses. Les résultats de ces évaluations sont présentés dans Barras (2018, y compris une description de la méthode de récolte des données) et Karges et al. (à paraître, sur la langue des questions et réponses).

Les réponses des participant·e·s aux tests ont été enregistrées et évaluées en partie automatiquement et en partie par des membres de l'équipe de projet. Deux personnes ont évalué les réponses courtes aux tests de compréhension de l'écrit ; ces évaluations ont ensuite été harmonisées lors d'une discussion. En ce qui concerne les tests supplémentaires, certains résultats ont été analysés à double afin de vérifier la qualité des évaluations humaines.

Les résultats obtenus ont été soumis à des contrôles de qualité (on a par exemple identifié des types de réponse saillants) et reportés sur une échelle (notamment à partir de la théorie des réponses aux items, voir encadré ci-après). Enfin, les résultats ainsi classés de tous les élèves à tous les tests et les informations du questionnaire ont été combinés en un seul ensemble de données.



La théorie des réponses aux items

Lorsque l'on analyse des résultats de tests à partir de la théorie des réponses aux items (TRI), la difficulté des tâches et la compétence des participants, exprimées en termes de probabilités, sont reportées sur une même échelle. Ceci permet de comparer les éléments du test et les participant·e·s au test de manière plus fiable, cela même si tous les participant·e·s n'ont pas réalisé les mêmes tâches d'un test.

Il existe toute une série de modèles mathématiques permettant d'évaluer les caractéristiques des tâches de test. Le modèle de Rasch, un modèle logistique à un paramètre (modèle 1-PLM) dans lequel « seule » la difficulté de la tâche est estimée, est très répandu. Le modèle logistique à deux paramètres (modèle 2-PLM) permet d'évaluer en plus le pouvoir discriminant de chaque tâche de test. Les écarts constatés au niveau du pouvoir discriminant signifient que certaines tâches permettent de distinguer les candidat·e·s fort·e·s des candidat·e·s faibles mieux que d'autres ; elles ont donc un pouvoir discriminant plus élevé.

Cela a servi de base à des analyses plus approfondies, à l'instar de celle de Lenz et al. (2019) qui se penche en détail sur les effets du format des tâches. → encadré

Toutes les données brutes, les entretiens qualitatifs annotés ainsi que les résultats des tests mis à l'échelle se trouvent dans les archives des données de recherche du Centre scientifique de compétence sur le plurilinguisme et peuvent être utilisés pour des analyses plus approfondies. Les tests utilisés y sont également disponibles et l'on peut y recourir, par exemple, dans le cadre d'études sur l'influence des caractéristiques des textes et des tâches sur les résultats des tests.[12]

12 https://portailplurilingue.unifr.ch/starweb/KFM/k.skca-catalog/servlet.starweb?path=KFM/k.skca-catalog/FastLink.web&inum=000011889.

Résultats choisis

Pour une épreuve simple de compréhension de l'écrit, les questions et réponses devraient-elles être rédigées dans la langue étrangère ?

Plutôt pas. Les résultats de notre étude principale et des entretiens indiquent clairement que les tests visant à évaluer les compétences en compréhension de l'écrit des élèves dans une langue étrangère sont plus fiables si l'on utilise une langue de scolarisation commune. Nous recommandons donc de rédiger les tâches dans la langue de scolarisation, tout au moins pour le contexte que nous étudions (apprenant·e·s ayant un faible niveau de maîtrise de la langue étrangère, tests standardisés largement applicables, existence d'une langue commune[13]).

Dans le cadre du projet, il a été largement démontré que les apprenant·e·s dont le niveau de langue se situait entre A1 et A2 ne comprennent souvent pas assez précisément les tâches dans la langue étrangère. Ils-elles ont donc recours à des stratégies de compensation plus ou moins efficaces ou se contentent de deviner. Nous avons même constaté que les apprenant·e·s auraient effectivement trouvé l'information recherchée dans le texte si la question ou les options de réponse avaient été suffisamment compris. Ainsi, l'échec n'était pas dû à un manque de compétence en compréhension de l'écrit, mais à bien d'autres raisons.

Une enquête réalisée auprès des enfants qui ont participé à l'étude le montre : la plupart d'entre eux-elles préfèrent que les questions soient écrites en allemand, leur langue de scolarisation, car ils-elles sont ainsi certain·e·s de les comprendre. Seulement environ 10 % des apprenant·e·s interrogé·e·s déclarent préférer des questions formulées en français. Bon nombre d'entre eux-elles estiment que cela leur permet de trouver dans le texte des mots ou des phrases pouvant les mettre sur la piste de la réponse (supposée) correcte. Il s'agit là d'une stratégie de résolution de tâche qui ne traduit pas nécessairement de bonnes compétences de compréhension de l'écrit et qui échoue souvent pour des tâches élaborées de façon professionnelle.

Du point de vue du construit du test, de nombreux arguments plaident en faveur de la langue de scolarisation. Lorsqu'on lit, on le fait généralement dans un but précis (par exemple : trouver une certaine information, pour le plaisir, etc.). Un test ou une tâche de test précise généralement l'objectif dans lequel on doit lire un texte. Cependant, si cet objectif échappe aux personnes qui passent le test, elles ne peuvent effectuer la tâche de lecture de la manière dont le test l'exige. De cela peuvent découler des résultats qui ne correspondent pas à l'objectif du test et – dans le pire des cas - qui ne sont pas directement liés à la compétence de compréhension de l'écrit des personnes testées.

13 Dans les tests de langue internationaux où il n'existe pas de langue commune, d'autres moyens doivent être mis en œuvre pour permettre aux participant·e·s de comprendre les questions (par exemple, des formulations très simples, des images et la possibilité d'une formation préalable au test). Toutefois, cela limite la diversité des questions.

Le choix des types de tâches utilisés dans un test de compréhension de l'écrit est-il important ?

Oui et non. Les analyses statistiques des résultats aux tests suggèrent que les deux formats de test, à savoir le choix multiple (QCM) et la question à réponse ouverte courte (QROC), mesurent largement la même chose. Du point de vue du contenu, aucun format n'est donc à préférer à l'autre.

Cependant, on remarque aussi que les QCM sont en moyenne a) sensiblement plus simples et b) moins discriminants que les QROC ayant le même libellé. Cela signifie que les apprenant·e·s doivent « en savoir plus » pour répondre correctement à un QROC que ce ne serait le cas pour un QCM équivalent. En outre, les QROC différencient plus précisément les apprenant·e·s faibles des apprenant·e·s fort·e·s. Ces deux effets peuvent s'expliquer par le fait que les processus suivis pour répondre aux QCM et QROC sont différents : dans le cas du second, il faut s'aider du texte pour trouver une réponse que l'on doit ensuite rédiger. Ceci constitue une source d'erreur supplémentaire, qui réduit la probabilité d'une réponse correcte et rend donc la tâche plus difficile et plus sélective. En revanche, dans un QCM, trois options de réponse augmentent la probabilité de trouver la bonne solution sans même rien lire du texte. Il est donc possible de donner une réponse correcte en devinant intelligemment plutôt qu'en comprenant le texte. De ce point de vue, les items du QCM constituent des instruments moins efficaces de mesure des compétences linguistiques.

En termes de résultats de test, cela signifie qu'une réponse correcte à un QROC en dit en moyenne davantage sur le niveau de compétence linguistique de la personne testée qu'une réponse correcte à un QCM. Si les deux formats de tâches sont utilisés, il faut en tenir compte : les candidat·e·s devraient ainsi marquer plus de points pour avoir répondu correctement aux questions du QROC (dans notre étude, il aurait été approprié de doter les QROC d'un coefficient 2). S'il s'agit de constituer une échelle psychométrique à partir des résultats du test en utilisant la théorie des réponses aux items (TRI), il est recommandé d'utiliser un modèle 2-PL. Ce dernier permet de calculer le pouvoir discriminant de chaque item et d'en déduire sa pondération, laquelle sera utilisée pour évaluer la capacité des candidat·e·s au test. → encadré

Comment les jeunes élèves résolvent-ils-elles des tâches simples de compréhension de l'écrit dans une langue étrangère ?

En plus des tâches de compréhension de l'écrit, d'autres tests ont été utilisés pour évaluer les sous-compétences liées à la compréhension de l'écrit. Dans l'ensemble, les résultats des différentes analyses corroborent les liens postulés dans la littérature. Il a notamment été constaté qu'au niveau élémentaire de langue testé (niveaux A du CECR), le vocabulaire joue un rôle important :

les connaissances réceptives du vocabulaire sont un prédicteur relativement précis des résultats du test de compréhension de l'écrit.

Un lien clair a pu être établi entre les tests de compréhension de l'écrit et les deux tests courts portant sur la conscience phonologique et la reconnaissance (automatisée) des mots, tout autant qu'avec le plus complexe C-Test et la tâche de segmentation des mots pour lesquels des connaissances textuelles entrent également en ligne de compte. Ces résultats montrent que même des tests courts et moins contextualisés peuvent donner des indications sur le niveau de compétence linguistique des débutant·e·s dont les connaissances textuelles ne sont pas encore très stables et donc difficiles à mesurer au travers de tests classiques de compréhension de l'écrit.

Quelles ont été les retombées de Task Lab ?

Les enquêtes COFO 2017 et 2020 → note de bas de page 3 sur les langues étrangères et de scolarisation ont bénéficié de manière significative des connaissances et de l'expérience acquises dans Task Lab, notamment sur le plan technique. La conception et la structure de base des tâches n'ont eu besoin que de légères adaptations pour être utilisées dans les enquêtes COFO. De même, les matériaux et l'expérience issus de Task Lab ont aussi pu être utilisés pour le développement des tâches. Enfin, les questions et réponses portant sur les textes sont par principe rédigées dans la langue de scolarisation pour les besoins des enquêtes COFO.

Pour le projet « Formes innovantes d'évaluation » (IFB[14]), qui s'inscrit dans la suite de Task Lab, certaines parties de la procédure de récolte des données ont pu être reprises et développées. La structure technique de base des tâches de Task Lab a également profité au nouveau projet ainsi qu'à d'autres.

En matière de recherche sur les tests, soulignons que Task Lab est l'un des rares projets à avoir étudié, dans une autre langue que l'anglais (à savoir le français), la structure de la compétence de compréhension de l'écrit en langues étrangères chez de jeunes apprenant·e·s (12 ans) ayant un niveau élémentaire (niveau A du CECR). Il est également plutôt rare que l'on étudie de façon systématique les effets des caractéristiques des tâches de compréhension de l'écrit, en particulier la langue des questions et des réponses (langue scolaire : allemand ou langue cible : français) et les formats de réponse (QCM ou QROC).

14 https://centre-plurilinguisme.ch/fr/recherche/formes-innovantes-devaluation.

Où trouver davantage d'informations ?

Les deux articles suivants proposent des présentations et des discussions détaillées autour des résultats du projet de recherche décrit ci-dessus (tous deux en anglais) :

- Lenz, P., Karges, K. & Barras, M. (2019). Investigating test method effects in French L2 reading items for young learners. In A. Huhta, G. Erickson & N. Figueras (eds), *Developments in language education: a memorial volume in honour of Sauli Takala*. Jyväskylä: University of Jyväskylä & EALTA, 182-202.
- Karges, K., Barras, M. & Lenz, P. (forthcoming). 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*. Berlin: Peter Lang.

L'article suivant présente la méthode qualitative utilisée pour la validation des tâches de test. Il présente également une sélection de résultats du projet Task Lab :

- Barras, M. (2018). „Stimulated Recall" in der Sprachtestforschung. Ein praktisches Beispiel aus der Erprobung eines computerbasierten Leseverstehenstests. In K. Aguado, C. Finkbeiner & B. Tesch (Hrsg.), *Lautes Denken, „Stimulated Recall" und Dokumentarische Methode. Rekonstruktive Verfahren in der Fremdsprachenlehr- und -lernforschung*. Frankfurt: Peter Lang, 69-86.

Enfin, voici quelques articles de synthèse :

- Barras, M., Karges, K. & Lenz, P. (2016). Leseverstehen überprüfen: Welche Sprache für die Fragen und Antworten in den Testitems? *Babylonia, 2016*(2), 13-18.
- Karges, K., Barras, M. & Lenz, P. (2016). Task Lab: Untersuchungen zum besseren Verständnis von computerbasierten kommunikativen Testaufgaben zum Leseverstehen in Französisch. *Babylonia, 2016*(3), 56.

De plus, divers posters et supports de conférence sont accessibles sur le site du Centre scientifique de compétence sur le plurilinguisme.[15]

15 https://centre-plurilinguisme.ch/fr/recherche/task-lab-recherches-pour-une-meilleure-comprehension-et-une-augmentation-de-la-validite.

Test di comprensione scritta in una lingua straniera

Task Lab – uno studio empirico sugli esercizi al computer in francese

—

Katharina Karges, Malgorzata Barras, Peter Lenz

Da dove nasce questo studio?[1]

Nella primavera 2017, la Conferenza svizzera dei direttori cantonali della pubblica educazione (CDPE) ha condotto per la prima volta un rilevamento volto a verificare il raggiungimento delle competenze fondamentali nella prima lingua straniera alla fine della scuola elementare (VeCoF[2]). In vista della primavera 2020, invece, era stato preparato un secondo studio,[3] questa volta dedicato alla prima *e* alla seconda lingua straniera alla fine della scuola dell'obbligo.[4] Il Centro di competenza per il plurilinguismo ha partecipato a entrambi i rilevamenti elaborando esercizi per i test.

In modo indipendente dalla CDPE, nel quadro del progetto di ricerca Task Lab sono stati sviluppati esercizi in formato elettronico di comprensione scritta e ne è stato esaminato il funzionamento. Per il primo sviluppo di esercizi VeCoF è stato così possibile prendere una serie di decisioni basate su esperienze empiriche con il gruppo mirato. Dal punto di vista della ricerca, Task Lab contribuisce a descrivere più in dettaglio quale ruolo rivestono determinate conoscenze e capacità nel risolvere esercizi di comprensione scritta di una lingua straniera insegnata a scuola.

1 Ringraziamo le/i numerosi insegnanti e direttrici/direttori di scuola per il sostegno al nostro progetto e tutte/i le/gli allievi che ci hanno aperto le porte delle loro aule per condurre questo studio. Il rilevamento dei dati non sarebbe inoltre stato possibile senza il prezioso supporto di Thomas Aeppli e delle/dei nostri assistenti tra studentesse e studenti.

2 Maggiori informazioni: https://vecof-svizzera.ch/.

3 A causa delle misure adottate per far fronte al Covid-19, il rilevamento principale è stato rinviato.

4 Nella Svizzera romanda la prima lingua straniera è il tedesco, nella Svizzera tedesca lungo la frontiera linguistica con la Romandia è il francese. In entrambe le regioni, l'inglese è la seconda lingua straniera. Nel resto della Svizzera tedesca, l'inglese è invece la prima lingua straniera, il francese la seconda. In Ticino, il francese è la prima lingua straniera, il tedesco la seconda. Il Canton Grigioni non ha partecipato allo studio del 2017 e solo in parte a quello previsto nel 2020 (nei Grigioni tedescofoni e italofoni con l'inglese come lingua straniera).

Quali erano gli obiettivi dello studio?

Nel quadro di Task Lab, sono stati elaborati esercizi di comprensione scritta analoghi a quelli che si potrebbero utilizzare in un rilevamento nazionale delle competenze nelle lingue straniere come la VeCoF. Il progetto di ricerca era volto in particolare a esaminare il funzionamento di tre diversi formati (scelta multipla, risposta breve, abbinamento). Gli esercizi sono stati preparati in due versioni linguistiche al fine di valutare se domande e risposte andassero formulate nella lingua straniera, quindi nella lingua dei testi da leggere (francese), o in quella scolastica (tedesco).

Si trattava altresì di esplorare più in dettaglio ciò che costituisce la competenza nella comprensione scritta. A tale scopo, tra gli allievi partecipanti sono state rilevate altre capacità di cui si sa che sono legate alla competenza nella comprensione scritta nella lingua straniera, per esempio il lessico ricettivo, la disponibilità di un lessico di "parole a vista" nella lingua straniera, l'utilizzo della memoria di lavoro o la motivazione per l'apprendimento linguistico (cfr. Alderson et al., 2015; Harsch & Hartig, 2016; Sabatini et al., 2013).

L'obiettivo principale era quello di analizzare più approfonditamente i fattori di successo nella comprensione scritta così da interpretare meglio i risultati e le scale dei test nel quadro di studi più ampi.

Chi ha partecipato allo studio?

Gli esercizi sono stati sottoposti ad allievi che hanno frequentato l'ultimo anno di elementari. Per questioni pratiche, nel quadro del progetto Task Lab ci si è limitati agli esercizi relativi all'apprendimento del francese. Lo studio è stato condotto nei Cantoni di Basilea Campagna, Soletta, Berna, Friburgo e Vallese.

Come di consueto nell'elaborazione di test, tutti gli esercizi sono stati dapprima testati e poi sottoposti a una fase pilota (cfr. p.es. Kenyon & MacGregor, 2012). Durante la fase di prova, gli esercizi e i processi per arrivare alla soluzione sono stati discussi con singoli allievi nel quadro di cosiddette interviste *stimulated recall*. Le conoscenze così acquisite sono servite a ottimizzare gli esercizi. A queste interviste hanno partecipato 34 allievi di due classi, poi alla fase pilota 97 allievi di cinque classi. Queste ragazze e questi ragazzi hanno svolto l'intera batteria di test in condizioni normali in classe. Al termine della fase pilota, si è proceduto agli ultimi adeguamenti.

Allo studio principale durante l'estate 2015 hanno partecipato in totale 623 allievi di 34 classi (309 ragazzi, 314 ragazze, il 20% circa con passato migratorio). Le scuole hanno aderito volontariamente al progetto, per cui non si tratta di un campione rappresentativo. Lo studio è stato condotto durante i normali orari di lezione dalle/dai collaboratrici/collaboratori del progetto e da ausiliari appositamente formati.

Quali tipi di esercizi sono stati utilizzati?

Tutti gli esercizi prevedevano la lettura di uno o più testi in francese, seguiti da tre domande. Per l'elaborazione degli esercizi ci si è avvalsi di importanti conoscenze scientifiche sulla comprensione scritta e sulla creazione di prove scritte (p.es. Alderson, 2000; Alderson et al., 2015; Grabe, 2009; Khalifa & Weir, 2009; Lutjeharms & Schmidt, 2010), di descrizioni rilevanti degli obiettivi di apprendimento (p.es. Bersinger et al., 2005; D-EDK, 2013; CDPE, 2011; Consiglio d'Europa, 2001; Passepartout, 2013) e del manuale didattico *Mille feuilles* (Bertschy et al., 2011ff.), utilizzato nella regione mirata. Tutti gli esercizi si situavano al livello A1 del quadro comune europeo di riferimento per la conoscenza delle lingue (Consiglio d'Europa, 2001), un livello che in Svizzera corrisponde alle competenze di base per la comprensione scritta nella prima lingua straniera alla fine della scuola elementare.

Il test è stato creato con il software CBA ItemBuilder[5] e richiamato sui computer delle scuole mediante un browser internet.[6]

Dodici dei diciotto esercizi erano disponibili sia nel formato a scelta multipla (MCQ) sia nel formato a risposta breve (SAQ). Entrambi i formati esistevano in due lingue: uno con domande e risposte in francese e uno in tedesco (nel formato a risposta breve, gli allievi dovevano formulare la risposta nella relativa lingua). Nel complesso, ognuno di questi dodici esercizi è stato quindi utilizzato in quattro versioni. I singoli allievi erano chiamati a risolverne una. → Fig. 1

5 Maggiori informazioni sul software: https://tba.dipf.de/en/infrastructure/software-development/cba-itembuilder-1?set_language=en.

6 Ringraziamo il Leibniz-Institut für Bildungsforschung und Bildungsinformation (DIPF), e in particolare Helge Einspanier, per il sostegno e la messa a disposizione del server per il test.

Figura 1: Un esercizio Task Lab in quattro versioni

Altri sei esercizi sono stati realizzati secondo il formato dell'abbinamento: tra diversi brevi testi, gli allievi dovevano sceglierne uno contenente una determinata informazione. Anche questo formato era disponibile in due lingue (domande in tedesco e in francese). Un breve questionario chiedeva agli allievi quale formato, rispettivamente quale lingua preferivano.

Oltre agli esercizi di comprensione scritta, sono stati utilizzati test su competenze notoriamente legate alla comprensione scritta e un questionario. Questi esercizi supplementari erano volti a rilevare altri aspetti della competenza linguistica, tra cui la competenza lessico-grammaticale degli allievi (C-Test [7] e segmentazione di vocaboli[8] in tedesco e francese). Altri test vertevano sull'estensione del vocabolario, sul riconoscimento automatico di vocaboli e sulla consapevolezza fonologica nella lingua straniera (lessico ricettivo, lessico visivo,

7 In diversi brevi testi, a una parola su due viene cancellata la metà della parola (con l'eccezione della prima e dell'ultima frase). Gli allievi devono ricostruire queste parole. Questo formato di test consente di effettuare in breve tempo una prima valutazione della competenza linguistica degli allievi, chiamati a mettere in pratica contemporaneamente le loro conoscenze lessicali, grammaticali e ortografiche.

8 In un testo senza spazi vuoti, gli allievi devono segnare i confini tra i vocaboli.

associazione di suoni e vocaboli[9]), come pure sulla memoria di lavoro (cosiddetti *digit-span task*[10]). Nel questionario, gli allievi venivano interpellati sulla loro provenienza e sulla/e loro prima/e lingua/e, ma soprattutto sulle loro abitudini di lettura e la loro motivazione ad apprendere il francese. →Fig. 2

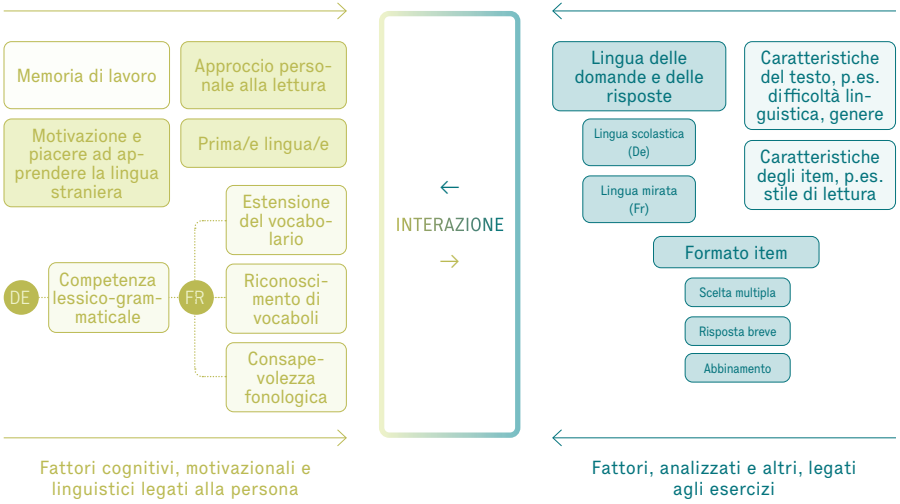


Figura 2: Rappresentazione schematica della batteria di test Task Lab[11]

9 *Lessico ricettivo:* in un elenco, gli allievi segnano i vocaboli che conoscono. Una parte è costituita da "pseudovocaboli", ossia sequenze di lettere che potrebbero parere una parola nella lingua straniera, ma non lo sono. *Lessico visivo*: gli allievi pronunciano un vocabolo ad alta voce che appare su uno schermo per qualche millisecondo. *Associazione di suoni e vocaboli:* gli allievi leggono "pseudovocaboli" ad alta voce.

10 Gli allievi ripetono sequenze di numeri sempre più lunghe, anche in senso inverso.

11 Il grafico riporta fattori che per mancanza di risorse non sono stati analizzati nel quadro del progetto. Dato però che i dati sono registrati nell'archivio della ricerca, è possibile analizzare anche il loro influsso sui risultati del test. Eventuali ricercatori interessati sono caldamente invitati a procedere.

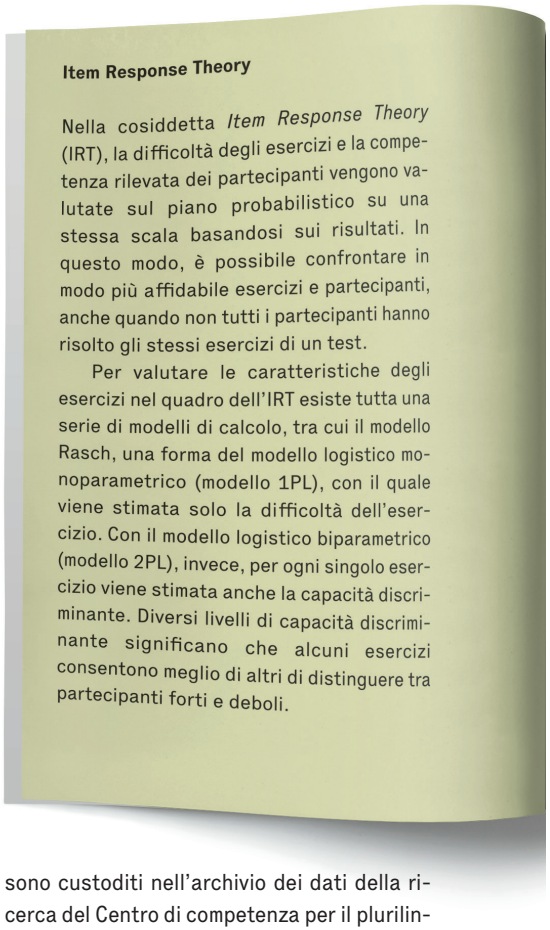
Come sono stati analizzati i dati?

Le interviste durante la fase di test e le risposte al questionario breve sono state trascritte e sottoposte a un'analisi quantitativa dei contenuti (Mayring, 2010) concentrandosi in particolare sul processo adottato dagli allievi per arrivare alla soluzione e la lingua delle domande e delle risposte. I risultati di queste analisi possono essere consultati in Barras (2018, compresa una discussione sul metodo di rilevamento) e in Karges et al. (*forthcoming*, lingua delle domande e delle risposte).

Le risposte dei partecipanti al test sono state rilevate e valutate in parte automaticamente, in parte da membri del team di progetto. Le risposte brevi alla prova di comprensione scritta sono state sottoposte a doppia valutazione e poi uniformate tramite discussione. I risultati degli altri test sono stati in parte valutati da due persone per verificare la qualità delle valutazioni umane.

I risultati così ottenuti sono poi stati oggetto di controlli della qualità (sono p.es. stati individuati modelli di risposta anomali) e classificati su una scala (p.es. avvalendosi della *Item Response Theory* [IRT], cfr. riquadro). Infine, i risultati così classificati di tutti gli allievi e di tutti i test, come pure i dati tratti dal questionario sono stati riuniti in un unico set di dati, il quale ha funto da base per analisi più approfondite, in particolare in Lenz et al. (2019), in cui è stato indagato nei dettagli l'effetto del formato dell'esercizio.

Tutti i dati grezzi, le interviste qualitative annotate e i risultati classificati su una scala sono custoditi nell'archivio dei dati della ricerca del Centro di competenza per il plurilinguismo e possono essere utilizzati per altre analisi. L'archivio contiene anche i test effettuati, utili per esempio per indagini sull'influsso delle caratteristiche del testo e degli esercizi sui risultati.[12]



Item Response Theory

Nella cosiddetta *Item Response Theory* (IRT), la difficoltà degli esercizi e la competenza rilevata dei partecipanti vengono valutate sul piano probabilistico su una stessa scala basandosi sui risultati. In questo modo, è possibile confrontare in modo più affidabile esercizi e partecipanti, anche quando non tutti i partecipanti hanno risolto gli stessi esercizi di un test.

Per valutare le caratteristiche degli esercizi nel quadro dell'IRT esiste tutta una serie di modelli di calcolo, tra cui il modello Rasch, una forma del modello logistico monoparametrico (modello 1PL), con il quale viene stimata solo la difficoltà dell'esercizio. Con il modello logistico biparametrico (modello 2PL), invece, per ogni singolo esercizio viene stimata anche la capacità discriminante. Diversi livelli di capacità discriminante significano che alcuni esercizi consentono meglio di altri di distinguere tra partecipanti forti e deboli.

12 https://portailplurilingue.unifr.ch/starweb/KFM/k.skca-catalog/servlet.starweb?path=KFM/k.skca-catalog/FastLink.web&inum=000011889.

Risultati selezionati

Domande e risposte di una semplice prova di comprensione scritta dovrebbero essere redatte nella lingua straniera?

Piuttosto no. I risultati del nostro studio principale e quanto emerso dalle interviste dimostrano chiaramente che l'utilizzo di una lingua scolastica comune contribuisce a testare in modo più affidabile la competenza degli allievi nella comprensione scritta. Per lo meno nel caso del contesto da noi approfondito (allievi con un livello linguistico basso e test standard a bassa soglia, esistenza di una lingua comune[13]), raccomandiamo di redigere gli esercizi nella lingua scolastica.

Il progetto ha evidenziato che gli allievi con un livello linguistico A1-A2 capiscono spesso in modo insufficiente le consegne redatte nella lingua straniera e si affidano quindi a strategie di compensazione più o meno efficaci oppure, semplicemente, tirano a indovinare. In varie occasioni, abbiamo osservato che gli allievi avrebbero trovato l'informazione cercata nel testo, ma non avevano capito la domanda o le opzioni di risposta. La mancata risoluzione dell'esercizio non era quindi dovuta a carenze a livello di comprensione del testo, bensì ad altri motivi.

Anche un sondaggio tra i bambini che hanno partecipato allo studio porta allo stesso risultato: la maggior parte di loro afferma che sarebbe più facile se le domande fossero redatte in tedesco. Solo il 10% circa degli interpellati dichiara di preferire il francese. Molti di questi allievi hanno spiegato di prediligere le domande nella lingua straniera in quanto ciò consente loro di ritrovare nel testo singoli vocaboli o locuzioni che forniscono loro indizi sulla (presunta) risposta corretta. Si tratta in questo caso di una strategia non legata necessariamente a buone competenze di lettura e sovente inefficace in esercizi strutturati professionalmente.

Anche dal punto di vista del costrutto del test, molto parla a favore dell'impiego della lingua scolastica: chi legge lo fa di regola con un obiettivo, per esempio per trovare una determinata informazione o per divertirsi. Di solito, un esercizio dichiara tale obiettivo, ma se i partecipanti non lo capiscono non possono procedere alla lettura nel modo richiesto. Tutto ciò può portare a risultati che non corrispondono all'obiettivo del test e, nel peggiore dei casi, hanno ben poco a che vedere con la capacità di lettura dei partecipanti.

13 Nel caso di test linguistici internazionali in cui i partecipanti non condividono una stessa lingua, la comprensibilità delle domande deve essere garantita in altro modo (p.es. con formulazioni molto semplici, immagini o la possibilità di svolgere in precedenza un'esercitazione). Ciò riduce tuttavia la varietà delle possibili domande.

Il formato dell'esercizio influisce in una prova di comprensione scritta?

Sì e no. Analisi statistiche dei risultati lasciano presupporre che il formato a scelta multipla o quello a risposta breve misurano per lo più lo stesso costrutto, e che quindi dal punto di vista del contenuto non ce ne sia uno preferibile all'altro.

Si constata tuttavia che il formato a scelta multipla in media è a) significativamente più semplice e b) meno discriminante se paragonato al formato a risposta breve a *parità* di formulazione. Ciò significa che le competenze degli allievi devono essere superiori per svolgere correttamente un esercizio secondo il formato della risposta breve. Quest'ultimo, inoltre, differenzia in modo più preciso gli allievi più deboli da quelli più forti. Entrambi gli effetti sono spiegabili con i diversi processi di risposta: nel formato a risposta breve, a partire dal testo occorre trovare una risposta che poi va formulata, il che costituisce una fonte di errore supplementare e riduce quindi la probabilità di una risposta corretta. L'esercizio è quindi più difficile e, dunque, più selettivo. Nel formato con tre opzioni di risposta, invece, esiste già una certa probabilità di trovare la soluzione giusta senza nemmeno aver letto il testo. Anche la possibilità di indovinare abilmente la risposta invece di trovarla leggendo e capendo il testo riduce la qualità del formato a scelta multipla quale strumento di misurazione delle competenze linguistiche.

In genere, pertanto, una risposta breve corretta dice più sulle competenze linguistiche di un allievo di quanto non faccia una selezione corretta nel formato a scelta multipla. In presenza di entrambi i formati, i partecipanti dovrebbero quindi ricevere più punti per una risposta breve corretta. Nel caso specifico del nostro studio, sarebbe stata adeguata, in media, una ponderazione doppia di questo formato rispetto a quello a scelta multipla. Se i risultati del test devono essere classificati su una scala dal punto di vista psicometrico mediante *Item Response Theory* (IRT), si raccomanda l'utilizzo di un modello logistico biparametrico in grado di stimare direttamente il potere discriminante e da cui ricavare la ponderazione per la valutazione delle capacità dei partecipanti.

→ riquadro IRT

Come risolvono gli allievi semplici esercizi di comprensione scritta in una lingua straniera?

Oltre agli esercizi di comprensione scritta, sono stati effettuati altri test volti a rilevare competenze parziali di cui si sa che sono legate alla comprensione scritta. Nel complesso, i risultati delle varie analisi rispecchiano le conclusioni riportate dalla letteratura specializzata. Si constata in particolare che testando un livello linguistico elementare (livello A del QCER) il vocabolario riveste un ruolo importante, tant'è vero che le conoscenze lessicali ricettive permettono di prevedere con una relativa precisione il risultato del test.

Anche i due brevi test sulla consapevolezza fonologica e sul riconoscimento (automatizzato) di vocaboli sono strettamente legati alla prova di comprensione scritta, tanto quanto il più complesso C-Test e l'esercizio sulla segmentazione di vocaboli (in questi ultimi due esercizi rivestono una certa importanza anche le conoscenze a livello di testo). Ciò dimostra che anche brevi test meno contestualizzati possono fornire indicazioni sulla competenza linguistica di principianti, la cui comprensione scritta a livello di testo è ancora poco stabile e quindi difficilmente misurabile con prove di comprensione scritta.

Quali conseguenze ha avuto Task Lab?

I rilevamenti VeCoF del 2017 e del 2020 → vedi nota 3 nelle lingue scolastiche e straniere hanno beneficiato in misura determinante delle conoscenze e delle esperienze acquisite con Task Lab, soprattutto dal punto di vista tecnico. Il design e la struttura di base degli esercizi sono stati ripresi in forma solo leggermente modificata. In linea di principio, nei test VeCoF di lingua straniera, domande e risposte vengono formulate nella lingua scolastica. Anche materiali ed esperienze della fase di test di Task Lab sono stati utilizzati per lo sviluppo di esercizi nel quadro della VeCoF.

Per il progetto successore di Task Lab (forme di valutazione innovative, IFB[14]), sono state riprese, rispettivamente perfezionate parti del design del rilevamento, e anche la struttura tecnica di base degli esercizi è stata sfruttata per questo nuovo progetto e per altri ancora.

Dal punto di vista della ricerca, va evidenziato che Task Lab è uno dei pochi progetti ad aver approfondito la struttura delle competenze di comprensione scritta in una lingua straniera che non sia l'inglese (nella fattispecie il francese) tra giovani allievi (dodicenni) con conoscenze linguistiche elementari (livello A del QCER). Piuttosto raro è anche l'approccio sistematico nell'analisi degli effetti delle caratteristiche degli esercizi di comprensione scritta, in particolare la lingua delle domande e delle risposte (lingua scolastica tedesca o lingua mirata francese) e i formati delle risposte (a scelta multipla o a risposta breve).

14 Cfr. anche https://centre-plurilinguisme.ch/it/ricerca/forme-di-valutazione-innovative.

Dove trovo maggiori informazioni?

I due articoli seguenti (in inglese) contengono dettagli e discussioni su quanto presentato in questo resoconto:

- Lenz, P., Karges, K. & Barras, M. (2019). Investigating test method effects in French L2 reading items for young learners. In A. Huhta, G. Erickson & N. Figueras (eds), *Developments in language education: a memorial volume in honour of Sauli Takala*. Jyväskylä: University of Jyväskylä & EALTA, 182-202.
- Karges, K., Barras, M. & Lenz, P. (forthcoming). 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*. Berlin: Peter Lang.

Questo articolo presenta invece (in tedesco) la procedura qualitativo-metodica utilizzata per testare gli esercizi e riporta risultati selezionati del progetto Task Lab:

- Barras, M. (2018). „Stimulated Recall" in der Sprachtestforschung. Ein praktisches Beispiel aus der Erprobung eines computerbasierten Leseverstehenstests. In K. Aguado, C. Finkbeiner & B. Tesch (Hrsg.), *Lautes Denken, „Stimulated Recall" und Dokumentarische Methode. Rekonstruktive Verfahren in der Fremdsprachenlehr- und -lernforschung*. Frankfurt: Peter Lang, 69-86.

Infine, proponiamo due brevi articoli riassuntivi (in tedesco):

- Barras, M., Karges, K. & Lenz, P. (2016). Leseverstehen überprüfen: Welche Sprache für die Fragen und Antworten in den Testitems? *Babylonia, 16*(2), 13-18.
- Karges, K., Barras, M. & Lenz, P. (2016). Task Lab: Untersuchungen zum besseren Verständnis von computerbasierten kommunikativen Testaufgaben zum Leseverstehen in Französisch. *Babylonia, 2016*(3), 56.

Sul sito del Centro di competenza per il plurilinguismo sono inoltre disponibili diversi manifesti e documenti PDF di conferenze presentati in occasione di vari convegni specialistici.[15]

15 https://centre-plurilinguisme.ch/it/ricerca/task-lab-studi-migliorare-la-comprensione-e-la-validita-degli-esercizi-comunicativi.

Assessing reading comprehension in a foreign language

Task Lab – an empirical study of computer-based tasks in French

—

Katharina Karges, Malgorzata Barras, Peter Lenz

Origins of the study[1]

In the spring of 2017, the Swiss Conference of Cantonal Ministers of Education (EDK) conducted its first Swiss-wide assessment of how well schoolchildren have mastered the basic competencies in the first foreign language taught at school, specifically at the end of their primary schooling. The second EDK assessment[2] to verify the basic competencies was planned for the spring of 2020,[3] this time at the end of compulsory schooling (secondary school) and in both the first *and* second foreign language.[4] For both assessments, the Research Centre on Multilingualism (RCM) was charged in the preliminary stages with designing test tasks.

Independent of the EDK, the "Task Lab" research project developed computer-based reading comprehension tasks and explored their functioning in detail. Following the findings, a series of decisions was taken for the EDK assessment that were based on empirical experiences with the target group. From a language assessment research perspective, Task Lab contributes to a more precise identification of the role various skills and abilities play when students solve reading comprehension tasks in a foreign language.

1 We would like to thank the many teachers and school directors who supported our project. We also thank all students who, for the sake of this study, allowed us to look into their classrooms and into how they learn. In addition, without the active support of Thomas Aeppli and our student assistants, collecting data would have never progressed beyond the planning stage.

2 More information at https://uegk-schweiz.ch/.

3 Due to Covid-19 measures, the main assessment could not be conducted and has been postponed to a later time.

4 In Western Switzerland, German is the first foreign language taught at schools, whereas in areas of German-speaking Switzerland bordering on the French language region, French is the first foreign language taught. In both regions, English is the second foreign language taught; in the rest of German-speaking Switzerland, English is taught first, then French. In the Canton of Ticino, French is the first foreign language and German the second. The Canton of Graubünden was not part of the 2017 assessment and is only partially represented in the planned 2020 edition, specifically in the German- and Italian-speaking regions, where English is taught.

Goals of the study

The Task Lab team developed reading comprehension tasks that are suitable for use in a Swiss-wide large-scale assessment of foreign language skills. During the project, three different types of task format were studied: multiple-choice, short answer and matching tasks. These tasks were formulated in two language versions to determine whether questions and answers about a text passage should be framed in the foreign language, i.e. the language of the passage (French) or in the language of schooling (German).

The project also aimed to achieve a more precise definition of what reading comprehension skills entail. For this reason, participating students were tested for other skills and abilities that are known to be related to reading comprehension skills in a foreign language (e.g. receptive vocabulary, sight word recognition in the foreign language, working memory capacity and language learning motivation; cf. Alderson et al., 2015; Harsch & Hartig, 2016; Sabatini et al., 2013).

The overarching aim was to shed light on success factors for reading comprehension and, in the end, to attain more valid interpretations of test results and assessment scales in large-scale assessments.

Participants in the study

The tasks were given to schoolchildren who were attending their final year of primary school. Practical considerations dictated that Task Lab tasks were used to examine skills only in one foreign language, French. The study was conducted in five cantons: Basel-Landschaft, Solothurn, Bern, Fribourg and Valais.

As is standard when designing tests, all tasks were first pre-tested and then piloted (cf. e.g. Kenyon & MacGregor, 2012). During the pre-testing stage, the individual tasks and different approaches to finding solutions were discussed in detail in stimulated recall interviews with individual students. The resulting findings were used to further optimise the tasks. 34 students from two classes participated in these interviews. 97 students from 5 classes then took part in the pilot phase; they completed the entire set of tests as a class in real conditions. After the pilot phase, final modifications were made to the tasks.

The main study was carried out in the early summer of 2015 with a total of 623 schoolchildren from 34 classes (309 boys, 314 girls, approximately 20% with an immigration background). The participating schools volunteered to take part in the study; as such, the sample cannot be viewed as representative. The project team and trained assistants conducted the study at the participating schools during regular class time.

Task types used

All reading comprehension tasks were made up of one or more passages in French, and three questions. To design the tasks, the researchers consulted specialist literature for key findings in the area of reading comprehension and language assessment (for instance, Alderson, 2000; Alderson et al., 2015; Grabe, 2009; Khalifa & Weir, 2009; Lutjeharms & Schmidt, 2010) and drew on relevant learning outcome descriptions (Bersinger et al., 2005; D-EDK, 2013; EDK, 2011; Council of Europe, 2001; Passepartout, 2013 and others) as well as on the *Mille feuilles* textbook used in the region selected for the study (Bertschy et al., 2011ff.). All tasks were situated in the general vicinity of level A1 according to the Common European Framework of Reference for Languages (Council of Europe, 2001). In Switzerland, this level corresponds to the basic competencies in reading comprehension that should be mastered in the first foreign language at the end of primary schooling.

The entire test was developed as a computer-based test using CBA ItemBuilder software;[5] the test could be accessed at the schools via an Internet browser.[6]

12 of the overall 18 tasks developed were designed as multiple-choice questions (MCQ) and as short-answer questions (SAQ). In addition, both item formats were available in two language versions: one with questions and answers in French, and one with questions and answers in German. For the SAQ, the students were asked to write their answers in the same language as the question. Overall, four different variations were used for each of the 12 reading comprehension tasks, with the individual students each solving one of the four tasks.

→ Image 1

5 The latest information on this software is available at: https://tba.dipf.de/en/infrastructure/software-development/cba-itembuilder-1?set_language=en.

6 We are grateful to the Leibniz Institute for Research and Information in Education (DIPF), especially Helge Einspanier, for the friendly support and for providing the test server.

Image 1: The four variations of a Task Lab task

Six additional tasks were developed using a matching format in which students were asked to read several short passages and select the passage containing the required information. This format, too, was developed in two language versions, with questions posed in German and French. The students were asked to fill in a short questionnaire asking which task format and which language they preferred.

In addition to the reading comprehension tasks, tasks on competences related to reading comprehension skills were used and a questionnaire was given to the students. These additional tasks were used mainly to identify other aspects of language competence such as general lexical and grammatical competence (C-Tests[7] and word segmentation[8] in German and French) as well as vocabulary tests, automatic sight word recognition and phonological awareness in the foreign language (receptive vocabulary test, sight word recognition, spelling-to-sound mapping[9]). Moreover, tests to determine working memory abilities were used (so-called digit-span tasks[10]). In the questionnaire, the students were asked about their family origins and their first language(s), but the questions focused mainly on their reading habits and their motivation to learn French at school. →Image 2

7 C-Tests comprise several short text passages in which half of every second word is deleted (except in the first and last sentences). Students are asked to reconstruct these words. This test format is useful for quickly making an approximate assessment of language competence, as test takers are required to simultaneously apply their knowledge of vocabulary, grammar and orthography.

8 Test takers are asked to mark where words begin and end in a text passage that has no spaces between words.

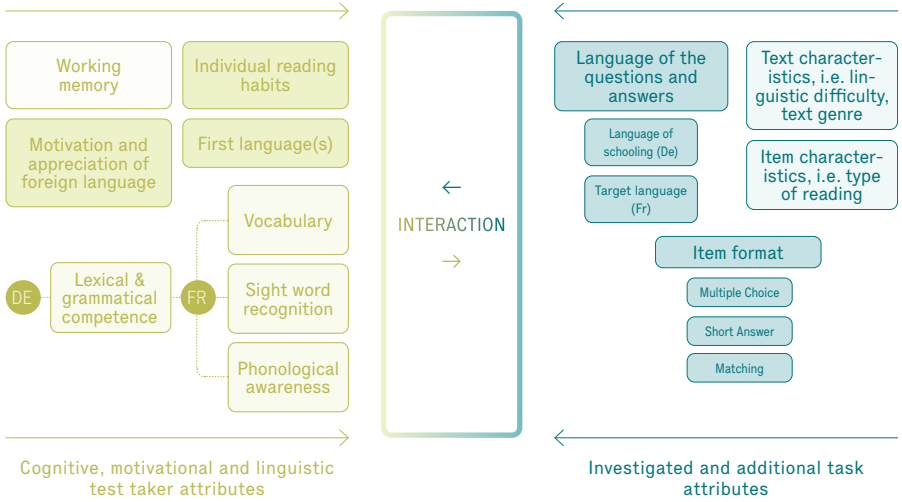


Image 2: Diagram of Task Lab tests[11]

9 *Receptive vocabulary test:* After being given a list of words, students are asked to tick the words they are familiar with. Some of the words are pseudowords, i.e. series of letters that theoretically could be a word in the foreign language, but that actually are nonsense words (this test format is also called a Yes-No test). *Sight word recognition:* Students are required to say a word aloud that appears on the computer screen for only a few milliseconds. *Spelling-to-sound mapping:* Students read pseudowords aloud.

10 Students are asked to repeat numerical sequences of increasing length, partially in reverse order.

11 The diagram includes task factors that were not analysed in the project due to a lack of resources. Nevertheless, because the data are stored in the research database of the Research Centre on Multilingualism, their influence on test results can also be examined. Interested researchers are cordially invited to explore these aspects.

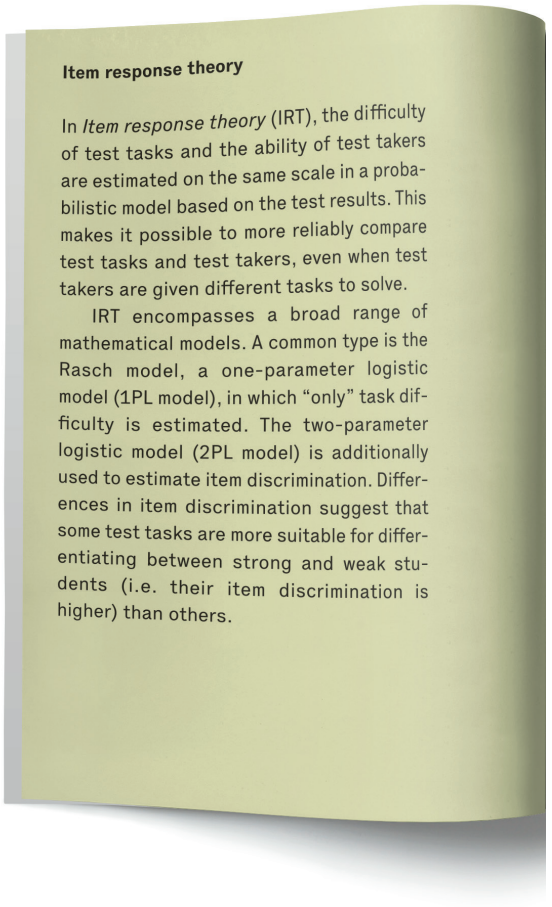
Interpretation of the data

Interviews from the testing phase as well as the answers from the short questionnaires were transcribed and interpreted in the sense of a structuring qualitative content analysis (Mayring, 2010) with reference to the following two issues: how young learners approach solving the tasks, and the language used in the questions and answers. The results of these analyses are presented in Barras (2018, including a discussion of research design and methods) and Karges et al. (forthcoming, on the language used in questions and answers).

The answers of the test takers were recorded and interpreted in part by the project team and in part automatically. The short answers from the reading comprehension tests were assessed first by two raters individually who then agreed to one rating in a discussion. The results from the additional tasks were in part assessed both manually and automatically in order to monitor the quality of the human ratings.

The subsequent test results were controlled for quality (for instance, unusual answer patterns were identified) and scaled (by using instruments such as Item response theory [IRT], cf. the box "Item response theory"). Lastly, the scaled test results of all students across all tests as well as information from the questionnaire were consolidated into a single dataset that served as the basis for further analyses, particularly in Lenz et al. (2019), where the impact of the task format was examined in detail. → Box IRT

All raw data, the annotated qualitative interviews and the scaled test results have been archived in the research database of the Research Centre on Multilingualism and can be used for further analyses. The Task Lab tasks are also stored in the archives and would be suitable for use in, for instance, studies on the impact of text and task characteristics on test results.[12]



Item response theory

In *Item response theory* (IRT), the difficulty of test tasks and the ability of test takers are estimated on the same scale in a probabilistic model based on the test results. This makes it possible to more reliably compare test tasks and test takers, even when test takers are given different tasks to solve.

IRT encompasses a broad range of mathematical models. A common type is the Rasch model, a one-parameter logistic model (1PL model), in which "only" task difficulty is estimated. The two-parameter logistic model (2PL model) is additionally used to estimate item discrimination. Differences in item discrimination suggest that some test tasks are more suitable for differentiating between strong and weak students (i.e. their item discrimination is higher) than others.

Selected findings

Should questions and answers in a simple reading comprehension test be formulated in the foreign language?

Probably not. The results from our main study and the findings from the interviews in the pre-testing phase provide a clear indication that tasks written in the common language of schooling contribute to achieving a more reliable assessment of young learners' reading comprehension skills in the foreign language. We would therefore recommend formulating tasks in the test takers' language of schooling, at least in a context similar to the one we analysed (low level of language skills, standardised tests with a broad range of application, presence of a common language[13]).

The project team found a great deal of evidence indicating that students with language skills ranging between levels A1 and A2 often have a poor understanding of the task questions in the foreign language and thus resort to more or less successful compensation strategies, or they simply guess. Indeed, we observed various situations in which students would have correctly identified the information in the passage if they had sufficiently understood the question or the possible answers. In such cases, wrong answers were unrelated to a lack of reading comprehension skills in the foreign language and instead were caused by factors that have nothing to do with understanding the text passage.

A survey of the children participating in the study addressed this question explicitly, and the responses underscore our finding: most students agree that it is easier for them when the questions are formulated in their language of schooling (German) because they are certain to understand the task. Only some 10 percent of the students said they would prefer to have questions written in French. However, many of these latter students claimed this makes it easier for them to find words or phrases in the text, which helps them to identify what they believe is the right answer – a strategy that is not necessarily indicative of good reading comprehension skills and that often fails in professionally designed tasks.

Considering the test construct, there are also strong arguments for using the language of schooling: when we read, we generally have a goal (for instance, we read for information or for pleasure). A test, or a test task, generally formulates this goal, and if test takers fail to understand what they are supposed to do, they are unable to read a

12 https://portailplurilingue.unifr.ch/starweb/KFM/k.skca-catalog/servlet.starweb?path=KFM/k.skca-catalog/FastLink.web&inum=000011889.

13 In the case of international language tests, where test takers have no common language, other methods must be developed to ensure that the questions are clear and comprehensible, for instance, very simply formulated questions, pictures or mock exams as preparation. This, however, limits the range of possible questions.

passage the way the test intends. This can give rise to test results that relate poorly to the test's objectives and – in the worst case – do not reflect the actual reading comprehension skills of the test taker.

Does it matter which task format is used in a reading comprehension test?

Yes and no. Statistical analyses of the test results suggest that both multiple-choice (MCQ) and short-answer (SAQ) formats largely measure the same construct, and that there are no content-related reasons to prefer one format over another.

Nevertheless, studies have also shown that MCQ are on average a) significantly easier and b) less discriminating than SAQ having the *same* wording. This means that students must have better skills to correctly answer an SAQ compared to the same MCQ. Moreover, SAQ allow a clearer distinction to be made between stronger and weaker students. Both effects can be explained by the differing approaches to answering SAQ and MCQ: in the case of SAQ, the text itself must be consulted to find an answer, which the test taker must then write down. This type of task contains an additional source of error, lowering the probability of answering correctly and thus making the task both more difficult and more discriminating. In the case of MCQ with three possible answers, by contrast, there is already a certain probability that the test taker will find the correct solution – even without having read the text. In addition, the possibility that correct answers are the result of clever guessing strategies rather than evidence of good reading comprehension skills lowers the quality of MCQ items as instruments to measure language skills.

In terms of test results, this means that a correct short answer on average generates more information about the level of language competence in the test taker than a correct multiple-choice answer. This should be taken into account when both formats are used. Test takers should actually be given more points for correct answers to SAQ tasks (in our study, double weighting of SAQ answers compared to MCQ tasks would have been appropriate on average). If the test results are to be scaled psychometrically using Item response theory (IRT), a 2PL model is advisable, as it can directly estimate discrimination and use these findings to derive the weighting for estimating the test takers' abilities. → Box IRT

How do young learners solve simple reading comprehension tasks in a foreign language?

In addition to reading comprehension tasks, other tests were given to identify partial competences that are known to be connected to reading comprehension. Overall, the results of the various analyses reflect the connections postulated in the research literature. In particular, it is clear that a learner's vocabulary plays an instrumental role in the elementary language level tested (level A of CEFR): receptive vocabulary skills are a relatively good predictor of the results of a reading comprehension test.

In addition, the two short tests on phonological awareness and on (automatic) sight word recognition have a clear correlation to reading comprehension tests – to an extent that is approximately as strong as the more complex C-Test and word segmentation tasks (in which knowledge at the text level factors in). This suggests that short, less contextualised tests can reveal quite a bit about the language competence of beginning language learners whose reading comprehension skills at the text level are not yet consolidated and that are thus difficult to measure using reading comprehension tests.

What impact has Task Lab had?

The EDK language assessments in 2017 and 2020 → footnote 3 benefited to a great extent from Task Lab findings and experiences, particularly with regard to technical aspects. The design and basic structure of the Task Lab tasks were adopted for the EDK assessments with only minor modifications made. As a rule, questions and answers in the EDK tests on reading and listening comprehension in the foreign languages were formulated in the languages of schooling. In addition, material and experiences from the Task Lab pre-testing were used during task development for the EDK assessments.

The direct follow-up project to Task Lab – Innovative forms of assessment[14] – was able to adopt and further develop parts of Task Lab's assessment design. Moreover, both this new project as well as other projects profited from the technical architecture of Task Lab tasks.

With regards to language assessment research, it should be emphasised that Task Lab is one of very few research projects dedicated to examining general patterns of reading comprehension competencies in a foreign language other than English (namely French) among young language learners (12-year old schoolchildren) with elementary language skills. Also relatively rare in this area of research is the systematic approach to examining the impact of reading comprehension tasks characteristics, in particular the language used for questions and answers (language of schooling German or target language French) and the type of answer formats (MCQ or SAQ).

14 https://centre-plurilinguisme.ch/en/research/innovative-forms-assessment.

Where can I find more information?

The following articles provide detailed descriptions and discussions of the findings (presented above) from the research project:

- Lenz, P., Karges, K. & Barras, M. (2019). Investigating test method effects in French L2 reading items for young learners. In A. Huhta, G. Erickson & N. Figueras (eds), *Developments in language education: a memorial volume in honour of Sauli Takala*. Jyväskylä: University of Jyväskylä & EALTA, 182-202.
- Karges, K., Barras, M. & Lenz, P. (forthcoming). 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*. Berlin: Peter Lang.

The following article (in German) describes the qualitative methodology applied in testing tasks. In addition, it presents selected findings from the Task Lab project:

- Barras, M. (2018). „Stimulated Recall" in der Sprachtestforschung. Ein praktisches Beispiel aus der Erprobung eines computerbasierten Leseverstehenstests. In K. Aguado, C. Finkbeiner & B. Tesch (Hrsg.), *Lautes Denken, „Stimulated Recall" und Dokumentarische Methode. Rekonstruktive Verfahren in der Fremdsprachenlehr- und -lernforschung*. Frankfurt: Peter Lang, 69-86.

Lastly, the following short summaries (in German) have been published:

- Barras, M., Karges, K. & Lenz, P. (2016). Leseverstehen überprüfen: Welche Sprache für die Fragen und Antworten in den Testitems? *Babylonia, 16*(2), 13-18.
- Karges, K., Barras, M. & Lenz, P. (2016). Task Lab: Untersuchungen zum besseren Verständnis von computerbasierten kommunikativen Testaufgaben zum Leseverstehen in Französisch. *Babylonia, 2016*(3), 56.

Various posters and presentations (in PDF format) from diverse conferences can be accessed via the website of the Research Centre on Multilingualism.[15]

15 https://centre-plurilinguisme.ch/en/research/task-lab-studies-better-understanding-and-higher-validity-communicative-test-tasks.

Bibliographie | Bibliografia | Bibliography

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

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.

Barras, M. (2018). „Stimulated Recall" in der Sprachtestforschung. Ein praktisches Beispiel aus der Erprobung eines computerbasierten Leseverstehenstests. In K. Aguado, C. Finkbeiner & B. Tesch (Hrsg.), *Lautes Denken, „Stimulated Recall" und Dokumentarische Methode. Rekonstruktive Verfahren in der Fremdsprachenlehr- und -lehrnforschung*. Frankfurt: Peter Lang, 69-86.

Barras, M., Karges, K. & Lenz, P. (2016). Leseverstehen überprüfen: Welche Sprache für die Fragen und Antworten in den Testitems? *Babylonia, 16*(2), 13-18.

Bersinger, S., Jordi, U. & Tchang, M. (2005). *Europäisches Sprachenportfolio. Version für Kinder und Jugendliche von 11 bis 15 Jahren (ESP II)*. Bern: Schulverlag.

Bertschy, I., Grossenbacher, B., Sauer, E., Thommen, A., Cavelti, S., Keller, M. et al. (2011ff.). *Mille feuilles 3-6*. Bern: Schulverlag Plus.

Conseil de l'Europe (2002). *Quadro comune europeo di riferimento per le lingue: apprendimento insegnamento valutazione*. Milano: La Nuova Italia – Oxford.

Conseil de l'Europe (2001). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Strasbourg : Conseil de l'Europe.

Council of Europe (2001). *Common European framework of reference for languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

D-EDK (Hrsg.) (2013). Lehrplan 21 – Konsultationsfassung.

EDK (Hrsg.) (2011). Grundkompetenzen für die Fremdsprachen. Nationale Bildungsstandards. EDK. http://edudoc.ch/record/96780/files/grundkomp_fremdsprachen_d.pdf.

Europarat (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*. Berlin: Langenscheidt.

Grabe, W. (2009). *Reading in a second language: moving from theory to practice*. New York: Cambridge University Press.

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.

Karges, K., Barras, M. & Lenz, P. (forthcoming). 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*. Berlin: Peter Lang.

Karges, K., Barras, M. & Lenz, P. (2016). Task Lab: Untersuchungen zum besseren Verständnis von computerbasierten kommunikativen Testaufgaben zum Leseverstehen in Französisch. *Babylonia, 2016*(3), 56.

Kenyon, D. M. & MacGregor, D. (2012). Pre-operational testing. In G. Fulcher & F. Davidson (eds), *The Routledge handbook of language testing*. London, New York: Routledge, 295-306.

Khalifa, H. & Weir, C. J. (2009). *Examining reading: research and practice in assessing second language reading*. Cambridge, New York: Cambridge University Press.

Lenz, P., Karges, K. & Barras, M. (2019). Investigating test method effects in French L2 reading items for young learners. In A. Huhta, G. Erickson & N. Figueras (eds), *Developments in language education: a memorial volume in honour of Sauli Takala*. Jyväskylä: University of Jyväskylä & EALTA, 182-202.

Lutjeharms, M. & Schmidt, C. (Hrsg.) (2010). *Lesekompetenz in Erst-, Zweit- und Fremdsprache*. Tübingen: Gunter Narr.

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

Passepartout (2013). Lehrplan Französisch und Englisch. Projektversion.

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).

Verhelst, N. D. (2011). Profile Analysis: a closer look at the PISA 2000 reading data. *Scandinavian Journal of Educational Research*, *56*(3), 315-332.