DE

Fremdsprachenkompetenzen nahe an der Realität testen

Szenariobasierte Testaufgaben für den Computer – eine Vertiefungsstudie

—

FR

Évaluer les compétences en langues étrangères au plus proche de la réalité

Tâches de test informatisées basées sur des scénarios – une étude approfondie

—

IT

Testare le competenze nelle lingue straniere vicino alla realtà

Esercizi computerizzati basati su scenari – uno studio di approfondimento

—

EN

Assessing foreign language skills in near-authentic settings

Scenario-based test tasks on the computer – an in-depth study

Katharina Karges, Peter Lenz, Thomas Aeppli, Malgorzata Barras

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** Societé 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, Peter Lenz, Thomas Aeppli, Malgorzata Barras

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
pro-verbial, Joël Rey - Traduzioni e redazioni, tran-scribe (Mary Carozza)

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

2021.003-IFB

Fremdsprachenkompetenzen nahe an der Realität testen

Szenariobasierte Testaufgaben für den Computer – eine Vertiefungsstudie

—

Évaluer les compétences en langues étrangères au plus proche de la réalité

Tâches de test informatisées basées sur des scénarios – une étude approfondie

—

Testare le competenze nelle lingue straniere vicino alla realtà

Esercizi computerizzati basati su scenari – uno studio di approfondimento

—

Assessing foreign language skills in near-authentic settings

Scenario-based test tasks on the computer – an in-depth study

Katharina Karges, Peter Lenz, Thomas Aeppli, Malgorzata Barras

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

Fremdsprachenkompetenzen nahe an der Realität testen

Szenariobasierte Testaufgaben für den Computer – eine Vertiefungsstudie

Katharina Karges, Peter Lenz, Thomas Aeppli, Malgorzata Barras

Überblick

Mit der Erfindung des Computers und der Entwicklung des Internets, spätestens aber mit der Verbreitung von Tablets und Smartphones haben Lese- und Hörverstehen neue Dimensionen gewonnen. Aus dem linearen Text, egal ob im Buch oder auf der Kassette, ist ein multimedialer Hypertext geworden, mit Verlinkungen und eingebetteten Medien – Bildern, Ton, Videos und interaktiven Grafiken. Auch die Grenze zwischen dem geschriebenen und dem gesprochenen Wort wird durch Sprachnachrichten in Messengerdiensten, Tweets und Podcasts mehr und mehr aufgelöst. Diese Veränderungen sind nicht nur im Alltag der Schülerinnen und Schüler längst angekommen, sie wirken sich auch auf den schulischen Fremdsprachenunterricht und damit auf die Beurteilung von sprachlicher Kommunikationsfähigkeit aus. Mit der Digitalisierung der Kommunikation sind auch die Möglichkeiten für die Beurteilung der Lese- und Hörverstehenskompetenzen von Lernenden breiter geworden: Neben den neuen digitalen Lese- und Hörtextsorten ist in computerbasierten Tests auch der Einsatz von neuen Aufgabenformaten vorstellbar. Solche neu entwickelten Testaufgaben müssen jedoch erprobt und erforscht werden, bevor sie für eine breitere Verwendung empfohlen werden können.

Diese Überlegungen waren ein Ausgangspunkt für das Forschungsprojekt „Innovative Formen der Beurteilung“ (IFB), welches von 2016 bis 2019 am Kompetenzzentrum für Mehrsprachigkeit (KFM) durchgeführt wurde. Im Projekt wurden Testaufgaben entwickelt, die in sogenannte Szenarien eingebettet waren: inhaltliche Rahmenhandlungen, die durch ihre Nähe zu realweltlichen Aufgabenstellungen sowie durch die Verwendung von digitalen Textsorten eine möglichst authentische Sprachverwendung in der Fremdsprache simulieren und die Testteilnehmenden dadurch motivieren sollen, ihre sprachlichen Fähigkeiten vollumfänglich einzusetzen. So mussten die Schülerinnen und Schüler zum Beispiel im Rahmen eines solchen Szenarios Aufgaben zu Lese- und Hörtexten bearbeiten, die sich auf die Planung eines Ausflugs bezogen.

Mit Blick auf die Aufgabenentwicklung für die zweite nationale Leistungsmessung zur „Überprüfung des Erreichens der Grundkompetenzen“ (ÜGK),[1] an der das IFM später beteiligt war, wurden im Projekt IFB Hör- und Leseverstehensaufgaben in den Fremdsprachen entwickelt, die mittels geschlossener,[2] computerbasierter Aufgaben Kompetenzen im Niveaubereich A2/B1 des Gemeinsamen europäischen Referenzrahmens für Sprachen (GeR, Europarat, 2001) erfassen sollten. Mit der Bündelung der Aufgaben in Szenarien ging das Projekt IFB allerdings über den Ansatz der ÜGK hinaus.

Die so entstandenen Aufgaben wurden in einer Validierungsstudie mithilfe eines Mixed-Methods-Designs untersucht. In der *qualitativen* Studie wurden introspektive Verfahren (Lautdenkprotokolle und *Stimulated-Recall*-Interviews) angewandt, um herauszufinden, wie einzelne Lernende die Aufgaben wahrnahmen und bearbeiteten.[3] Im Rahmen der *quantitativen* Studie kamen neben den szenariobasierten Aufgaben auch Tests zu Teilkompetenzen (u.a. zu Wortschatz, Geschwindigkeit der Sprachverarbeitung und Grammatik) sowie Fragebögen zum Einsatz.

Alle Aufgaben wurden in Deutschschweizer Schulen in Fassungen für die beiden Fremdsprachen Französisch und Englisch eingesetzt, wobei der Fokus der qualitativen Studie auf dem Französischen lag. Die Datenerhebung erstreckte sich in mehreren Phasen über das Jahr 2017 und die ersten Monate des Jahres 2018.

Ein übergeordnetes, praktisches Ziel des Projekts bestand darin, zu erkunden, inwiefern es möglich und sinnvoll ist, szenariobasierte Aufgaben für die Überprüfung fremdsprachlicher Verstehenskompetenzen auf einem niedrigen Sprachniveau einzusetzen. In der Validierungsstudie sollten daher u.a. folgende Fragen beantwortet werden:

1. Welche Teilkompetenzen nutzen die Lernenden beim Lösen szenariobasierter Aufgaben in den Schulfremdsprachen?

2. Wird die angenommene höhere Authentizität der Aufgaben von den Lernenden wahrgenommen bzw. geschätzt?

3. Genügen die Aufgaben den psychometrischen Ansprüchen an Aufgaben in Leistungsmessungen (*large scale assessments*)?

1 Die ÜGK ist Teil des schweizweiten Bildungsmonitorings. Im Jahr 2016 wurden erstmals Kompetenzen in Mathematik und 2017 in den Schulsprachen und der ersten schulischen Fremdsprache gemessen. Für 2020 war geplant, am Ende der obligatorischen Schulzeit die Lese- und Hörverstehenskompetenzen der Schülerinnen und Schüler in ihrer Schulsprache und in ihren beiden Schulfremdsprachen zu erheben. Das IFM war an der Entwicklung der Aufgaben beteiligt, wegen der COVID-19-Massnahmen konnte die Haupterhebung allerdings nicht stattfinden und wurde auf 2023 verschoben. Näheres unter https://uegk-schweiz.ch/.

2 Bei geschlossenen Aufgabenformaten wählen die Testteilnehmenden ihre Antworten aus mindestens zwei vorgegebenen Antwortmöglichkeiten aus.

3 Grosse Teile der qualitativen Teilstudie wurden durch Malgorzata Barras für ihr assoziiertes Dissertationsprojekt geleistet (Barras, i.V.).

Was ist szenariobasierte Beurteilung?

Szenariobasierte Beurteilung (auf Englisch *scenario-based assessment*) ist ein Teil des „Reading for Understanding"-Frameworks (Sabatini et al., 2014a; Sabatini & O'Reilly, 2013), welches auch in internationalen Studien zur Anwendung kommt, z.B. im Rahmen von PISA 2018 (OECD, 2019, S. 41). „Reading for Understanding" ist das Ergebnis einer Reihe von Forschungsprojekten des US-amerikanischen Educational Testing Service (ETS), in denen Erkenntnisse über die Natur des Lesens aus verschiedenen Disziplinen zusammengetragen wurden. Es ging dabei darum, ein Modell zu entwerfen, das den unterschiedlichen Facetten des Lesens gerecht wird. Textverstehen kann sich demnach in fünf Bereichen abspielen (Sabatini et al., 2013, S. 14ff.). → Tabelle 1

Aus diesem Verständnis von Textverstehen ergibt sich eine breit gefasste Definition von Lesekompetenz, die auch die Ansprüche an ihre Beurteilung verändert: Neben Aufgaben, die das Verstehen von Texten erfassen, müsste ein Beurteilungsinstrument z.B. auch Aufgaben enthalten, die Lerngelegenheiten schaffen und Ergebnisse des Gelernten erfassen, die Integration von neuem in bereits vorhandenes Wissen überprüfen, oder die Glaubwürdigkeit von Quellen einschätzen lassen (konkrete Beispiele für derartige Aufgaben finden sich bei O'Reilly et al., 2014; Sabatini et al., 2014b). Im Projekt IFB wurde versucht, dieses erweiterte Verständnis von Sprachverstehen auch für die Beurteilung von fremdsprachlichen Kompetenzen umzusetzen. Aufgrund der zunehmenden Vermischung von mündlichen und schriftlichen Quellen in den digitalen Medien wurden im Projekt IFB Aufgaben zum Lese- und Hörverstehen entwickelt.

Bereiche des Textverstehens	Beispiele
Visuelle Elemente und Struktur von gedruckten Texten	Buchstaben, Wörter, Satzzeichen, Funktion von z.B. Absätzen oder Überschriften, aber auch Grafiken, Hyperlinks, Emoticons usw.
Verbale Elemente einer Sprache	Wortbedeutungen, Morphologie, Syntax usw.
Diskursstrukturen und Textgenres	Textorganisation, inter- und intratextuelle Verweise, Einbezug von Weltwissen
Konzeptuelle Inhalte und Bedeutungen	Integration von vorhandenem und neuem Wissen, kritische Bewertung von Inhalten
Soziale Inhalte und Bedeutungen	Auseinandersetzung mit den Ideen des Autors/der Autorin bzw. den Protagonist*innen eines Textes, Verständnis für die Entstehungssituation eines Textes, Einordnung der sozialen Rolle(n) von Texten und Akteuren

Tabelle 1: Die fünf Bereiche des Textverstehens nach Sabatini et al. (2013)

Welche Aufgaben wurden eingesetzt?

Das zentrale Element der IFB-Studie sind sechs Szenarien, die aus vier inhaltlich zusammenhängenden, aber unabhängig voneinander lösbaren Aufgaben (A-D) im Niveaubereich A2/B1 bestehen, welche alle selbstständig am Computer gelöst werden können. Jedes Szenario liegt in einer englischsprachigen und einer französischsprachigen Version vor, die inhaltlich identisch sind. Alle Szenariobeschreibungen, Anweisungen und Aufgabenstellungen sind in der Schulsprache verfasst, um Schülerinnen und Schüler mit schwachen Verstehenskompetenzen in der Fremdsprache nicht zu benachteiligen (für eine ausführlichere Begründung siehe auch Barras et al., 2016).

Die Entwicklung der Szenarien erfolgte iterativ durch das Projektteam. Nachdem die Themen und Aufgabenformate festgelegt waren, wurden die Szenarien nach und nach von mehreren Personen entwickelt, mit einzelnen Lernenden erprobt und meist mehrmals überarbeitet. Sämtliche Lesetexte wurden vom Projektteam verfasst, wobei realweltliche Texte als Modelle dienten. Die Hörtexte wurden ebenfalls speziell für die Szenarien konzipiert, allerdings wurden den Sprecherinnen und Sprechern im Studio nur Eckpunkte und der allgemeine Gesprächsverlauf vorgegeben. Die eigentliche Wortwahl und Ausgestaltung der Texte wurden ihnen überlassen, um eine höhere Authentizität der gesprochenen Sprache zu erreichen. Alle Szenarien wurden parallel in englischer und französischer Sprache entwickelt, sodass keine der beiden Versionen eine eigentliche Übersetzung der anderen ist.

Im Folgenden wird als Beispiel das Szenario „Walkies" vorgestellt, in dem sich die Lernenden mit einer fiktiven Volksabstimmung zur Einführung von Robotern befassen sollten, die mit Hunden Gassi gehen können. Das gewählte Szenario vermittelt einen guten Eindruck von der Bandbreite der Textsorten und Aufgabenformate. Es umfasst eine Hör- (B) und zwei Leseverstehensaufgaben (A und C) sowie eine Aufgabe, in der beide Fertigkeiten eingesetzt werden mussten (D). Insgesamt sind die Texte in diesem Szenario vergleichsweise lang und die Aufgabenschwierigkeit der Aufgaben C und D liegt eher im oberen Mittelfeld.

Zu Beginn des Szenarios werden die Lernenden über den Kontext informiert, in den die Aufgaben eingebettet sind: „Du machst einen Sprachaufenthalt in der kanadischen Stadt Vancouver[4] und gehst dort zur Schule. Im Fach „Social Studies" besprecht ihr gerade eine aktuelle Volksabstimmung. Das Thema: Sollen Roboter mit Hunden spazieren gehen dürfen?" Danach lösen die Lernenden die erste Aufgabe. Sie lesen auf einer nachgebauten Internetseite einen Informationstext und wählen aus mehreren Vorschlägen die wichtigsten Informationen aus, die im Text gegeben werden. → Abbildungen 1 und 2

4 In der französischen Version wurde statt Vancouver Québec gewählt und in allen Aufgaben entsprechend ersetzt.

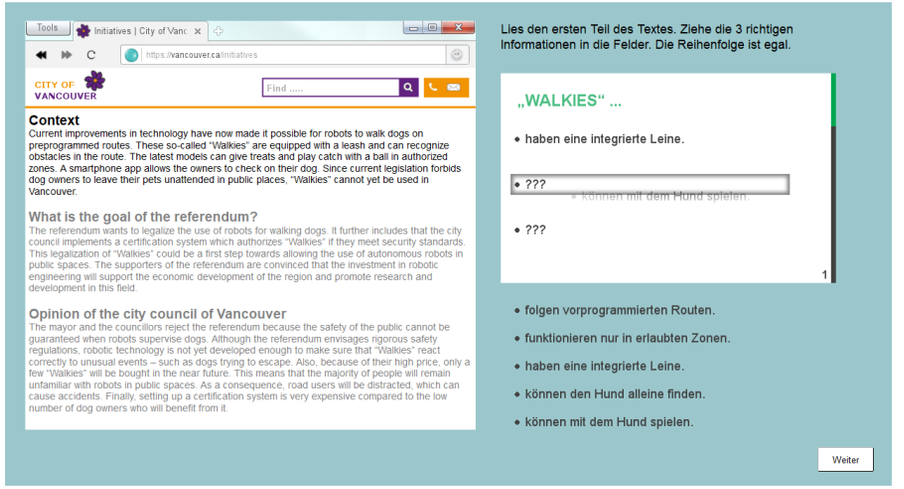

Abbildung 1: Die erste Seite der Aufgabe A im Szenario „Walkies" (englische Version)[5]

Abbildung 2: Die erste Seite der Aufgabe A im Szenario „Walkies" (französische Version)[5]

5 Im Text ist der Teil dunkler gedruckt, der für diese Bildschirmseite relevant ist. Auf den zwei folgenden Seiten ist dann jeweils ein anderer Abschnitt hervorgehoben. Rechts ist die Aufgabe zu sehen: Aus den fünf Optionen unten können die Lernenden drei auswählen und per Drag-and-Drop in die darüberstehende Liste ziehen.

In der zweiten Aufgabe hören die Lernenden dann persönliche Meinungen von fiktiven Mitschülerinnen und Mitschülern zur Einführung dieser Roboter und beantworten dazu „klassische" Multiple-Choice-Fragen. Auch in der dritten Aufgabe werden Multiple-Choice-Aufgaben gestellt, allerdings geben die Lernenden zusätzlich an, wo sie ihre Antwort gefunden haben. Textgrundlage ist hier ein Gruppenchat in einer Messenger-App, in der die fiktive Gruppenarbeit geplant wird. Die Lernenden lesen die Nachrichten in diesem Chat und beantworten die Frage. Die Nachricht, in der sie die Antwort gefunden haben, ziehen sie per Drag-and-Drop in das dafür vorgesehene Feld. → Abbildung 3

Im letzten Teil des Szenarios (Aufgabe D) vervollständigen die Lernenden schliesslich einen Zeitstrahl über das Leben der fiktiven Erfinderin der „Walkies". Grundlage dafür sind Teile eines Wikitextes sowie Auszüge aus einem Interview, in dem die Erfinderin mündlich über sich selbst Auskunft gibt. → Abbildung 4

Die anderen Szenarien drehen sich um die folgenden Themen: ein Kinobesuch, ein Besuch in einem Berufsinformationszentrum, die Planung eines Schulfests, ein Wochenendausflug in eine Stadt sowie Recherchen für einen Vortrag im Geografieunterricht. Die beschriebenen Aufgabenformate und Textsorten treten dabei in verschiedenen Varianten auf, z.B. soll aus Ergebnissen einer Suchmaschine ein passendes Resultat gewählt oder aufgrund von Audionachrichten ein Zeitplan vervollständigt werden. In jedem Szenario gibt es mindestens eine Hör- und eine Leseverstehensaufgabe sowie eine Aufgabe, in der die Lernenden beide Fertigkeiten einsetzen müssen.

Neben diesen Szenarien wurden im Rahmen der Hauptstudie weitere Tests und Fragebögen eingesetzt, mit denen untersucht wurde, welche sprachlichen Teilkompetenzen und weiteren individuellen Voraussetzungen aufseiten der Schülerinnen und Schüler beim Lösen der Aufgaben eine Rolle spielen. → Tabelle 2

Die so entstandenen Daten ermöglichen es, genauer zu beschreiben, wie die Lernenden die szenariobasierten Aufgaben bearbeiteten sowie – ganz allgemein – empirisch basierte Aussagen über die Voraussetzungen und Prozesse beim Verstehen von Texten in einer Fremdsprache zu machen.

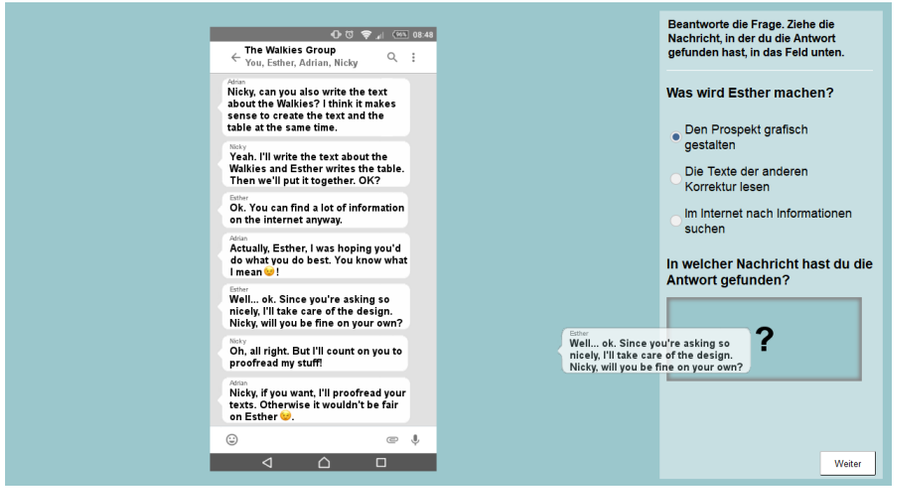

Abbildung 3: Die zweite Seite der Aufgabe C im Szenario „Walkies“ (englische Version)

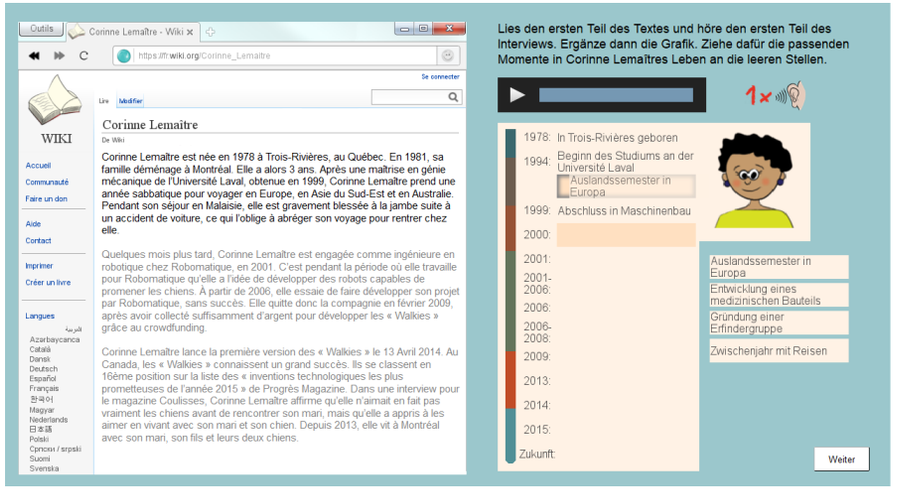

Abbildung 4: Die erste Seite der Aufgabe D im Szenario „Walkies“ (französische Version)[6]

6 Die Lernenden lesen einen Wikitext und hören ein Interview mit ähnlichem Inhalt. Dann ziehen sie zwei der vier ganz rechts gegebenen Optionen in den „Lebenslauf“.

Lese- und Hörverstehen	6 Szenarien mit je 4 Aufgaben	40 Items Leseverstehen 32 Items Hörverstehen 12 Items kombiniertes Hör- und Leseverstehen
	6 Aufgaben zum Textverstehen mit bekannten Formaten	36 Items zum Leseverstehen (Multiple Choice, Richtig/Falsch) 56 Items zum Hörverstehen (Multiple Choice, Vervollständigen)
Sprachliche Teilkompetenzen	2 Wortschatztests	28 Items (Multiple Choice mit geschriebenen Wörtern) 22 Items (Multiple Choice mit gesprochenen Wörtern)
	2 Tests zur Segmentierung von Wörtern in einem Text	2 geschriebene Texte ohne Leerzeichen (max. 326 Wörter) 28 Items (Erkennen der Wortanzahl in gesprochenen Äusserungen)
	Grammatische Kompetenz	24 Items (Einschätzung der grammatischen Korrektheit einer Äusserung)
	Sichtwortschatz	30 Items (Erkennen eines Worts, das nur kurz eingeblendet wird)
Individuelle Voraussetzungen	Fluide Intelligenz	20 Raven-Matrizen (ein nonverbaler Intelligenztest, bei dem eine Folge von Mustern vervollständigt werden muss)
	Individuelle Einstellung zum Erlernen der Fremdsprache	11 Fragebogen-Items zur Motivation 7 Fragebogen-Items zur Sprachlernangst
	Strategien beim Lösen von Hör- bzw. Leseverstehensaufgaben	15 Fragebogen-Items zum Leseverstehen (aktives Problemlösen, Fokus auf Wörter und Details) 10 Fragebogen-Items zum Hörverstehen (aktiv-konzentriertes Zuhören, Planen und Evaluieren)
	Sprachliche Kompetenzen	10 Fragebogen-Items (individuelle Sprachgeschichte) 7 Fragebogen-Items (Einsatz der Fremdsprache im Unterricht)
	Weitere individuelle Voraussetzungen	Fragebogen-Items zu: Herkunft, sozioökonomischer Hintergrund, technische Ausstattung, Nutzung digitaler Geräte und Kanäle

Tabelle 2: Übersicht über die Tests und Fragebögen, die im Projekt IFB eingesetzt wurden

Wer hat an der Studie teilgenommen?[7]

Vor der eigentlichen Hauptstudie wurden die in der Entwicklung befindlichen Szenarien zunächst mit einzelnen Lernenden (n=53) ausprobiert, im Detail besprochen und laufend überarbeitet. Danach wurden auch die weiteren Tests und Fragebögen in Einzelgesprächen und Gruppeninterviews mit Lernenden thematisiert (n≈50). Ausserdem wurden alle Aufgaben im Klassenverband erprobt (n=85), um die technische Machbarkeit der Haupterhebung zu überprüfen. Auch die qualitative Haupterhebung wurde in dieser Phase mit sechs Lernenden erprobt und verfeinert.

An der Haupterhebung nahmen insgesamt 631 Schülerinnen und Schüler aus 39 Klassen teil, davon 30 an der qualitativen Studie. In den meisten Klassen löste etwa die Hälfte der Lernenden die Aufgaben in englischer Sprache (insg. 292 Lernende), die anderen (darunter alle Lernenden aus der qualitativen Studie) bearbeiteten die französischsprachige Version des Tests. Die Schülerinnen und Schüler aus den Kantonen Bern, Zürich, Freiburg, Luzern und Obwalden besuchten zweite und dritte Sekundarschulklassen (Durchschnittsalter: 15 Jahre) aller Leistungsniveaus. Ungefähr zwei Drittel der beteiligten Schülerinnen und Schüler nahmen in Bern bzw. Freiburg an der Studie teil, also in Kantonen, in denen zuerst Französisch und dann Englisch unterrichtet wird. Insgesamt dauerte die Teilnahme an der Haupterhebung über zwei Tage verteilt 6 Lektionen. Dabei lösten die Schülerinnen und Schüler alle Aufgaben weitgehend selbstständig an mitgebrachten Laptops unter Aufsicht einer oder eines Projektmitarbeitenden. → Abbildung 5

Lernende, die an der qualitativen Studie teilnahmen, bearbeiteten während 3 Lektionen ein Szenario bzw. fünf einzelne Szenarioaufgaben in Einzelarbeit mit der Forscherin und verbalisierten dabei ihre Gedanken (Methode des Lauten Denkens, vgl. z. B. Bowles, 2010; Knorr & Schramm, 2012). Anschliessend besprachen sie die bearbeiteten Aufgaben zusätzlich in einem *Stimulated-Recall*-Interview, in dem sie alle Aufgaben und ihre Lösungen mit der Forscherin besprachen (Näheres zur Methode z.B. in Barras, 2018; Gass & Mackey, 2017). Am anderen Tag bearbeiteten diese Lernenden die restlichen Tests (mit Ausnahme der anderen Szenarien) und Fragebögen im Klassenverband.

7 Das Team möchte sich an dieser Stelle bei den Schülerinnen und Schülern und den Vertreterinnen und Vertretern der Schulen für die Unterstützung bedanken. Ohne ihre Zeit und ihr Engagement wäre diese Untersuchung so nicht möglich gewesen. Auch unseren studentischen Hilfskräften, die mitten im Winter früh morgens aufgestanden sind, Koffer durch die halbe Schweiz transportiert und geduldig den Lernenden beim Arbeiten zugesehen haben, gebührt ein herzliches Dankeschön.

Abbildung 5: Ein Klassenraum ist für die Haupterhebung vorbereitet (Foto: Malgorzata Barras)

Wie wurden die Daten ausgewertet?

Während der Erprobungen wurden die Interviews zusammengefasst und wichtige Informationen für die Überarbeitung der Aufgaben bzw. die Durchführung der Datenerhebungen notiert. Ein Teil der Interviews aus der Erprobung wurde zusätzlich transkribiert und steht für Analysen zur Verfügung. Die Testergebnisse aus der Pilotierung wurden aus den vom Testsystem erzeugten Rohdaten extrahiert und beschreibend ausgewertet.

Auch die Testergebnisse aus der Haupterhebung wurden zunächst aufbereitet und skaliert, in den meisten Fällen mittels Item Response Theory (IRT). In einem weiteren Schritt wurden dann alle quantitativen Daten zusammen imputiert. → Kasten

Alle 30 Lautdenkprotokolle und *Stimulated-Recall*-Interviews aus der Haupterhebung der qualitativen Studie wurden vollständig transkribiert. Für die vertiefende Analyse der Strategien der Lernenden beim Lösen der Testaufgaben wurden anschliessend kriterienbasiert 20 Transkripte ausgewählt. Über die Resultate kann in Barras (i.V.) nachgelesen werden.

Alle Aufgaben, Testergebnisse und Transkripte (darunter die 20 codierten) sind im Sinne des Open-Data-Gedankens über das Forschungsdatenarchiv des Instituts für Mehrsprachigkeit zugänglich und können im Hinblick auf weitere Fragestellungen untersucht werden.

Item Response Theory

In der sogenannten *Item Response Theory* (IRT) werden die Schwierigkeit von Testaufgaben (sog. Items) 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 IRT-Modellen, die jeweils unterschiedliche Bedingungen an den Test und die Testresultate stellen. 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, was insbesondere in Tests, die verschiedene Aufgabenformate umfassen, sinnvoll ist. Am Ende einer IRT-Modellierung stehen für eine Testdurchführung jeweils die geschätzten Itemmerkmale (insb. die Itemschwierigkeit) sowie die geschätzten Fähigkeitswerte der Testteilnehmenden. Ausserdem ist es möglich, zu überprüfen, ob sich Items und Personen so verhalten wie erwartet, also z.B. schwierige Items nur von guten Schülerinnen und Schülern gelöst werden.

Imputation

Ziel einer *Imputation* ist es, ein unvollständiges Datenset zu vervollständigen, um statistische Analysen zu ermöglichen bzw. zu verbessern. Fehlende Daten entstehen z.B., wenn ein Schüler wegen eines Zahnarzttermins nicht den gesamten Test bearbeitet hat. Bei der Imputation werden diese fehlenden Daten durch ein statistisches Verfahren geschätzt und alle Datenpunkte mit bekannter Messfehlerverteilung (z.B. die Ergebnisse von IRT-Skalierungen) durch Werte ersetzt, die sich im Bereich des Messfehlers bewegen. Das hierfür verwendete MICE[8]-Verfahren nutzt dafür einerseits die vorhandenen Daten einer Variablen und andererseits die mehr oder weniger starken Korrelationen zwischen den Variablen im Datensatz (z.B. zwischen dem Ergebnis im Wortschatztest und im Leseverstehenstest).

Die dabei entstehenden imputierten Datensätze sind vollständig und messfehlerfrei. Um dann statistische Analysen durchzuführen, werden mehrere Datensätze produziert, in denen die geschätzten Werte jeweils leicht voneinander abweichen. Die Analysen werden dann für jeden einzelnen Datensatz durchgeführt und deren Ergebnisse nach bestimmten Regeln zusammengefasst, um Aussagen über die Daten zu machen.

8 Multiple imputation by chained equations

Ausgewählte Resultate

Die entwickelten Lese- und Hörverstehensaufgaben eignen sich prinzipiell für den Einsatz in einem computerbasierten, handlungsorientierten Test

Aufgrund der Beobachtungen während der Datenerhebungen, qualitativer Feedbacks von den Lernenden und der psychometrischen Analysen der Testergebnisse konnten wir feststellen, dass sich die verschiedenen Lese- und Hörverstehensaufgaben in den Szenarien prinzipiell dazu eignen, im Klassenzimmer im Rahmen eines computerbasierten Fremdsprachentests eingesetzt zu werden.

Diese Feststellung hat zum einen mit der Struktur und dem Design der Tests zu tun. Die Aufgaben konnten von den Lernenden selbstständig und in der erwarteten Zeit gelöst werden. Der Einsatz der Computer erwies sich als unproblematisch. Die Touchscreens bzw. Touchpads der Laptops und die Headsets wurden ohne Anleitung bedient. Auch die Elemente auf dem Bildschirm stellten keine Hürde dar – praktisch allen Schülerinnen und Schülern gelang es, durch die Testumgebung zu navigieren, die Hörtexte selbstständig zu starten und die Aufgaben mittels Anklicken und Drag-and-Drop zu bearbeiten. All dies lässt sich durch die Vertrautheit der Lernenden mit digitalen Geräten erklären (im Fragebogen gaben 60% der befragten Jugendlichen an, mindestens einmal täglich einen Computer zu nutzen, 90% nutzen täglich ein Smartphone), hat aber auch mit dem sorgfältigen Design der Testaufgaben zu tun. Dieses stand neben der inhaltlichen Aufgabenentwicklung immer wieder im Fokus der Erprobungen und folgte dabei vor allem den Prinzipien der authentischen Darstellung und der einfachen, fehlerfreien Bedienung.

Zum anderen zeigte sich in der statistischen Analyse, dass die Testresultate der Hör- und Leseverstehensaufgaben[9] gute psychometrische Eigenschaften aufweisen. Im Projekt IFB erwies sich aufgrund der unterschiedlichen Aufgabenformate im gleichen Test ein 2PL-IRT-Modell als zielführend. → Kasten IRT

Eine kleine Zahl von Aufgaben fiel statistisch auf (d.h. einzelne Items wiesen einen schlechten Modellfit auf) und war meist aus plausiblen Gründen problematisch. Beispielsweise war eine zur Auswahl stehende Antwort nicht eindeutig oder missverständlich, oder eine eigentlich falsche Multiple-Choice-Option übermässig attraktiv. In wenigen Fällen fielen auch einzelne Aufgaben auf, die für die Zielgruppe zu anspruchsvoll waren. Die Mehrheit der Aufgaben erfüllte aber die Qualitätsansprüche an einen Hör- und Leseverstehenstest für den Einsatz in einem Large-Scale-Assessment.

9 Gemeint sind hier nur die Aufgaben A-C der Szenarien, also nicht die jeweils letzte Aufgabe, in der Hör- und Leseverstehen in einer komplexen Aufgabenstellung kombiniert wurden. Diese erwiesen sich als problematischer: Sie waren sowohl inhaltlich als auch bezüglich der Aufgabenstellung anspruchsvoller und wurden daher oft nur von sehr fortgeschrittenen Lernenden der Zielgruppe korrekt bearbeitet.

Die statistischen Analysen zeigen, dass die Lernenden in erster Linie ihre Sprachkenntnisse einsetzen, um die Verstehensaufgaben zu lösen

Wie erwähnt kamen neben den Szenarien verschiedene weitere Tests und Fragebögen zum Einsatz, die verschiedene Aspekte erhoben, die für den Erfolg bei den szenariobasierten Aufgaben potenziell relevant sind. Dies sollte es ermöglichen, die Ressourcen zu identifizieren, welche die Schülerinnen und Schüler mobilisieren, um die Testaufgaben zu lösen. → Tabelle 2

Die Analysen zeigen, dass die Lernenden in erster Linie ihre Sprachkenntnisse und ihre allgemeinen kognitiven Fähigkeiten einsetzten, um die szenariobasierten Aufgaben zu lösen (insbesondere Wortschatz- und Grammatikkenntnisse, aber auch nichtverbale fluide Intelligenz spielten eine Rolle).[10] Aspekte wie Sprachlernmotivation, Sprachlernangst sowie der Einsatz von Teststrategien waren hingegen weniger entscheidend. Dies gilt sowohl für stärkere als auch für schwächere Lernende. Das Ergebnis ist im Hinblick auf die Validität der Testaufgaben erfreulich, weil es aufzeigt, dass auch die szenariobasierten Aufgaben in erster Linie spezifisch Sprachkenntnisse testen und in geringerem Mass bloss allgemeine Kompetenzen. Dafür spricht auch, dass die Ergebnisse der „traditionelleren“ Hör- und Leseverstehensaufgaben stark mit den Resultaten aus den szenariobasierten Aufgaben korrelieren. Die Rolle, welche die fluide Intelligenz spielt, kann als Hinweis dafür gewertet werden, dass die Aufgaben über die sprachlichen Anforderungen hinaus kognitiv einigermassen komplex waren.

Die Szenarien werden von den Schülerinnen und Schülern weniger wahrgenommen als vermutet, aber die digitalen Textsorten kommen gut an

Ergebnisse aus der qualitativen Untersuchung, aber auch Äusserungen der Lernenden in den Fragebögen lassen vermuten, dass die Einbettung der Aufgaben in Szenarien von den Lernenden nur am Rande wahrgenommen wurde. Einiges deutet darauf hin, dass die Schülerinnen und Schüler in der Testsituation, die in der Schule häufig mit Leistungsdruck verbunden ist, nur bedingt Interesse an den Inhalten von Lese- und Hörtexten aufbringen und stattdessen ihre Aufmerksamkeit auf die Bearbeitung der gerade vor ihnen liegenden Fragestellungen richten. Besonders eingängig ist in diesem Zusammenhang die Äusserung der Schülerin „Sofia“,[11] die an der

10 Zur Analyse des Englischtests wurden auf der Basis der potenziell relevanten Variablen Strukturgleichungsmodelle und Zwei-Ebenen-Regressionsmodelle aufgestellt. Die hier berichteten Ergebnisse gelten nur für die Hör- und Leseverstehensaufgaben in den Szenarien, nicht aber für die kombinierten Aufgaben am Ende jedes Szenarios.

11 Selbstgewähltes Pseudonym.

qualitativen Datenerhebung teilgenommen hat:

Sofia:

(…) ich glaube, es hat noch nie so ein Thema gegeben, das mich an einem Test interessieren würde, weil dann denke ich nicht mal so richtig dran. Dann ist es nur so der Stress (…) dass ich die Lösungen finden muss und so. (…)

Forscherin:

Ähm, das heisst, dass du beim, bei einem Test nicht wirklich das Thema mitbekommst.

Sofia:

Doch, schon. Aber ich kann es nicht geniessen sozusagen, auch wenn es jetzt ein Thema wäre, das ich gern hätte. Ich bin gestresst und ich (…) kann mich nicht so auf den Text sozusagen konzentrieren, (…) also das einzige, woran ich denke, ist ja, dass ich die Lösungen finden muss und so. Und dann denke ich nicht „Oh cool" und so.

Die im Test simulierten Textsorten wie Smartphone-Chats, Webseiten und Podcasts wurden von den Lernenden deutlich stärker wahrgenommen und mehrheitlich positiv eingeschätzt. So äusserte „Billy" die folgende Überlegung:

Also ich finde es noch gut, weil (…) die heutige Zeit ist ja auch sehr viel auf die Medien konzentriert, und die Medien sind ein fester Bestandteil. Ich denke, dass man jetzt auch für die Leute, die noch kommen, also die Kinder (…) ich finde das gut, (…) dass man es ihnen durch den Chat vielleicht ein bisschen näher bringt (…) weil das wie eine alltägliche Situation ist.

Es konnte allerdings auch beobachtet werden, dass einige Lernende vor allem auch deshalb positiv reagierten, weil sie im Fremdsprachenunterricht eher selten mit solchen Texten zu tun haben. In diesem Fall war also das „Neue" für kurze Zeit motivierend. „Omega" bringt dies auf den Punkt:

Ja, äh nein, (…) also es ist zuerst eine Motivation und dann irgendwie kommt dir in den Sinn, dass es ein Test ist. Dann denkst du „Ja, ist doch witzig, aber jetzt muss ich mich wieder konzentrieren."

Fazit

Diese Beobachtungen, die praktischen Erfahrungen aus den Testdurchführungen, die beschriebenen Resultate der ersten statistischen Analysen und die Ergebnisse der qualitativen Studie (Barras, i.V.) geben einen Eindruck davon, welches Potential und welche Komplexität im Projekt IFB und in der Sprachtestforschung allgemein stecken. Im Forschungsdatenarchiv des KFM lagert nun reichhaltige empirische Evidenz, um den Testlösungsprozessen von Schülerinnen und Schülern in zwei Fremdsprachen nachzugehen. Mögliche Fragestellungen sind zum Beispiel die folgenden: Wie nehmen Schülerinnen und Schüler nah-authentische Aufgabenstellungen und Texte in einer Testsituation wahr? Welche Bedeutung hat Wortschatzkenntnis für das Leseverstehen in der Fremdsprache? Warum konkret sind bestimmte Hörverstehensaufgaben schwieriger als andere? Was denken die Lernenden über die Aufgabenstellungen in ihrer Schulsprache? Spielt es eine Rolle, ob die Lernenden Französisch oder Englisch als erste Fremdsprache gelernt haben?

Unsere Analysen lassen den Schluss zu, dass der Einsatz von Szenarien auch im niedrigeren Niveaubereich zahlreiche Möglichkeiten bietet, um realer Sprachverwendung im Fremdsprachenunterricht näher zu kommen.

Es bleibt festzuhalten, dass das Potential des *Reading for Understanding*-Frameworks durch geschlossene Aufgabenformate und eine kurze Intervention im Klassenraum, wie sie im Projekt IFB umgesetzt wurde, noch nicht ausgeschöpft ist: Es wäre spannend, auch produktive und interaktive Sprachverwendung (d.h. Sprechen und Schreiben) in ein szenariobasiertes Beurteilungsinstrument miteinzubeziehen. Ebenso wäre es, im Sinne der Lernförderung, wünschenswert, Beurteilungen und Unterricht stärker aufeinander abstimmen zu können. Beides ist im *Reading for Understanding*-Konzept angelegt und wäre eine lohnende Fortsetzung des Projektes „Innovative Formen der Beurteilung".

Wo finde ich mehr Informationen?

Auf der Webseite des Kompetenzzentrums für Mehrsprachigkeit können verschiedene Poster und PDFs von Vorträgen eingesehen werden, die an Fachkonferenzen präsentiert wurden: https://tinyurl.com/268uewnc.

Évaluer les compétences en langues étrangères au plus proche de la réalité

Tâches de test informatisées basées sur des scénarios – une étude approfondie

Katharina Karges, Peter Lenz, Thomas Aeppli, Malgorzata Barras

Vue d'ensemble

Avec l'invention de l'ordinateur, le développement d'internet et surtout l'essor des tablettes et des smartphones, la compréhension de l'oral et de l'écrit a pris de nouvelles dimensions. Le texte linéaire – qu'il soit imprimé ou enregistré – a cédé la place à l'hypertexte multimédia, comportant des liens ainsi que des images, du son, des vidéos et des graphiques interactifs intégrés. De même, les services de messagerie vocale, les tweets et les podcasts brouillent toujours davantage la frontière entre l'écrit et l'oral. Si ces nouveautés font depuis longtemps partie de la vie quotidienne des élèves, elles ont également eu un impact sur l'enseignement des langues étrangères à l'école et donc sur l'évaluation de la capacité des élèves à communiquer dans une langue. La numérisation de la communication a aussi permis d'élargir la palette des instruments destinés à évaluer les compétences des apprenant·e·s en matière de compréhension de l'écrit et de l'oral : outre les nouveaux types de textes numériques à lire ou écouter, il est désormais envisageable de recourir à des formats de tâches inédits dans le cadre de tests informatisés. Ces nouvelles tâches doivent cependant elles-mêmes faire l'objet de tests et d'études avant de pouvoir être recommandées pour un usage à large échelle. Ces considérations ont constitué le point de départ du projet de recherche « Formes innovantes d'évaluation » (titre original « *Innovative Formen der Beurteilung* », ci-après IFB), mené par le Centre scientifique de compétence sur le plurilinguisme (CSP) de 2016 à 2019. Ce projet a consisté à élaborer des tâches de test intégrées dans ce que l'on appelle des scénarios. Il s'agit d'activités qui se déroulent dans un cadre donné et qui, de par leur proximité avec des tâches de la vie quotidienne et le recours à des textes de type numérique, simulent un usage de la langue étrangère aussi authentique que possible, censé motiver les participant·e·s à utiliser pleinement leurs capacités linguistiques. Dans l'un de ces scénarios par exemple, les élèves devaient accomplir des tâches de lecture et d'écoute de textes liés à la planification d'une excursion.

En vue d'élaborer les tâches destinées à être utilisées dans le cadre de la seconde enquête nationale « Vérification de l'atteinte des compétences fondamentales (COFO) »,[1] un travail dans lequel l'Institut de plurilinguisme (IDP) a été impliqué par la suite, le projet IFB a développé des tâches informatisées fermées.[2] Ces dernières ont été conçues

1 La COFO fait partie du monitorage de l'éducation à l'échelle nationale. En 2016, les compétences ont été mesurées pour la première fois en mathématiques puis, en 2017, dans la langue de scolarisation et dans la première langue étrangère enseignée à l'école. Il était prévu d'enquêter en 2020 sur les compétences en matière de compréhension de l'oral et de l'écrit des élèves dans leur langue de scolarisation et dans leurs deux langues étrangères à la fin de la scolarité obligatoire. L'IDP a participé à l'élaboration des tâches, mais en raison des mesures liées à la COVID-19, la récolte de données principale n'a pas pu avoir lieu et a été reportée à 2023. Pour plus d'informations https://cofo-suisse.ch.

2 Dans les formats de tâches fermées, les participant·e·s choisissent leurs réponses parmi au moins deux options de réponse données.

pour évaluer les compétences en compréhension de l'oral et de l'écrit dans les langues étrangères pour les niveaux A2/B1 du Cadre européen commun de référence pour les langues (CECRL ; Conseil de l'Europe, 2001). En regroupant les tâches en scénarios, le projet IFB est toutefois allé au-delà de l'approche choisie pour la COFO.

Les tâches ainsi développées ont été examinées dans le cadre d'une étude de validation par méthodes mixtes. L'étude *qualitative* a fait appel à des procédures introspectives (méthode de la pensée à voix haute et entretiens de rappel stimulé) pour observer comment les apprenant·e·s percevaient et traitaient les tâches.[3] Dans l'étude *quantitative*, des tests sur les compétences partielles (vocabulaire, vitesse de traitement de la langue et grammaire notamment) ainsi que des questionnaires se sont ajoutés aux tâches basées sur des scénarios.

Alors que l'étude qualitative était axée sur le français, toutes les tâches développées ont été proposées dans des écoles suisses-allemandes en français et anglais, soit les deux langues étrangères enseignées. La récolte de données s'est déroulée en plusieurs phases tout au long de l'année 2017 et au début de l'année 2018.

Le projet avait pour but principal et pratique d'étudier dans quelle mesure il est possible et judicieux d'utiliser des tâches basées sur des scénarios pour évaluer les compétences en matière de compréhension d'une langue étrangère à un faible niveau de maîtrise. L'étude de validation devait donc répondre, entre autres, aux questions suivantes :

1. Quelles sont les compétences partielles que les apprenant·e·s utilisent lorsqu'elles et ils résolvent des tâches dans les langues étrangères enseignées ?
2. La plus grande authenticité supposée des tâches est-elle perçue voire appréciée par les apprenant·e·s ?
3. Les tâches répondent-elles aux exigences psychométriques des évaluations standardisées (en anglais, *large scale assessments*) ?

3 Une grande partie de l'étude qualitative a été réalisée par Malgorzata Barras dans le cadre de son projet de thèse associé (Barras, i.V.).

Qu'est-ce que l'évaluation basée sur des scénarios ?

L'évaluation basée sur des scénarios (en anglais, *scenario-based assessment*) fait partie de la démarche « Reading for Understanding » (« Lire pour comprendre », Sabatini et al., 2014a ; Sabatini & O'Reilly, 2013) également utilisé dans certaines études internationales comme PISA 2018 (OCDE 2019, p. 41). « Reading for Understanding » est le résultat d'une série de projets de recherche menés par l'Educational Testing Service (ETS) aux États-Unis, qui ont permis de réunir les conclusions de diverses disciplines sur la nature de la lecture. L'objectif était de concevoir un modèle qui reflète les différentes dimensions de la lecture. Selon cette approche, la compréhension d'un texte peut relever de cinq domaines (Sabatini et al., 2013, p. 14 ss). → Tableau 1

Cette conception de la compréhension de texte conduit à une définition large de la compréhension de l'écrit et modifie en conséquence les exigences relatives à son évaluation : outre les tâches servant à évaluer le niveau de compréhension des textes, tout instrument d'évaluation devrait notamment inclure des tâches qui créent des opportunités d'apprentissage et en saisissent les résultats, testent l'intégration de nouvelles connaissances dans un savoir existant ou permettent d'évaluer la crédibilité des sources (pour des exemples concrets de telles tâches, voir O'Reilly et al., 2014 ; Sabatini et al., 2014b). Le projet IFB s'est attaché à appliquer cette conception de la compréhension des langues à l'évaluation des compétences en langues étrangères. La frontière entre sources orales et écrites dans les médias numériques devenant toujours plus floue, le projet IFB a développé des tâches de compréhension de l'écrit comme de l'oral.

Domaines de compréhension	Exemples
Éléments visuels et structure des textes imprimés	Par exemple lettres de l'alphabet, mots, signes de ponctuation, fonction des paragraphes ou des titres, mais aussi graphiques, hyperliens, émoticônes, etc.
Éléments verbaux d'une langue	Sens des mots, morphologie, syntaxe, etc.
Structures discursives et genres de textes	Organisation du texte, références inter- et intra-textuelles, prise en compte de la culture générale
Contenus et sens conceptuels	Intégration du savoir existant et des nouvelles connaissances, évaluation critique du contenu
Contenus et sens sociaux	Examen des idées de l'autrice/de l'auteur ou des protagonistes d'un texte, compréhension de la situation dans laquelle un texte est créé, classification du ou des rôles sociaux des textes et des parties prenantes

Tableau 1 : Les cinq domaines de compréhension de texte adaptés de Sabatini et al. (2013)

À quelles tâches a-t-on eu recours ?

La partie principale de l'étude IFB se compose de six scénarios de quatre tâches (A à D) de niveaux A2 à B1, dont le contenu est lié mais qui peuvent être résolues indépendamment les unes des autres et sans aide extérieure en utilisant un ordinateur. Chaque scénario existe dans deux versions (anglaise et française) au contenu identique. Toutes les descriptions de scénarios, les instructions et les consignes de tâches sont rédigées dans la langue de scolarisation afin de ne pas désavantager les élèves ayant de faibles compétences en compréhension de la langue étrangère (pour plus de détails, voir Barras et al., 2016).

Les scénarios ont été élaborés de manière itérative par l'équipe du projet. Une fois les sujets et les formats de tâches définis, les scénarios ont été développés progressivement par plusieurs membres de l'équipe, testés individuellement avec des apprenant·e·s et en général révisés plusieurs fois. Tous les textes à lire ont été rédigés par l'équipe du projet sur la base de modèles tirés de situations réelles. Les textes audio ont également été spécialement conçus pour les scénarios, mais les locutrices et locuteurs en studio ne se sont vu indiquer que les éléments clés et grandes lignes de la conversation. Le choix des mots proprement dit et la mise en forme des textes ont été laissés à leur appréciation afin que la langue parlée soit plus authentique. Tous les scénarios ont été élaborés en parallèle en anglais et en français, de sorte qu'aucune des deux versions n'est une traduction de l'autre.

Le scénario « Walkies » présenté ci-dessous à titre d'exemple invite les apprenant·e·s à réfléchir à un référendum fictif sur l'introduction de robots capables de promener les chiens. Le scénario choisi illustre bien tout l'éventail des types de textes et des formats de tâches. Il comprend une tâche de compréhension de l'oral (B) et deux tâches de compréhension de l'écrit (A et C), ainsi qu'une tâche pour laquelle les deux compétences étaient requises (D). Dans l'ensemble, les textes de ce scénario sont relativement longs et la difficulté des tâches C et D se situe plutôt dans la moyenne supérieure.

Au début de chaque scénario, les apprenant·e·s sont informé·e·s du contexte dans lequel les tâches s'inscrivent : « Tu es en séjour linguistique dans la ville canadienne de Québec[4] où tu vas à l'école. En cours de sciences sociales, vous discutez du référendum populaire en cours sur le thème : « Les robots doivent-ils être autorisés à promener les chiens ? » Ensuite, les apprenant·e·s résolvent la première tâche. Elles et ils lisent un texte d'information sur une page internet fictive et sélectionnent les informations les plus importantes données dans le texte parmi plusieurs suggestions. → Figures 1 et 2

4 Dans la version anglaise, la ville de Québec de la version française a été remplacée par la ville de Vancouver dans toutes les tâches.

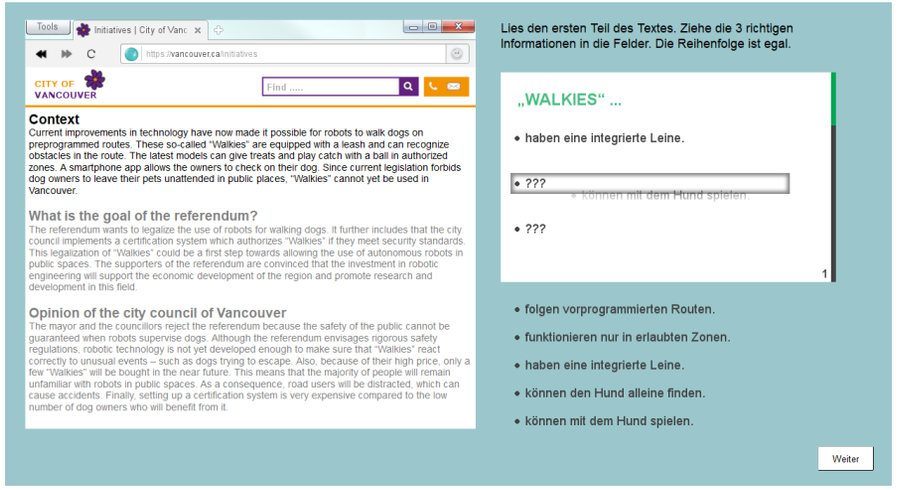

Figure 1 : La première page de la tâche A du scénario « Walkies » (version anglaise)[5]

Figure 2 : La première page de la tâche A du scénario « Walkies » (version française)[5]

5 Dans le texte, la partie qui est en rapport avec cette page d'écran est plus foncée. Un paragraphe différent est ensuite mis en évidence sur chacune des deux pages suivantes. La tâche à réaliser est affichée à droite : parmi les cinq options d'informations proposées en dessous, les apprenant·e·s peuvent en sélectionner trois et les déposer avec l'outil « drag & drop » dans la liste au-dessus.

Dans la deuxième tâche, les apprenant·e·s écoutent ce que des camarades de classe fictifs pensent de l'introduction de ces robots et répondent à des questions à choix multiples « classiques » à leur sujet. La troisième tâche consiste également en des questions à choix multiple, mais les apprenant·e·s indiquent en plus où elles et ils ont trouvé leur réponse. Ici, la base textuelle est un chat de groupe dans une application de messagerie dans laquelle le travail d'un groupe fictif est planifié. Les apprenant·e·s lisent les messages de ce chat et répondent aux questions en déposant le message qui contient la réponse dans le champ prévu à cet effet avec l'outil « drag & drop ». → Figure 3

Enfin, dans la dernière partie du scénario (tâche D), les apprenant·e·s complètent une chronologie de la vie de l'inventrice imaginaire des « Walkies » en se basant sur une entrée Wiki fictive et une interview dans laquelle l'inventrice parle d'elle-même. → Figure 4

Les autres scénarios abordent les thèmes suivants : une sortie au cinéma, la visite d'un centre d'orientation professionnelle, l'organisation d'une fête d'école, une excursion d'un week-end dans une ville et des recherches pour un exposé en cours de géographie. Les formats de tâche et les types de texte décrits sont variés. Il s'agira par exemple de sélectionner un résultat pertinent dans la liste des résultats proposés par un moteur de recherche ou de compléter un emploi du temps sur la base de messages audio. Chaque scénario comprend au moins une tâche de compréhension de l'oral, une tâche de compréhension de l'écrit et une tâche pour laquelle les apprenant·e·s doivent mobiliser les deux compétences.

En plus de ces scénarios, d'autres tests et questionnaires ont été utilisés dans l'étude principale afin de déterminer quelles sont les compétences linguistiques partielles et autres caractéristiques individuelles des élèves qui jouent un rôle dans la résolution des tâches. → Tableau 2

Les données ainsi obtenues permettent de décrire plus précisément la façon dont les apprenant·e·s ont travaillé sur les tâches basées sur des scénarios et, plus généralement, de tirer des conclusions empiriques sur les caractéristiques individuelles et les processus impliqués dans la compréhension de textes dans une langue étrangère.

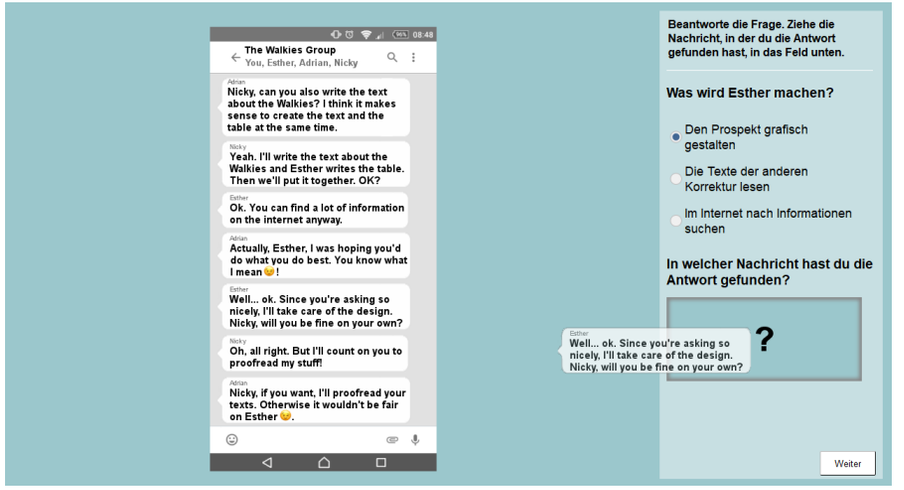

Figure 3 : La seconde page de la tâche C du scénario « Walkies » (version anglaise)

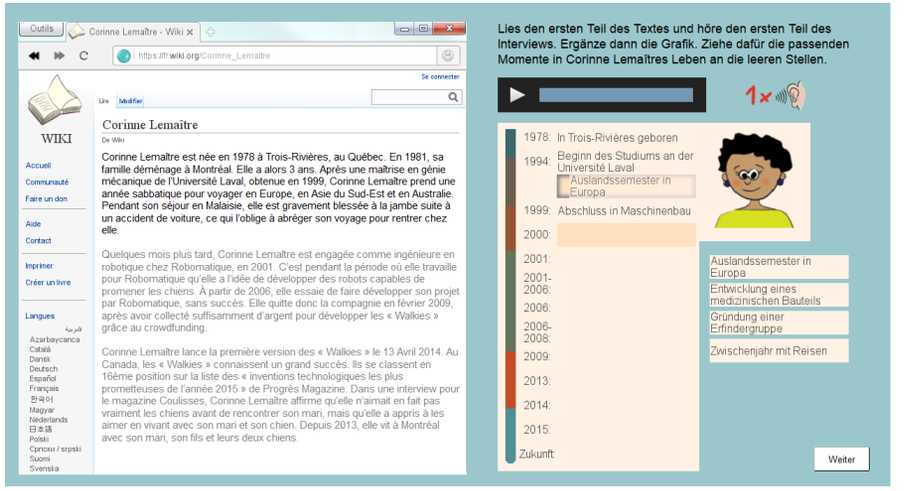

Figure 4 : La première page de la tâche D du scénario « Walkies » (version française)[6]

6 Les apprenant·e·s lisent une entrée wiki et écoutent une interview au contenu similaire. Ensuite, elles et ils font glisser deux des quatre options proposées à l'extrême droite dans le « CV ».

Compréhension de l'oral et de l'écrit	6 scénarios de 4 tâches chacun	40 items de compréhension de l'écrit 32 items de compréhension de l'oral 12 items de compréhension de l'oral et de l'écrit combinés
	6 tâches de compréhension de texte dans des formats connus	36 items de compréhension de l'écrit (choix multiple, vrai/faux) 56 items de compréhension de l'oral (choix multiple, texte à compléter)
Compétences linguistiques partielles	2 tests de vocabulaire	28 items (choix multiple avec mots écrits) 22 items (choix multiple avec mots énoncés à l'oral)
	2 tests de segmentation de mots dans un texte	2 textes écrits sans espaces à reconstituer (max. 326 mots) 28 items (reconnaître le nombre de mots dans les énoncés oraux)
	Compétences grammaticales	24 items (évaluer l'exactitude grammaticale d'un énoncé)
	Vocabulaire visuel	30 items (reconnaître un mot qui n'est que brièvement affiché à l'écran)
Prérequis individuels	Intelligence fluide	20 matrices de Raven (test d'intelligence non verbale dans lequel une séquence de motifs doit être complétée)
	Attitude personnelle vis-à-vis de l'apprentissage d'une langue étrangère	11 items de questionnaire sur la motivation 7 items de questionnaire sur l'anxiété langagière
	Stratégies pour résoudre des tâches de compréhension	15 items de questionnaire sur la compréhension de l'écrit (résolution active de problèmes axée sur les mots et les détails) 10 items de questionnaire sur la compréhension de l'oral (écoute active-concentrée, planification et évaluation)
	Compétences linguistiques	10 items de questionnaire (histoire linguistique personnelle) 7 items de questionnaire (utilisation de la langue étrangère en classe)
	Autres caractéristiques individuelles	Items de questionnaire sur l'origine, le contexte socio-économique, l'équipement technique, l'utilisation des appareils et des canaux numériques

Tableau 2 : Aperçu des tests et questionnaires utilisés dans le cadre du projet IFB

Qui a participé à l'étude ?[7]

Avant de procéder à l'étude principale, les scénarios ont été testés au cours de leur phase de développement avec des apprenant·e·s pris individuellement (n=53), discutés dans le détail et continuellement améliorés. Les autres tests et questionnaires ont également été discutés avec des apprenant·e·s (n≈50) dans le cadre d'entretiens individuels et collectifs. En outre, toutes les tâches ont été testées en classe (n=85) afin de vérifier la faisabilité technique de la récolte de données principale. La récolte de données qualitative principale a également été testée et affinée au cours de cette phase avec six apprenant·e·s.

Au total, 631 élèves (soit 39 classes) ont participé à la récolte de données principale, dont 30 ont pris part à l'étude qualitative. Dans la plupart des classes, environ la moitié des apprenant·e·s ont résolu les tâches en anglais (292 apprenant·e·s au total), tandis que les autres (parmi lesquels l'ensemble des apprenant·e·s participant à l'étude qualitative) ont passé le test dans sa version française. Dans les cantons de Berne, Zurich, Fribourg, Lucerne et Obwald, les élèves fréquentaient la deuxième et troisième année du secondaire I (âge moyen : 15 ans) et représentaient tous les types de classes. Environ deux tiers des élèves ayant participé étaient issus des cantons de Berne et Fribourg, soit des cantons dans lesquels le français est enseigné en première langue étrangère et l'anglais en second. Au total, la participation à l'enquête principale a nécessité six leçons réparties sur deux jours. Les élèves ont résolu toutes les tâches de manière largement autonome sur les ordinateurs portables fournis, sous la supervision d'un membre de l'équipe projet. → Figure 5

Les apprenant·e·s impliqués dans l'étude qualitative ont travaillé individuellement avec la chercheuse chargée de cette partie de l'étude. Pendant trois leçons, ils et elles ont travaillé sur un scénario ou sur cinq tâches individuelles et ont verbalisé leurs réflexions (méthode de la pensée à voix haute, cf. p. ex. Bowles, 2010 ; Knorr & Schramm, 2012). Ensuite, elles et ils ont encore commenté les tâches accomplies dans le cadre d'un entretien de rappel stimulé (en anglais, *stimulated recall interview*) au cours duquel toutes les tâches et leurs solutions ont été passées en revue avec la chercheuse (pour plus de détails sur la méthode, voir p. ex. Barras, 2018 ; Gass & Mackey, 2017). Le second jour, ces apprenant·e·s ont effectué le reste des tests (à l'exception des autres scénarios) et des questionnaires en classe.

7 L'équipe tient à remercier les élèves et les responsables d'établissements pour leur soutien. Cette étude n'aurait pas pu être menée sans leur disponibilité et leur engagement. Nos assistant·e·s scientifiques, qui se sont levé·e·s de bonne heure en plein hiver, ont transporté des valises à travers la moitié de la Suisse et patiemment regardé les élèves travailler, méritent également nos plus sincères remerciements.

Figure 5 : Une salle de classe est prête pour la récolte de données principale (photo : Malgorzata Barras)

Comment les données ont-elles été analysées ?

Durant la phase de test, les entretiens ont été résumés et l'on en a ressorti les informations qui étaient importantes pour la révision des tâches et la récolte des données. Une partie des entretiens a également été transcrite et a été mise à disposition pour d'autres analyses. Les résultats des tests issus de la phase pilote ont été extraits des données brutes générées par le système de test avant de faire l'objet d'une évaluation descriptive.

Les résultats des tests provenant de la récolte de données principale ont également été traités et reportés sur une échelle, dans la plupart des cas en utilisant la théorie des réponses aux items (en anglais, *item response theory* IRT). Au cours d'une étape ultérieure, toutes les données quantitatives ont été soumises à une procédure d'imputation.

→ Encadrés

Les 30 protocoles de pensée à voix haute et entretiens de rappel stimulé de la récolte de données principale ont été entièrement transcrits. 20 transcriptions ont ensuite été sélectionnées sur la base de critères précis pour servir à l'analyse approfondie des stratégies appliquées par les apprenant·e·s dans la résolution des items du test. Les résultats sont décrits dans Barras (i.V.).

Selon les principes des données ouvertes, l'ensemble des tâches, résultats des tests et les 30 transcriptions (y compris les 20 transcriptions codées) sont accessibles via les archives de données de recherche de l'Institut de plurilinguisme et peuvent être consultés en vue d'autres recherches.

Théorie des réponses aux items

Lorsque l'on analyse des résultats de tests à partir de la *théorie des réponses aux items* (IRT), la difficulté des tâches (appelées items) et les compétences mesurées des participant·e·s, 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 l'ensemble des participant·e·s n'ont pas réalisé les mêmes tâches d'un test.

Il existe toute une série de modèles mathématiques de l'IRT, chacun fixant des conditions différentes au test et à ses résultats. 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, ce qui est particulièrement utile pour les tests qui regroupent différents formats de tâches.

Ainsi, pour un test donné, la modélisation IRT permet d'obtenir une estimation des caractéristiques des items (en particulier leur niveau de difficulté) et une estimation des compétences des participant·e·s. Il est également possible de vérifier si les items et les participant·e·s se comportent comme prévu, par exemple si les items difficiles ne sont résolus que par les élèves possédant un bon niveau.

Imputation

L'objectif d'une *imputation* est de compléter un ensemble de données lacunaire afin de pouvoir réaliser des analyses statistiques ou d'en améliorer la fiabilité. Des données peuvent manquer, par exemple, lorsqu'une ou un élève n'a pas terminé le test en raison d'un rendez-vous chez le dentiste. Lors de l'imputation, ces données manquantes sont estimées par un procédé statistique par lequel tous les points de données dont la distribution des erreurs de mesure est connue (p. ex. les résultats des mises à l'échelle IRT) sont remplacés par des valeurs qui se situent dans la fourchette de l'erreur de mesure. La méthode MICE[8] adoptée à cette fin utilise d'une part des données existantes d'une variable et d'autre part, les corrélations plus ou moins fortes entre les variables de l'ensemble de données (p. ex. entre le résultat du test de vocabulaire et celui du test de compréhension de l'écrit).

Les ensembles de données imputées qui en résultent sont complets et exempts d'erreurs de mesure. Afin d'effectuer ensuite des analyses statistiques, plusieurs ensembles de données sont générés dans lesquels les valeurs estimées diffèrent légèrement les unes des autres. Chaque ensemble de données fait à son tour l'objet d'analyses, puis les résultats sont résumés selon des règles précises afin de tirer des conclusions sur ces données.

8 Multiple imputation by chained equations

Résultats choisis

Les tâches de compréhension de l'oral et de l'écrit développées sont adaptées aux tests informatisés et orientés vers l'action

Les observations effectuées lors de la récolte des données, les commentaires qualitatifs des apprenant·e·s et les analyses psychométriques des résultats des tests ont montré que les tâches de compréhension de l'oral et de l'écrit contenues dans les scénarios se prêtent en principe aux tests de langues étrangères informatisés, réalisés en classe.

D'une part, ceci est lié à la structure et à la conception des tests. Les tâches ont pu être réalisées par les apprenant·e·s sans aide extérieure et dans les délais impartis. Le recours aux ordinateurs n'a pas posé de problème. Les écrans tactiles ou touchpads des ordinateurs portables et les casques audio ont été utilisés sans instructions. Les éléments affichés à l'écran n'ont pas non plus présenté de difficulté : la grande majorité des élèves a pu naviguer dans l'environnement de test, lancer les audios sans aide extérieure et travailler sur les tâches en cliquant dessus ou en utilisant l'outil « drag & drop ». Cela est dû à la familiarité des apprenant·e·s avec les appareils numériques (dans le questionnaire, 60 % des jeunes interrogé·e·s ont déclaré utiliser un ordinateur au moins une fois par jour et 90 % un smartphone tous les jours), mais aussi au soin apporté au design des tâches de test. Si, lors de la phase test, le contenu a été constamment amélioré, le côté visuel a lui aussi fait l'objet d'une attention particulière afin de respecter les principes d'une présentation réaliste et de permettre un maniement aisé minimisant le risque d'erreur.

D'autre part, l'analyse statistique a mis en évidence les bonnes propriétés psychométriques des résultats des tests de compréhension de l'oral et de l'écrit.[9] En raison des différences entre les formats de tâches utilisés dans le même test, le recours à un modèle IRT 2-PLM s'est avéré utile dans le projet IFB. → Encadré IRT

Il apparaît dans les statistiques qu'un petit nombre de tâches posaient problème (certains items étaient mal ajustés au modèle), souvent pour des raisons tout à fait plausibles ; par exemple, une option de réponse ambigüe ou trompeuse, ou une option à choix multiple incorrecte et trop attrayante. Il est aussi arrivé que certaines tâches individuelles se soient avérées trop exigeantes pour le groupe cible. Cependant, la majorité des tâches satisfaisait aux critères de qualité requis pour un test de compréhension de l'oral et de l'écrit destiné à être utilisé dans une évaluation standardisée (en anglais, *large scale assessment*).

9 Il s'agit uniquement des tâches A à C des scénarios, et non pas des dernières tâches, dont les consignes étaient plus complexes et qui combinaient la compréhension de l'oral et de l'écrit. Plus exigeantes au niveau du contenu autant que des consignes, ces dernières se sont avérées plus problématiques et n'ont souvent été réalisées correctement que par les élèves très avancé·e·s du groupe cible.

Les analyses statistiques montrent que les apprenant·e·s mobilisent principalement leurs compétences linguistiques pour résoudre les tâches de compréhension

Comme indiqué plus haut, divers autres tests et questionnaires ont été utilisés à côté des scénarios pour étudier des aspects pouvant potentiellement jouer un rôle dans la réussite des tâches basées sur les scénarios. Il s'agissait d'identifier quelles ressources les élèves mobilisaient pour résoudre les items du test. → Tableau 2

Les analyses montrent que les apprenant·e·s ont principalement fait appel à leurs connaissances linguistiques et leurs capacités cognitives générales pour résoudre les tâches basées sur des scénarios (en particulier les connaissances en vocabulaire et en grammaire, mais aussi l'intelligence fluide non verbale).[10] Les aspects tels que la motivation, l'anxiété langagière et l'utilisation de stratégies de tests sont moins déterminants, ce qui s'applique aux élèves les plus fort·e·s comme aux plus faibles. Du point de vue de la validité des items du test, ce résultat est encourageant. Il montre en effet que les tâches basées sur des scénarios mesurent avant tout spécifiquement les connaissances linguistiques et dans une moindre mesure les compétences générales. Ceci se trouve confirmé par le fait que les résultats des tâches « plus traditionnelles » de compréhension orale et écrite sont fortement corrélés aux résultats des tâches basées sur des scénarios. Le rôle joué par l'intelligence fluide indique probablement, qu'au-delà des exigences linguistiques, les tâches étaient d'une certaine complexité cognitive.

Les élèves ont moins conscience que prévu des scénarios mais les textes de type numérique sont bien accueillis

Les résultats de la recherche qualitative mais aussi les réponses aux questionnaires indiquent que l'intégration des tâches dans des scénarios n'a été remarquée que de façon très marginale par les apprenant·e·s. Il semble que dans une situation de test, qui dans le cadre scolaire est souvent associée à une pression à la performance, les élèves ne montrent qu'un intérêt limité pour le contenu des textes de lecture et d'écoute et concentrent plutôt leur attention sur les questions à traiter. La déclaration de l'élève « Sofia »,[11] qui a participé à la collecte de données qualitatives, est particulièrement révélatrice à cet égard :

10 Des modèles d'équations structurelles et de régression à deux niveaux ont été mis en place pour analyser le test d'anglais en fonction des variables potentiellement pertinentes. Les résultats rapportés ici s'appliquent uniquement aux tâches de compréhension de l'oral et de l'écrit des scénarios, et non aux tâches combinées à la fin de chaque scénario.

11 Pseudonyme choisi par l'intéressée.

Sofia :

(…) je pense qu'il n'y a jamais eu un sujet qui m'intéresse dans un test, parce qu'en fait je n'y pense même pas vraiment. C'est tellement le stress (…) que je dois trouver les réponses et tout ça. (…)

Chercheuse :

Hum, ça veut dire que quand tu passes un test, tu ne comprends pas vraiment de quoi il s'agit.

Sofia :

Si, quand même. Mais je ne peux pas l'apprécier, comment dire, même si c'était un sujet qui pourrait m'intéresser. Je suis stressée et je (…) ne peux pas vraiment me concentrer sur le texte, (…) et la seule chose à laquelle je pense, c'est que je dois trouver les réponses, etc. Et puis je ne pense pas «oh cool» et tout ça.

Les types de textes simulés dans le test, tels que les chats sur smartphone, les sites internet et les podcasts, ont été nettement mieux perçus par les apprenant·e·s qui les ont pour la plupart évalués positivement. Par exemple, «Billy» a exprimé la réflexion suivante :

Alors je trouve que c'est plutôt bien, parce que (…) à notre époque, on est très axés sur les médias, on vit avec les médias. Je pense que pour les personnes qui viennent après nous, donc les enfants, (…) c'est bien (…) qu'on puisse peut-être les rapprocher un peu plus de tout ça par des chats (…) parce que c'est comme ce qu'on vit tous les jours.

On a également pu observer que certain·e·s apprenant·e·s ont réagi positivement justement parce que ce genre de textes est rarement utilisé en classe de langue étrangère. Dans ce cas, la «nouveauté» devient momentanément un élément de motivation. «Omega» l'exprime bien :

Oui, euh non, (…) donc c'est d'abord motivant et puis tout à coup, on se rappelle que c'est un test. Puis on se dit «ok, c'est marrant, mais maintenant il faut que je me concentre de nouveau.»

Conclusion

Ces observations, l'expérience pratique acquise lors des sessions de tests, les descriptions des premières analyses statistiques et les résultats de l'étude qualitative (Barras, i.V.) laissent apparaître le potentiel et la complexité du projet IFB et de la recherche sur les tests de langue en général. Les archives des données de recherche du CSP contiennent désormais de nombreuses preuves empiriques à partir desquelles il est possible d'étudier les processus de résolution de tests des élèves dans deux langues étrangères. Les questions de recherche pourraient être les suivantes : comment les élèves perçoivent-elles et ils les consignes de tâches et les textes quasi-authentiques dans une situation de test ? Quelle est l'importance de la connaissance du vocabulaire pour la compréhension de l'écrit en langue étrangère ? Pourquoi, concrètement, certaines tâches de compréhension de l'oral sont-elles plus difficiles que d'autres ? Que pensent les apprenant·e·s des consignes de tâches dans leur langue de scolarisation ? Le fait d'apprendre le français ou l'anglais comme première langue étrangère a-t-il son importance ?

Nos analyses permettent de conclure que l'utilisation de scénarios, même à des niveaux de maîtrise plus faibles, offre de nombreuses possibilités de se rapprocher d'un usage réel de la langue dans l'enseignement des langues étrangères.

Reste à souligner que le potentiel de l'approche «Reading for Understanding» qui, comme c'est le cas dans le projet IFB, recourt à des tâches fermées accompagnées d'une brève intervention en classe, n'a pas encore été épuisé : il serait intéressant d'inclure également l'utilisation productive et interactive de la langue (c.-à-d. l'expression orale et écrite) dans un instrument d'évaluation basé sur des scénarios. Dans l'intérêt de la promotion de l'apprentissage, il serait également souhaitable que les évaluations et l'enseignement fassent l'objet d'une coordination plus étroite. Ces deux éléments sont inhérents à l'approche «Reading for Understanding» et constitueraient une suite intéressante au projet IFB.

Où trouver plus d'informations ?

Divers posters et PDF de conférences présentées lors de congrès spécialisés peuvent être consultés sur le site internet du Centre scientifique de compétence sur le plurilinguisme : https://tinyurl.com/268uewnc.

Testare le competenze nelle lingue straniere vicino alla realtà

Esercizi computerizzati basati su scenari – uno studio di approfondimento

Katharina Karges, Peter Lenz, Thomas Aeppli, Malgorzata Barras

Panoramica

Con l'invenzione del computer e lo sviluppo di internet, o al più tardi con la diffusione dei tablet e degli smartphone, la comprensione orale e scritta ha assunto nuove dimensioni. Da un testo lineare – su un libro o su una cassetta – si è passati a un ipertesto multimediale, con rimandi e media integrati (immagini, audio, video e grafici interattivi). Anche il confine tra parola scritta e parlata è reso vieppiù labile da notizie vocali trasmesse attraverso servizi di messaggistica, tweet e podcast. Questi cambiamenti non riguardano solo la quotidianità di allieve e allievi, influenzano anche l'insegnamento scolastico delle lingue straniere e, quindi, la valutazione della capacità di comunicazione linguistica. Con la digitalizzazione della comunicazione, sono aumentate anche le possibilità di valutazione della comprensione orale e scritta: oltre a nuove forme digitali di testo scritto o audio, i test computerizzati consentono di proporre nuovi formati di esercizi, i quali devono tuttavia essere valutati e studiati prima di raccomandarli per un uso più ampio.

Queste riflessioni sono state il punto di partenza per il progetto di ricerca "Forme di valutazione innovative", svolto tra il 2016 e il 2019 presso il Centro di competenza per il plurilinguismo. Nell'ambito del progetto, sono stati sviluppati esercizi integrati in cosiddetti scenari, ossia cornici narrative che, attraverso la loro vicinanza alla realtà e all'utilizzo di testi digitali, simulano un impiego quanto più possibile autentico della lingua straniera e motivano i partecipanti a fare ricorso alle loro competenze linguistiche. In uno di questi scenari, per esempio, le allieve e gli allievi dovevano risolvere esercizi di lettura e ascolto di testi che si riferivano all'organizzazione di una gita scolastica.

Nell'ottica dello sviluppo di esercizi per la seconda "Verifica delle Competenze Fondamentali" (VeCoF),[1] nel quale è stato poi coinvolto l'Istituto di plurilinguismo, sono stati concepiti compiti di comprensione orale e scritta nelle lingue straniere che, tramite esercizi chiusi[2] e computerizzati rilevano le competenze al livello A2/B1 del Quadro comune europeo di riferimento per la conoscenza delle lingue (Consiglio d'Europa, 2002). Poiché gli esercizi erano riuniti negli scenari, il progetto andava tuttavia oltre l'approccio della VeCoF.

Gli esercizi così concepiti sono stati passati al vaglio in uno studio di validazione con l'ausilio di un design *mixed-method*.

1 La VeCoF è parte del monitoraggio dell'educazione in Svizzera. Nel 2016, si è svolto il primo rilevamento, incentrato sulle competenze in matematica, nel 2017 è stata la volta delle lingue scolastiche e della prima lingua straniera appresa a scuola. Nel 2020, era previsto il rilevamento, tra le allieve e gli allievi giunti alla fine della scuola dell'obbligo, delle competenze in comprensione orale e della lettura nella lingua scolastica e nelle due lingue straniere. L'Istituto di plurilinguismo era coinvolto nello sviluppo degli esercizi, ma le misure antipandemiche hanno impedito lo svolgimento del rilevamento principale, rinviato al 2023. Maggiori informazioni: https://vecof-svizzera.ch.

2 I formati chiusi prevedono che i partecipanti scelgano le loro risposte tra almeno due possibilità predefinite.

Nella parte *qualitativa*, sono state applicate procedure introspettive (pensiero ad alta voce e interviste *stimulated recall*) per capire come i singoli partecipanti percepiscono ed elaborano gli esercizi.[3] Nella parte *quantitativa*, invece, ci si è avvalsi, oltre che di esercizi basati su scenari, anche di test di competenze parziali (p.es. lessico, rapidità dell'elaborazione linguistica e grammatica), nonché di questionari.

Tutti gli esercizi sono stati attuati in scuole svizzero-tedesche per verificare le competenze in francese e inglese. Lo studio qualitativo era focalizzato sul francese. Il rilevamento dei dati si è esteso su più fasi tra il 2017 e i primi mesi del 2018.

Uno degli obiettivi pratici principali del progetto era indagare in che misura sia possibile e ragionevole impiegare a un livello linguistico basso esercizi basati su scenari per la verifica delle competenze di comprensione in una lingua straniera. Lo studio di validazione doveva pertanto rispondere alle domande seguenti:

1. Quali competenze parziali utilizzano le allieve e gli allievi per risolvere esercizi basati su scenari nelle lingue straniere imparate a scuola?

2. La presunta maggiore autenticità degli esercizi è percepita e apprezzata dalle allieve e dagli allievi?

3. Gli esercizi soddisfano i requisiti psicometrici per la misurazione delle prestazioni (*large scale assessment*)?

3 Della parte qualitativa dello studio si è occupata per lo più Malgorzata Barras per il suo progetto associato di tesi di dottorato (Barras, i.V.).

Che cos'è una valutazione basata su scenari?

La valutazione basata su scenari (in inglese *scenario-based assessment*) è una parte del quadro strutturale "Reading for Understanding" (Sabatini et al., 2014a; Sabatini & O'Reilly, 2013) che trova applicazione anche in studi internazionali, per esempio nell'ambito del PISA 2018 (OECD, 2019, pag. 41). Il "Reading for Understanding" è il risultato di una serie di progetti di ricerca dello statunitense Educational Testing Service (ETS), i quali riuniscono conclusioni scientifiche sulla natura della lettura acquisite in varie discipline. Si tratta quindi di un modello che considera le differenti sfaccettature dell'atto di leggere. Secondo questo modello, la comprensione di testi può avvenire in cinque ambiti (Sabatini et al., 2013, pag. 14 segg.). → Tabella 1

Questo approccio alla comprensione di testi si traduce in un'ampia definizione della competenza di lettura, la quale modifica anche i requisiti a livello di valutazione: oltre a esercizi che misurano il grado di comprensione di testi, uno strumento di valutazione dovrebbe prevedere per esempio compiti che creino opportunità di apprendimento, rilevino i risultati di quanto appreso, verifichino l'integrazione di nuove conoscenze e permettano di ponderare la credibilità delle fonti (esempi concreti di compiti di questo genere si trovano in O'Reilly et al., 2014; Sabatini et al., 2014b). Nel quadro del progetto "Forme di valutazione innovative", si è tentato di estendere questo approccio alla comprensione linguistica alla valutazione delle competenze nelle lingue straniere. In considerazione della crescente commistione di fonti orali e scritte nei media digitali, sono stati sviluppati esercizi di comprensione orale e scritta.

Ambiti della comprensione di testi	Esempi
Elementi visuali e struttura di testi stampati	Lettere, parole, segni di punteggiatura, funzione (p.es. di capoversi o titoli), ma anche grafici, collegamenti ipertestuali, emoticon ecc.
Elementi verbali di una lingua	Significato delle parole, morfologia, sintassi ecc.
Strutture del discorso e generi testuali	Organizzazione del testo, rimandi inter e intratestuali, integrazione di conoscenze generali
Contenuti e significati concettuali	Integrazione di conoscenze esistenti e nuove, valutazione critica dei contenuti
Contenuti e significati sociali	Confronto con le idee dell'autore/autrice, risp. dei protagonisti di un testo, comprensione delle condizioni in cui è nato un testo, valutazione dei ruoli sociali di testi e attori

Tabella 1: I cinque ambiti della comprensione di testi secondo Sabatini et al. (2013)

Quali esercizi sono stati utilizzati?

L'elemento centrale del progetto "Forme di valutazione innovative" comprende sei scenari al livello A2/B1 costituiti da quattro esercizi (A-D). Gli esercizi di uno scenario sono legati in quanto a contenuto ma risolvibili separatamente e sono stati concepiti in modo ad essere risolvibili autonomamente al computer. Ogni scenario è disponibile in versioni identiche in inglese e in francese. Tutte le descrizioni, le istruzioni e le consegne sono redatte nella lingua scolastica per non svantaggiare le allieve e gli allievi più deboli a livello di competenze di comprensione nella lingua straniera (per una motivazione dettagliata vedi anche Barras et al., 2016).

Il team di progetto ha sviluppato gli scenari seguendo un procedimento iterativo. Una volta stabiliti i temi e i formati degli esercizi, gli scenari sono stati concepiti progressivamente da più membri del team, testati con singoli allievi e rielaborati più volte. Tutti i testi da leggere sono stati scritti dal team. Testi già esistenti relativi al mondo reale hanno funto da modello. Anche i testi da ascoltare sono stati preparati in funzione degli scenari, ma alle persone che hanno prestato la loro voce in studio sono stati forniti solo i punti chiave e lo svolgimento generale della conversazione. Onde raggiungere l'autenticità della lingua parlata, si è lasciato che scegliessero loro il lessico e la struttura dei testi. Tutti gli scenari sono stati sviluppati parallelamente in inglese e in francese. Nessuna versione è dunque una traduzione dell'altra.

Di seguito, presentiamo lo scenario "Walkies", nel quale le allieve e gli allievi sono confrontati con il tema di una votazione popolare fittizia per decidere in merito all'introduzione di robot che portino a passeggio i cani. Si tratta di uno scenario che consente di farsi un'idea della gamma dei tipi di testo e dei formati di esercizi. Esso comprende un esercizio di comprensione orale (B), due esercizi di comprensione scritta (A e C) e un esercizio in cui occorre mettere alla prova entrambe le capacità (D). Nel complesso, i testi sono piuttosto lunghi, e il grado di difficoltà degli esercizi C e D si situa nella fascia superiore.

All'inizio dello scenario, le allieve e gli allievi vengono informati sul contesto degli esercizi: "Stai facendo un soggiorno linguistico a Vancouver,[4] dove frequenti la scuola. A lezione di scienze sociali, discutete di una votazione popolare. L'argomento: "I robot dovrebbero essere autorizzati a portare a spasso i cani?". In seguito, le allieve e gli allievi risolvono il primo esercizio: su una pagina internet simulata leggono un testo e, tra varie proposte, scelgono le informazioni più importanti. → Figure 1 e 2

4 Nella versione francese, in tutti gli esercizi la città è Québec.

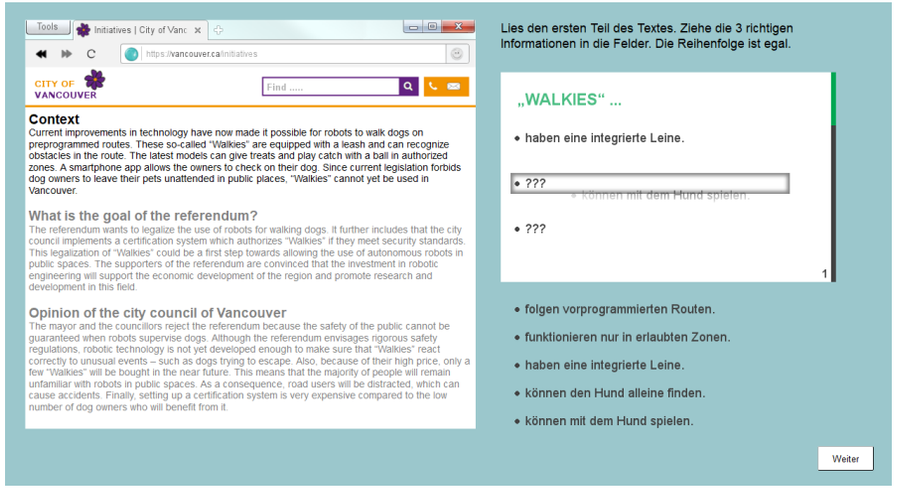

Figura 1: La prima pagina dell'esercizio A nello scenario "Walkies" (versione inglese)[5]

Figura 2: La prima pagina dell'esercizio A nello scenario "Walkies" (versione francese)[5]

5 La parte di testo rilevante è quella in grassetto. Nelle due schermate successive sono messi in evidenza altri paragrafi. A destra, si vede l'esercizio: tra le cinque opzioni proposte, le allieve e gli allievi possono sceglierne tre tra trascinare nell'elenco sopra.

Nel secondo esercizio, le allieve e gli allievi ascoltano opinioni personali di compagne e compagni fittizi in merito all'introduzione di questi robot, e rispondono a classiche domande a scelta multipla. Anche il terzo esercizio prevede domande a scelta multipla, ma in più occorre indicare dove sono state trovate le risposte. La base di testo è in questo caso una chat in un'applicazione di messaggistica in cui viene organizzato il lavoro di gruppo fittizio. Le allieve e gli allievi leggono i messaggi e rispondono alla domanda. Il messaggio in cui hanno trovato la risposta deve poi essere trascinato nell'apposito campo. → Figura 3

Nell'ultima parte dello scenario (esercizio D), le allieve e gli allievi completano infine una tavola cronologica sulla vita dell'inventrice immaginaria dei robot basandosi su estratti di un testo Wiki fittizio e di un'intervista orale in cui la donna parla di sé. → Figura 4

Gli altri scenari ruotano attorno a un'uscita al cinema, a una visita presso un centro di orientamento professionale, alla pianificazione di una festa scolastica, a un'escursione in una città per un fine settimana e a ricerche per una presentazione nel quadro delle lezioni di geografia. I formati degli esercizi e i tipi di testo sono proposti in diverse varianti: occorre per esempio trovare la risposta corretta tra i risultati forniti da un motore di ricerca oppure completare un programma orario sulla base di messaggi audio. Ogni scenario prevede almeno un compito di comprensione orale e uno di comprensione scritta, nonché un esercizio in cui le allieve e gli allievi devono mettere alla prova entrambe le capacità.

Nell'ambito dello studio principale, oltre a questi scenari sono stati impiegati altri test e questionari volti a indagare quali competenze linguistiche parziali e quali condizioni individuali influiscano sulla capacità di risolvere gli esercizi. → Tabella 2

I dati così acquisiti consentono di descrivere più dettagliatamente come le allieve e gli allievi hanno elaborato gli esercizi basati su scenari e, in generale, di formulare conclusioni empiricamente fondate sulle premesse e i processi legati alla comprensione di testi in una lingua straniera.

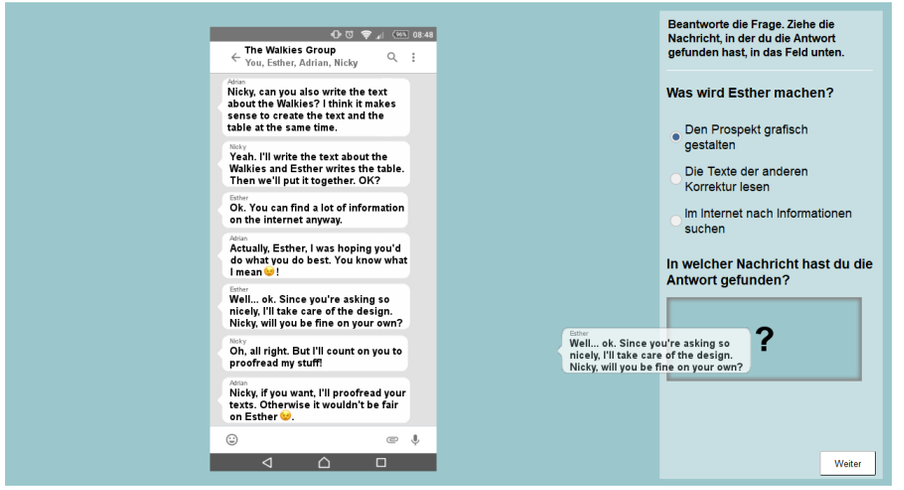

Figura 3: La seconda pagina dell'esercizio C nello scenario "Walkies" (versione inglese)

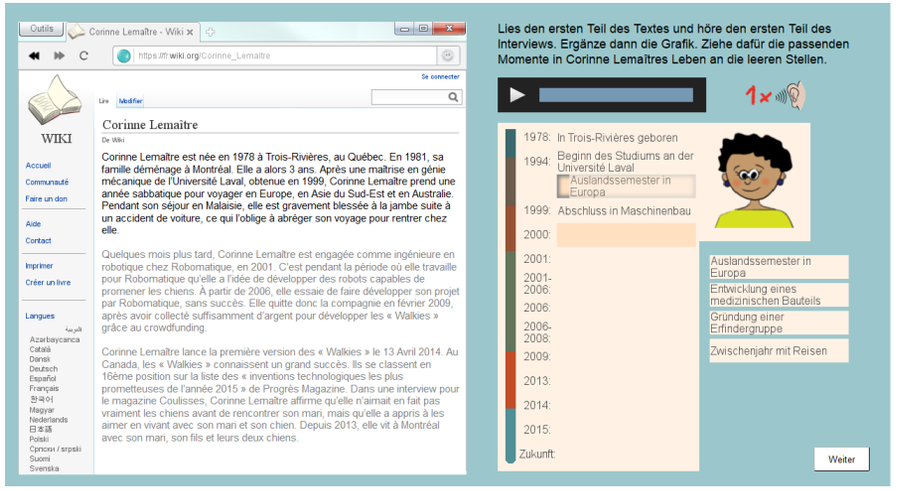

Figura 4: La prima pagina dell'esercizio D nello scenario "Walkies" (versione francese)[6]

6 Le allieve e gli allievi leggono un testo Wiki e ascoltano un'intervista dal contenuto simile, e in seguito trascinano nella cronologia a destra due delle quattro opzioni proposte.

Comprensione orale e scritta	6 scenari, ciascuno con 4 esercizi	40 item di comprensione scritta 32 item di comprensione orale 12 item combinati
	6 esercizi di comprensione di testi con formati noti	36 item di comprensione scritta (scelta multipla, vero/falso) 56 item di comprensione orale (scelta multipla, completare)
Competenze linguistiche parziali	2 test lessicali	28 item (scelta multipla con parole scritte) 22 item (scelta multipla con parole dettate)
	2 test di segmentazione di parole in un testo	2 testi scritti senza spazi vuoti (mass. 326 parole) 28 item (riconoscere il numero di parole in dichiarazioni orali)
	Competenza grammaticale	24 item (valutazione della correttezza grammaticale di un'affermazione)
	Parole a vista	30 item (riconoscere una parola mostrata per un breve lasso di tempo)
Condizioni individuali	Intelligenza fluida	20 matrici di Raven (test cognitivo non verbale in cui occorre completare una sequenza di figure)
	Approccio individuale all'apprendimento di una lingua straniera	11 item sotto forma di questionario sulla motivazione 7 item sotto forma di questionario sul timore di apprendere una lingua
	Strategie per risolvere esercizi di comprensione orale e scritta	15 item sotto forma di questionario sulla comprensione scritta (risoluzione attiva di problemi, con attenzione particolare su parole e dettagli) 10 item sotto forma di questionario sulla comprensione orale (ascolto attivo concentrato, pianificazione e valutazione)
	Competenze linguistiche	10 item sotto forma di questionario (passato linguistico personale) 7 item sotto forma di questionario (utilizzo della lingua straniera a lezione)
	Altre condizioni individuali	Item sotto forma di questionario su: origine, retroscena socioeconomico, dotazione tecnica, utilizzo di dispositivi e canali digitali

Tabella 2: Panoramica dei test e dei questionari utilizzati nel progetto "Forme di valutazione innovative"

Chi ha partecipato allo studio?[7]

Prima dello studio principale, gli scenari in fase di sviluppo sono stati testati con singole allieve e singoli allievi (n=53), discussi dettagliatamente ed elaborati man mano. In seguito, nel quadro di colloqui individuali e di gruppo (n≈50) sono stati tematizzati anche gli altri testi e questionari. Tutti gli esercizi sono inoltre stati testati con intere classi (n=85) per verificare la fattibilità tecnica del rilevamento principale. Il rilevamento qualitativo principale è stato a sua volta esaminato e affinato durante questa fase con sei allieve e allievi.

Al rilevamento principale hanno partecipato in totale 631 allieve e allievi (di 39 classi). Lo studio qualitativo ne ha coinvolti trenta. Nella maggior parte delle classi, la metà circa delle allieve e degli allievi (292) ha risolto gli esercizi in inglese, gli altri (tra cui le allieve e gli allievi dello studio qualitativo) in francese. Le allieve e gli allievi provenienti dai Cantoni Berna, Zurigo, Friburgo, Lucerna e Obvaldo frequentavano la seconda e la terza secondaria (età media: 15 anni) e rappresentavano tutti i livelli di prestazione. Circa due terzi hanno partecipato allo studio a Berna, rispettivamente a Friburgo, quindi in Cantoni in cui si incomincia a imparare il francese prima dell'inglese. Il rilevamento principale è durato in totale sei lezioni ripartite su due giorni. Le allieve e gli allievi hanno svolto, per lo più autonomamente, tutti gli esercizi su laptop sotto la sorveglianza di una collaboratrice o di un collaboratore al progetto. → Figura 5

Le allieve e gli allievi partecipanti allo studio qualitativo hanno elaborato individualmente nell'arco di tre lezioni uno scenario o cinque esercizi singoli in presenza della ricercatrice, verbalizzando le loro riflessioni (metodo del pensiero ad alta voce, cfr. p.es. Bowles, 2010; Knorr & Schramm, 2012). In seguito, hanno discusso con la ricercatrice degli esercizi elaborati e delle soluzioni nel quadro di interviste *stimulated recall* (maggiori ragguagli sul metodo p.es. in Barras, 2018; Gass & Mackey, 2017). Un altro giorno, le stesse allieve e gli stessi allievi hanno elaborato in classe i test restanti (con l'eccezione degli scenari restanti) e i questionari.

7 Il team desidera cogliere l'occasione per ringraziare per il sostegno le allieve e gli allievi, nonché i rappresentanti delle scuole. Senza il loro tempo e il loro impegno, il presente studio non sarebbe stato possibile. La nostra riconoscenza va anche alle studentesse e agli studenti che in pieno inverno si sono alzati alle prime ore del mattino, hanno trasportato scatole in giro per la Svizzera e hanno seguito pazientemente il lavoro delle allieve e degli allievi.

Figura 5: Un'aula preparata per il rilevamento principale (foto: Malgorzata Barras)

Come sono stati analizzati i dati?

Durante la fase di test, si è proceduto a riassumere le interviste e a prendere nota delle informazioni rilevanti per la rielaborazione degli esercizi e per lo svolgimento dei rilevamenti. Una parte delle interviste è inoltre stata trascritta ed è a disposizione per analisi. I risultati dei test di pilotaggio sono stati estratti dai dati grezzi generati dal sistema e analizzati in modo descrittivo.

Anche i risultati dei test del rilevamento principale sono stati dapprima elaborati e classificati su una scala, nella maggior parte dei casi mediante Item Response Theory (IRT). In una fase successiva, tutti i dati quantitativi sono stati imputati. → Riquadri

Tutti i trenta verbali di pensiero ad alta voce e le interviste *stimulated recall* del rilevamento principale dello studio qualitativo sono stati trascritti per intero. Per l'analisi approfondita delle strategie delle allieve e degli allievi sono poi state selezionate venti trascrizioni sulla base di determinati criteri. I risultatit sono riportati in Barras (i.V.).

> *Tutti gli esercizi, i risultati dei test e le trenta trascrizioni (tra cui le venti codificate) sono accessibili nell'archivio di ricerca dell'Istituto di plurilinguismo e possono essere consultati per approfondire altre fattispecie.*

Item Response Theory

Nella cosiddetta *Item Response Theory* (IRT), la difficoltà degli esercizi e la competenza rilevata dei partecipanti vengono stimate 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.

Esiste tutta una serie di modelli di calcolo IRT che pongono condizioni diverse ai test e ai risultati. Piuttosto diffuso è il modello Rasch, una forma del modello logistico monoparametrico (1PL), con il quale viene stimata solo la difficoltà dell'esercizio. Con il modello logistico biparametrico (2PL), invece, per ogni singolo esercizio viene stimata anche il potere discriminante, il che è utile soprattutto per i test che comprendono esercizi di diverso formato.

Con la modellizzazione IRT, si ottengono le caratteristiche stimate degli item (in particolare la difficoltà) e i valori legati alla capacità stimata dei partecipanti da usare per un test. È altresì possibile verificare se item e persone si comportano secondo le aspettative, per esempio se gli item complessi vengono risolti solo da allieve e allievi bravi.

Imputazione

L'obiettivo di un'*imputazione* è quello di completare un set di dati incompleto per rendere possibili e più esatti le analisi statistiche. Mancano dati per esempio se un allievo non può elaborare tutto il test a causa di un appuntamento dal dentista. I dati mancanti vengono stimati mediante una procedura statistica e tutti i punti dato con una ripartizione nota degli errori di misurazione (p.es. risultati delle classificazioni sulla scala IRT) sostituiti con valori che si situano nella fascia dell'errore di misurazione. La procedura MICE[8] applicata a tale scopo utilizza i dati disponibili di una variabile e le correlazioni più o meno forti tra le variabili nel set di dati (p.es. tra il risultato del test lessicale e di quello di comprensione scritta).

I set di dati imputati sono completi e privi di errori di misurazione. Per procedere ad analisi statistiche, vengono prodotti più set di dati in cui i valori stimati divergono ogni volta leggermente l'uno dall'altro. Le analisi vengono poi svolte per ogni singolo set di dati e i risultati riassunti secondo determinate regole per formulare conclusioni sui dati.

8 Multiple imputation by chained equations

Risultati selezionati

Gli esercizi di comprensione orale e scritta si prestano in primis a un impiego nel quadro di un test computerizzato e orientato all'azione

Sulla scorta delle osservazioni effettuate durante i rilevamenti dei dati, dei riscontri qualitativi delle allieve e degli allievi, e delle analisi psicometriche dei risultati, è stato possibile constatare che i diversi esercizi di comprensione orale e scritta proposti negli scenari si prestano in primis a condurre test al computer svolti nell'aula scolastica.

Questa constatazione è legata alla struttura e al design dei test. Gli esercizi sono stati svolti dalle allieve e dagli allievi autonomamente e nei tempi previsti. L'utilizzo del computer non ha comportato problemi. Gli schermi tattili e le cuffie sono stati utilizzati senza istruzioni, e neppure la navigazione nell'ambiente dei test presentato a schermo ha posto ostacoli: praticamente tutte le allieve e tutti gli allievi sono riusciti ad avviare da soli le registrazioni dei testi orali e a svolgere gli esercizi che richiedevano di cliccare e trascinare elementi. Tutto ciò è spiegabile con l'abitudine dei giovani a utilizzare i dispositivi digitali (nel questionario, il 60% degli interpellati ha dichiarato di utilizzare almeno una volta al giorno un computer, il 90% di utilizzare quotidianamente uno smartphone), ma anche con la buona qualità del design degli esercizi. Questo è sempre stato al centro degli sforzi, al pari dello sviluppo dei contenuti degli esercizi, per il quale l'attenzione è sempre posta all'obiettivo di dare un'impressione di autenticità e di garantire la facilità d'uso senza errori.

L'analisi statistica ha dal canto suo rivelato che i risultati degli esercizi di comprensione orale e scritta[9] presentano buone caratteristiche psicometriche. A causa dei diversi formati di esercizi in uno stesso test, per il progetto "Forme di valutazione innovative" si è dimostrato essere opportuno il modello 2PL-IRT. → Riquadro IRT

Un numero contenuto di item si è distinto statisticamente (singoli item non erano adeguati al modello) e si è rivelato problematico, per lo più per motivi plausibili. Una risposta tra quelle selezionabili, per esempio, si prestava a malintesi, o un'opzione errata suonava eccessivamente giusta. In singoli casi, determinati esercizi sono stati troppo impegnativi per il gruppo mirato. La maggioranza degli esercizi soddisfaceva però i requisiti di qualità di un test di comprensione orale e scritta nell'ottica di un impiego su vasta scala (in inglese, *large scale assessment*).

9 Si intendono solo gli esercizi A-C degli scenari, quindi non l'ultimo esercizio, il quale combina comprensione orale e scritta. Quest'ultimo si è rivelato più problematico, in quanto più impegnativo sia dal punto di vista del contenuto sia da quello delle istruzioni, ed è stato spesso risolto correttamente solo da allieve e allievi particolarmente dotati.

Le analisi statistiche dimostrano che le allieve e gli allievi ricorrono in primis alle loro conoscenze linguistiche per risolvere gli esercizi di comprensione

Come menzionato, oltre agli scenari sono stati impiegati diversi altri test e questionari volti a rilevare svariati aspetti potenzialmente pertinenti per la riuscita degli esercizi. Lo scopo era di individuare le risorse alle quali le allieve e gli allievi attingono per risolvere gli esercizi degli scenari. → Tabella 2

Le analisi rivelano che le allieve e gli allievi ricorrono in primis alle loro conoscenze linguistiche e alle loro capacità cognitive generali, in particolare alle nozioni lessicali e grammaticali, ma anche all'intelligenza fluida non verbale.[10] Fattori come la motivazione, il timore e l'impiego di strategie sono invece in generale meno decisivi, a prescindere dalla forza, rispettivamente dalla debolezza delle allieve e degli allievi. Questo risultato è positivo nell'ottica della validità dei test, in quanto dimostra che gli esercizi basati su scenari testano in primo luogo le conoscenze linguistiche e solo in misura minore le competenze generali. Ciò è confermato dal fatto che i risultati degli esercizi "tradizionali" di comprensione orale e scritta corrispondono a quelli degli esercizi basati su scenari. Il ruolo giocato dall'intelligenza fluida può essere interpretato come segnale che, al di là dei requisiti linguistici, gli esercizi presentavano una certa complessità sul piano cognitivo.

Le allieve e gli allievi fanno meno caso del previsto agli scenari, ma i testi di tipo digitale sono accolti con favore

I risultati dell'analisi qualitativa e i riscontri individuali nei questionari fanno supporre che le allieve e gli allievi prestano solo marginalmente attenzione al fatto che gli esercizi sono inseriti in scenari. Nelle situazioni di test, spesso fonte di pressione a scuola, le allieve e gli allievi mostrano solo un interesse limitato per il contenuto dei testi e concentrano la loro attenzione soprattutto sull'elaborazione degli esercizi. È esemplare in tal senso la dichiarazione di "Sofia",[11] che ha partecipato al rilevamento qualitativo (libera traduzione).

Sofia:
(…) credo di non aver ancora incontrato un tema che mi interessi, non ci do molto peso quando risolvo un test, sono presa dallo stress di dover trovare le soluzioni. (…)
Ricercatrice:
Significa che svolgendo un test non ti rendi bene conto del tema trattato?
Sofia:
Sì, me ne rendo conto, ma non riesco a

10 Per l'analisi del test inglese sono stati creati modelli a equazioni strutturali e modelli di regressione a due livelli sulla base delle variabili potenzialmente rilevanti. I risultati qui presentati riguardano soltanto gli esercizi di comprensione orale e scritta, non quelli combinati alla fine di ogni scenario.

11 Pseudonimo scelto dal o dalla partecipante.

godermelo, anche se magari è un tema che mi potrebbe piacere. Sono stressata e non riesco a concentrarmi sul testo. L'unica cosa a cui penso è trovare le soluzioni. Non penso "Com'è interessante questo tema" o cose così.

I testi fittizi di tipo digitale, come chat, siti internet e podcast, hanno invece attirato molto più l'attenzione delle allieve e degli allievi, che li hanno per lo più accolti positivamente. "Billy" ha commentato:

È una bella cosa, in fondo al giorno d'oggi molto ruota attorno ai media, che sono parte integrante della nostra vita. Trovo sia una buona soluzione per le nuove generazioni, quindi i bambini, utilizzare tipi di testo come le chat, sono cose con le quali hanno a che fare tutti i giorni.

È stato pure osservato che alcune allieve e alcuni allievi hanno apprezzato i testi soprattutto perché erano diversi da quelli abitualmente incontrati nelle lezioni di lingue straniere. La momentanea motivazione è quindi stata data dal fattore della novità, come spiega "Omega":

Sì, no, insomma all'inizio è stato motivante, ma poi capisci che si tratta comunque di un test e pensi che, sì, sarà anche divertente, ma ora è meglio concentrarsi.

Conclusione

Queste osservazioni, le esperienze pratiche con i test, i risultati delle prime analisi statistiche e quelli dello studio qualitativo (Barras, i.V.) danno un'idea del potenziale e della complessità del progetto "Forme di valutazione innovative" e della ricerca sui test di lingua in generale. L'archivio di ricerca del Centro di competenza per il plurilinguismo contiene ora una buona quantità di risultati empirici per indagare i processi di risoluzione dei test di allieve e allievi in due lingue straniere. Possibili interrogativi sono i seguenti: come percepiscono le allieve e gli allievi gli esercizi e i testi pressoché autentici in una situazione di test? Quale importanza rivestono le conoscenze lessicali per la comprensione della lettura in una lingua straniera? Come mai determinati esercizi di comprensione orale sono più difficili di altri? Che cosa pensano le allieve e gli allievi delle consegne nella loro lingua scolastica? Cambia qualcosa se la prima lingua straniera appresa è il francese o l'inglese?

Dalle nostre analisi emerge che l'impiego di scenari offre numerose possibilità anche a livelli più bassi per simulare un utilizzo realistico della lingua nelle lezioni di lingue straniere.

Il potenziale del quadro strutturale *Reading for Understanding* con formati di esercizio chiusi e un breve intervento in classe, come si è fatto in questo progetto, non è ancora esaurito: sarebbe interessante includere in uno strumento di valutazione basato su scenari anche l'utilizzo produttivo e interattivo delle lingue (ossia parlare e scrivere). Sarebbe altresì auspicabile, nell'ottica della promozione dell'apprendimento, armonizzare meglio valutazioni e lezioni. Entrambe le cose sono inerenti al quadro strutturale *Reading for Understanding* e varrebbe la pena studiarle in una tappa futura "Forme di valutazione innovative".

Dove trovo maggiori informazioni?

Sul sito del Centro di competenza per il plurilinguismo si trovano diversi manifesti e documenti PDF di relazioni presentate durante convegni specialistici: https://tinyurl.com/268uewnc.

Assessing foreign language skills in near-authentic settings

Scenario-based test tasks on the computer – an in-depth study

Katharina Karges, Peter Lenz, Thomas Aeppli, Malgorzata Barras

Overview

With the invention of computers and the emergence of the Internet, but at the latest since the advent of tablets and smartphones, listening and reading comprehension has taken on new dimensions. What was once linear text – whether in a book or on a cassette – has evolved into multimedia hypertexts with links and embedded media: images, sound, videos, and interactive graphics. In addition, voice messages in messaging apps as well as tweets and podcasts are increasingly blurring the boundaries between spoken and written language. These developments have long since become a fixed component in students' every-day life, but they also impact how foreign languages are taught and thus how communicative competence is assessed. Indeed, the increasing digitisation of communication also means the possibilities for assessing learners' listening and reading skills have expanded: in addition to incorporating new digital listening and reading text types into computer-based tests, new task formats have also become feasible. It is important that these new test tasks are researched and tested before being recommended for general use.

These considerations served as a starting point for the research project "Innovative forms of assessment" (IFB, in German: Innovative Formen der Beurteilung), which was conducted at the Research Centre on Multilingualism (RCM) from 2016 to 2019. In the IFB project, test tasks were designed and embedded in so-called scenarios: narrative frameworks that incorporate highly realistic task questions as well as the use of digital text types to simulate near-authentic language use in the foreign language and thus motivate test takers to mobilise their entire linguistic repertoire. For example, in one scenario, students were asked to work with written and audio texts to plan a school outing.

In consideration of task development for the second Swiss-wide assessment of basic competencies conducted by the Swiss Conference of Cantonal Ministers of Education (EDK)[1] – in which IOM participated at a later stage – reading and listening comprehension tasks for foreign languages were developed in the IFB project; these tasks used closed,[2] computer-based tasks to assess language skills at the A2/B1 level as defined by the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001). By incorporating the tasks in scenarios, however, the IFB project went beyond the standardised approach later adapted in the national assessment.

1 The national assessment is a part of the Swiss-wide education monitoring scheme. In 2016, maths skills were measured for the first time and, in 2017, skills in the language of schooling and the first foreign language. Mapping the students' listening and reading comprehension skills in the language of schooling and in both foreign languages taught at school at the end of compulsory schooling was originally planned for 2020. IOM was involved in the task development; due to Covid-19 measures, however, the main survey could not be conducted and has been postponed to 2023. Details at https://uegk-schweiz.ch/ (information in German, French and Italian).

2 In the closed task formats, the test takers select their answers from at least two possible answers.

The tasks developed in this context were examined in a validation study using a mixed-methods design. In the *qualitative* study, introspective methods were used (think aloud protocols and stimulated recall interviews) with the aim of understanding how individual learners experience and complete the tasks.[3] In the *quantitative* study, tests on partial competences (e.g. vocabulary, speed of processing language and grammar) and questionnaires were used in addition to the scenario-based tasks.

All tasks were used in versions for both foreign languages (French and English) at schools in German-speaking Switzerland, with the focus of the qualitative study placed on French. Data were collected during several phases in 2017 and the early months of 2018.

An overarching, practice-oriented goal of the project was to explore the extent to which it is feasible and desirable to use scenario-based tasks to assess foreign language comprehension skills in a target group with a low level of language skills. The validation study therefore aimed to answer the following questions:

1. What partial competences do learners draw on when completing scenario-based tasks in the foreign languages taught at school?

2. Do learners notice and appreciate the greater degree of intended authenticity in the tasks?

3. Do the tasks fulfil psychometric requirements for large-scale assessments?

3 Large sections of the qualitative study were developed and conducted by Malgorzata Barras in the scope of her doctoral thesis (Barras, i.V.).

What is scenario-based assessment?

Scenario-based assessment forms part of the "Reading for Understanding" framework (Sabatini et al., 2014a; Sabatini & O'Reilly, 2013), which is also used in international studies, for instance, as part of PISA 2018 (OECD, 2019, p. 41). Reading for Understanding is the result of a series of research projects conducted by the US Educational Testing Service (ETS), in which findings from various disciplines on the nature of reading were compiled. The main goal was to design a model able to accommodate the different facets of reading; according to the model, text comprehension unfolds in five dimensions (Sabatini et al., 2013, p. 14ff.). → Table 1

This understanding of text comprehension gives rise to a broad definition of reading literacy, which also impacts assessment standards: in addition to tasks designed to assess the level of text comprehension, an assessment instrument should also contain tasks that can create learning opportunities, measure the results of learning, monitor the integration of new and existing knowledge, or gauge the credibility of sources (concrete examples for such tasks are found in O'Reilly et al., 2014; Sabatini et al., 2014b). In the IFB project, this expanded understanding of language comprehension was applied to the assessment of foreign language skills. Because oral and written language is increasingly blended in digital media, the IFB project designed tasks for both listening and reading comprehension.

Dimensions of text comprehension	Examples
Visual elements and structure of printed texts	Letters, words, punctuation, function of elements like paragraphs or headers as well as graphics, hyperlinks, emojis etc.
Verbal elements of a language	Vocabulary, morphology, syntax etc.
Discourse and text genres	Text structure, inter- and intratextual references, incorporation of general knowledge
Conceptual content and meaning	Integration of existing and new knowledge, critical evaluation of content
Social content and meaning	Engaging with the ideas of the author or protagonist in a text, appreciation for how a text was written, understanding the social role(s) of texts and characters

Table 1: The five dimensions of text comprehension adapted from Sabatini et al. (2013)

Which tasks were used?

The main element of the IFB project are six scenarios consisting of four separate tasks with related content (tasks A-D) at the A2/B1 level that can be completed independently on the computer. A French and English version with identical content were developed for each scenario. All scenario descriptions, instructions and task questions were written in the language of schooling so that students with low reading comprehension skills in the foreign language would not be disadvantaged (for a detailed rationale, cf. Barras et al., 2016).

The project team developed the scenarios using iterative processes. After the topics and task formats were determined, the scenarios were developed over time by several members of the team tested on individual learners, and generally revised many times. All reading texts were written by the project team based on real-world texts. The listening texts were also designed specifically for the scenarios, although the speakers in the recording studio were given only general points of reference for the conversation. To achieve a higher level of authenticity of spoken language, the actual choice of words and details in the texts was left to the discretion of the speakers. The French and English versions for all scenarios were developed in tandem; therefore, one version is not really a translation of the other.

The following section presents the scenario "Walkies", in which learners were required to explore a fictive vote on using robots for walking the dog. The selected scenario is a good example of the range of text types and task formats. It comprises one listening (B) and two reading comprehension tasks (A and C) as well as a task that requires learners to use both skills (D). The texts in this scenario are somewhat longer than average; the difficulty of tasks C and D is situated in the upper average range.

The scenario opens by informing learners about the context of the tasks: "You are on a language stay in Vancouver, Canada,[4] and are going to school there. In the subject 'social studies', you are discussing a current political issue: Should robots be allowed to take dogs on walks?" The learners are then asked to complete the first task. They read an informative text on a simulated Internet page and select the correct information from several suggestions. → Figures 1 and 2

4 Québec was the city chosen in the French version.

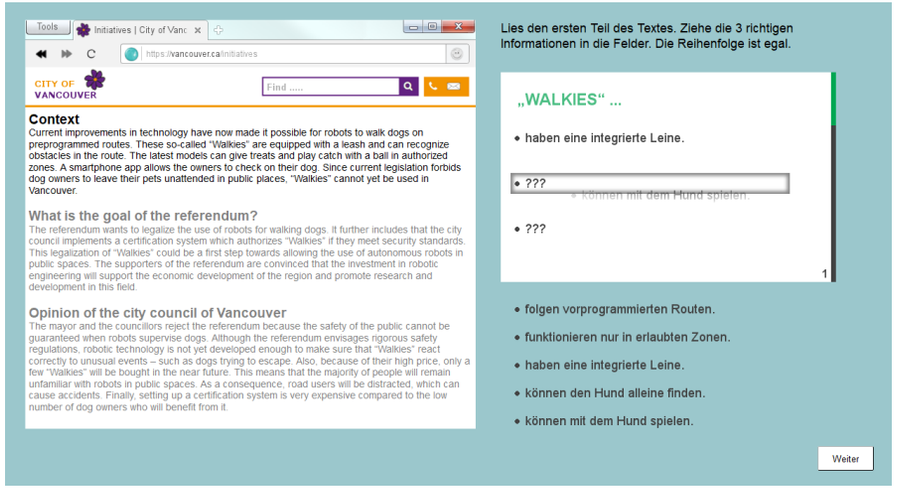

Figure 1: The first page of task A in the "Walkies" scenario (English version)[5]

Figure 2: The first page of task A in the "Walkies" scenario (French version)[5]

5 The part of the text relevant for the task on this screen page is highlighted in a darker font. On each of the next two pages, a different section is highlighted. The task is on the right-hand side of the screen: the learners must select three correct answers from the five options and drag and drop them to the list above the answer options.

In the second task, learners listen to the opinions their fictive classmates have on using robots; they are then asked to answer classic multiple-choice questions. In the third task, multiple choice questions are used, but the learners are also asked to state where they found their answers. The input text is a group chat in a messaging app used to plan a fictive group project. Learners read the messages in the chat and answer the question by dragging and dropping the message they believe has the correct answer into the designated field. → Figure 3

In the final section of the scenario (task D), learners are asked to complete a timeline for the life of the fictive inventor of the "Walkies" based on excerpts from a simulated Wiki entry and an interview with the inventor in which she talks about herself. → Figure 4

The other scenarios deal with the following topics: a visit to the cinema, a visit to a career centre, planning a school party, a weekend excursion to a city, and research for a presentation in geography class. There are different variants for the described task formats and text types, for instance, an appropriate answer must be chosen from the results generated by a search machine, or students are asked to fill in missing information for a schedule on the basis of audio messages. Each scenario contains at least one listening and one reading comprehension task as well as a task requiring learners to use both skills.

In addition to these scenarios, other tests and questionnaires were used in the main study to identify which linguistic partial competences and additional individual characteristics impact how students complete the tasks. → Table 2

The data thus generated make it possible to provide a more detailed description of how learners approach the scenario-based tasks and – more generally – to make empirical statements on both prerequisites for and processes in understanding texts in a foreign language.

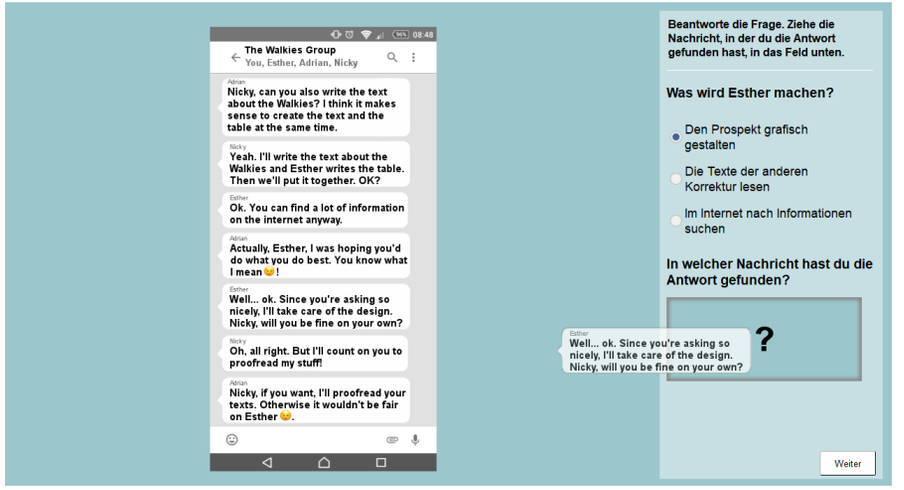

Figure 3: The second page of task C in the “Walkies” scenario (English version)

Figure 4: The first page of task D in the “Walkies” scenario (French version)[6]

6 The learners read a Wiki entry and listen to an interview with related content. They then drag and drop two of the four options given on the far right into the timeline.

Listening and reading comprehension	6 scenarios with 4 tasks each	40 items on reading comprehension 32 items on listening comprehension 12 items on combined listening and reading comprehension
	6 tasks on text comprehension with known formats	36 items on reading comprehension (multiple choice, true/false) 56 items on listening comprehension (multiple choice, fill in the blanks)
Linguistic partial competences	2 vocabulary tests	28 items (multiple choice with written words) 22 items (multiple choice with spoken words)
	2 tests on word segmentation in a text	2 written texts without spaces (max. 326 words) 28 items (recognising the number of words in spoken utterances)
	Grammar competence	24 items (assessing the grammaticality of a statement)
	Sight vocabulary	30 items (recognising a word that appears briefly on a screen)
Individual characteristics	Fluid intelligence	20 Raven's Matrices (a nonverbal intelligence test in which a series of patterns with missing elements must be completed)
	Personal attitude towards learning a foreign language	11 questionnaire items on motivation 7 questionnaire items on language learning anxiety
	Strategies used to complete listening and reading comprehension tasks	15 questionnaire items on reading comprehension (active problem solving, focus on words and details) 10 questionnaire items on listening comprehension (active, concentrated listening; planning and evaluating)
	Linguistic competences	10 questionnaire items (personal language background) 7 questionnaire items (use of the foreign language in the classroom)
	Other individual characteristics	Questionnaire items on linguistic heritage, socio-economic background, technical equipment, use of digital devices and channels

Table 2: Overview of tests and questionnaires used in the IFB project

Who participated in the study?[7]

Before the main study began, the scenarios being developed were first discussed in detail with individual learners (n=53) and repeatedly revised. Afterwards, the other tests and questionnaires were the topic of one-on-one discussions as well as group interviews with learners (n≈50). In addition, all tasks were tested in entire school classes (n=85) to assess the technical feasibility of the main survey. During this phase, the qualitative main survey was also tested and improved with six learners.

A total of 631 students from 39 classes took part in the main survey; 30 of these students participated in the qualitative study. In most classes, roughly half the learners completed the tasks in English (292 students in all), while the others (including the students participating in the qualitative study) worked on the French version of the test. The students from the cantons of Bern, Zurich, Fribourg, Lucerne, and Obwalden were in their second and third year of secondary school attending classes of all levels (average age: 15 years). About two-thirds of the participating students were learning in Bern or Fribourg – cantons, where French is taught before English. In total, the main survey was held during six lessons distributed over two days. As a rule, the students completed all tasks on their own using laptops provided by the project team and under the supervision of one member of the project team. → Figure 5

Students who took part in the qualitative study worked alone with the researcher on a scenario or five individual scenario tasks during three lessons; first, they verbalised their thoughts (thinking-aloud method, cf. e.g. Bowles, 2010; Knorr & Schramm, 2012). Afterwards, they discussed the completed tasks in a stimulated recall interview in which they discussed all tasks and their answers with the researcher (for more on this method, c.f. e.g. Barras, 2018; Gass & Mackey, 2017). On the other day, these learners worked on the remaining tests items (but not the other scenarios) and questionnaires with the entire school class.

7 The research team would like to take this opportunity to thank all students and school representatives for their support. Without their time and commitment, this study would not have been possible in its current form. We also offer our heartfelt thanks to our student assistants who, in the middle of winter, got up early and transported materials across half Switzerland and patiently watched students work.

Figure 5: A classroom prepared for the main survey (photo: Malgorzata Barras)

How were the data analysed?

During the pre-piloting phases, the interviews were summarised and key information for revising the tasks or collecting the data was recorded. Some of the interviews from the testing phase were also transcribed and are available for future analyses. The test results from the pilot phase were extracted from the raw data generated by the test system and interpreted using descriptive methods.

In addition, the test results from the main survey were prepared and then scaled, mainly using item response theory (IRT). In a further step, all quantitative data were imputed. → Boxes

All 30 think-aloud protocols and stimulated-recall interviews from the main survey of the qualitative study were transcribed in full. On the basis of set criteria, 20 of these transcripts were selected for the in-depth analysis of strategies learners use to complete the test tasks. The results are described in Barras (i.V.).

In the spirit of open research data, all tasks, test results and all 30 transcripts (including the 20 encoded transcripts) can be accessed via the research data archives of the Research Centre on Multilingualism (RCM) and used for follow-up lines of inquiry.

Item Response Theory

In item response theory (IRT), the difficulty of test tasks (items) 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.

IRT encompasses a broad range of mathematical models that assume different conditions for tests and test results. A common type is the Rasch model, a one-parameter logistic model (1PL model) in which "only" task difficulty is estimated. The two-parameter model (2PL model) is used to estimate item discrimination, which is particularly useful in tests that have various task formats. IRT modelling is used to estimate item characteristics (in particular item difficulty) as well as test taker abilities. In addition, it is possible to check whether items and people behave as the model expects, for instance, that only good students are able to solve difficult items.

Imputation

The aim of imputation is to complete an incomplete dataset in order to enable or improve statistical analyses. Missing data arise when, for instance, a student is unable to finish an entire test due to a dentist appointment. Through imputation, the missing data are estimated on the basis of a statistical procedure, and all data points with a known measurement error distribution (e.g. results from an IRT scaling) are replaced by values that fall within the margin of error. The MICE[8] approach selected for this project uses all available data on a variable as well as the more or less strong correlations between the variables in the dataset (e.g. the correlation between the results of a vocabulary and a reading comprehension test).

The datasets created using imputation are complete and measurement-error free. To conduct statistical analyses, multiple datasets in which the estimated values differ slightly are generated. The analyses are then performed on each individual dataset and the results summarised according to predefined rules, thus enabling the data to be interpreted.

8 Multiple imputation by chained equations

Selected findings

The listening and reading comprehension tasks developed are in general suitable for use in computer-based, action-oriented assessments

The observations made during data collection as well as the qualitative feedback from learners and psychometric analyses of the test results all indicate that the different reading and listening comprehension tasks in the scenarios are in general suitable for computer-based foreign language assessments in the classroom setting.

This observation is related to both the structure and the design of the tests. Students were able to complete the tasks independently and in the given timeframe. In addition, using computers posed no problems: the students worked with the touchscreens and touchpads on the laptops and used the headsets without requiring instructions. The on-screen elements also caused no difficulties – nearly all students succeeded in navigating the test environment, starting the listening texts on their own, and using drag and drop to answer questions. Although this can be partly explained by the students' familiarity with digital devices (in the questionnaire, 60% of all learners said they used a computer at least once a day and 90% said they used a smartphone every day), it is also related to the carefully designed test tasks. Indeed, task design was, in addition to task content, repeatedly a focus during the trial phase, with tasks being structured mainly according to principles of authentic language use and simple, error-free usability.

The statistical analysis also revealed that the results from the listening and reading comprehension tasks[9] have good psychometric properties. Due to the different task formats in the same test in the IFB project, a 2PL-IRT model proved most effective. → Box IRT

A small number of tasks had statistical anomalies (i.e. individual items had a poor model fit), for which there was generally a plausible explanation. Examples of this include answer options that were unclear or misleading, or a wrong answer in a multiple-choice question that was made too attractive. In individual cases, it became apparent that some tasks were too difficult for the target group. Most of the tasks, however, met the quality demands of a listening and reading comprehension test for use in a large-scale assessment.

9 The discussion here pertains solely to tasks A-C in the scenarios – and not the final task D – in which listening and reading comprehension elements were combined in a complex task question. These tasks proved to be more problematic: they were more complicated in terms of content as well as in terms of the task's assignment, and often only very advanced learners in the target group solved them correctly.

The statistical analyses reveal that learners mainly apply their language skills to complete comprehension tasks

In addition to the scenarios, various other tests and questionnaires were used to gather data on other aspects that are potentially relevant for success in completing scenario-based tasks. Here, the aim was to identify the resources students mobilise when working on test tasks. → Table 2

The analyses show that the students primarily draw on their language skills and their general cognitive abilities when working on the scenario-based tasks – in general vocabulary and grammar knowledge, but also nonverbal fluid intelligence.[10] By contrast, aspects such as language learning motivation, foreign language anxiety and the use of test strategies were less decisive. These observations apply to both strong and weak learners. This result is encouraging with respect to the validity of test tasks, as it demonstrates that the scenario-based tasks, too, primarily target specific language skills and test general competencies only to a lesser extent. This is corroborated by the fact that the results of the "traditional" listening and reading comprehension tasks correlate strongly with the results from the scenario-based tasks. The role played by fluid intelligence indicates that the tasks have a certain degree of cognitive complexity beyond their language-specific demands.

Students take less note of the scenarios than expected but appreciate digital text types

Results from the qualitative study as well as student statements in the questionnaire suggest that learners took little note of the fact that the tasks were embedded in scenarios. Indeed, it became apparent that – in testing situations at school, with the associated pressure to perform – students are not particularly interested in the content of the listening and reading texts and instead focus their attention on answering the questions. This disconnect becomes particularly noticeable in statements made by "Sofia",[11] who participated in the qualitative part of the study:

Sofia:
(…) I think, well, there never was a topic that would interest me on a test, because I don't really think about the topic. Then it's just more, like, the pressure (…) that I just have to find the answers. (…)
Researcher:
Er, that means that you don't really notice the topic on the, on the test?
Sofia:
Well, I do. But I can't, you know, really enjoy it, even if it were a topic that I'm interested in. I'm stressed and I (…) can't, you know, concentrate on the text, (…) so, the only thing I think about

10 To analyse the English test, structural equation models and two-level regression models were created based on potentially relevant variables. The results reported here pertain solely to the listening and reading comprehension tasks in the scenarios and not to the combined tasks at the end of each scenario.

11 Pseudonym chosen by the student.

is that I have to find the answers. And then I don't think, "oh, cool" or anything.

However, students did notice the simulated digital text types like smartphone chats, websites, and podcasts much more frequently, with most students giving a favourable assessment. "Billy" had the following thoughts:

So, I think it's quite good because (…) today a lot of attention is paid to the media and the media are a fixed part of life. I think now you have to, for the people that come next, I mean for children (…) I think it's good (…) that you maybe make it a little more like real life for them, in a chat, (…) because then it's like an everyday situation.

It was also observed that several learners reacted positively because these types of texts are rarely used in their foreign language classes. In such cases, the aspect of "new" was motivating for a short time. "Omega" had the following to say:

Yes, um, no (…) I mean, at first, it's motivating, and then at some point in time you remember that it's a test. And then you think: "Yes, that's clever, but now I have to concentrate again."

Conclusion

The observations discussed above, the practical experiences from conducting the tests, the results from the first statistical analyses, and the results from the qualitative study (Barras, i.V.) offer a general impression of both the potential and the complexity of the IFB project and of conducting research on language assessment in general. In the RCM research data archives, there is now an abundance of empirical evidence to enable further study of students' task-solving processes in two foreign languages at school. Examples of possible research questions include the following: How do students perceive the near-authentic task questions and texts in a test situation? What significance does vocabulary have for reading comprehension in the foreign language? Why exactly are certain listening comprehension tasks more difficult than others? What do students think about having the task questions formulated in the language of schooling? Does it matter whether French or English is the first foreign language taught at school?

The analyses performed in this study indicate that using scenarios even for lower language levels provide numerous possibilities to achieve a more authentic language use in foreign language education.

As a final observation: the potential of the Reading for Understanding framework through closed-question task formats and a brief intervention in the classroom – as carried out in the IFB project – has not yet been exhausted. It would be interesting to also incorporate productive and interactive language use (i.e. speaking and writing) in a scenario-based assessment tool. And in the spirit of fostering learning, it would be desirable to better align assessments and teaching. Both lines of inquiry are situated in the Reading for Understanding concept and pursuing them would be a worthwhile continuation of the project "Innovative forms of assessment".

Further information

Various posters and presentations (in PDF format) from conferences can be accessed via the website of the Research Centre on Multilingualism:
https://tinyurl.com/268uewnc.

Bibliographie
Bibliografia
Bibliography

Barras, M. (i.V.). *Strategien beim Lösen rezeptiver Sprachtestaufgaben* (Dissertation). Freiburg, CH: Universität Freiburg.

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* (Bd. 10). 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.

Bowles, M. A. (2010). *The think-aloud controversy in second language research*. New York: Routledge.

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

Consiglio d’Europa (2002). *Quadro comune europeo di riferimento per le lingue: apprendimento insegnamento valutazione*. Milano: La Nuova Italia – Oxford.

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

Europarat. (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*. (J. Quetz & G. Schneider, Übers.). Berlin: Langenscheidt.

Gass, S. M. & Mackey, A. (2017). *Stimulated Recall methodology in applied linguistics and L2 research* (2. Aufl.). New York & London: Routledge.

Knorr, P. & Schramm, K. (2012). Datenerhebung durch Lautes Denken und Lautes Erinnern in der fremdsprachendidaktischen Empirie. Grundlagenbeitrag. In S. Doff (Hrsg.), *Fremdsprachenunterricht empirisch erforschen. Grundlagen – Methoden – Anwendung*). Tübingen: Narr Francke Attempto, 184-201.

OECD. (2019). *PISA 2018 assessment and analytical framework*. Paris: OECD Publishing.

O’Reilly, T., Weeks, J., Sabatini, J., Halderman, L. & Steinberg, J. (2014). Designing reading comprehension assessments for reading interventions: How a theoretically motivated assessment can serve as an outcome measure. *Educational Psychology Review*, *26*(3), 403-424.

Sabatini, J. P. & O’Reilly, T. (2013). Rationale for a new generation of reading comprehension assessments. In B. Miller, P. McCardle & R. Long (Hrsg.), *Teaching reading and writing: improving instruction and student achievement*. Baltimore: Brookes Publishing, 100-111.

Sabatini, J. P., O’Reilly, T. & Deane, P. (2013). *Preliminary reading literacy assessment framework: foundation and rationale for assessment and system design* (Research Report No. RR-13-30). ETS. http://www.ets.org/research/policy_research_reports/publications/report/2013/jrmh

Sabatini, J. P., O’Reilly, T., Halderman, L. & Bruce, K. (2014a). Broadening the scope of reading comprehension using scenario-based assessments: preliminary findings and challenges. *L’Année psychologique*, *114*(4), 693-723.

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