frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 16 October 2025
DOI 10.3389/fcomm.2025.1537384

Check for updates

OPEN ACCESS

EDITED BY
Jacopo Torregrossa,
Goethe University Frankfurt, Germany

REVIEWED BY
Despina Papadopoulou,
Aristotle University of Thessaloniki, Greece
Ulrike Jessner-Schmid,
University of Innsbruck, Austria

*CORRESPONDENCE
Christoph Gantefort
✉ christoph.gantefort@mercator.uni-koeln.de

RECEIVED 30 November 2024
ACCEPTED 28 August 2025
PUBLISHED 16 October 2025

CITATION
Gantefort C, Barberio T, Busch L and Goltsev E (2025) Tackling the challenge of construct validity in assessing newly arrived learners' receptive skills: investigations on a multilingual online-based diagnostic tool. *Front. Commun.* 10:1537384.
doi: 10.3389/fcomm.2025.1537384

COPYRIGHT
© 2025 Gantefort, Barberio, Busch and Goltsev. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Tackling the challenge of construct validity in assessing newly arrived learners' receptive skills: investigations on a multilingual online-based diagnostic tool

Christoph Gantefort[1]*, Teresa Barberio[2], Lukas Busch[1] and Evghenia Goltsev[3]

[1]Mercator Institute for Literacy and Language Education, University of Cologne, Cologne, Germany, [2]Department of German Language and Literature, University of Münster, Münster, Germany, [3]Institute for Germanistics, University of Koblenz, Koblenz, Germany

Introduction: This study presents the theoretical rationale and first empirical findings on Allrad-M, a newly developed multilingual procedure for the assessment of receptive linguistic skills. The tool is designed to enhance test fairness and construct validity when evaluating newly arrived learners' listening and reading comprehension. Unlike monolingual diagnostic instruments, Allrad-M enables learners to switch flexibly between German, Ukrainian, Russian, and English, allowing for a more accurate assessment of comprehension skills regardless of their proficiency in German as a second language.

Methods: The exploratory study is based on two data sources. First, ten screen recordings of learners' interactions with Allrad-Mwere analyzed to examine how participants used their linguistic repertoires when processing texts and responding to test items. Second, a semi-structured interview was conducted with a teacher who implemented the tool in classroom practice, providing professional insights into its diagnostic potential.

Results: The analysis of the screen recordings shows that learners actively mobilized their multilingual resources while working with the tool. Language choices were shaped by context (reception vs. assessment) and modality (reading vs. listening). Case analyses further highlight individual strategies in the use of multiple languages. The teacher interview indicates that Allrad-M reveals aspects of learners' potential that often remain undetected in monolingual assessments.

Discussion: The findings suggest that Allrad-M can strengthen formative assessment practices for newly arrived learners by recognizing multilingual repertoires as resources rather than obstacles. Future development should focus on integrating additional languages and providing targeted teacher training to support the implementation of multilingual diagnostic tools.

KEYWORDS

newly arrived learners, multilingualism, assessment, construct validity, translanguaging

1 Introduction

Given the monolingual norms in schools and lessons, assessing the language and subject-specific skills of multilingual learners poses significant challenges. Preventing these students from using their entire linguistic repertoire in assessments risks underestimating their abilities. From a pedagogical perspective, the Translanguaging approach underscores the need to distinguish between skills in a named language and general linguistic skills (Otheguy et al., 2015). This distinction is in particular relevant for newly arrived learners who acquire the language of instruction as a second language and therefore initially have limited proficiency. Nevertheless, their broader communicative abilities may remain hidden if assessments are restricted to the language of instruction. Hence, it is essential to provide teachers with diagnostic tools that evaluate learners' general language skills independently of their limitations in specific languages. These tools can help educators recognize learners' learning potential and design translingual learning environments aligned with their zone of proximal development.

This study introduces a diagnostic tool designed to meet the above-mentioned criteria. Allrad-M (Allgemeine rezeptive sprachliche Fähigkeiten diagnostizieren – Mehrsprachig/Diagnosing General Receptive Language Skills - Multilingually) is an online-based diagnostic tool that enables newly arrived learners with a Ukrainian/Russian language background to demonstrate their receptive skills using their entire linguistic repertoire. To this purpose, learners engage with two subtitled audio texts, allowing them to fluidly switch between German, Ukrainian, Russian, and English. After the listening phase, they complete test-items based on Germany's competence-level model for listening comprehension (Kultusministerkonferenz, 2014). This approach enables a criterion-referenced evaluation of their general comprehension ability, with the option to use all four languages when navigating questions and selecting answers.

First, we illustrate the theoretical foundation of Allrad-M and outline the details of this diagnostic tool. The exploratory and preliminary findings are then examined from two perspectives: First we analyze screen recordings to investigate how learners use their linguistic repertoire to construct meaning, and second insights from a guided interview with a teacher provide a practical perspective on the pedagogical applicability in school contexts.

1.1 Theoretical background

Holistic approaches to multilingualism and multilingual learners argue that languages are not sharply defined, enumerable units represented independently at a cognitive level (Herdina and Jessner, 2002; Krulatz et al., 2022 for an overview). Concepts such as multicompetence (Cook, 2007) and the Dynamic Model of Multilingualism (Herdina and Jessner, 2002) exemplify a broader *multilingual turn*, which marks a shift in perspective from monolingual to multilingual conceptualizations of language and language education (May, 2014). Within this framework, the term Translanguaging has become particularly influential and can be understood as an umbrella term encompassing both multilingual practices and pedagogical strategies (Cenoz and Gorter, 2022).

Otheguy et al. (2015) emphasize that, from a linguistic perspective (i.e., in terms of lexicon and grammatical structure), languages are not enumerable and distinctive entities. This has pedagogical implications: If multilingual learners cannot use their entire linguistic repertoire, but are instead constrained by rigid language separations, they are disadvantaged compared to "monolingual learners." Consequently, "pedagogical translanguaging" approaches have been developed (García et al., 2017; Cenoz and Gorter, 2017), expanding traditional models of multilingual education. With regard to assessment and testing as crucial dimensions of pedagogical work, translanguaging introduces key implications. To validly assess general linguistic skills, learners should be able to draw on their entire linguistic repertoire when completing tasks. As Otheguy et al. (2015) assert:

> […] testing the proficiency of children *in a language* must be kept separate from testing their proficiency *in language*. Assessing the size, development, flexibility, richness, complexity, and agility of deployment of an idiolect must be kept separate from testing the ability to recognize and adhere to politically defined boundaries in the deployment of the idiolect (Otheguy et al., 2015, p. 299).

This principle aligns with the test-theoretical concept of bias. Van de Vijver and Tanzer (2004) distinguish between three categories of bias in psychological testing: construct bias, method bias, and item bias. Construct bias can occur when the construct being operationalized is not universal across populations, and the test instrument fails to account for differences that may be rooted in cultural background. Item bias refers to situations in which individual items on a measurement scale have a different probability of being solved correctly by two groups, although both groups have identical proficiency levels ("differential item functioning"). With regard to multilingualism, we here focus on the relevance of method bias. Method bias arises when distortions in test results are caused by the characteristics of the testing instruments or the conditions under which they are administered. When learners are assessed on general linguistic abilities, but are required to respond exclusively in a codified standard language, their full linguistic competence is not represented. This restriction can lead to an underestimation of their true level of competence (García et al., 2017).

To illustrate method bias with an example: In a standardized and norm-based test of German language skills, a subtest may assess conceptual knowledge. Test takers might be asked to identify the overarching category for four shown objects (e.g., "plum," "pineapple," "banana," and "orange") by naming the superordinate concept ("fruit"). This task requires the ability to categorize concepts hierarchically, a skill that is considered language-independent and cross-linguistic. However, if the test is conducted in German and scoring requires verbalization of the category in German, a learner who has the concept of "fruit" but not the expression in German ("Obst") is penalized. This represents a case of method bias, as defined by van de Vijver and Tanzer (2004), since the test does not account for conceptual scoring (Pearson et al., 1993). As a result, emerging multilinguals' ability to form categories cannot be assessed with construct validity.

In the following, we argue that higher-order receptive skills are primarily based on cognitive processes that are not language-specific but rather cross-linguistic. To illustrate this point, we turn to reading comprehension, where cognitive processes can be divided into lower-order and higher-order processes (Lenhard, 2019). Lower-order processes, such as decoding and constructing sentence-level coherence, are largely bound to being proficient in a named language, as they depend on phoneme-grapheme correspondences and syntactic structures based on the specific conventions of that language. The cognitive processes of decoding and constructing local coherence (i.e., making sense on the sentence level) are processed very close to the linguistic surface structures (Rosebrock et al., 2011) which indicate, for instance, which constituent of an utterance functions as the subject and which as the object. However, these linguistic surface structures play a much smaller role in higher-level cognitive processes, such as activating and using text-type knowledge and reading strategies, drawing inferences and constructing global coherence (Rosebrock et al., 2011). Empirical evidence supporting the idea that higher-level linguistic skills, as described by Marx (2020), are "transversal," i.e., not dependent on linguistic surface structures, can be found, for example, in the studies by Vanhove and Berthele (2018) and Gebauer et al. (2013). These studies demonstrated significant longitudinal cross-linguistic effects in reading comprehension across different languages. Similarly, Marx and Steinhoff (2021) observed that a text-type schema taught in German (in the given case: an abstract standard solution for describing persons) was also evident in schoolchildrens' written texts in Turkish. In line with these findings, Barberio (2021) identified similar patterns in the written texts of bilingual Italian-German students, further reinforcing the cross-linguistic applicability of text-type schemata.

Moreover, according to the "Simple View of Reading" (Gough and Tunmer, 1986), there is a strong connection between listening comprehension and reading comprehension. Reading comprehension is thereby modeled as the arithmetic product of decoding ability (a lower-order skill) and listening comprehension (which requires both lower-order and higher-order skills). While research has shown that this relationship is not strictly multiplicative (Knoepke et al., 2013), listening comprehension nonetheless explains a substantial amount of variance in reading comprehension when decoding ability is controlled for (Tunmer and Chapman, 2012). Crucially, within the framework of the Simple View of Reading, reading comprehension cannot exceed if decoding ability is zero. Analogously, the measured comprehension performance of newly arrived learners will remain close to zero if assessments rely exclusively on linguistic surface structures in a language that learners are only beginning to acquire. In other words, just as individuals who have not been instructed in reading and writing cannot construct meaning based on written texts (though they might from spoken language), emerging bilinguals are unable to fully realize their potential for meaning-making when they are forced to process information exclusively in the language of instruction. However, they might succeed if allowed to listen to and/or read using linguistic resources they already command. From this perspective, method-bias is likely to occur when multilingual learners' linguistic repertoires collide with monolingual testing conditions. Bias arises when given higher-order processing skills cannot be applied. In line with Otheguy et al. (2015) higher-order receptive skills represent a case of proficiency "in language" and should not be assessed based on "a language" but rather through the learners' full linguistic repertoire. In a translingual and multimodal context, text comprehension can thus be modeled as the product of the accessible portion of a learner's linguistic repertoire and their higher-order receptive skills:

Comprehension (C) = (individual repertoire (I) – fraction of the individual repertoire that cannot be employed due to monolingual testing (I′)) * higher-order receptive skills (H).

Given that C as well as I, I′ and H can attain values between 0 and 1, consider two individuals, A and B, with both a comparable repertoire of linguistic means ($I = 0.8$) necessary for the conceptual comprehension of a given text. However, Persons' A and B linguistic repertoire is not equally distributed across named languages. Furthermore, both persons do have comparable higher order receptive skills ($H = 0.8$). Person A is "monolingual" in the test language, while 60% of Person B's repertoire consists of linguistic means outside the test language. In a monolingual test setting, Person A's comprehension score is calculated as: C = (0.8–0) * 0.8 = 0.64, whereas Person B's comprehension score is calculated as: C = (0.8 – (0.6 * 0.8)) * 0.8 = 0.26. This illustrative numerical quantification of the bias to be expected in monolingual test settings (here: $\Delta = 0.38$) is intended to underline the need to operationalize general linguistic skills (as opposed to skills "in *a* language") in a multilingual design.

1.2 Multilingual testing

De Angelis (2021) distinguishes between summative, formative and diagnostic assessment. While summative assessment evaluates learning progress over a specific period of time, formative assessment is characterized by criterion-reference and process orientation. This enables educators to design learning opportunities tailored to students' zones of proximal development. In contrast, diagnostic assessment places less emphasis on the learning process and is often applied in situations such as assessing newly arrived learners transitioning from preparatory to regular classes.

In recent years, significant efforts have been made to develop both general theories and specific instruments for assessing multilingual competencies (Shohamy et al., 2017; Melo-Pfeifer and Ollivier, 2024). An early framework for multilingual testing was presented by Shohamy (2011), organizing approaches on a continuum. At one end, multilingual individuals are assessed with strict language separation, while, at the other, they are allowed to use linguistic resources from their entire repertoire in a fluid and integrated manner. Seed (2020) identifies four categories describing how linguistic repertoires can be employed in summative and formative assessment for foreign language learning and teaching: (a) assessment of one named language, (b) assessment of several named languages, (c) assessment in subject-specific contexts, and (d) assessment in contexts involving languages unfamiliar to the learner.

We here employ the framework of de Angelis (2021) when contextualizing the developed procedure ("Allrad-M"). In that framework, a distinction between traditional and holistic approaches in assessing multilingual competencies is made, aligning roughly with the poles of Shohamy's continuum. Specific test procedures can therefore be categorized as "monolingual," "multilingual by translation" or "multilingual by design" (de Angelis, 2021, p. 24). While "multilingual by translation" refers to a test available in different language versions that maintain language separation, "multilingual by design" incorporates heteroglossic principles, allowing test-takers to fluidly switch between the languages in their repertoire during testing. Examples of the multilingual by translation" approach are the diagnostic tools developed within the German "FörMig"-program, which assessed language skills separately in German and other languages, such as Russian and Turkish (Gogolin et al., 2011). In contrast, examples of the "multilingual by design"-approach are rare. One example is a digital tool used to assess learners' mathematical skills fluidly in English and Spanish (Lopez et al., 2019). An exploratory study on this tool showed that learners used it to process and respond to mathematical tasks multimodally (both orally and in writing) and multilingually (English and Spanish) (Lopez et al., 2019).

Whether diagnostic procedures based on the principle of "multilingual by design" principles reduce method bias or improve test fairness and construct validity remains an open question. Shohamy (2011) found that Israeli students whose family language is Russian performed better in math assessments when allowed to use both Hebrew and Russian compared to a monolingual Hebrew condition. However, contrasting results were observed by de Backer et al. (2024) in an experimental study using test materials from the Trends in International Mathematics and Science Study, a large-scale international assessment designed to evaluate math and science competencies of students worldwide. In this study multilingual accommodations, such as providing test materials in multiple languages or offering a "read aloud" function, had no measurable effect on the results of multilingual students. Neither the language factor nor the additional support improved overall performance. From the learners' perspective, however, these accommodations are reported as helpful for comprehension (de Backer et al., 2019). Yet, the effectiveness of such accommodations appears to be limited by the learners' proficiency in their family languages. Lower language skills in these languages negatively impacted the purposeful effect of the accommodations (de Backer et al., 2020). In another study, Schissel et al. (2018) investigate the effect of multilingual materials on the quality of writing products in English. Here, a positive effect of the multilingual condition was seen, as the task context in which the participants had access to materials in both English and Spanish led to better text quality than when materials were only available in English (see also Hinger, 2024).

According to de Angelis (2021, p. 25), assessment procedures based on "multilingual by design" principles must meet quality criteria known as 'VIVA': validity, inclusivity, viability, and accessibility. In the following section, we outline the "Allrad-M" procedure and evaluate its design against these criteria.

1.3 Allrad-M

Allrad-M ("Allgemeine rezeptive sprachliche Fähigkeiten diagnostizieren – Mehrsprachig"[1]) is an online-based tool being developed to help teachers assess the receptive skills of newly arrived students in Germany with a Ukrainian-Russian language background without bias—that is, by evaluating their overall language repertoire. The acronym "Allrad" was chosen deliberately as it means "4-Wheel Drive" in German. Drawing on García's (2009) metaphor of individual multilingualism as an "all-terrain vehicle," the name reflects the concept of enabling learners to "drive on four wheels" during text comprehension. This means they are not subjected to artificial restrictions in using their available linguistic resources.

The learners successively receive two subtitled audio texts ("At the Airport" and "In the Museum") in an online environment[2]. The texts are presented in a video window with an audio track narrating the text aloud while the subtitles and the current chapter title are displayed. The texts, which are taken from Goltsev (2019), have narrative structural features, are each divided into five sections and are deliberately uniform in terms of length, composition and temporal structure, linguistic complexity, protagonists and referenced objects. Learners can switch fluidly between German, Ukrainian, Russian, and English by clicking buttons below the video window, which display corresponding video layers in the selected language (Figure 1).

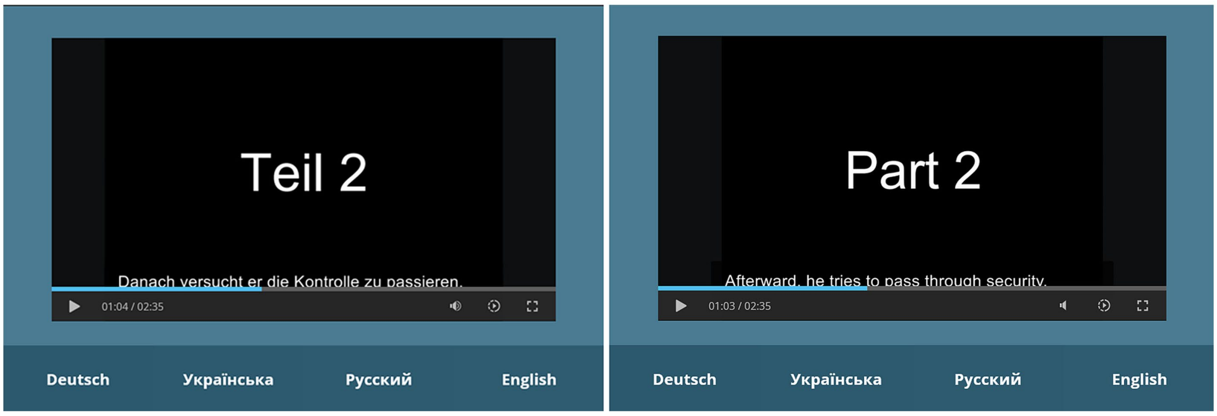
After each reception phase, learners proceed to the assessment section, which involves working on closed-task formats. For each text, 12 items were developed based on the KMK competence model, a framework relevant to educational standards in Germany (Kultusministerkonferenz, 2014). This model defines levels for listening comprehension as shown in Table 1. With regard to our considerations on construct validity, the definitions of the competence levels clearly refer to higher-order skills as a cross-linguistic construct. A newly arrived learner with still limited skills in German, but capable of establishing global coherence based on other languages, can only be assessed validly within the KMK-framework if the entire linguistic repertoire is taken into account. We illustrate competence-levels 1a to 3 each with an item taken from Allrad-M. Since only levels 1a to 3 can be operationalized through closed-task formats (Kultusministerkonferenz, 2014), levels 4 and 5 could not be included in Allrad-M as an online tool.

Three items per level were constructed for each text, resulting in a total of six items per competence-level. During these assessment-sections as well, learners can fully utilize their linguistic repertoire and multimodal options. They can read the questions and answer options in all four languages using a mouseover function on the left-hand side of the screen or listen to them in any of the four languages via a clickable audio icon on the right-hand side (Figure 2). However, since the construct to be operationalized in Allrad-M is text-comprehension according to the model presented above, it should be noted that metalinguistic awareness and cross-linguistic awareness and ability - as described by Hofer and Jessner (2019) - may be activated through Allrad-M's multilingual ecology, but they are not the primary focus of measurement.

In the assessment section, the default view is German for practical reasons. This implies that (a) the reading function in other languages is activated when the cursor hovers over the corresponding language button and deactivates when the cursor exits the button; (b) the audio functions for German and other languages are accessible only from this default level and (c) answering options can only be selected from this view. In the current version of Allrad-M, the central functions (reception and assessment) are framed by a language-free tutorial at the beginning and a feedback section at the end. The tutorial introduces users to the tool's functionality, particularly the language-switching options. The feedback section provides learners and teachers with information on the achieved competence level. A level is considered achieved if at least four out of six items are answered correctly. In line with de Angelis (2021), Allrad-M is primarily designed for formative assessment. As a criterion-referenced assessment tool, it enables teachers to interpret results in terms of the learners' current and proximal zones of development. This is meant to allow educators to create both monolingual and multilingual learning opportunities to enhance receptive skills. Allrad-M thus aims to recognize the full potential of

[1] "Diagnosing general receptive linguistic skills - multilingually"

[2] The tool was created using the "Articulate Storyline" software, a platform for developing interactive e-learning content.


Teil 2

Danach versucht er die Kontrolle zu passieren.

Part 2

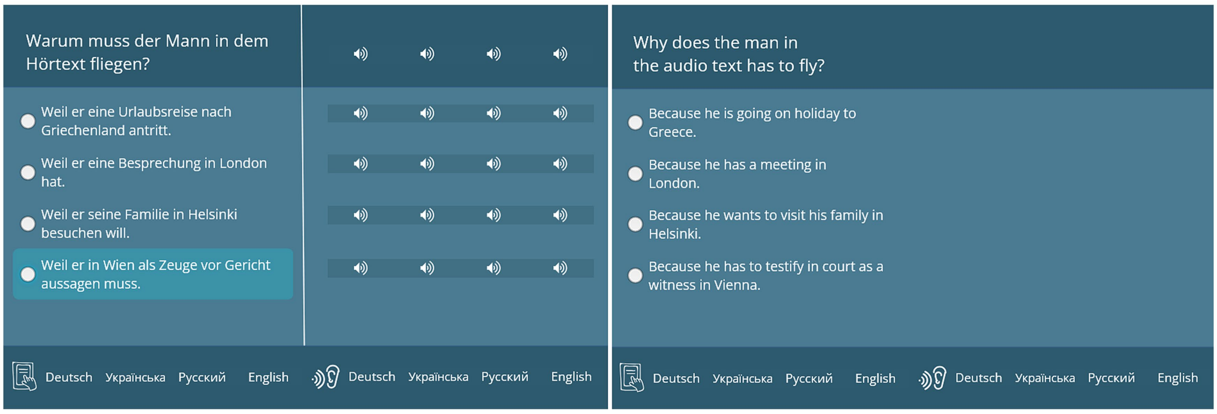
Afterward, he tries to pass through security.

FIGURE 1
Two screenshots from the Allrad-M Tool (reception section).

TABLE 1 Competence levels for listening comprehension.

Competence-level	Illustrating item
Level 1a: Processing and remembering important information	Which statement is true? • A pocket knife was found during the security. • The man in the listening text goes into a duty-free shop.
Level 1b: Combining related information	Why does the man in the audio text has to fly? • Because he is going on holiday to Greece. • Because he has a meeting in London. • Because he wants to visit his family in Helsinki. • Because he has to testify in court as a witness in Vienna.
Level 2: Combining dispersed information and assigning it to a genre	Which title is least fitting to the listening text? • Fear of Flying and a Storm • Stress at the Airport • The Friendly Pilot • Where is my flight?!
Level 3: Combining dispersed information and gaining a rough comprehension of the text	What is the biggest problem for the man in the listening text? • Something beeps at the security control. • It is raining. • He does not find a nice watch. • He is at the wrong gate.
Level 4: Recognizing essential textual relationships, reflecting on structure and recalling less prominent details	/
Level 5: Advanced reception, interpretation, argumentation and evaluation	/

learners, particularly those who have newly arrived and are just beginning to learn German. Additionally, Allrad-M exemplifies the "multilingual by design" approach, as it allows test-takers to fluidly select languages during both the reception and assessment phases based on their individual needs. This is the core feature of its diagnostic concept. However, as a prototype, Allrad-M currently lacks data on its core quality criteria and psychometric properties. Nonetheless, the concept aligns with de Angelis's (2021) "VIVA criteria," which focus on validity, inclusivity, viability, and accessibility: At the heart of Allrad-M's design is the aim to assess listening and reading comprehension in a construct-valid and bias-free manner by leveraging a fluid multilingual approach (validity). de Angelis (2021, p. 25) states that "A test is inclusive when it is designed for the multilingual population in general, not for a subset of the multilingual population such as immigrants or minority language speakers with poor language proficiency in the language(s) of testing." While Allrad-M currently supports only German, Ukrainian, Russian, and English, its design allows for additional languages to be implemented, potentially enhancing inclusivity in the future. The digital environment supports scalability by allowing multiple languages to be integrated without increasing interface complexity (viability and accessibility). However, the tool does require access to digital devices and headphones, which may pose logistical challenges in some contexts.



Warum muss der Mann in dem Hörtext fliegen?

○ Weil er eine Urlaubsreise nach Griechenland antritt.

○ Weil er eine Besprechung in London hat.

○ Weil er seine Familie in Helsinki besuchen will.

○ Weil er in Wien als Zeuge vor Gericht aussagen muss.

Why does the man in the audio text has to fly?

○ Because he is going on holiday to Greece.

○ Because he has a meeting in London.

○ Because he wants to visit his family in Helsinki.

○ Because he has to testify in court as a witness in Vienna.

FIGURE 2
Two screenshots from the Allrad-M Tool (assessment section).

2 Materials and methods

We first outline the research questions and subsequently present the procedure for evaluating the screen recordings and analyzing the interview separately.

2.1 Research questions

Based on 10 screen recordings of newly arrived learners working on Allrad-M and an interview with a teacher who explored the use of Allrad-M in practice, this study addresses the following research questions:

a. How do newly arrived learners with a Ukrainian-Russian linguistic background utilize the multilingual options provided in Allrad-M to leverage their entire linguistic repertoire? Does the degree of multilingual use correlate with the score achieved in the assessment?
b. How does the pedagogical practicability of Allrad-M appear from the perspective of a teacher working with newly arrived students?

2.2 Language choice behavior of learners

2.2.1 Sample and data collection

Data was collected in 2023 as part of two master's theses conducted by pre-service teachers. The aims of these theses were: (a) to perform initial exploratory analyses of how learners use the multilingual options offered by Allrad-M (Busch, 2024) and (b) to examine the effects of these multilingual options on the level of competence achieved (Cwalina, 2024).

For this study, a secondary analysis of the screen recordings collected during these two projects was carried out. The sample consists of 10 newly arrived learners from two schools in the Cologne-Region, North Rhine-Westphalia, Germany. Among these students, 4 are female and 6 are male, the average age is 12.4 years (ranging from 11 to 16 years). The learners' language skills were recorded using questionnaires during the two studies. As the questionnaires were slightly different, we report here the learners' linguistic self-assessments as the mean value of two 5-level scales which were implemented in both questionnaires and measure their self-reported reading and writing skills in the languages included in Allrad-M. The results are summarized in Table 2.

TABLE 2 Self-assessment of linguistic skills (N = 10).

Language	*M*	Max	Min
German	3.40	5	2
Russian	4.25	5	1
Ukrainian	3.15	5	1
English	3.20	5	2

The learners' average length of exposure to the German language is 12.89 months (ranging from 4 to 25 months). Screen recordings were made under different conditions: at School 1, recordings (N = 6) were conducted individually on a laptop PC, while at School 2, recordings (N = 4) were made in a group setting using tablet PCs.

2.2.2 Data preparation and analysis

The screen recordings were analyzed using MAXQDA software to evaluate the use of the multilingual reception options based on the duration of the respective sequences. Learner behavior was coded according to the following categories:

- Language used: German, Ukrainian, Russian, English
- Modality (only in the assessment section): Reading, listening
- Answering behavior in the assessment section: Task solved/ not solved

The internal structure of the tool was coded according to the following categories:

- Text: "At the Airport" vs. "In the Museum"
- Section: "Reception" vs. "Assessment"

By utilizing MAXQDA's complex segment search function, detailed outputs were generated, such as the duration of language usage within specific sections. These outputs were exported as frequency tables. SPSS software was then used to aggregate these tables at the case level, enabling analysis of total language usage durations by text, section, and modality.

2.3 Interview with a teacher

The Allrad-M procedure was implemented by a teacher at a *Realschule* in the Ruhr-Region, North Rhine-Westphalia, involving three newly arrived students at the lower secondary level (*Sekundarstufe I*). The implementation was structured into three distinct phases:

a. In the first phase, the teacher received training through an online meeting, during which two researchers from the project team introduced the procedure and provided detailed instructions for its application.
b. Subsequently, the procedure was piloted during a regular classroom session. Throughout this phase, detailed observational notes were taken by the teacher to document the process.
c. Approximately one week after the pilot session, a semi-structured one-hour interview (Niebert and Gropengießer, 2014) was conducted with the teacher, with the participation of two researchers from the project team.

The interview was recorded and later transcribed verbatim, with minor adjustments made to language and punctuation for clarity and coherence (Kuckartz, 2010). All sensitive data were anonymized to ensure privacy.

The interview aimed to address the following research questions:

- Can the implementation of Allrad-M help derive targeted interventions and (multilingual) learning opportunities?
- Does it highlight the full potential of multilingual learners more effectively?
- Is it possible to formulate differentiated learning objectives, particularly in terms of German as a second language and overall language competence, based on the model of listening comprehension levels?
- Does observing students' language choices offer valuable insights?
- Can Allrad-M be effectively integrated as a diagnostic tool in everyday teaching practices?
- Is the interface user-friendly and appropriate for students at different proficiency levels?
- How do students react to the multilingual content? Does it serve as a motivational factor in their learning process?

The teacher tested the tool with three newly arrived students whose heritage languages were Ukrainian, Russian, Arabic, and Polish. These students, aged 11, 13, and 15, provided a diverse sample for exploring the procedure's effectiveness across different age groups. Each session lasted approximately 15–30 minutes, providing sufficient time to observe how the students engaged with the multilingual tool.

3 Results

3.1 Language choice behavior

The total average time for completing the procedure was 1193.84 s (*SD*: 260.45). Differences emerged between the two device types:

- Laptop PC: *M* = 1297.55 s
- Tablet PC: *M* = 1038.28 s

Learners need more time for the assessment section (672.13 s) than for the reception section (521.47 s). However, the time required for reception and assessment of the two texts "At the Airport" (603.4 s) and "In the Museum" (590.2 s) is almost equal.

The language usage reported in the following section is presented as proportions. For each learner, a quotient was calculated by dividing the duration of time spent using each language by the total time of language usage.

Figure 3 shows the overall language choices made by learners throughout the Allrad-M procedure. All four languages were employed by the learners. The use of German predominates, accounting for nearly 50% of the total usage. However, it should be noted that German serves as the default-mode in the assessment, i.e., if learners did not actively select another language using the "mouseover" or clicking functions, German was automatically coded. This might include time spent thinking rather than actively engaging with the material.

Figure 4 shows the use of languages in contrast between the two parts "At the Airport" and "In the Museum." While the proportions of Ukrainian and English remain roughly constant in both parts, the proportion of German declines quite sharply in favor of Russian in the second part. Taking the uniformity of the both texts into account, this suggests a possible "practice effect," where the learners become more familiar with the multilingual options and increasingly leverage their full linguistic repertoire during the second task.

A comparison of the use of the four languages between the reception and assessment phases (Figure 5) reveals that the learners employed languages other than German more frequently during reception. This difference is likely due to the assessment default mode, where learners must actively select another language, potentially distorting the results. Additionally, there is a marked difference in the frequency of language switches between the two phases: learners switched languages 2.14 times per minute during reception compared to 12.39 times per minute during assessment. This reflects the cognitive demands of each phase: Reception involves processing longer text sequences, while assessment focuses on shorter, discrete tasks such as answering questions.

The analysis of assessment language use was further broken down into reading and listening modalities. Reading (average duration: 557 s) dominates compared to listening (average duration: 139.39 s). As already mentioned, it should be noted here that the coding does not differentiate between actual reading and other type of processing. German was the dominant language for reading, while listening showed a broader utilization of the learners' entire linguistic repertoire (Figure 6). A deeper analysis of the coded sequences allows a reconstruction of how modality and multilingualism are intertwined: The learners appear to have followed a strategy whereby the questions and answer options were first read in German and then heard in another language to ensure comprehension. On the one hand, listening in other languages outweighed listening in German (253 occurrences vs. 78 occurrences). On the other hand, 76 occurrences were identified in which reading in German was followed by listening in German. This contrasts with a frequency of 253 occurrences in which reading in German is followed by listening in one of the other languages.

Regarding the underlying concept of Allrad-M, which posits that listening and reading comprehension can only be assessed with a diagnostic tool that is construct-valid and bias-free

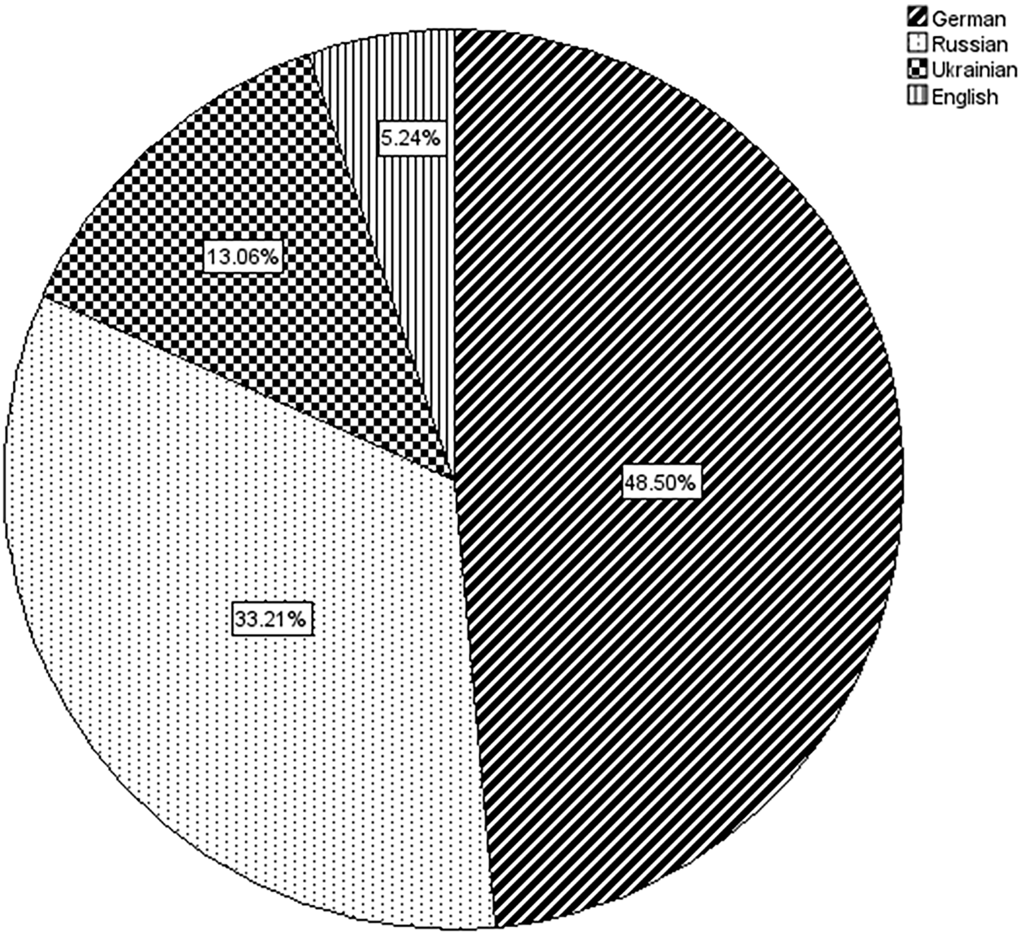


FIGURE 3
Language choice in total when working on Allrad-M (*N* = 10).

('multilingual by design'), we expect the number of items solved correctly in the assessment sections to be associated with a language choice behavior matching with the individual linguistic repertoire. To explore this, we calculated an index that serves as an approximate measure of the alignment between the linguistic repertoire and the choice of languages when working on Allrad-M. This measure was calculated in four steps: First, the sum of the variable values reduced by 1 was determined for the self-assessments in German, Ukrainian, Russian and English. Then the proportions of the individual languages in the entire linguistic repertoires were determined by calculating the quotient of the variable values of the linguistic self-assessment and the sum of all languages for each language. Then, for each language, the difference between the proportion of the respective language in the use of Allrad-M and the repertoire proportion of this language was calculated. This value expresses the fit between repertoire and usage for the respective language. If this value was negative, the sign was inverted. Finally, the mean value of all four variables calculated in this way was formed. This final value serves as a measure of the fit between the individual multilingual repertoire and the use of the languages in Allrad-M and can reach a maximum of 1. In the latter case, there is a full correspondence between the repertoire and the use of languages. A value of 0 would indicate that only languages for which no skills were specified in the self-assessment were selected when using Allrad-M. Figure 7 visualizes the relationship between the fit of repertoire and language choice and the number of items solved correctly.

A visual inspection indicates a positive association between the score achieved in the assessment sections and the fit between repertoire and language choice. One case takes an outlier position (high fit index with a low score in the assessment). This may be related to the total length of use of this case, which is about 1.5 standard deviations below the mean value of all cases and represents the minimum for this group.

Two cases were selected to highlight contrasting language choice behaviors based on the principle of maximum contrast. An visual inspection of sequential diagrams displaying language choice behavior taken from the preliminary work of Busch (2024) served as the basis for the selection of two maximally contrasting cases. Table 3 summarizes the background characteristics of these cases.

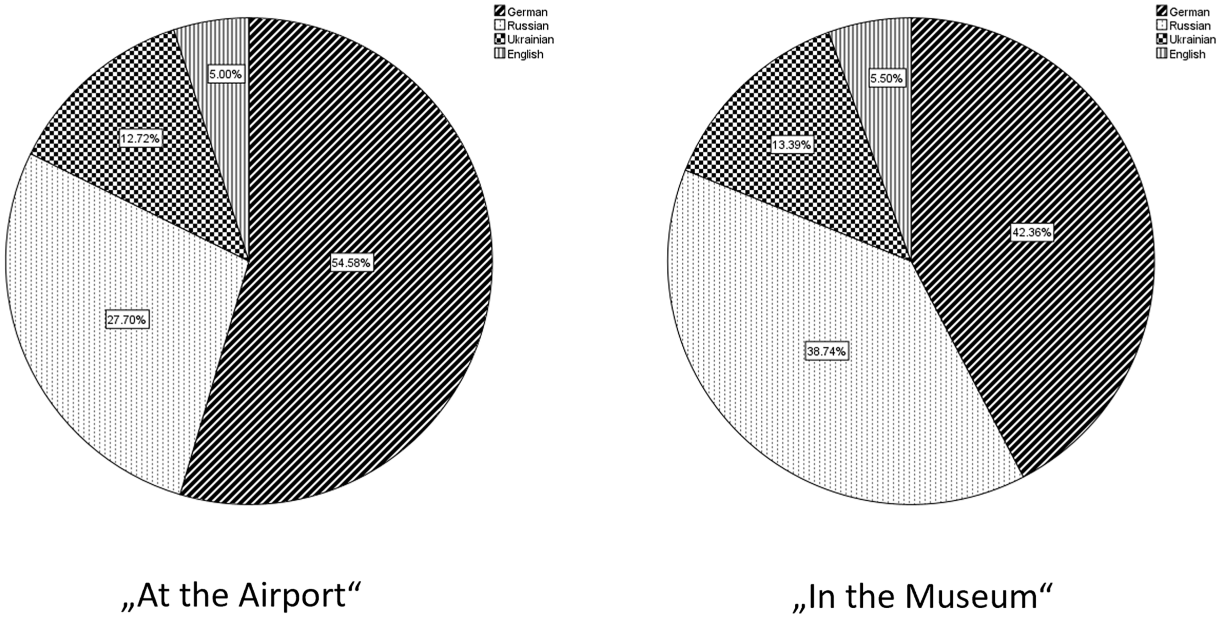


FIGURE 4
Language choice in the parts "At the Airport" and "In the Museum" (*N* = 10).

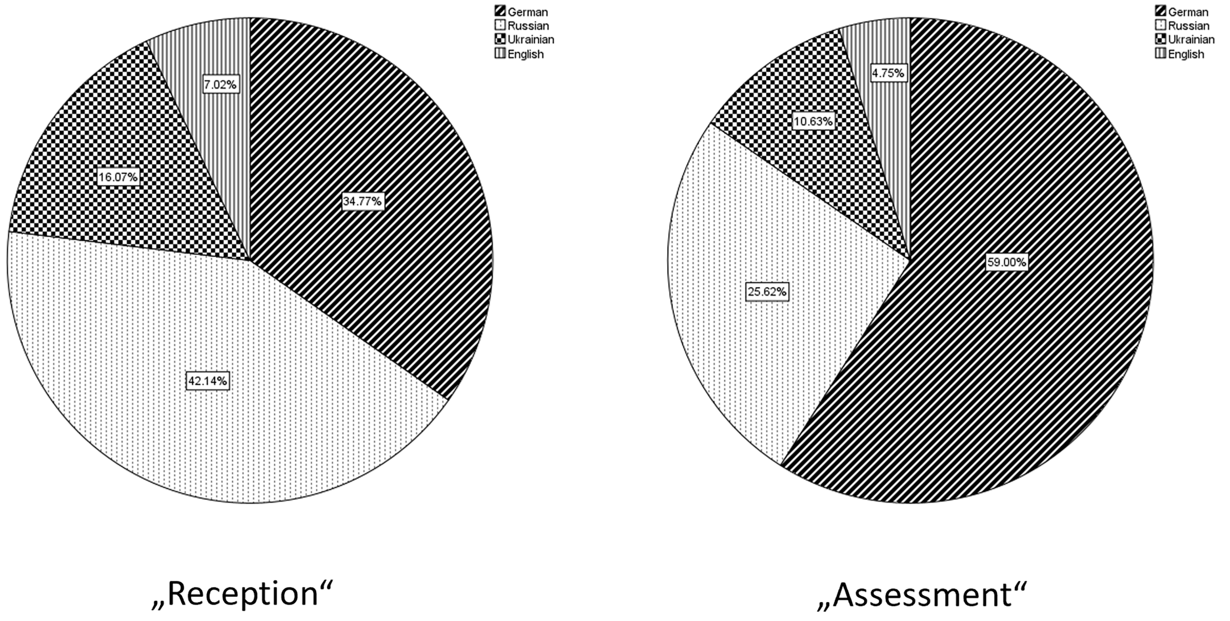


FIGURE 5
Language choice during "Reception" and "Assessment" (*N* = 10).

For the analysis, the sequences in which the languages German, Ukrainian, Russian and English were used during reception and assessment were first coded to the nearest tenth of a second in the screen recordings. For the assessment phase, it was also coded whether the languages were used in the "listening" or "reading" modality. In a second step, this data was transformed into sequence diagrams, which illustrate the use of the languages (and, in the case of the assessment, the modality) throughout the entire period of usage. Figure 8 shows the use of the languages during the reception of the text "In the Museum."

Case 1 exclusively relies on its two strongest languages, Ukrainian and Russian, to construct meaning during the reception

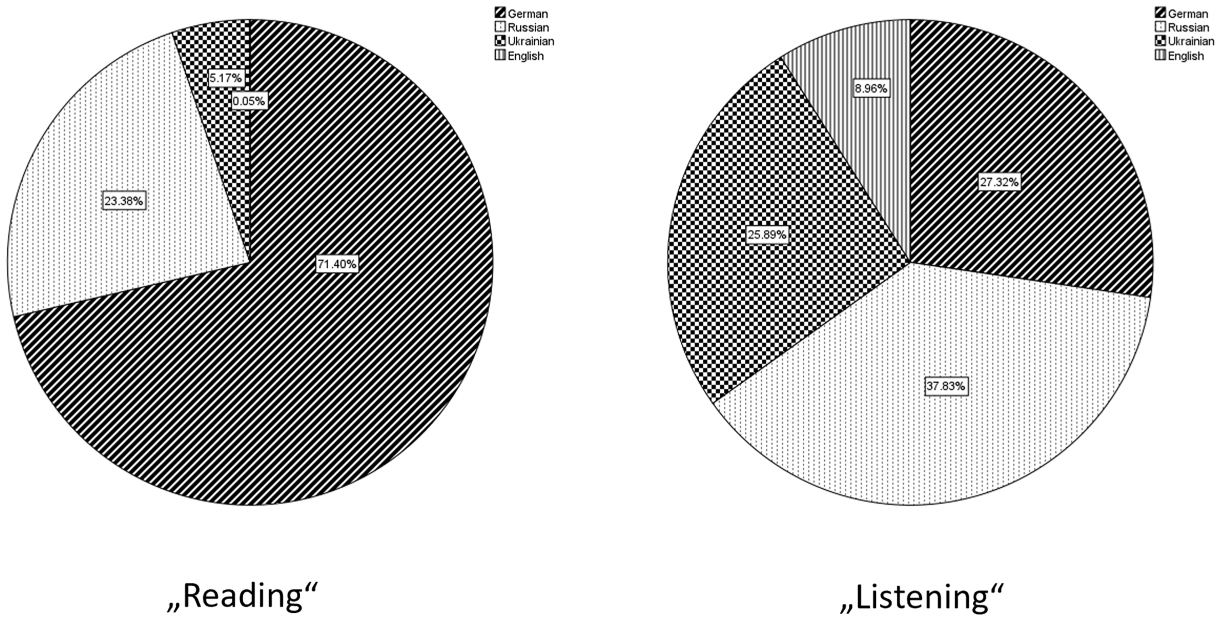


FIGURE 6
Language choice in assessment by “Reading” and “Listening” (*N* = 10).

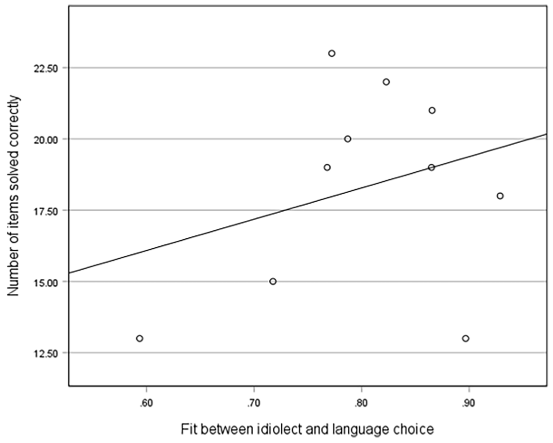


FIGURE 7
Association between test scores and alignment between language choice and linguistic repertoire (*N* = 10).

of the text. This strategy involves two complete runs through the text, beginning with Ukrainian and followed by Russian. The reception approach can therefore, as there is only one switch from language to language and the text is received in one go in each case, be characterized as *linearly-macroalternating*. In contrast, the reception behavior of Case 2 is more complex: This student also uses his two strongest languages (Russian and German), but interrupts the reception initiated in German several times in favor of a sequential renewed reception in Russian. This pattern occurs in Chapter 2, then Chapters 3 and 4 and finally Chapters 3 and 4 are revisited after completing Chapter 5 in German. The strategy employed by Case 2 can be described as *discontinuously micro-alternating*.

Figure 9 contrasts the choice of languages during the assessment phase for the text “In the Museum.” The visualization confirms the prior observation that language changes occur more frequently per unit of time during assessment compared to reception. In contrast to reception, Case 1 makes greater use of his entire linguistic repertoire, with a notably higher frequency of German usage. Within the Ukrainian language, the proportion and frequency of events are significantly higher for listening than for reading. Furthermore, the proportion of Russian usage increases with increasing level of difficulty of the items. This behavior suggests a possible relationship between cognitive demands and utilization of multilingual resources. The usage behavior of Case 1 can be summarized as *increasingly multimodal-multilingual*.

Case 2 uses two languages (German and Russian, as in reception) to complete the tasks. The first four items are completed almost exclusively in German, with a high reliance on the listening modality. As the difficulty of the items increases, the proportion of listening in German decreases, while the use of Russian rises, but only in the reading modality. This suggests that listening in German was initially used as a tool to aid comprehension. However, as comprehension demands grew, reading in Russian—the learner’s strongest language—became the preferred strategy. The language usage behavior of Case 2 can be summarized as transitioning *from monolingual-multimodal to bilingual-monomodal*.

3.2 Interview

The interview with the teacher was analyzed using qualitative content analysis (Mayring, 2022), which allowed for the development of categories grounded in the data’s content.

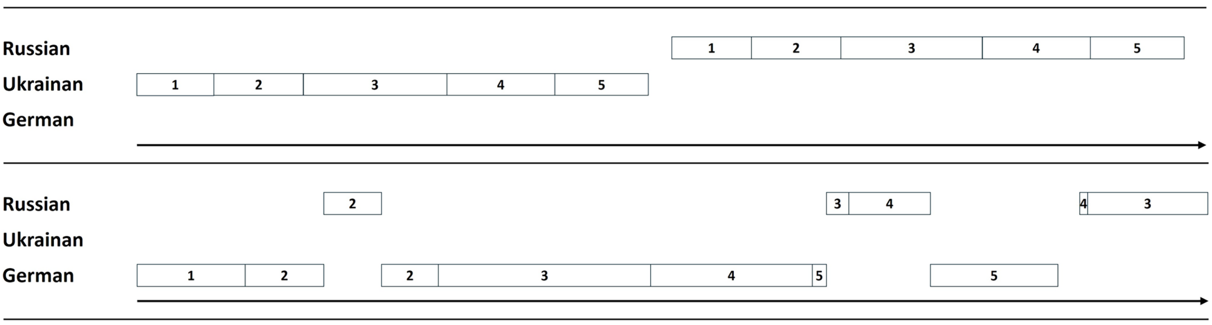


FIGURE 8
Language choice diagrams for reception of the text "In the Museum" (on top: case 1; below: case 2; numbers indicate the respective chapters of the text; total time case 1: 376.3 s; total time case 2: 254 s).

TABLE 3 Background characteristics of cases.

Feature	Case 1	Case 2
Age	11	11
Gender	Male	Male
Time since arrival	12 months	18 months
Self-assessment German	3.5	3.5
Self-assessment Russian	4	5
Self-assessment Ukrainian	5	1
Self-assessment English	3.5	3.5
Items solved	19	15

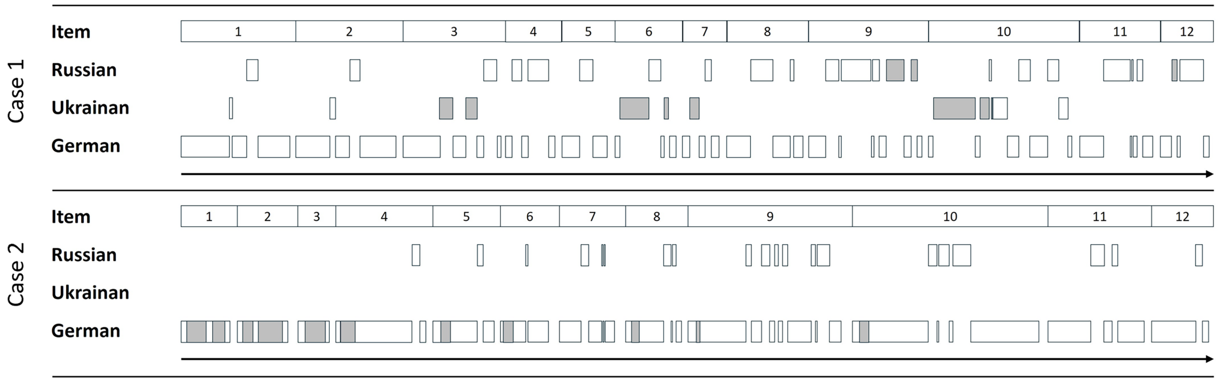


FIGURE 9
Language choice diagrams for assessment of the text "In the Museum" (on top: case 1; below: case 2; numbers indicate the respective items of the text "In the Museum"; grey background indicates sequences of listening; total time case 1: 456.4 s; total time case 2: 434.3 s).

For this analysis, four main categories were defined:

a. The teacher's perspective on the tool and assessment of students' competencies
b. The student's use of the tool's multilingual features
c. Practicality of the tool's application in the classroom, and
d. Challenges with the tool's interface.

The findings are presented according to these categories, with relevant examples provided to illustrate key insights.

In terms of the teacher's perspective on the tool and the assessment of competencies, the tool demonstrated its ability to capture students' competencies in their family languages. A notable observation was the teacher's ability to evaluate a student's proficiency in Ukrainian and Russian despite not speaking these languages. The tool enabled the teacher to recognize the student's linguistic potential, as highlighted in this excerpt from the interview:

"[…] And it was interesting for us to see how he went through this diagnostic process and what he can actually do in his heritage language. From this, we could easily determine whether we should offer him something again in his heritage language, or whether he was unsure about it. You can see that, even if you do not speak the language yourself. So, it was really helpful to see how well he knows Ukrainian or Russian, because normally, we would always write down sentences and try to get him to conjugate in his heritage language, so that even without knowing the language, we could understand how well he can do it. And now, this went pretty quickly in digital form. That was definitely helpful. […]"	*"[…] Und es war für uns interessant zu sehen, wie er diese Diagnostik durchlaufen hat und was er eigentlich in seiner Herkunftssprache kann. Daraus ableitend einfach zu wissen, ob wir für ihn nochmal in der Herkunftssprache was anbieten sollten. Oder ob er da unsicher ist. Das kann man ja auch, obwohl man die Sprache nicht selber spricht, dann sehen. Das heißt, das war schon hilfreich zu sehen, wie er Ukrainisch oder Russisch kann, weil wir da ja normalerweise immer auch Sätze aufschreiben und ihn versuchen, in seiner Herkunftssprache konjugieren zu lassen, um auch ohne Sprachkenntnisse zu wissen, wie er das eigentlich kann. Und jetzt ging das aber recht flott digital. Das war auf jeden Fall hilfreich. […]"*

Furthermore, the teacher emphasized that the digital format of the tool allowed faster documentation of students' proficiency levels. This was particularly valuable for a student with limited German proficiency, previously evidenced by his production of only single-word utterances in German. However, the tool revealed the student's significantly more advanced abilities on the general linguistic level, as noted in the following comment:

"[…] We had tried the Grießhaber method with him before, and it was really the case (pauses to think) that we could not get beyond one-word utterances. But through the diagnostic process, I could see that he is, of course, much more advanced in his heritage language than in German, which was naturally to be expected. But it was still nice for us to see once again just how far along he actually is in his heritage language. […]"	*"[…] Also wir hatten nach Grießshaber das mal versucht bei ihm, war schon so, dass (überlegt kurz) wir eigentlich über Ein-Wort-Äußerungen nicht hinauskamen tatsächlich. So habe ich über die Diagnostik gesehen, dass er natürlich in seiner Herkunftssprache viel weiter ist als auf Deutsch, ist aber auch eigentlich natürlich gewesen. Aber es war für uns einfach schön nochmal zu sehen, wie weit er in seiner Herkunftssprache eigentlich ist. […]"*

The teacher observed that such insights are particularly valuable for educators with limited experience working with multilingual students, who are typically only assessed in German:

"[…] So, I think that for teachers who might not have direct experiences with students who are proficient in a second or third language, this could also be a learning experience for them, showing that students can be more advanced in their heritage language while perhaps not in German. […] It certainly offers great potential and can help us in the future. *[…] I think mainstream class teachers simply lack the knowledge to assess what these students can do. Of course, they also have bilingual students, but they do not have the understanding of what a student might be capable of after being here for 3 months. Usually, these students have been here for at least two years before a mainstream class teacher, according to regulations, even has the student in their classroom. So, knowledge about students with less than two years in the country is virtually non-existent. In this sense, diagnostic assessments or evaluations are often (pauses to think) difficult. Many teachers are also hesitant to make judgments about what these students can do. And here, it might be helpful to show through diagnostic tools what a student is capable of in their heritage language. Depending on the level of competence, you could see that the student is familiar with certain genres, and maybe it's possible to make connections between different text types, showing that it's entirely feasible. And here, instead of a competence level 1 in their heritage language, the student might actually be at level 3, suggesting that we should encourage more integrative learning. […]"*	*"[…] Also ich denke mal, die Lehrkräfte, die vielleicht nicht den direkten Umgang mit Schülern haben, die noch eine zweite oder dritte Sprache beherrschen, dass das vielleicht auch ein Lerneffekt für Lehrkräfte ist, zu sehen, dass die natürlich in der Herkunftssprache auch fortgeschritten sein können und eben auf Deutsch vielleicht dann nicht. […] Es bietet halt ein großes Potenzial und das kann auf jeden Fall in der Zukunft uns helfen.* *[…] Ich denke, den Lehrern der Regelklasse fehlt einfach, das Knowledge zu gucken, was können diese Schüler. Natürlich haben die auch bilinguale Schüler, aber es fehlt einfach die Kenntnis, was kann jetzt ein Schüler können, wenn er drei Monate hier ist. Normalerweise sind die dann mindestens zwei Jahre schon hier, bevor ein Lehrer der Regelklasse laut Erlass den Schüler bei sich sitzen hat. Das heißt, die Kenntnis unter zwei Jahre ist eigentlich gar nicht gegeben. Insofern ist auch die Diagnostik oder die Einschätzung meistens (überlegt) schwierig. Also dann trauen sich auch viele nicht einzuschätzen, was dieser Schüler kann. Und hier wäre es vielleicht hilfreich, anhand der Diagnostik zu zeigen, was dieser Schüler in der Herkunftssprache kann. Und je nach Kompetenzstufe eben, dass er bestimmte Genre eben auch kennt, dass man eine Übertragung nochmal vielleicht herstellen kann zwischen den Textsorten, dass das durchaus möglich ist. Und hier nicht Kompetenzstufe 1 in der Herkunftssprache vorliegt, sondern 3, dass man vielleicht mehr vernetztes Lernen eigentlich fördern sollte. […]"*

These observations highlight the significance of using tools that assess students in their heritage languages, allowing for a more accurate evaluation of their competencies.

While the tool provided a clear and detailed picture of students' multilingual abilities, translating these insights into actionable educational strategies proved challenging. The main difficulties arose from the time required to work individually with students during the assessment process and the complexity of interpreting the competency levels achieved by the students:

"[…] So, we teachers are somewhat familiar with the competence level framework. This is due to our training and personal interest. I engage with it myself. But it's not something that can be taken for granted. In daily life, it's simply not feasible to go back and deal with academic models that were developed after one's training. This is especially true for older teachers who just do not have the time for it. […]"	*"[…] Also wir Lehrkräfte kennen das Kompetenzstufenniveau teilweise. Das liegt an der Ausbildung und an dem persönlichen Interesse. Also ich setze mich damit auseinander. Aber dass das jetzt nicht selbstverständlich ist. Es ist einfach im Alltag erstens nicht machbar, sich nochmal rückwirkend mit fachwissenschaftlichen Modellen auseinanderzusetzen, wenn sie dann nach der Ausbildung irgendwie entwickelt sind. Also gerade bei älteren Lehrkräften, wenn man die Zeit einfach nicht dafür hat. […]"*

In terms of the second category, „use of the tool's multilingual options by the students", a notable emotional aspect emerged when students encountered materials in their heritage languages. One student, for instance, hesitated and sought permission before engaging with the Ukrainian and Russian options:

"[…] The students [were] also quite surprised and happy to suddenly be able to click on something in their heritage language. It was just nice to see. But also a little (searching for words) sad in a way, because he looked at us and asked for permission to really do it now and click on something in his own language. Which shows us that, at first, he simply did not want to do it and needed permission. He's been with us for almost half a year now, and we had always thought it was clear that it was okay to do so. *[…] Then it was very, very interesting for me that he kept switching back and forth between Russian and Ukrainian. […] He ended up staying more with Ukrainian.* *[…] And then, it was interesting for us what the legal guardians had to say, who told us that he preferred to use Russian.* *[…] He seemed more relaxed because he was in his own language, and he really took a deep breath and then approached everything a bit more calmly. […]"*	*"[…] Die Schüler [waren] auch relativ überrascht und glücklich, in der Herkunftssprache plötzlich etwas anklicken zu können. Das war einfach schön zu sehen. Aber auch ein bisschen (sucht nach Worten) traurig irgendwo, weil er uns angeguckt hat und um Erlaubnis gebeten hat, wirklich das jetzt machen zu dürfen und in seiner Sprache etwas anklicken zu dürfen. Was uns zeigt, dass er das einfach auch erstmal nicht machen wollte und die Erlaubnis gebraucht hat. Und bei uns ist er jetzt seit knapp einem halben Jahr und eigentlich haben wir für uns selber immer gedacht, dass das klar wäre, dass das in Ordnung ist.* *[…] Dann war es für mich sehr, sehr interessant, dass er zwischen Russisch und Ukrainisch immer wieder hin und her switchte. […] Er ist dann mehr bei Ukrainisch geblieben. […] Und dann ist es wiederum für uns interessant, was die Sorgeberechtigten sagen, die dann wiederum uns gesagt haben, dass er lieber Russisch nimmt.* *[….]. Also er wirkte entspannter, weil er dann eben in seiner Sprache und er hat dann wirklich einmal durchgeatmet und sich dann dem Ganzen ein bisschen entspannter zugewandt. […]"*

This emotional reaction highlights the positive response to the tool's multilingual features, but also suggests underlying discomfort, likely due to previous educational experiences where the use of heritage languages was discouraged. However, this feature was only available to students whose heritage languages were Ukrainian or Russian. Students with other heritage languages, such as Polish or Arabic, had to use English, which led to frustration, particularly among younger students with limited English proficiency:

"[…] With the younger student, who wasn't very advanced in English, she visibly became frustrated quite quickly and at some point looked very stressed. I even asked if she wanted to stop, but she said that she wanted to finish it. […]"	*"[…] Bei der jüngeren Schülerin, die tatsächlich im Englischen jetzt nicht so weit war, die war relativ schnell frustriert sichtlich und sah auch sehr gestresst aus irgendwann. Also ich habe auch gefragt, ob sie abbrechen möchte. Und dann hat sie gesagt, dass sie das noch zu Ende führen möchte. […]"*

The tool's flexibility in allowing students to switch between languages was seen as a strength, but this also highlighted the need for further development, such as integrating additional languages and better accommodating students' linguistic preferences. Regarding the practicality of the tool's use, the teacher pointed out the cognitive challenges associated with using the tool during testing. Managing multiple tasks, like monitoring student progress, taking notes, and assessing competence levels, proved demanding:

"[…] [it] was quite challenging for me to assess everything at once, looking at the interface. Does he get it? Does he understand it? Can he click on it? At the same time, I had to take notes and keep an eye on the competence levels, evaluating what he was doing at that moment. Is he there now? Ah, okay, he did not understand that. Then I had to take notes. That was cognitively quite demanding. […]"	*"[…] [es] war für mich schon auch schwierig, gleichzeitig einzuschätzen, die Oberfläche sich anzugucken. Kommt er damit klar? Versteht er das? Kann er das anklicken, gleichzeitig Notizen zu machen und die Kompetenzstufen im Blick zu haben und einzuschätzen, was macht er jetzt gerade? Ist er jetzt da? Ah, okay, das hat er nicht verstanden. Dann sich Notizen zu machen. Das war kognitiv schon nicht so einfach. […]"*

Finally, regarding the category of „challenges with the tool's interface", the multilingual options and the ability to both read and listen to texts were noted as strengths. However, students found it difficult to gage the test's length, which negatively affected their motivation. The teacher suggested incorporating clearer indicators to inform students when they are nearing the end of the test. In conclusion, the teacher's interview highlights several key insights into the use of the Allrad-M tool. First, the tool effectively captures multilingual students' linguistic competencies, particularly in their heritage languages, providing valuable diagnostic information even when the teacher does not speak those languages. Second, its ability to accommodate multiple languages was beneficial, though challenges remain, such as the limited availability of certain languages and the need for further development to support a broader range of linguistic backgrounds. Third, the practical application of the tool revealed cognitive challenges for the teacher, particularly in managing simultaneous tasks like monitoring student progress, taking notes, and assessing competence levels. Finally, feedback on the user interface underscored the importance of clear guidance on test duration to help students maintain motivation. Overall, the Allrad-M tool demonstrates great potential as a diagnostic resource, though further refinements are needed for smoother integration into everyday classroom practices.

4 Discussion

We first presented the concept and theoretical basis of the Allrad-M diagnostic procedure, which is currently under development and is based on the principle of "multilingual by design" (de Angelis, 2021). Using a prototype, initial exploratory studies were conducted to examine learners' language choice behavior during the reception of two subtitled listening texts and while answering closed-task formats in the assessment. For this purpose, screen recordings of learners' usage behavior were analyzed using content analysis software to identify which languages were used in specific sequences. Additionally, an interview was conducted with a teacher who tested Allrad-M in practice to explore its pedagogical applicability. The analysis of the language choice behavior among the group of 10 newly arrived learners shows that they extensively utilized their entire linguistic repertoires. From an overall perspective, the proportions of language usage appeared in the following order: German, Russian, Ukrainian, and English. This pattern diverges slightly from the learners' linguistic self-assessments, where Russian is the strongest language on average, followed by German, Ukrainian and English. However, more detailed analysis revealed that: (a) the order of languages from the self-assessment is more accurately reflected during the reception section (while distorted in the assessment section), and (b) the proportion of German usage decreases in the second part of the procedure ("In the Museum") compared to the first part ("At the Airport"). This suggests a potential learning effect: with prolonged exposure to Allrad-M, learners become better able to exploit their entire linguistic repertoires, both conceptually and in terms of navigating the digital environment. Moreover, the observed shifts between the two texts may indicate that learners initially need to overcome internal barriers to abandon a "monolingual habitus" (Gogolin, 2008). This interpretation is supported by the teacher's observation that one student explicitly asked for permission to use languages other than German. Overall, our findings are exploratory and preliminary due to the small sample size, but align with those of Lopez et al. (2019), who showed that multilingual learners integrate their own multilingualism in a fluid multilingual assessment environment. However, unlike Lopez et al., where the family language was preferred over English as the language of instruction, our study reflects a broader usage of German as language of instruction. Indirectly, the extensive use of multilingual options also confirms the results of de Backer et al. (2019), according to which learners perceive multilingual test accomodations as both positive and helpful. The comparison of language use between reception and assessment phases reveals that German plays a more prominent role in the assessment phase than in reception. However, this observation is influenced by the technical characteristics of Allrad-M, where German serves as the default language in the assessment section. As this study does not employ an eye-tracking, it was not possible to separate processes of reading from processes of thinking about the correct choice (both cases lead to a coding of "German" and "reading"). Nonetheless, there are significantly more language switches per time unit in the assessment sections, which is plausible in view of the shorter language units (questions and answer options) to be processed cognitively. Further analysis of language choice during assessment revealed more multilingual usage in listening compared to reading. A recurring pattern was identified: Segments coded as "reading in German" were frequently followed by "listening in another language," more so than by "listening in German." This suggests a general strategy where learners first read in German and then listen in another language to ensure comprehension. The interplay between multilingualism and multimodality observed here echoes findings by Lopez et al. (2019) regarding the use of "read-aloud" functions. Future research should investigate how different strategies for managing multilingual reception complexity might affect test fairness and construct validity (see also de Backer, 2020, p. 149). In the present study, the two qualitative case studies on language choice behavior highlighted contrasting strategies: a linearly-macroalternating approach versus a discontinuously-microalternating approach to multilingual usage. These contrasting practices likely affect comprehension in different ways, underscoring the need for further research into individual reception strategies and their impact on assessment outcomes.

An exploratory investigation into the relationship between the fit of learners' repertoire with their language choice and their success in the assessment revealed a potential association. This was visualized through a scatterplot, where the alignment of a language choice according to the individual repertoire and test scores suggested a positive correlation. Due to the small sample size, this relationship should be interpreted as hypothetical. Nonetheless, the findings lend preliminary support to the theoretical assumption that the construct-valid assessment of higher-order competencies in listening and reading comprehension requires enabling multilingual learners to use their full linguistic repertoire. However, our preliminary findings somehow contrast to the study of de Backer et al. (2024), where no effects were found for multilingual test accommodations. The authors link this finding to the learners' different abilities in their family languages (de Backer et al., 2024, p. 98). As newly arrived learners, the participants in our study all reported a high level of proficiency in their family language, so that they were apparently able to draw on this as a resource for constructing meaning. Hence, future research could expand on these findings through larger-scale correlation studies or experimental designs comparing performance under monolingual versus multilingual testing conditions and additionally contrasting newly arrived with resident multilingual learners.

Complementing the analysis of language choice behavior, an interview with a teacher provided additional insights consistent with translanguaging pedagogy principles (García, 2009; García and Wei, 2014). The teacher reported that Allrad-M effectively uncovered linguistic competencies on the general linguistic level that would have remained hidden in a monolingual assessment in German. This aligns with the objective of Allrad-M to distinguish between zones of proximal development in German and general linguistic skills. Furthermore, the teacher observed positive socio-emotional effects among learners, which they attributed to the tool's multilingual design (García et al., 2017). However, the teacher also identified several challenges. First, formative, criterion-referenced assessments like Allrad-M require teachers to have a strong understanding of competence levels to interpret results effectively. Second, inclusivity (as defined by de Angelis, 2021) emerged as a challenge: Learners whose linguistic repertoires did not align with the implemented languages experienced frustration. Furthermore, the objective of a valid assessment cannot be met if the languages offered by Allrad-M and the linguistic repertoire of the learners do not match. Addressing this issue requires the inclusion of additional languages in the tool. Finally, the lack of a clear progress indicator during the test was highlighted as a potential source of demotivation for learners.

To further validate the tool and its theoretical assumptions, larger-scale studies are required to evaluate the psychometric properties of Allrad-M. These studies should also include correlation analyses to explore the relationship between the learners' exploitation of their linguistic repertoire and their achieved competency levels. This would empirically substantiate the assumption that "multilingual by design" assessments are construct-valid and bias-free. Moreover, in alignment with the "VIVA" framework by de Angelis (2021), future iterations of Allrad-M should focus on enhancing viability and accessibility. This includes integrating additional relevant heritage languages and disseminating the tool in an accessible, user-friendly format, ideally accompanied by online training resources for educators to ensure its low-barrier implementation.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical approval was not required for the study involving human samples in accordance with the local legislation and institutional requirements because, with regard to the adolescent participants, our study is a non-invasive, secondary analysis of completely anonymized data. In addition, we conducted and analyzed a semi-structured interview with a teacher who piloted the tool in classroom practice. The teacher provided informed consent prior to participation. All procedures were carried out in compliance with the ethical standards of the University of Cologne and with the Declaration of Helsinki. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin. Written informed consent was obtained from the minor(s)' legal guardian/next of kin for the publication of any potentially identifiable images or data included in this article.

Author contributions

CG: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. TB: Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. LB: Investigation, Resources, Visualization, Writing – review & editing. EG: Conceptualization, Methodology, Project administration, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

We would like to thank Natalia Cwalina for the opportunity to conduct a secondary analysis of the data collected as part of her Master's thesis.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. Generative AI (ChatGPT, OpenAI) and AI-based translation software (DeepL, DeepL SE) were employed exclusively for language editing (phrasing and style). They were not used for generating or interpreting content. The authors reviewed all outputs and take full responsibility for the final manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Barberio, T. (2021). Schreiben in zwei Sprachen: Argumentative und narrative Texte bilingualer italienisch-deutscher Schülerinnen und Schüler. Munich: LMU Open Publishing in the Humanities. doi: 10.5282/OPH.11

Busch, L. (2024). Allrad-Antrieb für das Hör- und Leseverstehen? Wie neu zugewanderte Lernende ihre Mehrsprachigkeit für die Konstruktion von Bedeutung nutzen. [master's thesis]. [Köln]: Universität zu Köln.

Cenoz, J., and Gorter, D. (2017). "Translanguaging as a pedagogical tool in multilingual education," in Language awareness and multilingualism, eds. J. Cenoz, D. Gorter and S. May (Cham: Springer), 309–321. doi: 10.1007/978-3-319-02240-6_20

Cenoz, J., and Gorter, D. (2022). Pedagogical translanguaging. Cambridge: Cambridge University Press (Elements in Language Teaching). doi: 10.1017/9781009029384

Cook, V. (2007). "Multi-competence: black hole or wormhole for second language acquisition research?," in Understanding second language process, eds. Z. Han, E. S. Park, A. Révész, C. Combs and J. H. Kim (Clevedon, Buffalo, Toronto: Multilingual Matters Ltd), 16–26.

Cwalina, N. (2024). Assessment von Hör- und Leseverstehen bei neu zugewanderten Schüler*innen: Eine experimentelle Studie zu den Effekten ein- und mehrsprachiger Rezeptionsbedingungen. [master's thesis]. Köln: Universität zu Köln.

de Angelis, G. (2021). Multilingual testing and assessment. Bristol, Blue Ridge Summit: Multilingual Matters. doi: 10.21832/9781800410558

de Backer, F. (2020). Multilingual assessment in education: a mixed-methods study of the effectiveness of assessment accommodations and user perspectives on them. [dissertation]. [Gent]: Universiteit Gent. Faculteit Letteren en Wijsbegeerte.

de Backer, F., Slembrouck, S., and Aevermaet van, P. (2020). Functional use of multilingualism in assessment: opportunities and challenges. *Res Notes* 78, 35–41.

de Backer, F., Slembrouck, S., and van Avermaet, P. (2019). Assessment accommodations for multilingual learners: pupils' perceptions of fairness. *J. Multiling. Multicult. Dev.* 40, 833–846. doi: 10.1080/01434632.2019.1571596

de Backer, F., Vantieghem, W., and van Avermaet, P. (2024). "Functional multilingualism in educational assessment," in Assessment of plurilingual competence and plurilingual learners in educational settings, eds. S. Melo-Pfeifer and C. Ollivier (London: Routledge), 92–105.

García, O. (2009). Bilingual education in the 21st century: A global perspective. Chichester: Wiley-Blackwell.

García, O., Ibarra Johnson, S., and Seltzer, K. (2017). The translanguaging classroom: Leveraging student bilingualism for learning. Philadelphia: Caslon. doi: 10.1057/9781137385765

García, O., and Wei, L. (2014). Translanguaging. Language, Bilingualism and Education. New York: Palgrave Macmillan.

Gebauer, S. K., Zaunbauer, A. C. M., and Möller, J. (2013). Cross-language transfer in English immersion programs in Germany: reading comprehension and reading fluency. *Contemp. Educ. Psychol.* 38, 64–74. doi: 10.1016/j.cedpsych.2012.09.002

Gogolin, I. (2008). Der monolinguale Habitus der multilingualen Schule. Münster: Waxmann.

Gogolin, I., Dirim, I., Klinger, T., Lange, I., Lengyel, D., Michel, U., et al. (2011). Förderung von Kindern und Jugendlichen mit Migrationshintergrund FörMig: Bilanz und Perspektiven eines Modellprogramms. Münster: Waxmann.

Goltsev, E. (2019). Typen und Frequenzen von L2-Merkmalen im Deutschen als Zweitsprache. Wahrnehmung, Bewertung und Verständlichkeit. Berlin, Boston: De Gruyter Mouton. doi: 10.1515/9783110633696

Gough, P. B., and Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial Spec. Educ.* 7, 6–10. doi: 10.1177/074193258600700104

Herdina, P., and Jessner, U. (2002). A dynamic model of multilingualism: perspectives of change in psycholinguistics. Clevedon, Buffalo, Toronto, Sydney: Multilingual Matters Ltd. doi: 10.21832/9781853595547

Hinger, B. (2024). "Assessing language competences of multilingual speakers – a brief review of two test approaches: C-tests and TBLA (task based language assessment)," in Assessment of Plurilingual competence and Plurilingual learners in educational settings, eds. S. Melo-Pfeifer and C. Ollivier (London: Routledge), 106–115. doi: 10.4324/9781003177197-8

Hofer, B., and Jessner, U. (2019). Mehr-Sprachig-Kompetent MSK 9–12. Mehrsprachige Kompetenzen fördern und evaluieren. Innsbruck: Studia Universitätsverlag.

Knoepke, J., Richter, T., Isberner, M.-B., Neeb, Y., and Naumann, J. (2013). "Leseverstehen = Hörverstehen X Dekodieren? Ein stringenter Test der Simple View of Reading bei deutschsprachigen Grundschulkindern," in Sprachförderung und Sprachdiagnostik: Interdisziplinäre Perspektiven, eds. A. Redder and S. Weinert (Münster: Waxmann), 256–276.

Krulatz, A., Neokleous, G., and Dahl, A. (2022). "Multilingual approaches to additional language teaching: bridging theory and practice," in Theoretical and applied perspectives on teaching foreign languages in multilingual settings: Pedagogical implications, eds. A. Krulatz, G. Neokleous and A. Dahl (Bristol, Blue Ridge Summit: Multilingual Matters), 15–29.

Kuckartz, U. (2010). Einführung in die computergestützte Analyse qualitativer Daten. Wiesbaden: VS Verlag für Sozialwissenschaften. doi: 10.1007/978-3-531-92126-6

Kultusministerkonferenz (2014). Integriertes Kompetenzstufenmodell zu den Bildungsstandards für den Hauptschulabschluss und den Mittleren Schulabschluss im Fach Deutsch für den Kompetenzbereich Sprechen und Zuhören, Teilbereich Zuhören. Available online at: https://www.iqb.hu-berlin.de/bista/ksm/iKSM_Zuhoeren_20_1.pdf [Accessed November 30, 2024].

Lenhard, W. (2019). Leseverständnis und Lesekompetenz: Grundlagen – Diagnostik – Förderung. Stuttgart: Kohlhammer Verlag. doi: 10.17433/978-3-17-035020-5

Lopez, A., Guzman-Orth, D., and Turkan, S. (2019). Exploring the use of translanguaging to measure the mathematics knowledge of emergent bilingual students. Transl. Translanguaging Multilingual Contexts 5, 143–164. doi: 10.1075/ttmc.00029.lop

Marx, N.. (2020). "Transfer oder Transversalität? – Designs zur Erforschung der Mehrschriftlichkeit," in Bulletin VALS/ASLA, (Neuchâtel: E-periodica), 15–33.

Marx, N., and Steinhoff, T. (2021). Können einzelsprachliche Interventionen sprachenübergreifende Effekte haben? Wie die schulische Majoritätssprache Herkunftssprachen fördern kann. *Z. Erziehungswiss.* 24, 819–839. doi: 10.1007/s11618-021-01032-5

May, S. (2014). The multilingual turn. Implications for SLA, TESOL and Bilingual Education. New York, London: Routledge. doi: 10.4324/9781003177197

Mayring, P. (2022). Qualitative Inhaltsanalyse: Grundlagen und Techniken. Weinheim, Basel: Beltz.

Melo-Pfeifer, S., and Ollivier, C. (2024). Assessment of plurilingual competence and plurilingual learners in educational settings. London: Routledge.

Niebert, K., and Gropengießer, H. (2014). "Leitfadengestützte Interviews", in Methoden in der naturwissenschaftsdidaktischen Forschung, ed. D. Krüger, I. Parchmann and H. Schecker (Berlin, Heidelberg: Springer Verlag), 121–132. doi: 10.1007/978-3-642-37827-0_10

Otheguy, R., García, O., and Reid, W. (2015). Clarifying translanguaging and deconstructing named languages: a perspective from linguistics. *Appl. Linguist. Rev.* 6, 281–307. doi: 10.1515/applirev-2015-0014

Pearson, B. Z., Fernández, S. C., and Oller, D. K. (1993). Lexical development in bilingual infants and toddlers: comparison to monolingual norms. *Lang. Learn.* 43, 93–120. doi: 10.1111/j.1467-1770.1993.tb00174.x

Rosebrock, A., Nix, D., Rieckmann, C., and Gold, A. (2011). Leseflüssigkeit fördern: Lautleseverfahren für die Primar- und Sekundarstufe. Seelze: Klett Kallmeyer.

Schissel, J. L., Leung, C., López-Gopar, M., and Davis, J. R. (2018). Multilingual learners in language assessment: assessment design for linguistically diverse communities. *Lang. Educ.* 32, 167–182. doi: 10.1080/09500782.2018.1429463

Seed, G. (2020). What is plurilingualism and what does it mean for language assessment? *Cambridge Assess Engl Res Notes.* 78, 5–15.

Shohamy, E. (2011). Assessing multilingual competencies: adopting construct valid assessment policies. *Mod. Lang. J.* 95, 418–429. doi: 10.1111/j.1540-4781.2011.01210.x

Shohamy, E., Or, I. G., and May, S. (2017). Language testing and assessment. Cham: Springer. doi: 10.1007/978-3-319-02261-1

Tunmer, W. E., and Chapman, J. W. (2012). The simple view of reading redux: vocabulary knowledge and the independent components hypothesis. J. Learn. Disabil. 45, 453–466. doi: 10.1177/0022219411432685

Van de Vijver, F., and Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: an overview. Eur. Rev. Appl. Psychol. 54, 119–135. doi: 10.1016/j.erap.2003.12.004

Vanhove, J., and Berthele, R. (2018). "Testing the interdependence of languages (HELASCOT project)," in Heritage and school language literacy development in migrant children: Interdependence or independence?, eds. R. Berthele and A. Lambelet (Bristol, Blue Ridge Summit: Multilingual Matters), 97–118. doi: 10.21832/9781783099054-007