DEVELOPMENT AND VALIDATION OF A TEST OF THINKING STRATEGIES IN GERMAN AS A FOREIGN LANGUAGE: LANGUAGE ASSESSMENT IN LOWER-SECONDARY SCHOOL STUDENTS

Martina Hoffmannová[1]; **Simona Tušicová**[2]

[1] Charles University, Faculty of Education, Department of Germanic Studies, Ovocný trh 560/5, 116 36 Praha 1, Czech Republic

[2] University of Hradec Králové, Faculty of Science, Department of Applied Cybernetics, Rokitanského 62, 500 03 Hradec Králové III., Czech Republic

e-mail: [1]martina.hoffmannova443@student.cuni.cz; [2]simona.tusicova@uhk.cz

Abstract

This study focuses on the development and validation of the test of thinking strategies in German as a foreign language. The main aim is to develop a valid and reliable instrument for assessing the thinking strategies of lower-secondary school children (aged 14–16) in the Hradec Králové region, Czech Republic. The test development process involved item analysis, pilot testing, and psychometric evaluation conducted between October and December 2024 on a sample of 69 respondents, to establish both item-level and overall test validity and reliability. The results indicate that the developed test is an appropriate assessment tool enabling professionals to assess and therefore better understand the needs of individual learners during the foreign language acquisition process.

Keywords

Educational measurement; Pilot study; Psychometric properties; Item analysis.

Introduction

Acquiring a foreign language is not only about the ability to communicate. In today's society, knowledge of a foreign language is also associated with numerous cognitive, economic, educational, and socio-cultural benefits. As globalization intensifies the need for intercultural interaction and international cooperation, understanding the value of foreign language learning has become a priority. Therefore, knowledge of a foreign language is a valuable asset that enhances communication, fosters cultural understanding, expands career opportunities, improves cognitive skills, and promotes personal growth.

However, teaching and learning a foreign language bring significant challenges due to the complexities of grammar, vocabulary, pronunciation, and cultural nuances. It is a multidimensional process influenced by biological, cognitive, social, and environmental factors. Unlike first-language acquisition, which typically occurs effortlessly in early childhood, learning a second or additional language often requires sustained effort, targeted instruction, and ongoing evaluation to ensure proficiency.

Assessing language proficiency is essential for providing a structured, effective, and personalized learning experience. It helps instructors guide their students more efficiently, ensuring steady progress and achieving desired language skills outcomes. Foreign language

© 2026 Author(s). This work is distributed under the Creative Commons Attribution-4.0 license (https://creativecommons.org/licenses/by-nc/4.0/).

knowledge assessment is traditionally based on measuring language skills in line with the communicative approach to language education. Receptive skills, such as reading and listening (Zhao & Aryadoust, 2024), and productive skills, such as speaking and writing, are most commonly assessed, with each area evaluated using specific assessment tools and criteria (Council of Europe, 2014). A new approach to implementing thinking strategies across various areas of education has emerged recently (Rottenhofer et al., 2022; Bilbao et al., 2018; Brečka & Valentová, 2017; Bauer et al., 2015). The new direction emphasizes multiple problem-solving strategies, and thinking strategies are becoming an important part, not only of research in the field of education, but also of education itself. However, there is a lack of research focused on the use of thinking strategies in German. Therefore, this study fills a gap in this research field.

The article describes the development and validation of a test of thinking strategies in German as a foreign language, presenting a new approach to assessing German language acquisition. Thinking strategies encompass a comprehensive set of cognitive and metacognitive skills and practices that enable individuals to plan, monitor, control, and evaluate their thinking and learning processes, including memory and comprehension (Schraw, 2011). For this text, they are understood as an essential component in meeting learning demands (Main, 2026). Focusing on assessment of thinking strategies provides a comprehensive overview of a student's abilities, including cognitive processes such as logical and algorithmic thinking, pattern recognition, decomposition, and creativity.

The primary objective of this study is to develop a valid and reliable language test for measuring thinking strategies in German as a foreign language among 9th-grade primary school students in the Hradec Králové region of the Czech Republic (aged 14-16). The article begins with a comprehensive literature review that synthesizes the main knowledge on test development, emphasizing psychometric properties such as content, construct, and criterion validity, as well as internal consistency and reliability. It then delineates the assessment domain using a theoretical framework that specifies targeted cognitive processes and knowledge dimensions. The methodology outlines research design, participant recruitment, and development of the measuring instrument. This is followed by the results section, which presents the data analysis procedures employed in the study and the findings of reliability. Finally, the discussion interprets the results in relation to existing literature and the study's aims. The conclusion summarizes the main findings, highlights implications for practice and future research, and discusses acknowledged limitations.

1 Literature Review

Although the role of individual differences in foreign language acquisition is well known, these signs of growing interest in the issue may reflect an overly optimistic view of the field's maturity (Teimouri et al., 2025). Research focused on individual learning approaches sometimes overlooks the understudied and underserved population. Plonsky (2023) examines the extent to which foreign language research draws samples from students of different ages, educational backgrounds, nationalities, native languages, and other characteristics, as indicated in the sample of meta-analyses and other types of synthetic research in applied linguistics. The issue of an individual approach in the language acquisition process plays a key role in the entire learning process and must receive sufficient attention. The development of testing tools that focus on individual differences plays a significant role in the individualization of the learning and teaching process; therefore, the test of thinking strategies was developed. Developing a testing tool presents numerous challenges, particularly regarding its validity and reliability.

The first important factor in the test development process is validating a testing tool. Validity is a crucial aspect that indicates whether a test measures what it is intended to measure. Validity

can be categorized into content, concurrent, predictive, and construct validity. It is essential not to rely solely on one type of validity during the test development process (Carr, 2011). In this study, content and construct validity were primarily verified. Expert judgment in language assessment is a long-established practice in test development, both in writing items and in estimating item difficulty, which affects item-level and scoring. In test construction, the use of an "expert" is critical (Coniam et al., 2022). It is a process of consulting with an expert in a particular field to obtain a qualified opinion. These can be external consultants or internal team members who have the necessary knowledge and experience.

Content validation through expert judgment, as defined by Escobar-Pérez and Cuervo-Martínez (2008), is an informed opinion from individuals with a track record in the field who are regarded by others as qualified experts and can provide information, evidence, judgments, and assessments. It involves ensuring that the test items correspond with the test's objectives. Content validity ensures that the assessed skills and competencies are accurately represented in the test, thereby generating valid scores (Asmar et al., 2025). In accordance with Galaczi (2020), the test items were designed to be diverse and cover all relevant language skills, thereby providing a comprehensive assessment of learners' abilities.

Construct validity is based on data collected during piloting, with a strong emphasis on the quality and quantity of the dataset. Construct validity is defined as

> "*the degree to which evidence and theory support the interpretation of test scores entailed by the proposed uses of the test*" (AERA et al., 2014, p. 11).

It combines theoretical and psychometric ideas, thereby verifying the validity of the theory underlying the test (Pelikán, 2011). In comparison to content validity, construct validity is established through the logical and mathematical analysis of test items (Finocchiaro & Sako, 1983). Construct validity plays a crucial role in research by ensuring that the measures developed accurately reflect the theoretical concepts they are intended to assess. Construct validity is theory-based and iterative, beginning with initial theoretical predictions and developing measures that are continually refined through empirical testing (Teimouri et al., 2025).

As part of the validation process, the difficulty and sensitivity of the test items are analyzed to determine their properties. The analysis of difficulty indicates the number of respondents in the sample who answered the item incorrectly or skipped it (Chráska, 2016). The sensitivity of a task reflects the extent to which the task favors candidates with higher scores over those with lower scores. Another important part of the validation process is expert judgment. Experts are chosen based on their proven engagement and depth of experience in the field, as evidenced by their academic background, professional achievements, or recognized authority within the community (Coniam et al., 2022).

The second important factor of a testing tool is its reliability. Reliability is understood as the degree to which a test can be trusted to provide consistent information about the value of a learning variable being measured (Finocchiaro & Sako, 1983). In language assessment, reliability refers to the extent to which a test consistently and accurately measures the same underlying construct over time. To determine the reliability of a language test, several methods can be employed (Shang et al., 2024). As recommended by Finocchiaro and Sako (1983), the split-half method and inter-item reliability were chosen. The split-half method involves comparing the scores of odd and even items of the same test. At the same time, inter-item reliability is based on comparing the scores of examinees who passed each test item with those who did not (Shang et al., 2024). Inter-item reliability (internal consistency) refers to the consistency of responses across items within a test or scale designed to measure the same construct. High inter-item reliability indicates that the items are highly correlated with each

other. Several factors influence test reliability, including test type, test length, number of items, item difficulty and discrimination, item type, interaction between test takers and tests, and test scoring (Shang et al., 2024). Comprehensiveness refers to the test containing an appropriate proportion of items relevant to its intended purpose. To ensure comprehensiveness, it is recommended to evaluate both the number of items and the test's overall length, ensuring that all relevant areas are adequately covered. Practicality of a test is determined by its suitability for large-scale use. Practical considerations include financial and personnel costs for distribution, ease of scoring, length of instructions, and the overall appearance of the test (Finocchiaro & Sako, 1983). A practical test can be efficiently administered and scored, even in large-scale settings, without imposing high costs or logistical burdens.

Educational diagnostics is a scientific discipline that requires professional assessment competence that can serve multiple purposes. In addition to the school's pedagogical responsibilities, the process of knowledge transmission, selection, socialization, and the learners' social situation must also be observed during instruction. In this article, this area is defined, in the words of Herzog et. al. (2012, pp. 3–4), as

> "*the clarification of issues related to diagnostic judgment in educational fields (particularly in schools), as well as the provision of criteria and tools to optimize the process of educational judgment*".

In this article, a diagnostic language test is defined as an assessment tool designed primarily to identify learners' overall strengths and weaknesses and pinpoint the root causes of these weaknesses. Its ultimate goal is to promote further learning and enhance learners' overall growth potential through diagnostic feedback and remedial activities (Lee, 2015). When developing a foreign language test, factors such as validity, reliability, comprehensiveness, and practicality must be considered to create a standardized assessment tool (Finocchiaro & Sako, 1983).

In this study, it is assumed that all thinking is linguistic thinking (Chomsky, 2016). Specifically, in foreign language acquisition, it is known that all learners employ different language learning strategies during the learning process[1] (Hismanoglu, 2000). From this perspective, language acquisition plays a significant role in enhancing individuals' cognitive abilities. This notion is supported, for example, by Willmes et al. (2014, as cited in Fritz et al., 2021, p. 57), who noted that

> "*language and number processing share a number of crucial functional characteristics at the brain level, such as an overlapping ventral network, at least for specific cognitive processes such as semantic classification*".

Given this interrelationship, foreign language acquisition needs to be examined within the context of diverse cognitive skills, not only in linguistics. Experimental investigations into language use can yield valuable insights into the representation and nature not only of grammatical knowledge (Wiese, 1989) but of the human thought process.

The American Psychological Association (APA, 2018a) defines cognitive ability as

> "*the skills involved in performing the tasks associated with perception, learning, memory, understanding, awareness, reasoning, judgment, intuition, and language*".

[1] Language learning strategies are defined as specific actions, behaviors, tactics, or techniques that facilitate the learning of the target language by the language learner.

For test development the components of cognitive skills based on Bloom's taxonomy[2] have been used.

Table 1 describes cognitive dimensions, and Table 2 describes knowledge dimensions.

Tab. 1 *Cognitive dimensions*

Cognitive dimension	Definition	Characteristics activity
Remember	Recognizing or recalling knowledge, facts, or concepts.	Define, describe, identify, know, label, list, match, outline, recall, recognize, reproduce, select, state, and locate.
Understand	Explaining ideas or concepts, determining the meaning of instructional messages, including oral, written, and graphic communication.	Illustrate, defend, compare, distinguish, estimate, explain, classify, generalize, interpret, paraphrase, predict, rewrite, summarize, and translate.
Apply	Carrying out or using a procedure in a given situation.	Implement, organize, solve, construct, demonstrate, discover, manipulate, modify, operate, predict, prepare, relate, show, solve, and choose.
Analyze	Breaking material into its constituent parts and detecting how the parts relate to one another and to an overall structure.	Analyze, break down, compare, select, contrast, deconstruct, discriminate, distinguish, identify, and outline.
Evaluate	Making judgments based on criteria and standards.	Rank, assess, monitor, check, test, and judge.
Create	Putting elements together to form a novel, coherent whole or make an original product.	Generate, plan, compose, develop, create, invent, organize, construct, produce, compile, design, and devise.

Source: Own based on (O'Reilly, 2021; Anderson & Krathwohl, 2001)

Tab. 2 *Knowledge dimensions*

Knowledge dimension	Definition
Factual knowledge	Basic elements used to communicate, understand, and organize a subject: terminology, scientific terms, labels, vocabulary, symbols, and specific details such as knowledge of events, people, dates, and sources of information.
Conceptual knowledge	Knowledge of classifications and categories, principles, theories, models, or structures of a subject.
Procedural knowledge	How to do something; methods of inquiry, and criteria for using skills, algorithms, techniques, and methods.
Metacognitive knowledge	The process or strategy of learning and thinking, an awareness of one's own cognition, and the ability to control, monitor, and regulate one's own cognitive process.

Source: Own based on (O'Reilly, 2021; Multhaup, 1997)

2 Methodology

This study employs a methodological approach focused on developing and pilot-testing a high-quality assessment tool. The methodology includes the definition of key constructs in educational assessment and psycholinguistics, which served as the basis for designing the test items. In developing the test, standard quality assessment requirements were considered, particularly validity, reliability, comprehensiveness, and practical applicability. Attention is

[2] The list of cognitive domains is often used to structure learning objectives, assessments, and activities. It was intended as a tool to facilitate the development of test items that measure the same educational objective. (Krathwohl, 2002)

also paid to the mental processes associated with foreign language acquisition and their placement within Bloom's Taxonomy of Educational Objectives (Krathwohl, 2002). This study aims to develop a robust diagnostic language test that can effectively identify students' strengths and weaknesses.

In this study, the authors have specifically selected cognitive processes relevant to foreign language learning at the school level. The individual exercises in the developed test are thematically divided into five categories, each focusing on a specific cognitive skill. Based on a comprehensive review of the literature, the authors selected five cognitive skills to represent these categories. The test categories are defined in Table 3.

Tab. 3 *Test category*

Test category	Definition
Logical thinking	A rational process of the brain by which people find correct conclusions. Logic reasoning, a central cognitive component, depends on theories of understanding, memory, learning, visual perception, planning, problem-solving, and decision-making.
Algorithmic thinking	A pool of abilities that are connected to constructing and understanding algorithms: the ability to analyze given problems; the ability to specify a problem precisely; the ability to find the basic actions that are adequate to the given problem; the ability to construct a correct algorithm to a given problem using the basic actions; the ability to think about all possible special and regular cases of a problem; the ability to improve the efficiency of an algorithm.
Pattern recognition	A process of identifying patterns in a data set to categorize, process, and resolve the information more effectively. Patterns are pieces or sequences of data that have one or multiple similarities.
Decomposition	Breaking down problems into smaller, more manageable parts and then focusing on solving each of these more minor problems.
Creativity	The ability to produce or develop original work, theories, techniques, or thoughts. It refers to the mental processes leading to a new invention or solution to a problem.

Source: Own based on (Serna & Serna, 2015; Futschek, 2006; Lcom Team, 2022; Bell & Lodi, 2019; APA, 2018b)

The research focused on a target group consisting of lower secondary school students aged 14–16, who were in 9th grade and learning German as a foreign language. A random sampling was conducted. Initially, 100 students in the Hradec Králové region of the Czech Republic were invited to participate, and 69 of them ultimately took part in the testing. Pilot testing, a key element in the test design phase that contributes to improving the validity and reliability of a testing tool (Zhao & Aryadoust, 2024), and data collection were carried out from October to December 2024. A conceptual framework for this research was initially developed and progressively refined and specified throughout the study. Table 3 outlines the five components and their definitions used to determine individual test sections.

With this framework in mind, an initial pool of 14 potential items was created, either self-created or inspired by teaching materials (e.g., exams, learning activities, textbooks) gathered through the experience in teaching foreign languages, for example, Momente (Evans et al., 2021), Beste Freunde (Georgiakaki et al., 2014), Team Deutsch Neu (Esterl et al., 2021), Sicher Übungsgrammatik (Hering et al., 2017), Passt schon! (Dusilová et al., 2014), Menschen (Evans et al., 2020).

It was ensured that all items met predetermined criteria. The criteria were as follows:

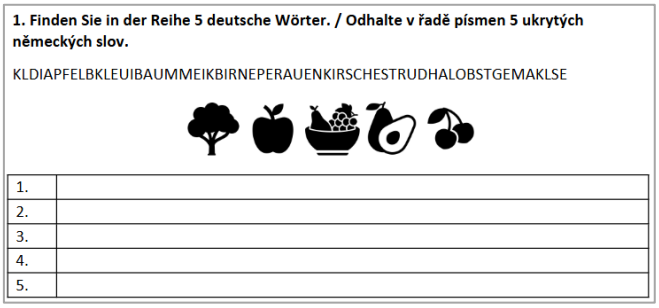
1. All items should fit into one or more test categories.

2. The difficulty should be optimal for foreign language learners at the A1 level (Council of Europe, 2014).
3. The pool of items should reflect the diversity of components and contents.
4. The items do not measure the level of foreign language proficiency; instead, their subtext measures the level of cognitive skills of the respondents.

When designing the test items, emphasis was placed on differentiating the items by type, including both open-ended and closed-ended items with one or more correct answers.

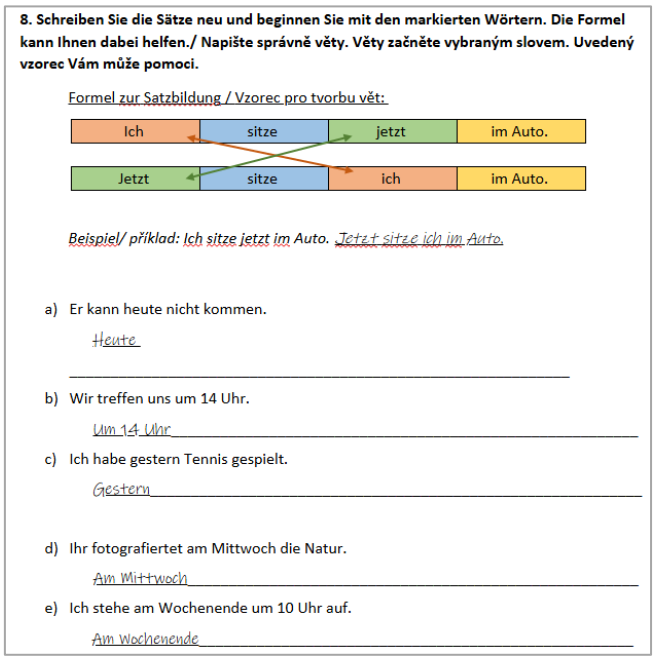
The items were designed with the target group in mind. Accuracy, appropriacy, range, flexibility, and size played an important role (Hughes & Hughes, 2020). Bloom's taxonomy was also considered when placing the test items in the test (Krathwohl, 2002).

Figures 1 and 2 illustrate examples of test items: Figure 1 shows decomposition, and Figure 2 demonstrates the use of an algorithm.



Source: Own

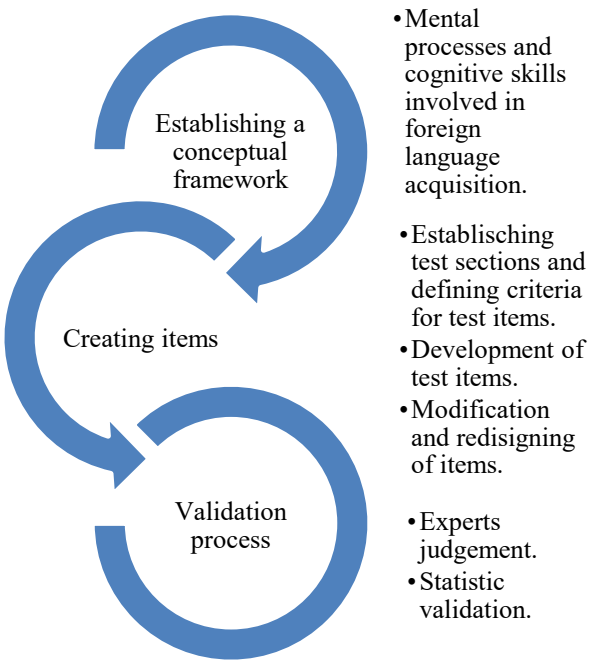
Fig. 1: *Example of a test item focusing on decomposition*



Source: Own

Fig. 2: *Example of a test item focused on the use of an algorithm*

Then, based on the above-mentioned criteria, 10 final items were selected. Once all the exercises and their items were finalized and the test was constructed, the test validation process was initiated, see Figure 3.



Source: Own

Fig. 3: *Study process schema*

The validation process was conducted in two sequential steps. Initially, the construct validity was assessed through expert evaluation of the Test. Following this, a pilot test was conducted to further validate the instrument. Pelikán (2011) emphasizes that pilot testing is a crucial component in establishing the robustness of a new research tool. In this phase, the Test was administered to a small sample that shared key characteristics with the target research population. The data obtained from this pilot testing were then analyzed to assess further and refine the construct validity of the Test, and subsequent expert-judgment validation followed.

During the pilot study three research questions (RQ) were defined and subsequently evaluated. The research questions are defined as follows:

RQ1: To what extent does the newly developed Test of thinking strategies in German as a foreign language meet fundamental psychometric standards required for educational assessment?

RQ2: How well does the Test of Thinking Strategies demonstrate psychometric validity and reliability in measuring the intended selected cognitive processes relevant to foreign language learning?

RQ3: How effectively does the Test of Thinking Strategies function as a suitable and valuable instrument for the assessment of students' cognitive processes relevant to foreign language learning?

In this study, eight experts in the field were asked to evaluate the measuring tool: educational psychologists, German language experts, a measurement expert, and subject-matter experts in the educational context. This type of validation was conducted during item design and piloting. Therefore, not only experts from academia but also experts from practice were selected. IBM SPSS Statistics and Microsoft Excel were used to analyze the data.

Cronbach's alpha and the Spearman-Brown formula were used to estimate the test's reliability. Cronbach's alpha was used to assess the degree to which the items consistently measured the same underlying construct. An exploratory factor analysis was conducted to assess the test's construct validity.

Additionally, split-half reliability was calculated using the Spearman-Brown formula to assess the consistency between two equivalent halves of the instrument and to account for test length. The joint application of these reliability estimates yields convergent evidence of internal consistency and contributes to a more robust and methodologically rigorous validation of the instrument. The reliability coefficient was calculated using the split-half method. The reliability coefficient, denoted by the letter *r*, is expressed as a value between 0 and 1, where *r* equal to 0 indicates no reliability and *r* equal to 1 indicates perfect reliability. The higher the value, the more stable, consistent, and free from random measurement errors the test results are (Chráska, 2016).

Item difficulty, expressed on a scale from 0 to 1, represents the proportion of respondents who correctly answer a given test item. Lower values indicate greater difficulty, whereas higher values indicate lower difficulty (Shang et al., 2024). For difficulty and sensibility analysis, the ULI index was calculated. This index is calculated based on the difference in task difficulty between respondents with higher and lower performance. The coefficient is recommended to take values of at least 0.25, respectively 0.15 (Chráska, 2016).

The test administration itself plays a crucial role in the testing process, including standardized conditions, clear instructions, a controlled environment, appropriate equipment and materials, monitoring and supervision, accommodations for special needs, and a post-test protocol. The guidelines before, during, and after testing cannot be forgotten. These guidelines enhance consistency, ensure test security, and safeguard the integrity of fair and reliable test results (Hughes & Hughes, 2020).

3 Results

The results of descriptive statistics are shown in Table 4.

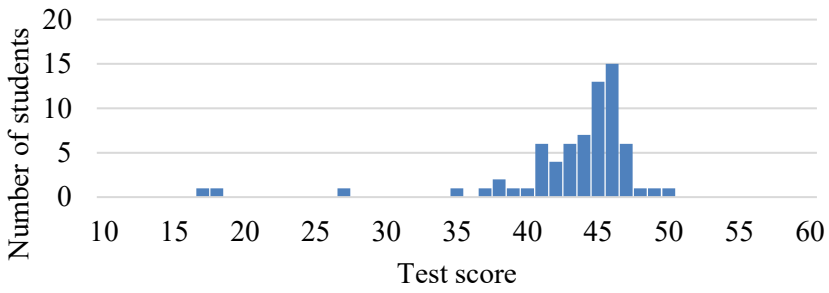
Tab. 4 *Descriptive statistics*

Number of respondents	**Valid**	69
	Missing	0
Mean		43.087
Median		45
Standard deviation		5.635
Skewness		-3.165
Standard error of skewness		0.289
Kurtosis		11.841
Standard error of kurtosis		0.570
Minimum score		17
Maximum score		50
Percentiles	**25**	42
	50	45
	75	46

Source: Own using IBM SPSS Statistics

The total number of respondents is 69. The minimum score is 17; the maximum is 50 out of 60. The mean of the test score is 43.087.

The histogram in Figure 4 shows the distribution of the test scores. Based on the data obtained, the test scores follow a normal distribution.



Source: Own using Microsoft Excel

Fig. 4: *Test distribution histogram*

Descriptive statistics for item means are described in Table 5. Descriptive statistics for each item are presented in Table 6. The item worth emphasizing is 4, which has a high number of missing values. Regarding the other items, the descriptive statistics results are acceptable.

Tab. 5 *Item means*

	Mean	Minimum	Maximum	Range	Maximum/ Minimum	Variance	Number of items
Item means	4.658	0.842	9.667	8.825	11.479	6.240	10

Source: Own using IBM SPSS Statistics

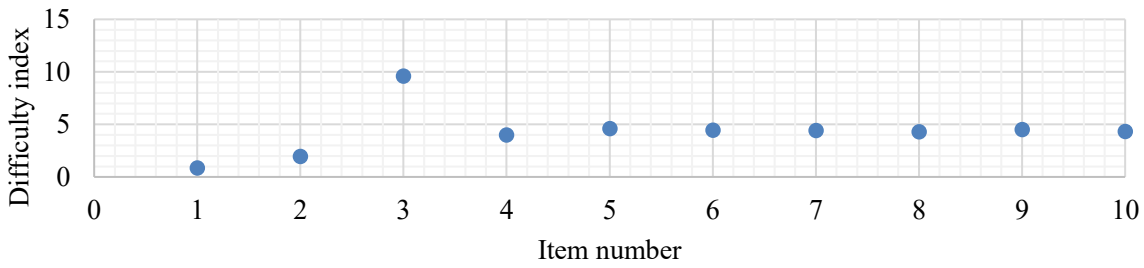
The item difficulty index is shown in Figure 5. The average success rate per item is 4.658, see Table 5. The minimum success rate is achieved in item 1 ($\bar{x}_1 = 0.855$; an item with very great difficulty), while item 3 is, on the other hand, the item with very low difficulty ($\bar{x}_3 = 0.623$). By items 5–10, the difficulty is comparable. For details see Table 6.

Tab. 6 *Descriptive statistics of each test item*

		Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Number of respondents	**Valid**	69	69	69	58	69	69	69	69	68	69
	Missing	5	5	5	16	5	5	5	5	6	5
Mean		0.855	1.971	9.623	3.897	4.609	4.449	4.435	4.304	7.441	4.333
Median		1	2	10	4	5	5	5	5	8	5
Standard deviation		0.355	0.84	0.824	0.583	1.191	1.119	1.356	1.102	2.229	1.291
Minimum		0	1	7	2	0	0	0	0	0	0
Maximum		1	6	10	5	5	5	5	5	9	5
Percentile	**25**	1	2	10	4	5	4	5	4	7	4
	50	1	2	10	4	5	5	5	5	8	5
	75	1	2	10	4	5	5	5	5	9	5

Source: Own using IBM SPSS Statistics

The higher level of missing responses for Item 4 may be related to its greater cognitive complexity, which may have led some respondents to omit it. In the statistical analyses, missing responses were handled as missing values and appropriately accounted for in the computation of item statistics and reliability coefficients.



Source: Own using Microsoft Excel

Fig. 5: *Item difficulty index*

The reliability coefficients indicate that the test exhibits good reliability, see Tables 7 and 8. The Cronbach's alpha of 0.696 and the Spearman-Brown coefficient of 0.796 are the results that can be considered satisfying in the context of test development.

Tab. 7 *Cronbach's Alpha of the test as a whole*

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	Number of Items
0.696	0.621	10

Source: Own using IBM SPSS Statistics

Tab. 8 *Reliability index of the test*

Cronbach's Alpha	**Part 1**	**Value**	0.591
		Number of items	5
	Part 2	**Value**	0.468
		Number of items	5
	Total number of items		10
Correlation between forms			0.661
Spearman-Brown coefficient	**Equal length**		0.796
	Unequal length		0.796
Guttman split-half coefficient			0.795

Source: Own using IBM SPSS Statistics

Before conducting the exploratory factor analysis, the data's suitability was assessed. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy reached 0.726, indicating acceptable sampling adequacy and confirming that the correlation matrix was appropriate for factor analysis.

Exploratory factor analysis was subsequently performed using Principal Component Analysis as the extraction method with Varimax rotation and Kaiser normalization. The analysis yielded a five-factor solution corresponding to the theoretical constructs defined during test development. These tests support the instrument's construct validity, as the identified factor structure aligns with the theoretical framework and indicates that the test measures the intended constructs.

Based on statistical analysis of the empirical data obtained, the newly developed Test of testing strategies in German as a foreign language meets the basic psychometric standards required for educational assessment. The analyses confirm the measurement tool's adequate reliability and the fulfillment of the criteria for content and construct validity.

The results provide preliminary evidence of acceptable internal consistency and measurement stability of the Test. They further suggest that the instrument captures certain aspects of thinking strategies involved in foreign language acquisition, as well as selected components of students'

cognitive processes relevant to foreign language learning. At this stage, the Test is a promising instrument with potential for exploratory use and possible screening in educational contexts. Further studies involving larger, more diverse samples will be necessary to confirm its psychometric properties and determine its broader applicability. Overall, the results offer tentative support for the research questions addressed in this study.

4 Discussion

The test presented in this publication was designed to support the principle of an individualized approach to education. Unlike traditional language tests that focus primarily on linguistic knowledge and performance, this instrument shifts attention to the cognitive processes that underpin language comprehension, production, and problem-solving. Over the past decade, numerous studies have examined thinking strategies and their application in education. Rottenhofer et al. (2022) examine the use of computational thinking to support language teaching, specifically English and Spanish; Bilbao et al. (2018) examine the use of thinking strategies in individuals with special educational needs; Brečka & Valentová (2017) deal with thinking strategies in the teaching of technical subjects; Bauer et al. (2015) describe the implementation of thinking strategies in educational games. These studies demonstrate that the role of thinking strategies in education is gaining recognition.

The developed test of strategic thinking processes in German as a foreign language meets the basic psychometric standards required for educational assessment which is the answer to RQ1.

The data analyses indicate that the Test of Thinking Strategies demonstrates adequate reliability and meets the criteria for content and construct validity, with sufficient internal consistency and stability, thereby enabling the valid and reliable assessment of thinking strategies in foreign language learning and related selected cognitive processes which is the answer to RQ2.

The presented test may represent a useful instrument for assessing selected cognitive processes relevant to foreign language learning. Beyond elements related to linguistic knowledge, it aims to capture higher-order processes such as reasoning, problem-solving, planning, and the use of cognitive strategies. The results suggest that the test may offer insights into aspects of students' cognitive functioning relevant to educational assessment, screening, and the individualization of teaching which is the answer to RQ3.

The results align with current theories of language acquisition, which emphasize the role of metacognition and strategic learning in the development of communication skills. From a broader perspective, this study contributes to the field of foreign language assessment by introducing a new approach that integrates the assessment of cognitive strategies into German-as-a-foreign-language testing. It also addresses a gap in current assessment tools available for secondary school students in non-German-speaking contexts. Typically, assessment tools used in German as a foreign language at the lower secondary level employ a combination of formative and summative tools. These include written tests that assess vocabulary, grammar, reading, and writing skills, as well as oral assessments and listening tasks. In addition, continuous classroom-based assessment, portfolios, project work, and self- and peer-assessment tools are used to monitor learners' progress and support communicative competence. However, there is currently no assessment tool that focuses on cognitive strategies. The innovative nature of this tool lies in its ability to link language-level measurement with the assessment of cognitive strategies, thereby enriching understanding of student variability.

The limitations during the test development and validation process may be associated with data sensitivity, test takers' ages, and deviations in statistical data processing. Also, given the exploratory nature of the study, the results should be interpreted with caution and require further empirical validation.

Conclusion

Modern education increasingly emphasizes the individual's needs, and society recognizes that every student has a unique learning pace and distinct abilities. Individualized teaching is therefore becoming a key element in enabling students to develop their full potential. Thanks to modern technology, teachers can better monitor each student's progress and create personalized lesson plans. Traditional classroom teaching no longer fully meets the diverse needs of today's students. The role of the teacher is changing; they are no longer just lecturers but guides on an educational journey. Project-based methods, teamwork, and independent thinking are becoming increasingly prevalent. Students are encouraged to take greater responsibility for their own education, which helps them prepare for real life.

This approach also includes the use of a test focused on thinking strategies, which may help better understand individual needs and ways of thinking. The information obtained encourages individual approaches to each person and can improve the effectiveness of the educational process. The aim is to create an environment where every student can feel respected, supported, and motivated to develop personally. Education conceived in this way not only develops knowledge and skills but also contributes to the formation of personality and prepares young people for life in an ever-changing world.

Based on the obtained and evaluated data, it can be claimed that a reliable instrument has been successfully developed to measure the level of thinking involved in foreign language acquisition and concrete strategies in German as a foreign language. The primary contribution of the test is to provide a deeper understanding of the mental processes involved in foreign language acquisition. The results of this study can be beneficial for investigating individual learning needs, thereby optimizing and enhancing the learning process. Further studies focusing on the implementation of thinking strategies in foreign language teaching may follow this research and examine their impact on the educational process.

Acknowledgements

The authors thank the experts for their valuable comments during the validation process, as well as the schools that participated willingly in both the pilot study and the primary research.

References

AERA, APA, & NCME. (2014). *The Standards for Educational and Psychological Testing*. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. ISBN 978-0-935302-35-6. https://www.testingstandards.net/open-access-files.html

Anderson, L. W., & Krathwohl, D. R. (2001). Remembering and understanding: The cognitive process dimension. In L. W. Anderson & D. R. Krathwohl (Eds.), *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (pp. 27–38). Longman. ISBN 0-8013-1903-X.

APA. (2018a). *Cognitive ability*. In APA Dictionary of Psychology. https://dictionary.apa.org/cognitive-ability

APA. (2018b). *Creativity*. In APA Dictionary of Psychology. https://dictionary.apa.org/creativity

Asmar, N., Youssef, N. Y., Malhab, S. B., Bahous, S., & Karam, V. G. Y. (2025). Evaluating construct validity of virtual osces in exceptional conditions. *BMC Medical Education*, *25*, 841. https://doi.org/10.1186/s12909-025-07383-5

Bauer, A., Butler, E., & Popović, Z. (2015). Approaches for teaching computational thinking strategies in an educational game: A position paper. In 2015 IEEE Blocks and Beyond Workshop (Blocks and Beyond) (pp. 121–123). IEEE. ISBN 978-1-4673-8367-7. https://doi.org/10.1109/blocks.2015.7369019

Bell, T., & Lodi, M. (2019). Constructing Computational Thinking Without Using Computers. *Constructivist Foundations*, 14(3), 342–351. https://123dok.net/document/q2nke36q-constructing-computational-thinking-without-using-computers.html

Bilbao, N., Goioaga, O., Latorre, I., & Perea, F. (2018). Thinking Strategies and Special Educational Needs. In *12th International Technology, Education and Development Conference Proceedings* (pp. 3817–3821). IATED. https://doi.org/10.21125/inted.2018.0742

Brečka, P., & Valentová, M. (2017). Critical and Creative Thinking Strategies in Teaching Technical Subjects. In *10th annual International Conference of Education, Research and Innovation Proceedings* (pp. 1565–1571). IATED Academy. https://doi.org/10.21125/iceri.2017.0490

Carr, N. T. (2011). *Designing and Analyzing Language Tests*. Oxford University Press. ISBN 978-0-19-442297-0.

Coniam, D., Lee, T., Milanovic, M., Pike, N., & Zhao, W. (2022). Role of expert judgement in language test validation. *Language Education & Assessment*, *5*, 18–33. https://doi.org/10.29140/lea.v5n1.769

Council of Europe. (2014). *Common European Framework of Reference for Languages (CEFR)*. https://www.coe.int/en/web/common-european-framework-reference-languages

Dusilová, D., Kolocová, V., Haupenthal, T., & Krüger, J. (2014). *Passt schon! 1*. Polyglot. ISBN 978-80-86195-89-6.

Escobar-Pérez, J., & Cuervo-Martínez, A. (2008). Validez de contenido y juicio de expertos: Una aproximación a su utilización. *Avances en Medición*, *6*, 27–36. https://www.humanas.unal.edu.co/lab_psicometria/application/files/9416/0463/3548/Vol_6._Articulo3_Juicio_de_expertos_27-36.pdf

Esterl, U., Körner, E., Einhorn, Á., Kubicka, A., Andrášová, H., Beňová, Z., & Jenkins, E.-M. (2021). *Team Deutsch neu*. Klett. ISBN 978-80-7397-370-4.

Evans, S., Pude, A., & Specht, F. (2020). *Menschen A1 Kursbuch*. Max Hueber Verlag. ISBN 978-3-19-211901-9.

Evans, S., Specht, F., & Pude, A. (2021). *Momente A1 Kursbuch plus interaktive Version*. Max Hueber Verlag. ISBN 978-3-19-501791-6.

Finocchiaro, M., & Sako, S. (1983). *Foreign Language Testing: A Practical Approach*. New York: Regents Publishing. ISBN 0-88345-362-2.

Fritz, A., Gürsoy E., & Herzog, M. (Eds.). (2021). *Diversity Dimensions in Mathematics and Language Learning*. De Gruyter. https://doi.org/10.1515/9783110661941

Futschek, G. (2006). Algorithmic Thinking: The Key for Understanding Computer Science. In Mittermeir R. T. (Ed.), *Informatics Education – The Bridge between Using and Understanding Computers*. ISSEP 2006. Lecture Notes in Computer Science, vol. 4226. (pp. 160–168). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11915355_15

Galaczi, E. (2020). *What is Validity?* Cambridge University Press & Assessment. https://www.cambridgeenglish.org/blog/what-is-validity/

Georgiakaki, M., Bovermann, M., Graf-Riemann, E., & Seuthe, C. (2014). *Beste Freunde A1.1*. Max Hueber Verlag. ISBN 978-3-19-101058-4.

Hering, A., Matussek, M., & Perlmann-Balme, M. (2017). *Sicher! Deutsch als Fremdsprache Übungsgrammatik*. Max Hueber Verlag. ISBN 978-3-19-301206-7.

Herzog W. et al. (2012). *Pädagogische Diagnostik*. https://www.walterherzog.ch/lehre/virtueller-campus/

Hismanoglu, M. (2000). Language Learning Strategies in Foreign Language Learning and Teaching. *The Internet TESL Journal*, *6*(8). http://iteslj.org/Articles/Hismanoglu-Strategies.html

Hughes, A., & Hughes, J. (2020). *Testing for Language Teachers* (3rd ed.). Cambridge University Press. https://doi.org/10.1017/9781009024723

Chomsky, N. (2016). The language capacity: architecture and evolution. *Psychonomic Bulletin & Review*, *24*, 200–203. https://doi.org/10.3758/s13423-016-1078-6

Chráska, M. (2016). *Metody pedagogického výzkumu: Základy kvantitativního výzkumu* (2nd ed.). Grada Publishing. ISBN 978-80-247-5326-3.

Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Lcom Team. (2022). *Pattern Recognition in Computational Thinking*. Learning.com. https://www.learning.com/blog/pattern-recognition-in-computational-thinking/

Lee, Y.-W. (2015). Diagnosing diagnostic language assessment. *Language Testing*, *32*(3), 299–316. https://doi.org/10.1177/0265532214565387

Main, P. (2026). *Thinking Strategies: A Teacher's Guide*. Structural Learning. https://www.structural-learning.com/post/thinking-strategies-a-teachers-guide

Multhaup, U. (1997). Mental networks, procedural knowledge and foreign language teaching. *Language Awareness*, *6*(2–3), 75–92. https://doi.org/10.1080/09658416.1997.9959919

O'Reilly, L. (2021). *Bloom's Revised Taxonomy Worksheet*. University of Colorado Denver. https://www.ucdenver.edu/Docs/Librariesprovider279/Default-Document-Library/Blooms_taxonomy_worksheet6cfe02e6302864d9a5bfff0a001ce385.Pdf

Pelikán, J. (2011). *Základy empirického výzkumu pedagogických jevů* (2nd ed.). Praha: Karolinum. ISBN 978-80-246-1916-3.

Plonsky, L. (2023). Sampling and Generalizability in Lx Research: A Second-Order Synthesis. *Languages*, *8*(1), 75. https://doi.org/10.3390/languages8010075

Rottenhofer, M., Kuka, L., Leitner, S., & Sabitzer, S. (2022). Using Computational Thinking to Facilitate Language Learning: A Survey of Students' Strategy Use in Austrian Secondary Schools. *IAFOR Journal of Education*, *10*(2), 51–70. https://doi.org/10.22492/ije.10.2.03

Serna, M. E. & Serna, A. A. (2015). Knowledge in Engineering: A View from the Logical Reasoning. *International Journal of Computer Theory and Engineering*, *7*(4), 325–331. https://doi.org/10.7763/ijcte.2015.v7.980

Shang, Y., Aryadoust, V., & Hou, Z. (2024). A Meta-Analysis of the Reliability of Second Language Listening Tests (1991–2022). *Brain Sciences*, *14*(8), 746. https://doi.org/10.3390/brainsci14080746

Schraw, G. (2011). *Assessment of Thinking in Educational Settings*. Oxford Bibliographies. https://doi.org/10.1093/obo/9780199828340-0062

Teimouri, Y., Sudina, E., & Plonsky, L. (2025). Individual differences in language learning and teaching: Growing pains at the theory-methods interface. *Learning and Individual Differences*, *124*, 102786. https://doi.org/10.1016/j.lindif.2025.102786

Wiese, R. (1989). Psycholinguistik der Sprachproduktion. In Antos, G. & Krings, H. P. (Eds.), *Textproduktion: Ein interdisziplinärer Forschungsüberblick* (pp. 197–219). Berlin, Boston: Max Niemeyer Verlag. https://doi.org/10.1515/9783110962109.197

Willmes, K., Moeller, K., & Klein, E. (2014). Where numbers meet words: A common ventral network for semantic classification. *Scandinavian Journal of Psychology*, *55*(3), 202–211. https://doi.org/10.1111/sjop.12098

Zhao, H., & Aryadoust, V. (2024). A meta-analysis of the reliability of second language reading comprehension assessment tools. *Studies in Second Language Acquisition*, *47*(1), 388–416. https://doi.org/10.1017/s0272263124000627

Vývoj a validace testu myšlenkových strategií v němčině jako cizím jazyce: Hodnocení jazyka u žáků druhého stupně základních škol

Tato studie se zaměřuje na tvorbu a validaci testu myšlenkových strategií v němčině jako cizím jazyce. Hlavním cílem předložené studie je vytvořit validní a spolehlivý nástroj pro hodnocení myšlenkových operací žáků druhého stupně základních škol (věk 14–16 let) v regionu Hradec Králové, v České republice. Proces vývoje testu zahrnoval analýzu testových položek, pilotní testování a psychometrické hodnocení, které proběhlo v období od října do prosince 2024 na vzorku 69 respondentů s cílem ověřit platnost a spolehlivost testu jak na úrovni jednotlivých položek, tak v celkovém měřítku. Výsledky ukazují, že vyvinutý test je vhodným hodnotícím nástrojem, který odborníkům umožňuje posoudit, a tím i lépe porozumět, potřebám jednotlivých žáků v procesu nabývání cizího jazyka.

Entwicklung und Validation von Test über Denkstrategien in Deutsch als Fremdsprache: Bewertung der Sprachkenntnisse von SchülerInnen der Sekundarstufe I

Die vorliegende Studie befasst sich mit der Entwicklung und Validierung vom Test der Denkstrategien in Deutsch als Fremdsprache. Das Hauptziel der präsentierten Studie ist die Entwicklung eines validen und reliablen Instruments zur Bewertung der Denkstrategien von Schülerinnen und Schülern der Sekundarstufe I (Alter 14-16 Jahre) in der Region Hradec Králové, Tschechische Republik. Der Testentwicklungsprozess umfasste eine Item-Analyse, Pilotversuche und eine psychometrische Auswertung, die zwischen Oktober und Dezember 2024 an einer Stichprobe von 69 Befragten durchgeführt wurden, um sowohl die Validität und Reliabilität auf Item-Ebene als auch die Gesamtvalidität und -reliabilität des Tests zu ermitteln. Die Ergebnisse zeigen, dass der Test ein geeignetes Bewertungsinstrument ist, das Lehrkräften ermöglicht, die Bedürfnisse einzelner Lernenden während des Fremdsprachenerwerbsprozesses zu bewerten und so besser zu verstehen.

Opracowanie i walidacja testu ukierunkowanego na strategie myślenia w języku niemieckim jako języku obcym: Ocena znajomości języka uczniów szkół wyższych roczników szkół podstawowych

Niniejsze badanie skupia się na opracowaniu i walidacji testu strategii myślenia w języku niemieckim jako języku obcym. Głównym celem przedstawionych badań jest opracowanie ważnego i rzetelnego narzędzia do oceny strategii myślenia uczniów wyższych roczników szkół podstawowych (w wieku 14–16 lat) w regionie Hradec Králové w Czechach. Proces opracowywania testu obejmował analizę pozycji, testy pilotażowe oraz ocenę psychometryczną przeprowadzoną w okresie od października do grudnia 2024 r. na próbie 69 respondentów, co miało na celu ustalenie zarówno trafności i rzetelności poszczególnych pozycji, jak i całego testu. Wyniki wskazują, że opracowany test jest odpowiednim narzędziem oceny, umożliwiającym specjalistom ocenę, a tym samym lepsze zrozumienie potrzeb poszczególnych uczniów w procesie przyswajania języka obcego.