6. *Griva E., Deligianni A. CLIL implementation in foreign language contexts: Exploring challenges and perspectives – Part II // Research Papers in Language Teaching and Learning. – 2017. – Vol. 8. – No 2. – P. 63-73.*

7. *Mehisto P. Criteria for Producing CLIL Learning Material // Encuento. – 2012. – 15-33 p.*

8. *Marsh D., Maljers A., Hartiala A. K. Profiling European CLIL Classrooms //Jyväskylä, Finland: Centre for Applied Language Studies. – 2011. – 253 p.*

9. *Banegas D. L. Teachers develop CLIL materials in Argentina: A workshop experience // Latin American Journal of Content & Language Integrated Learning. – 2016. – Vol. 9. – No 1. – P. 17-36. DOI: https://doi.org/10.5294/5460*

10. *Coyle D. Content and language integrated learning: Towards a connected research agenda for CLIL pedagogies // International journal of bilingual education and bilingualism. – 2007. – Vol. 10. – No 5. – P. 543-562. DOI: https://doi.org/10.2167/beb459.0*

Daurenbek S.,[1]* iD *Ivatov S.,[1]* iD *Abuov Zh.[1]* iD
[1]Abai Kazakh National Pedagogical University
Almaty, Kazakhstan

MAIN STAGES OF THE LANGUAGE TEST DEVELOPMENT PROCESS AND VALIDATION ANALYSIS METHODS: THE CASE OF GAZI TÖMER

Abstract

Testing is the most widely used assessment tool in measuring the development process of students' language competence. Therefore, the existence of valid and reliable language tests appropriate to student achievements in education is of great importance. Research on language test development in foreign language teaching in Kazakhstan is insufficient. The aim of this study is to practically demonstrate the basic stages of the test development process and how to conduct content validity analysis with the example of the proficiency test of foreign language (Turkish) prepared by Gazi TÖMER. The research method was based on a quantitative research design. The sample of the research for validation was consisted of 7 experts with experience in Gazi TÖMER. For reliability, there were a total of 66 students studying at Abai KazNPU. The first part of the study was examined for the content validity of the relevant test. For this purpose, based on the criteria determined by Webb for the evaluation of learning outcomes, the opinions received from field experts were evaluated using the Lawshe technique. The CVR value of the 45-question test was .96, and the CVI was "significant". The second part of the study was about reliability. For this purpose, the difficulty level of each item was examined based on the data obtained from the pilot application of the test. Data were analyzed using Excel and SPSS programs. Additionally, the KR-20 value was calculated and found to be .80, and the test was considered reliable. It is believed that the steps and results included in the study will guide the process of creating valid and reliable tests in various fields.

Keywords: content validity, reliability analysis, testing, Gazi TÖMER proficiency test.

Дәуренбек С.Б., [1]* *Иватов С.К.,[1]* *Абуов Ж.[1]*
[1] Казахский национальный педагогический университет имени Абая,
г.Алматы, Казахстан

ОСНОВНЫЕ ЭТАПЫ ПРОЦЕССА РАЗРАБОТКИ ЯЗЫКОВОГО ТЕСТА И МЕТОДЫ АНАЛИЗА ВАЛИДАЦИИ: КЕЙС ГАЗИ ТОМЕР

Аннотация

Тестирование является наиболее широко используемым инструментом оценки процесса развития языковой компетенции учащихся. Поэтому наличие валидных и надежных языковых тестов, которые соответствуют образовательным достижениям учащихся, имеет большое значение. Исследования по разработке языковых тестов для обучения иностранным языкам в Казахстане недостаточны. Целью данного исследования является демонстрация на практике основных этапов процесса разработки тестов и способа проведения анализа валидности содержания на примере теста на знание иностранного языка (турецкий), подготовленного Гази ТОМЕР. Метод исследования был основан на количественном дизайне исследования. Выборка исследования для валидации состояла из 7 экспертов, имеющих опыт работы в Гази ТОМЕР. Для исследования надежности

выборку составили 66 студентов, обучающихся в КазНПУ имени Абая. Первая часть исследования была проверена на предмет валидности соответствующего теста. С этой целью на основе критериев, определенных Уэббом для оценки результатов обучения, мнения, полученные от экспертов на местах, оценивались с использованием метода Lawshe. Значение CVR в тесте из 45 вопросов составило 0,96, а CVI был "значительным". Вторая часть исследования была посвящена надежности. Для этого уровень сложности каждого задания проверялся на основе данных, полученных в результате пилотного применения теста. Данные анализировались с использованием Excel и SPSS. Кроме того, было рассчитано значение KR-20, которое оказалось равным 0,80, и тест оказался надежным. Считается, что шаги и результаты, включенные в исследование, будут определять процесс создания валидных и надежных тестов в различных областях.

Ключевые слова: валидность содержания, анализ надежности, тестирование, тест Гази ТОМЕР.

С.Б.Дәуренбек, [1*] **С.К.Иватов,** [1] **Ж.Абуов** [1]
[1] *Абай атындағы Қазақ ұлттық педагогикалық университеті*
Алматы қ., Қазақстан

ТІЛДІК ТЕСТІЛЕУДІ ҚҰРАСТЫРУ ҮДЕРІСІНІҢ НЕГІЗГІ КЕЗЕҢДЕРІ ЖӘНЕ ЖАРАМДЫЛЫҚТЫ ТАЛДАУДЫҢ ӘДІСТЕРІ: ГАЗИ ТӨМЕР КЕЙСІ

Аннотация

Тестілеу – білім алушылардың тілдік құзыреттілігін дамыту үдерісін өлшеуде кеңінен қолданылатын өлшеу құралы. Сондықтан білім алушылардың оқу жетістіктеріне сәйкес келетін жарамды және сенімді тілдік тесттердің болуы өте маңызды. Қазақстанда шетел тілін оқытуда тілдік тестілеуді құрастыру бойынша зерттеулер жеткіліксіз. Бұл зерттеудің мақсаты — Гази ТӨМЕР дайындаған шетел тілі (түрік тілі) біліктілік тестінің үлгісін пайдалана отырып, тілдік тестілеуді әзірлеу үдерісінің негізгі кезеңдерін және мазмұндылық жағынан жарамдылық талдауын жүргізуді іс жүзінде көрсету. Зерттеу әдісі сандық зерттеу моделіне негізделген. Жарамдылықты талдауға арналған зерттеу үлгісі Гази ТӨМЕР-де тәжірибесі бар 7 сарапшыдан тұрды. Сенімділік талдауы үшін зерттеу үлгісі Абай атындағы ҚазҰПУ-да білім алған барлығы 66 студентті қамтиды. Зерттеудің бірінші бөлімінде сәйкес тесттің мазмұндылық жағынан жарамдылығы тексерілді. Осы мақсатта оқу нәтижелерін бағалау үшін Уэбб анықтаған критерийлер, осы саладағы сарапшылардан алынған пікірлер Lawshe әдісі негізінде бағаланды. 45 сұрақтан тұратын тесттің CVR көрсеткіші .96, ал CVI "маңызды" болып табылды. Зерттеудің екінші бөлімі сенімділік талдауын қамтиды. Осы мақсатта тесттің пилоттық қолдануынан алынған деректер негізінде тесттегі әрбір элементтің қиындық деңгейі зерттелді. Деректер Excel және SPSS бағдарламалары арқылы талданды. Сонымен қатар, KR-20 мәні .80 болып есептеліп, сәйкесінше аталмыш тест сенімді болып табылды. Зерттеудегі қадамдар мен нәтижелер әртүрлі салаларда жарамды және сенімді тестілеуді жасау процесіне басшылық етеді деп күтілуде.

Түйін сөздер: мазмұн жарамдылығы, сенімділікті талдау, тестілеу, Гази ТӨМЕР тесті.

Introduction. It is known that communicative competence within linguistic competence is an important component in the development of intercultural communicative competence. Therefore, as in teaching any foreign language, improving linguistic competence in teaching Turkish as a foreign language is crucial. There are studies on various communicative competence models Canale and Swain [1], Van Ek [2], Bachman [3], and Celce-Murcia [4], including linguistic competence. In these models, communicative competence is considered as a cluster of various competences and is associated with linguistic competence. Different views emerge in studies of Gauriyeva and Zhumabekova [5], Iudina [6], Mirolubov [7], Nurmukhanova, Sagyndykova, Līce and Pāvulēns, [8], Vyatyutnev [9] regarding above mentioned competence. The definitions given in relevant studies show that the relevant competence is the individual's ability to have knowledge of language and to use this knowledge consciously in the communication process. In other words, the language knowledge, skills and competencies represent a set of values of correct understanding and use by students of grammatical rules and language codes of the target language. For this reason, the test, which is an important assessment method, is used to assess and evaluate language knowledge and codes related to linguistic competence. Using language tests helps learners track their language errors, monitor their progress, and plan their next goals.

Test is a widely utilized assessment method in educational organizations, including schools, universities, and teaching centers. There are two primary reasons for using tests: to identify any

fundamental obstacles encountered during the training process, especially those related to the unit or module, and to assess students' proficiency within a specific topic or a set of units [10]. For this reason, it is essential to accurately identify the behaviors in the table of specifications that need to be assessed in the test being developed. Measuring and interpreting the effectiveness of curriculum or teaching, in general, is possible by initially crafting appropriate items that align with the expected learning outcomes. To do this, it is needed to examine and identify observable behaviors that serve as indicators. It is known that Bloom's taxonomy is widely used to identify cognitive levels in the item preparation stage despite being focused on teaching objectives [11]. Bloom's taxonomy consists of six main stages that provide a progression from simple to complex related to the level of cognitive development. As well known, these stages have changed by the current information flow and modern needs: Knowledge/ remembering, comprehension/ understanding, application/ applying, analysis/ analyzing, synthesis/ evaluating, evaluation/ creating [11]. In the comprehension/ understanding stage, the learner is expected to exemplify the behavior he has gained at the level of knowledge or remembering, relate this information to a similar situation, and compare two known situations. In the application stage, it is expected that an application will be made and a problem will be solved by using the information obtained. In the stage of analysis, it is the task of separating a body of information into its elements as they exist within that whole and seeing the relationships between these elements. The process of judging according to certain criteria is evaluation, bringing the pieces together to create an original work, and producing is the process of re-creating.

When creating language tests (aptitude tests, proficiency tests, placement tests, diagnostic tests, achievement tests), identify what will be tested and questions appropriate to the relevant cognitive levels. Students are expected to acquire a variety of skills and competencies within a given subject. Evaluating these aptitudes can be achieved through the use of various types of tests, including but not limited to surveys, open-ended questions, true-false statements, short answer questions, fill-in-the-blank exercises, and multiple-choice inquiries [12]. Considering language tests are used as an assessment tool to measure students' knowledge, skills, and competencies, these tests must be valid in terms of scope and structure. Validity is the principle that an assessment accurately measures the feature it is intended to measure without confounding other features [12, p.33]. If the questions in the test measure the behavior it is intended to measure, the test has content validity. Reliability, like validity, is an important principle of the assessment tool. Reliability refers to the consistency, stability and sensitivity of the assessment tool regarding the feature it measures [12, p.27]. The person or institution that will perform the test must ensure that the test is error-free. Errors that may occur during the test may reduce the reliability of the test. For this reason, the test to be applied or prepared must be examined by more than one expert. In addition, for the test to be valid and reliable, it must be planned.

When developing language tests, it is essential to carefully follow specific steps. However, these steps vary depending on the type of questions included in the test and learning outcomes. In general, any test should follow these steps: First of all, the aim and scope of the test must be determined. After that it is necessary to determine the outcomes and their cognitive levels before choosing the appropriate question types. At this point, it is important how many questions will be asked regarding the outcome. This ensures that the test accurately measures the skills and knowledge it is intended to assess, providing valuable insights into the test-taker's abilities. The last step is how to administer the test and use the scores [10, p.7]. When following these steps, it is critical to create a table of specification. The table of specifications shows the scope of the test in determining learning levels appropriate to the outcomes expected to be achieved within a certain period [10, p.14-15]. This allows for direct learning of the achievements that must be completed within a specific time frame.

When reviewing the literature in the context of Kazakhstan, it is clear that researchers have conducted studies on the development process of achievement tests, but these studies are not

enough. Mingisheva [13] examines the fundamental changes in the development process of standardized tests and the main problems in the creation of tests. As it is known, there are no studies that focus only on the development of valid and reliable language tests. In a foreign context there are studies by Rossi [14], Tavassoli and Sorat [15], Yastıbaş and Takkaç [16], and Korkmaz and Konukaldı [17] that show the theory and practice of test development in language teaching with good examples. Such studies are needed in the context of Kazakhstan to improve the scientific and practical basis of language testing development. The study aims to practically demonstrate the fundamental stages of test development and how to conduct content validity analysis. This will be illustrated using the example of the grammar section of the proficiency test prepared by Gazi TÖMER. This study offers valuable insights into the development of tests used in teaching, particularly achievement tests. It also provides practical, guidance on conducting reliability studies, using the example of the grammar section of Gazi TÖMER. In addition, this study aims to reveal the validity and reliability analysis of the ready-developed test after it was applied at Abai Kazakh National Pedagogical University and to introduce this test to the literature. It is believed that this study helps to assist teachers and researchers in identifying the fundamental steps of test creation and in refining the process of test development with greater precision.

Basic provisions. To measure the knowledge, skills and competencies acquired by students in foreign language education, assessment tool must be valid and reliable. The assessment tool used to measure the language knowledge, skills and competencies of students, especially those learning Turkish as a foreign language, should be developed by the assessment principles. For this reason, language tests used as assessment tools must be carried out in a planned manner and within the scope of certain steps to be valid and reliable.

In general, the following steps should be followed when developing language tests. Firstly, the aim and scope of the test must be identified. Before selecting the proper question kinds, the outcomes and cognitive levels must first be determined. At this point, it is critical to determine how many questions will be asked about the outcome. This guarantees that the exam effectively assesses the skills and knowledge it is designed to measure, delivering useful insights into the test-takers' capabilities. The last step is how to administer the test and use the scores. These steps are important for the test to be valid. Therefore, for the test to be valid in terms of scope, expert opinion is taken and the scope valid index is calculated.

The reliability of a test is as important as its validity. Reliability analyses vary depending on the structure of the test and evaluation criteria. In this study, KR-20 reliability analysis was performed. Additionally, the difficulty level of each item was calculated. A reliability analysis of the Gazi TÖMER proficiency test (grammar section) was conducted after the pilot application. Thus, the test that measures the development of linguistic competence of Kazakh students learning Turkish as a foreign language must be valid and reliable in terms of providing accurate results.

The steps in this study will contribute to how the test can be developed in other areas. Additionally, by taking advantage of the test development processes in this study, researchers can develop valid and reliable tests for other science subjects and different fields.

Materials and methods. The study was based on the quantitative research method. The research sample for validation consisted of 7 experts (2 female, 5 male) in the field of Assessment and Evaluation and Turkish Language Education from Gazi TÖMER. For reliability, the sample was drawn from 66 students enrolled at Abai Kazakh National Pedagogical University in the 2020-2021 academic year. The sample group was selected using an appropriate sampling method that was easily accessible to the researchers. The average age of the students forming the sample group was 18-19, 64 were female (97,0 %) and 2 were male (3,0 %).

Firstly, an evaluation was conducted to determine the suitability of the chosen test for grammar topics being taught. In the assessment and evaluation process, the preferred type of test is just as important as the subject itself. Valid and reliable tests play a significant role in assessing the success of studies. To create such tests, specific methods must be followed. These include examining the

outcomes, formulating a minimum of three questions for each outcome, organizing the items based on expert opinion, carrying out preliminary studies involving the revised items, and ultimately perfecting the test through item analysis [12]. For this reason, when creating test items, three of the criteria developed by Webb [18] should be taken into consideration to ensure alignment between program outcomes and questions. First, the categorical concurrence criterion experts review whether the questions in the assessment tool cover all the relevant aspects of the subject and verify the percentage of compliance. Secondly, the criterion of consistency in depth of knowledge: it examines the compatibility of the outcomes with the questions. The balance of representation criterion examines the distribution of questions among outcomes. To assess the appropriateness of this distribution, the Balance Index is calculated using the formula $1 - [\left(\frac{\sum_0 k=2| \ 1}{0} - I \frac{(K)}{H}\right|/2)$. If the coefficient obtained from the Balance Index formula is greater than 0.70, it can be concluded that the distribution of the gains to the questions is appropriate. Another approach is to develop a table of specifications. This table assesses the compatibility of the outcomes with the questions, providing valuable insights into the evaluation process.

In the study, the grammar section of the language proficiency test prepared by Gazi TÖMER was used and the test items were evaluated in terms of certain criteria. The test items were examined for content validity based on the specified criteria and the learning outcomes determined for the intended application. To ensure the test's content validity, experts in the field were invited to evaluate its content. In this context, the researcher prepared a form with options *"the item measures the targeted structure", "the item is related to the targeted structure but unnecessary"*, and *"the item does not measure the targeted structure".* Experts evaluated the test according to this form.

Firstly, the items of the grammar section of the Gazi TÖMER proficiency test were analyzed, and the accomplishments related to the unit were reviewed. The outcomes were examined systematically (Bloom taxonomy/ table of specification), classifying them according to the levels of remembering, understanding, applying, analyzing, evaluating, and creating. The evaluation focused on assessing the relevance of the questions in the Gazi TÖMER proficiency test (grammar section) about the specified outcomes. Secondly, a panel of experts examined pre-made test items. The size of the expert group can range from 5 to 40 individuals, although it is recommended to have a minimum of 5 people [19, p.91]. The study involved a group of 7 experts, which included one assessment and evaluation expert, and six Turkish language teachers. This stage was conducted in Ar-Ge (Research and Development) department of Gazi TÖMER. The demographic characteristics of this expert group are presented in Table 1.

Table 1. Demographic Characteristics of Experts

No	Gender	Degree	Branch
Expert 1	Female	Dr.	Turkish Language Education
Expert 2	Male	Dr.	Turkish Language Education
Expert 3	Male	Dr.	Assessment and Evaluation
Expert 4	Male	Teacher	Turkish Language Education
Expert 5	Female	Teacher	Turkish Language Education
Expert 6	Male	Dr.	Turkish Language Education
Expert 7	Male	Teacher	Turkish Language Education

Upon examination of the table above, it becomes evident that there were 5 male experts and 2 female experts. It appears that one expert specializes in assessment and evaluation, while the others are proficient in the field of Turkish Language Education. Furthermore, it is noteworthy that 4 of the experts hold doctoral degrees, while the remaining three are dedicated teachers who specialize in teaching Turkish to non-native speakers. Relevant experts evaluated whether the questions

adequately encompassed all outcomes and were compatible with the corresponding accomplishments. The determined group of experts was given a form with options. They were then asked to evaluate the questions based on this form. The evaluation results of the experts were quantitatively presented by obtaining a content validity index (CVI) for each item using the Lawshe technique [20]. The Lawshe technique [20] represents a crucial method for transitioning the study of obtaining expert opinions on the content validity of prepared test items from a qualitative dimension to a quantitative one. Expert evaluation, guided by criteria such as *"the item measures the targeted structure"*, *"the item is related to the targeted structure but unnecessary"*, and *"the item does not measure the targeted structure"*. The formula for the technique is as follows: CVI $= \frac{N_G}{N/2} - 1$. N_G= Number of experts who answered *"the item measures the targeted structure"* N = Total number of experts giving their opinions. The value of CVI varies depending on the number of experts and their opinions. The CVI index values range from -1 to +1. If more than half of the experts indicate that the substance is unsuitable, the CVI value becomes negative. If all the experts agree it is suitable, CVI =1; if half of the experts agree, CVI =0; if more than half of the experts agree, CVI >0; and if less than half of the experts agree, CVI <0 [19]. The values for the CVI, which will be examined in the evaluation of the findings are provided in Table 2.

Table 2. Minimum Values for Scope Validity Rates

Number of Experts	Minimum CVI Value
5	0,99
6	0,99
7	0,99
8	0,75
9	0,78
10	0,62
11	0,59
12	0,56
13	0,54
14	0,51
15	0,49
20	0,42
25	0,37
30	0,33
35	0,31
40	0,29

When analyzing this table, consider that if the CVI value, obtained from the opinions of 9 experts, exceeds 0.78, the corresponding item will be selected from the pool for inclusion in the draft scale. The average CVI value was calculated by considering the CVI values of all the items included in the draft scale, as indicated in the CVI values table. Then, item analysis was conducted to ensure the test's construct validity, which is then adjusted based on expert opinions. In this study, 45 questions that were initially prepared were selected based on specific criteria and expert opinions. Subsequently, a preliminary form consisting of these questions was administered to 66 students of Abai KazNPU.

Thus, the item analysis was completed, and a discrimination index was calculated for each item, which ranges from − 1 to +1. According to the analysis, items with a score of .40 and above exhibit

very high discrimination, those with .30 to .39 are considered pretty good, while items with .20 to .29 should be addressed. Any items scoring .19 and lower are recommended to be removed from the test or adjusted [19]. Based on this, the resulting item analysis revealed that items with a discrimination index of .40 and above fell within the .30 to .39 range, so they were retained in the test. Finally, the appropriateness of distributing questions to outcomes was evaluated.

Results and discussion. This section presents the findings from the validity and reliability analyses of the grammar section of the proficiency test in the case of the Gazi TÖMER. Firstly, findings regarding the content validity study of the test. As previously mentioned, expert opinions were sought to evaluate the appropriateness of the test for measuring the outcomes. To achieve this, the Lawshe technique [20] was employed, using a specific table for analysis. The results meeting the established criteria are detailed in Table 3 below, which includes the content validity rate based on the input of 7 expert opinions.

Table 3. Content Validity Criteria

Number of outcomes	Appropriate	Not appropriate	CVR
1	7	0	1
2	7	0	1
3	7	0	1
4	7	0	1
5	7	0	1
6	7	0	1
7	7	0	1
8	7	0	1
9	7	0	1
10	7	0	1
11	7	0	1
12	7	0	1
13	7	0	1
14	7	0	1
15	7	0	1
16	7	0	1
17	7	0	1
18	7	0	1
19	7	0	1
20	7	0	1
21	7	0	1
22	7	0	1
23	7	0	1
24	7	0	1

As shown in Table 3, the evaluation of the questions' relevance to the achievements was conducted. The findings reveal that the majority of the questions are fitting for the specified achievements. Consequently, with a calculated Content Validity Rate (CVR) of 1, it is evident that the items in the test hold substantial significance for the outcomes. The Lawshe technique [20] was employed to analyze the test items, and the Content Validity Rates (CVR) were calculated first. Subsequently, the Content Validity Indices (CVI) were determined. The results can be found in Table 4.

Table 4. Content Validity Rates and Content Validity Indices of the Test

Number of questions	It measures	It measures but is unnecessary	Does not measure	CVR	CVI	Decision
1	7	0	0	1,00	significant	appropriate
2	7	0	0	1,00	significant	appropriate
3	6	1	0	,85	significant	appropriate
4	7	0	0	1,00	significant	appropriate
5	7	0	0	1,00	significant	appropriate
6	7	0	0	1,00	significant	appropriate
7	7	0	0	1,00	significant	appropriate
8	6	1	0	,85	significant	appropriate
9	7	0	0	1,00	significant	appropriate
10	7	0	0	1,00	significant	appropriate
11	7	0	0	1,00	significant	appropriate
12	7	0	0	1,00	significant	appropriate
13	7	0	0	1,00	significant	appropriate
14	7	0	0	1,00	significant	appropriate
15	7	0	0	1,00	significant	appropriate
16	7	0	0	1,00	significant	appropriate
17	7	0	0	1,00	significant	appropriate
18	6	1	0	,85	significant	appropriate
19	6	1	0	,85	significant	appropriate
20	7	0	0	1,00	significant	appropriate
21	7	0	0	1,00	significant	appropriate
22	7	0	0	1,00	significant	appropriate
23	7	0	0	1,00	significant	appropriate
24	7	0	0	1,00	significant	appropriate
25	7	0	0	1,00	significant	appropriate
26	7	0	0	1,00	significant	appropriate
27	7	0	0	1,00	significant	appropriate
28	7	0	0	1,00	significant	appropriate
29	7	0	0	1,00	significant	appropriate
30	6	1	0	,85	significant	appropriate
31	7	0	0	1,00	significant	appropriate
32	7	0	0	1,00	significant	appropriate
33	7	0	0	1,00	significant	appropriate
34	6	1	0	,85	significant	appropriate
35	6	1	0	,85	significant	appropriate
36	7	0	0	1,00	significant	appropriate
37	7	0	0	1,00	significant	appropriate
38	6	1	0	,85	significant	appropriate
39	6	1	0	,85	significant	appropriate
40	6	1	0	,85	significant	appropriate
41	7	0	0	1,00	significant	appropriate
42	7	0	0	1,00	significant	appropriate
43	7	0	0	1,00	significant	appropriate
44	7	0	0	1,00	significant	appropriate
45	6	1	0	,85	significant	appropriate
The whole test (CVI)				,96	significant	appropriate

Table 4 shows that the content validity rate of the 45-question test was .96, and the content validity index was "significant". As can be seen, the test's questions had content validity rates of .85 or higher. Looking at these findings, it is clear that the test questions are appropriate for the claimed outcomes. For this reason, it was decided to use the Gazi TÖMER test without any changes in practice. In other words, it is accepted that the relevant test is suitable for use in measuring the linguistic competence of Kazakh students learning Turkish as a foreign language.

Since an assessment tool is both valid and reliable, reliability analyses were conducted on the test in this study. The research sample consisted of 66 students who were enrolled at Abai KazNPU during the 2020-2021 academic year. The average age of the students forming the sample group was 18-19, 64 were female (97,0 %) and 2 were male (3,0 %). In scoring the test, correct answers were evaluated as "1" and incorrect answers were evaluated as "0". The data collected after the application was analyzed using Microsoft Excel and SPSS programs. In the analysis, the total scores of the students from the test were listed. Additionally, the difficulty indices (pjx) of each question were calculated. The results obtained are given in Table 5 below.

Table 5. Summary of Pilot Test Items According to Difficulty Indices

Number of questions	The difficulty indices (pjx)
1	0.70
2	0.65
3	0.73
4	0.76
5	0.79
6	0.71
7	0.77
8	0.70
9	0.74
10	0.68
11	0.77
12	0.74
13	0.74
14	0.70
15	0.67
16	0.52
17	0.50
18	0.53
19	0.65
20	0.67
21	0.55
22	0.65
23	0.61
24	0.68
25	0.70
26	0.73
27	0.71
28	0.80
29	0.68
30	0.70
31	0.68
32	0.71
33	0.71
34	0.65

35	0.74
36	0.86
37	0.89
38	0.91
39	0.77
40	0.79
41	0.86
42	0.79
43	0.76
44	0.88
45	0.79
Overall	0.54

When Table 5 is examıned, it is seen that the questions in the test have a difficulty index of .50 and above. If the item difficulty index is $0 \leq p < 0.20$, the item is considered "very difficult"; if it is $0.20 \leq p < 0.35$, the item is "difficult"; if it is $0.35 \leq p < 0.65$, the item is "medium"; if it is between $0.65 \leq p < 0.80$, the item is "easy"; if it is $0.80 \leq p \leq 1$, the item is considered "very easy" [19]. As seen in Table 5, the item difficulty index of the questions ranges between .50 and .91. Additionally, as a result of the analysis, it is seen that the difficulty index of the test is .54. A good test should have both easy and difficult questions and the average difficulty level should be around .50 [19]. In this context, it can be said that the difficulty level of the test is "medium level". KR-20 analysis was used to determine the reliability of the entire test. The KR-20 value was calculated, and it was found to be .80. A KR-20 value above .70 in an assessment tool indicates that the reliability of the tool is high [19]. Accordingly, after the application, it is understood that the reliability of the relevant test is high.

Conclusion. It is known that language knowledge, skills and competencies are an important component in the development of communicative competence. For this reason, as in every foreign language education, improving linguistic competence in teaching Turkish as a foreign language is a priority task. Therefore, any language test must be accurate, valid, and reliable in measuring the development of linguistic competence. Therefore, this study aims to reveal the validity and reliability analysis stages in the example of the Gazi TÖMER proficiency test (grammar section), which is applied to measure the linguistic competence of Kazakh students learning Turkish. In this study, the suitability, content validity and reliability of the proficiency test (grammar section), prepared by Gazi TÖMER with the outcomes determined during the application were examined. Significantly, the test is valid and reliable in assessing knowledge, skills, and competencies in foreign language teaching. For this reason, during the study, Gazi TÖMER proficiency test (grammar section), was first examined in terms of content validity. The reliability of the KR-20 was then tested and the difficulty level of each question was analyzed.

The steps followed during the study are necessary in the development of tests used in foreign language education. Performing these steps is essential to maintain the validity and reliability of the test. Examining the outcomes and determining which cognitive steps these outcomes cover is very essential in terms of creating the structure of the question to be prepared. In addition, this situation is also used to analyze to what extent students can achieve the level of cognitive perception. Then, the test should be piloted and its reliability calculated. At this point, the difficulty level of each item should be calculated. A test's validity, reliability, and usefulness are dependent on the precise completion of the specified steps. It is thought that this study will provide detailed information about the content validity and reliability of the test prepared by Gazi TÖMER within the scope of certain steps. The steps taken with the test in the study will contribute not only to foreign language education but also to the development of any test used in education. Thus, researchers can develop valid and reliable tests for other subjects and different fields by using the test development processes in this study.

References:

1. Canale, M., & Swain, M. Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics. – 1980. – 1(1). p.1-48.

2. Van Ek, J. A. Objectives for foreign language learning. Vol.1: Scope. Strasbourg: Council of Europe, 1986.

3. Bachman, L. F. Fundamental considerations in language testing. Oxford: Oxford University,1990.

4. Celce-Murcia., M. Rethinking the role of communicative competence in language teaching. In E. A. Soler & M.P.S. Jordà (Eds.), Intercultural language use and language learning (pp. 41-57). Springer, 2007.

5. Gauriyeva, G. M., & Zhumabekova, A. A. Use of authentic texts for formation of students' linguistic competence. Bulletin of the L.N. Gumilyov Eurasian National University, Philology Series.-2018.-2(123). pp. 70-74.

6. Iudina, T. V. Obosnovanie struktury lingvisticheskoj kompetencii kak komponenta kommunikativnoj kompetencii studenta yazykovogo vuza. Pedagogy. Theory and Practice.-2021.- 6(4), 629-636.

7. Mirolubov, A. A. Voprosy kontrolya obuchennosti uchashchihsya inostrannomu yazyku. Obninsk: Titul, 1999.

8. Nurmukhanova, D., Sagyndykova, Z., Līce, L., & Pāvulēns, J. Competency–based approach to teaching foreign languages in Kazakhstan. Rural environment. Education. Personality.– Jelgava. – 2014.- p.116-122.

9. Vyatyutnev, M. S. Ponyatie yazykovoj kompetencii v lingvistike i metodike prepodavaniya inostrannyh yazykov. Inostrannye Yazyki v Shkole. – 1975- 6, p.55-65.

10. Özçelik, Ali D. Test hazırlama kılavuzu. Ankara: Pegem, 2013.

11. Güler, N. Eğitim sürecinde hedef davranışlar ve hedef davranışların ölçülmesi. In Eğitimde Ölçe ve Değerlendirme (Ed. Tekindal, S.). Ankara: Pegem, 2020.

12. Şahin, Y. Yabancı dilde ölçme ve değerlendirme. Ankara: Pegem, 2021.

13. Mingisheva, N.A. Development and challenges of standardized testing in Kazakhstan: transition from national to international standards. Bulletin of Al-Farabi Kazakh National University, Pedagogical Series. – 2023.- 3(76). pp.94-103.

14. Rossi, O. The language test development process: main principles and suggestions for classroom practice. Testing, Evaluation and Assessment Today. -2021.- 4. pp. 53-57.

15. Tavassoli, K., & Sorat, Z. Iranian EFL teachers' oral/aural skills language assessment literacy: instrument development and validation. International Journal of Language Testing. – 2023. – 2(13). pp. 56-76.

16. Yastıbaş, A.E., & Takkaç, M. Understanding the development of language assessment literacy. Journal of Social Science Institute. – 2018.- 8(15). pp.89-106.

17. Korkmaz, H. & Konukaldı, I. İlköğretim fen ve teknoloji eğitiminde disiplinlerarası tematik öğretim yaklaşımının öğrencilerin öğrenme ürünleri üzerine etkisi. Buca Eğitim Fakültesi Dergisi.-2015.- 39. pp.1-22.

18. Webb, N. L. Determining alignment of expectations and assessments in mathematics and science education. NISE brief 1(2). Madison, WI: University of Wisconsin-Madison, National Institute for Science Education. 1997.

19. Karagöz, Y. SPSS AMOS META uygulamalı nitel-nicel-karma bilimsel araştırma yöntemleri ve yayın etiği. Ankara: Nobel, 2021.

20. Lawshe, C. H. A quantitative approach to content validity. Personnel Psychology.- 1975.- 28. pp.563–575.

Сыздыкбаева А., ¹ ⓘ Саржанова Г.,² ⓘ Книсарина М.³ ⓘ*

¹Казахский национальный женский педагогический университет, г.Алматы, Казахстан
²Карагандинский университет имени Е.А. Букетова, г.Караганда, Казахстан
³Западно-Казахстанский медицинский университет имени М.Оспанова,
г.Актобе, Казахстан

ФАКТОРЫ ЛИЧНОСТНОГО РАЗВИТИЯ И ПСИХОЛОГИЧЕСКОГО БЛАГОПОЛУЧИЯ В СРЕДНЕМ ОБРАЗОВАНИИ: КОНЦЕПТУАЛЬНЫЙ АНАЛИЗ И КЛАСТЕРИЗАЦИЯ

Аннотация

Данное исследование посвящено концептуальному анализу факторов личностного развития и психологического благополучия учащихся в системе среднего образования. Рассматриваются ключевые концепции личностного развития (теория психосоциального развития Erikson, теория самодетерминации Ryan & Deci, теория множественного интеллекта Gardner) и теории психологического благополучия (модель психологического благополучия Ryff & Keyes, теория субъективного благополучия Diener, концепция потока