Bartın Üniversitesi
Eğitim Fakültesi Dergisi
Cilt 6, Sayı 2, s. 395-410, Haziran 2017
BARTIN – TÜRKİYE

ISSN: 1308-7177

Bartin University
Journal of Faculty of Education
Volume 6, Issue 2, p. 395-410, June 2017
BARTIN – TURKEY

Doi: 10.14686/buefad.281870

Üniversite İngilizce Hazırlık Sınıfında Küçük Ölçekli Bir Sınav için Görünüş Geçerliği Çalışması

Emrah CİNKARA, Yard. Doç. Dr., Gaziantep Üniversitesi Eğitim Fakültesi, emrahcinkara@gmail.com
Özlem ÖZEN TOSUN, Okutman, Gaziantep Üniversitesi Yabancı Diller Yüksekokulu, ozlemozen88@gmail.com

Öz: Dil becerilerinin ölçülmesi ve değerlendirilmesi eğitim öğretim çalışmaları kapsamında önemli bir alandır ve eğitim öğretim süreçlerini de şekillendiren önemli sonuçlara sahiptir. Sınava girecek öğrenciler için olumlu ve motive edici bir öğrenme ve sınav ortamı oluşturmak hem sınav görünüş geçerliği kavramının temeli hem de sınavı hazırlayanların amaçlarından bir tanesidir. Bu amaçla, bu çalışmada sınava giren öğrencilerin, sınavı hazırlayan ve derse giren öğretim elemanlarının sınavdaki maddelerin hazırlanma amaçları ile ilgili görüşleri araştırılmıştır. Bu çalışmada Gaziantep Üniversitesi Yabancı Diller Yüksekokulu İngilizce hazırlık sınıfında öğrenim gören 38 öğrenci, aynı birimden yedi öğretim elemanı ve dört sınav hazırlama komisyonu üyesi yer almıştır. Cevap kodlama formu ve bire bir görüşme yöntemleri kullanılarak nicel ve nitel veri toplanmıştır. Araştırma sonuçlarına göre sınav hazırlayanlar ile derse giren öğretim elemanlarının sınav sorularının amaçlarını eşleştirmede orta düzeyde bir uyuma sahip olduğu, öğrencilerle ise düşük seviyeli bir uyum olduğu belirlenmiştir.

Anahtar Kelimeler: görünüş geçerliği, küçük ölçekli sınavlar, uzman görüşü, sınava giren öğrenci algısı

Face Validity Study of a Small-Scale Test in a Tertiary-Level Intensive EFL Program

Abstract: The assessment of language abilities plays a crucial role in designing tests. In order to create a positive and motivating learning environment for test takers, one of the most important factors is the administration of tests with high levels of face validity as well as test taker cognizance of the abilities necessary for exam success. Therefore, this study examined the overlap in test takers' and language instructors' perceptions of the abilities being tested as well as test developers' intentions in preparing exam items. The study was conducted at Gaziantep University School of Foreign Languages with 38 test takers, seven language instructors, and four test designers. Qualitative and quantitative data was collected by means of a response sheet on test takers' perceptions of tests after being modified in accordance with the test to be analysed. The results revealed a moderate level of agreement between test designers and language instructors while a lower agreement percentage was obtained with test takers.

Key Words: face validity, small-scale tests, expert opinion, test taker perception

1. INTRODUCTION

In general, testing and assessment have played an important role in the attempt to prove examinees' and applicants' capabilities and qualifications for varied purposes in education (Kelecioğlu, & Şahin, 2014). As tests have been used for many different aims, each of these aims has affected teaching, learning and individuals in the process. Therefore, a great deal of research has been conducted on tests from various perspectives to analyse in-depth their possible impacts (Kim, & Elder, 2015; Schmitt, 2002; Watkins, Dahlin, & Ekholm, 2005). As in learning in general, tests are crucial due to the effects they have on foreign language learners. Therefore, numerous studies were implemented to clarify these effects are informative for test administrators, as well (Kim, & Elder, 2015; Schmitt, 2002; Watkins et al.).

Both the teaching and learning processes are profoundly affected by examination and assessment, an assumption that has led to further investigation from a test designers' perspective. However, test takers' own views on examination also might provide important information to test designers (Brown, 1993). Although tests are based on pre-determined content and standarts, when administered they may influence the content or syllabus of a course, so test takers tend to focus solely on topics on which they perceive they will be tested. As Hughes (2003) suggests, teachers should test what they teach because test takers will definitely ignore what is taught if not tested. This impact of tests on test takers' learning is called as the 'backwash' or 'washback' effect. Bachman (1990) defines the backwash effect as 'the effect of each test item on teacher's teaching and learner's learning in terms of positive and negative aspects'. Brown and Hudson (2002) state that if the items in the test are in line with the goals of the syllabus, a positive backwash effect can be observed. Paker (2013) supports this idea by suggesting that the items should be designed in a constructive rather than destructive way on the backwash effect of test items in the achievement tests of intensive English as a Foreign Language (EFL) program.

Considering the effects of such a delicate issue, it is vital to make use of test takers' perceptions of tests, which provide valuable feedback for theoretical and practical applications (Brown, 1993; Hill, 1998; Nevo, 1985). Test takers' perceptions regarding what kinds of skills are essential to answering test items are among the most significant concepts in language testing; however, it has been a neglected aspect of the validation process among other validity types (Xie, 2011). For a positive backwash effect of tests on learning to occur, there should be congruence between the skills and abilities which test takers think are tested and the intentions of test designers, which is the subject of this study. Therefore, the purpose of this study is to indicate the level of agreement among test takers, course instructors' perceptions and test designers' intentions. However, any analysis into learner perceptions of the test should take into account the fact that learners' understanding of the concepts tested might be too limited to even evaluate the test items.

Face validity is an undervalued aspect of validation, and it is defined as "the degree to which a test appears to measure the knowledge or abilities it claims to measure" as determined by an expert or learners as in our case (Davies, Brown, Elder, Hill, Lumley, & McNamara, 1999, p.59). The reason why face validity remains under-investigated is that test designers are concerned about the reliability of test takers' perceptions (Xie, 2011). Nevertheless, a number of researchers have claimed that test takers may perform poorly when they do not see the connection between test items and the subject point assessed through them. Consequently, the test does not seem to assess the skills which the test designers actually have intended to measure, test takers may not pay enough attention to the items, which will result in a negative backwash effect (Bachman, 1990; Brown & Abeywickrama, 2010; Hughes, 2003).

Face validity of a test is the characteristics of a test as a whole and individual items. The characteristics cover "the appropriateness, sensibility, or relevance of the test and its items as they appear to the persons answering the test. Do a test and its items seem valid and meaningful to the individuals taking the test?" (Holden, 2010, pp: 1-2). In a formal explanation, face validity can be described as the extent to which the experts and test takers, in our case, see the course content and the content of the test and its items as relevant. This construct has been studied in certain contexts. In one case, MacLellan (2001) investigated the perceptions of test takers and designers about the objectives of the tests. Some differences as to the perceived goals of tests between the test takers and designers were detected in this study. In a similar vein, Xie (2011) examined face validity of an EFL test and found certain mismatching points in the test takers' and test designers' perceptions of the aim of the test items. It was argued that these mismatches are areas which researchers and test designers should examine in detail. Further, Shohamy (2001) emphasized the importance of understanding test takers' perceptions of tests for a variety of purposes such as recognizing the use and aims of tests and their significance in test takers' lives. In a similar understanding, Tsaia and Tsou (2009) proffered that test takers may be a good source for understanding the effectiveness of tests.

In another study by Teemant (2010), test takers' perceptions of a test given in English to non-native speakers of English were investigated and poor performance of the test takers was found in relation to a lack of content knowledge and appropriate discourse rather than the absence of knowledge pertaining to the English which was tested. Therefore, it becomes prominent for a test to have high level of face validity and reflect it through its items. Interestingly enough, the perceived test competence was found to affect actual performance and behaviour in young learners. Actual ability, such as physical activity, fitness and lowered risk of overweight and obesity were found to be associated with perceived competence of a test (Barnett, Ridgers, Zask, & Salmon, 2015). In their study, Barnett et al. (2015) found that participating children, who could tag the picture showing the skill performed 'good' or 'poor' illustration of the skill tested, were better at performing those skills.

In a different context in which international tests were analysed, He and Shi (2008) scrutinised test takers' perceptions of the comparison between the written part of Test of English as a Foreign Language (TOEFL) exam and English Language Proficiency Index (LPI). They found that some of the problems test takers had faced while they were taking the test arose as a result of their habit of relying on their memories to memorize the writing structures, which is a sign of low reliability because it is not a reflection of real writing abilities.

Scouller and Prosser's study focused on test takers' perceptions regarding the skills being assessed by multiple choice examinations (1994). Their results indicated that the test designers' intentions and the test takers' perceptions did not match. They also analysed the relationships between ability and perception suggesting that high achievers are better at recognizing the skills to be measured than are low achievers.

Kim and Elder (2015) stressed the importance of investigating the opinions of field specialists on the validity of test construct and noted that their possible contributions to language requirements to meet their occupational needs can be regarded as evidence for construct validity of the test. Brown (1993) supported this idea with the findings of the study in which feedback on the relevance of test tasks to the language required in the industry was taken from the test-takers and the findings proved that the test assessed the relevant test skills, which was considered as evidence for construct validity. Elder (2007) also analysed the perceptions of test-takers as field specialist in deciding between two tests, Occupational English Test (OET) and International English Language Testing System (IELTS), in terms of appropriateness to determine preparedness for work-related communication and emphasized the feasibility of test-taker perception as evidence for validity. Knoch (2014) carried out a study on test-takers' perceptions from a similar perspective in which test-takers were requested to analyse discourse in speech samples from a variety of English tests in aviation. The results indicated that the test takers identified more criteria than the ones which were prescribed by the guidelines of the tests, which proves the value of getting test taker feedback as evidence of validity.

Finally, Sato and Ikeda (2015) studied the potential effects of face validity based on learner opinion on learning and investigated the overlap between test designers' intentions and test takers' understanding of what abilities are tested by each item. They found in this study that if the items in a test are not perceived as testing the same ability both by the test designers and test takers, it will not create the desired backwash effect on learning.

1.1. The purpose of the study

When all these studies and their implications are taken into consideration, it becomes essential to emphasize the need to understand test takers' perceptions of the test items since there should be an agreement between the test designers' purposes for preparing the items and test takers' understanding of what skills they are supposed to apply for effective learning to occur. All of the above mentioned studies were conducted to analyse the effects of test takers' perceptions of high-stakes tests. However, it is a neglected issue that success on local exams in most cases is a prerequisite for taking these high-stakes exams. For this reason, this study has been conducted to examine the level of agreement between test designers' intentions and those of test takers as well as content instructors' perceptions of quizzes prepared in an intensive EFL programme, adding a new dimension to previous studies on the same subject. In line with the aims of the current study, the following research questions were constructed:

1. To what extent do test takers', course instructors' and test designers' intentions about the items in the test agree?

2. Which parts of the test caused disagreement among the perceptions of test takers, course instructors and test designers?

2. METHOD

2.1. Context

This study was conducted in the School of Foreign Languages at a state university in south-eastern Turkey after necessary permission and consent were obtained. The medium of instruction in 10 programmes at this university is English; therefore, an intensive English immersion program is mandated for all students under the required English proficiency level. Around 1800 students were enrolled each year in the intensive EFL programme. After enrolment, students are required to take a placement test unless they are exempt from language instruction. Students take a 26-hour English course each week for two months in each module and there are four modules throughout the program ranging from levels A1 to B2; moreover, they can reach level C1 depending on the module in which they are initially placed. Each module has an individual test designer who composes six quizzes and one end-of-the-module test. The study was carried out in the spring semester of 2016.

2.2. Participants

The participants of this study included thirty-eight test takers, seven course instructors and four test designers. Test takers were all engineering students and were enrolled at the intermediate level in module B1 during the data collection process. There were twenty-five male and thirteen female test takers with ages ranging from eighteen to twenty-three. They were asked to indicate their perceptions of the items which they previously had answered, which allowed them to be able to think more in-depth about the skills being tested.

Table 1.

Participants

	Male	Female	Total	Response sheet	Interview
Test designer	0	4	4	1 (as group)	0
Instructor	2	5	7	7 (individual)	2
Test taker	25	13	38	38 (individual)	2

Instructors and test designers included two male and nine female teachers with experiences ranging from five to twenty-six years. There were four test designers and seven course instructors participating in the study. Three of the test designers have been preparing tests for three years while one of them has been designing tests for one year. Four of the course instructors, on the other hand, are module coordinators, and the other three course instructors voluntarily participating in the study were anticipating to work as future test designers in the institution. They were asked to record their intentions and perceptions concerning the items as well as the skills they expected of test takers for answering the questions.

2.3. Data collection tools

The purpose of this study was to investigate students' and teachers' perceptions of a test and test designers' intentions of preparing the same test. Another objective was to investigate the reasons for possible mismatch between these two. In order to collect data for studying this phenomenon, two data collection tools were employed: response sheets and interviews.

Response sheets: The response sheets were composed of two parts. One was the actual test (Quiz 11) and the second was the sheet of skills and abilities (Appendix). The quiz consisted of forty-three questions and eight parts. The types of items in the quiz were an array of selected responses, limited production and receptive vocabulary tasks including multiple-choice questions, gap-filling activities and a vocabulary-matching exercise. The skills sheet included a table in which participants were asked to match test parts and intentions as well as a list of skills and abilities such as understanding tenses, sentence structure, implications and references in a reading text alongside vocabulary knowledge. There were twenty-seven items in this list which was mainly adapted from another response sheet developed by Sato and Ikeda (2015) for a similar study. The items in the list were specially adapted to reflect the abilities being measured by the quiz administered and to represent the possible abilities and skills which test takers might use to indicate their perceptions about the quiz in question.

Interviews: Secondly, an interview protocol was used to collect data regarding the possible causes of mismatch in test designers' intentions and test takers' perceptions of test aims. For this purpose, two test takers and two course instructors were interviewed. The interviews commenced with a general overview of their perceptions of the test and the skills expected to be measured by the test. Then, test takers and course instructors were asked about the reasons for the incongruence between the intentions of test designers and the perceptions of test takers and course instructors. The researcher took notes during the interview to unearth possible causes of mismatch.

2.4. Procedures

In each module, students take five quizzes and one module exit exam. Necessary permissions having been obtained from the institution, the test was selected randomly among the five quizzes done in the given level. Then, the response sheet was constructed as a first step in this study. Four test designers were asked to complete a response sheet and report their common intentions about the skills to be measured by the test. Course instructors were requested to indicate the required skills and abilities for answering each item in the response sheet, as well. Thus, the level of agreement pertaining to the required skills was examined and the results suggested some implications for a positive backwash effect on test takers' learning.

Next, a sample group for the research was determined randomly. After test taker consent was obtained, the purpose of the study was explained to them to their understanding of how they were expected to analyse the test and reflect their ideas on the response sheet. The participants were asked to match the items in the test with the ability statements in the questionnaire. As they had already taken the same test in the previous module and knew the answers to the test questions, they did not experience difficulty in examining and pairing the items and the abilities on the list. However, they were assisted when they needed help in terms of the content of the items in the response sheet while analysing and reflecting upon the quiz. For the sake of reliability, participants were not restricted time-wise when completing the survey.

By the time the necessary data from student participants were collected, instructors whose opinions were also asked about the same quiz had already completed the questionnaires and returned them to the researcher.

2.5. Data analysis

Data was collected from test takers, course instructors and test designers to be analysed in terms of the agreement rate among participants and it included the responses of four test designers, seven language instructors and thirty-eight test takers. First, to determine the agreement rate among participants, the number of items on which test takers had agreed regarding skills being measured was determined for each test component. Then, percentages of agreement were calculated for each item. For instance, 20 out of 25 learners matched item 1 with skill number 4, then the agreement was calculated as 80%. The overall agreement was calculated for students and teachers, and compared with those of test designers. The mean for each part of the quiz was also analysed separately to identify internal consistency. Finally, the data set of both teacher and test taker responses was examined to determine the means independent of each other yet based on the intended skills decided by test designers to allow for more detailed analysis of results.

3. FINDINGS

3.1. Findings of the response sheet

To begin with, the data set was analysed to find an answer to the first research question pertaining to the agreement level of items measuring specific skills among participants. The findings are summarised in Table 2:

Table 2

Test Designer – Teacher and Test Designer - Test Taker Agreement

	Test designer - teacher	**Test designer - test taker**
Mean	61.10%	36.32%
Highest	82.91%	73.12%
Lowest	37.08%	7.5%

When the percentages of the agreement level were analysed for different groups of participants, 61.10% agreement was found between teachers and test designers while this agreement was calculated as 36.32% for test designer - test taker agreement. Among teachers, the highest individual percentage was 82.91% whereas the lowest one 37.08%. Test takers' overall percentage was 36.32% with a range from 7.5% to 73.12%. Another point worth mentioning is that an individual average of four out of seven teachers have an agreement rate below the mean percentage of the group. A similar finding was obtained for test takers, as well, with twenty-three test takers constituting 60.52% having an average under the mean percentage in the whole group.

Table 3
Teacher – Test Designer Agreement Rate

Particip ants	**Part a**	**Part b**	**Part c**	**Part d**	**Part e**	**Part f**	**Part g**	**Part h**	**Mean (%)**
1	80	80	75	50	25	33.33	60	20	52.91
2	80	100	50	50	100	83.33	100	100	82.91
3	20	60	25	50	25	16.66	40	60	37.08
4	60	80	0	50	75	50	80	60	56.87
5	40	100	100	50	50	66.66	100	80	73.33
6	100	100	50	50	100	66.66	100	40	75.83
7	60	60	25	50	25	50	60	60	48.75
Mean	62.85	82.85	46.42	50	57.14	52.38	77.14	60	61.10

Table 3.
Student – test designer agreement rate

Test takers #12-49	**Part a**	**Part b**	**Part c**	**Part d**	**Part e**	**Part f**	**Part g**	**Part h**	**Mean (%)**
12	60	40	25	0	25	16.66	80	40	35.83
13	60	80	75	50	50	33.33	100	100	68.54
29	60	60	100	100	75	50	80	60	73.12
30	60	80	75	50	100	66.66	100	40	71.45
37	0	20	0	0	0	0	20	20	7.50
49	60	0	25	50	25	33.33	60	20	34.16
Mean	44.73	38.94	45.39	27.63	29.60	28.50	47.89	27.89	36.3

As for the second research question in this study, the following findings guide us in our query. The average percentage of teachers' responses for each part was calculated and given in Table 2. The highest agreement rate between teachers and test developers was found to be 82.85% in Part B, in which test takers were asked to rewrite some sentences using the given conjunctions as a prompt. The items that teachers selected in the response sheet were 7, 24, 25 and 26 because they were mainly concerned with understanding tense structure in a context as well as the basic structure and rules of English sentences and understanding the connection among sentences, respectively. However, the lowest level of compromise between teachers and test designers was in Part C, with a percentage slightly below 50%. In Part C, test takers were invited to read a passage and decide if the given statements were true or false. Common items representing the abilities which teachers thought to be measured in this section was 10, 16, 20 and 25, which indicated the ability to grasp the main point of each paragraph, the ability to grasp the content quickly as well as objectively and the ability to understand the connection among sentences. In addition to these most frequently selected common items, this is the part which test takers perceived as testing significantly more abilities than other parts at the same time. This perception led to the lowest agreement among teachers. On the other hand, the same part has one of the highest mean percentages on the side of test takers, whereas Part D has the lowest percentage which is 27.63% compared to other parts. In this part, test takers are required to find references for the pronouns given in the questions. What is really surprising for this component is that half of the test takers could not match even one item intended by the test developers. This is why the overall mean percentage is the lowest for the part in question in comparison to the other ones. Another part, Part G, in which test takers were supposed to fill in the blanks with the words from the box has the highest mean score with test takers while it has the second highest mean with teachers. These results may suggest that this part was the most comprehensible for test takers since they all seemed to understand the given task although they may not have received the highest grade in the same part as this was another factor incorporating test takers' ability to learn or study regularly and so on. With this high mean percentage, it can be concluded that face validity for this part seems quite high and test takers' chances of performing well in this part was a lot higher than in other components.

3.2. Findings of the interviews

The findings of the interviews conducted with course instructors and test takers suggest some notable points regarding the determining factors of the face validity of a test as well as the reasons for incongruence between test designers' intentions and test takers' and course instructors' perceptions about the skills being measured in the test.

The interviewed course instructors agreed on the factors determining face validity. Both instructors argue that a test should look well-organized in terms of both layout and the organization. They thought that test components should be easily recognizable and distinguishable from other parts, with clear instructions written in bold and that a test may look valid by providing test takers with carefully designed examples. They suggested that this is one of the most important factors for test takers' understanding, which also contributes to the face validity of the test. Course instructors also felt that only one skill should be measured in each question. They claimed that this would enable students to recognize, without difficulty, what is being measured in each question, which is crucial for contributing to the face validity of a test. There was disagreement among the instructors for certain reasons. One of them supported the idea of writing the skills above each part to increase face validity while the other thought this might be confusing for test takers and suggested using clear terms while forming the root of the question. For instance, instead of writing the skill to be measured as 'understanding the implied information' at the top of the part, 'What is implied in line 5?' could be better for test takers' perception of face validity. As for the second aspect of the interview, the instructors asserted that the incongruence between instructors' and test takers' perceptions of the skills to be measured might have resulted from the difference in awareness level of participants. As asserted by the instructors, it is a clear fact that the knowledge and awareness levels of test takers and course instructors is not the same. Therefore, there existed a difference in the perceptions of both sides, which is supported by the results of this study.

Test takers, on the other hand, focused on several primary factors, which were also emphasized by course instructors. To begin with, test takers agreed on the importance of noting the skills to be measured explicitly for each part of the test. They believed this is especially vital when the parts seem similar because it will help them to recognize easily the skills being tested and to answer the questions accordingly. Familiarity is another factor they discussed. The test takers claimed that being familiar with the types of questions and terminology used in the tests increased face validity. The final point worth mentioning is the reason for the lack of agreement in their perceptions and test designers' intentions. They proposed that this incongruence might stem from their low levels of English as well as their lack of knowledge and low level of awareness regarding language learning, which was confirmed by course instructors. In summary, the findings of the interviews parallel the results of the study to a great extent, with instructors having a higher agreement rate than test takers with test designers' intentions.

The number of people interviewed seems inadequate. Why haven't you interviewed all of the instructors (7), test designers (4) and about 10 students? Interviewing only 4 people out of 45 or 49 may not produce reliable results.

4. DISCUSSION

First of all, the data set was analysed to find an answer to the first research question pertaining to the agreement level of items measuring specific skills among participants. The agreement rate was higher between test designers and teachers, which was normally expected and reported in the literature (Sato & Ikeda, 2015; Xie, 2011). However, differences in the perceptions of teachers and test takers might result from a variety of skills being measured within the same test component and some of the items in the response papers represent similar skills, so these teachers and test takers might not have chosen exactly the same items in the response sheet as did the test designers. However, this does not mean that they entirely do not recognize what the items measure because there is still agreement, to some extent, between test developers' intentions and the responses of instructors and test takers. In a similar vein, Grand et al. (2010) suggested that in their study face validity had nonsignificant effects on test performance and test perceptions and did not affect the psychometric properties of either test.

The highest agreement rate between teachers and test developers was in the part where test takers were asked to rewrite some sentences using the given conjunction as a prompt. The items were mainly concerned with understanding tense structure in a context as well as the basic structure and rules of English sentences and understanding the connection among sentences, respectively. However, the lowest level of compromise between teachers and test designers was in the part where test takers were invited to read a passage and decide if the given statements were true or false. These results may suggest that this part was the most comprehensible for test takers since they all seemed to understand the given task although they may not have received the highest grade in the same part as this was another factor incorporating test takers' ability to learn or study regularly and so on (Chan & Schmitt, 1997; Kafer & Hunter, 1997). With this high mean percentage, it can be concluded that face validity for this part seems quite high and test takers' chances of performing well in this part was a lot higher than in other components. This situation is also likely to result in a positive washback effect on test takers' learning (He & Shi, 2008).

5. CONCLUSION

The purpose of this study was to investigate the agreement between test takers and language instructors' perceptions and test designers' intentions about the skills to be measured by a small-scale quiz. The results indicate that the general agreement rate between test takers and test designers is quite low, while the agreement between language instructors and test designers is higher because they are more content-conscious than the students, with a percentage rate which is almost twice as much as test takers' agreement level. These low agreements contradict the findings of a former study (Sato & Ikeda, 2015), and this maybe because they investigated test takers' perceptions of high-stakes tests, whereas the test in this study was a small-scale test. The high agreement rate in the latter study might have resulted from the scale of the exam. The fact that the tests analysed were high-stakes test might have caused higher agreement rates since test takers had all been familiar with the exams conducted in their countries for many years, which possibly helped them to better perceive test designer intentions; hence, this might have led to a higher agreement rate. The advantage of familiarity by test takers with tests is also reported by the findings of a study on the perceptions of test takers by He and Shi (2008).

Although the overall agreement in this study seems low on the side of learners, some parts have relatively higher mean percentages, suggesting different levels of face validity for specific parts. The findings from the interview data reveal that the parts with high agreement rates are those in which test takers state to understand easily, which is a sign of better perceived intentions and skills being required. As Xie (2011) suggested, only if test takers perceive the skills being tested can they adopt and apply them to fulfil the requirements of the test in question. This assertion parallels the high agreement rate for specific parts which test takers have reported in the interviews as being easy to comprehend and answer appropriately. What can be concluded from test taker interview findings about the parts in which they could understand which skills they were supposed to apply clearly indicate that there is a close relationship between high agreement rates and higher face validity features.

Nevertheless, the gap in perception of test taker and test designers intentions results from low face validity and this may lead to unintended washback effects on test takers' learning, as suggested by Watkins et al. (2005). The findings of Grand, Ryan, Schmitt and Hmurovic (2010) are consistent with this in that more valid tests decrease the level of difference in performance between genders across different domains. That is, it was found out that test takers' performance was better when more valid texts were introduced in tests.

Another important point is that language instructors' guidance to test takers in the preparation for tests could be useful for avoiding the outcomes of negative washback effect and for increasing face validity. Having recognized the importance of the role of language instructors and their possible positive effects on test takers' learning, instructor perception was included into the study as a new dimension along with the ones in the literature (Sato & Ikeda, 2015). I was found that instructors' agreement rate almost doubled that of test takers. However, it can be said that this agreement with test designers still is not at a desirable level when the negative reflections of test takers were considered in terms of relatively low agreement rates. Hence, it should be noted that language instructors are essential for raising awareness in many aspects throughout the learning process in educational settings. Moreover, in addition to helping their test takers prepare for exams, it is important that teachers also perceive what test designers intended to measure to be able make their test takers more conscious about the test and help to obtain positive washback effects as a result of increased understanding of collaboration between test designers and course instructors (Grand et al., 2010).

6. IMPLICATIONS

This study has been conducted to examine the agreement in the perceptions of test takers, language instructors and test developers' intentions. It was based on the analysis of a small-scale test administered at a college with a small number of test takers. First implication would regard the effects of this overlap on test takers' actual learning and success, which has not yet been researched. The relationship between test takers' understanding of the face validity of tests and their language achievements could be examined in further studies. Researchers should also place more emphasis on the role of language instructors in this respect and investigate the effects of awareness-raising on test takers' perceptions of the skills measured in tests. Moreover, they should analyse outcomes in terms of face validity of the test administered.

7. LIMITATIONS

The main limitation of this study was the low number of participants. Although the qualitative nature of the study did not facilitate data collection and analysis from a great number of participants, it could yield better results with more participants. Another important point that might be improved in future research was about demographics of the participants. All the participants were students who will study at engineering departments. Though they did not start studying at their departments yet, it is always safe to assume this unity in learner profiles might prove to be a limitation for this research.

REFERENCES

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Barnett, L. M., Ridgers, N. D., Zask, A., & Salmon, J. (2015). Face validity and reliability of a pictorial instrument for assessing fundamental movement skill perceived competence in young children. *Journal of Science and Medicine in Sport*, 18, 98–102. https://doi.org/10.1016/j.jsams.2013.12.004

Brown, A. (1993). The role of test-taker feedback in the test development process: test- takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing, 10,* 277-303.

Brown, HD, & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (2nd ed.). White Plains, NY: Pearson Education.

Brown, J. D. & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge: Cambridge University Press.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology*, 82(1), 143–159. https://doi.org/10.1037/0021-9010.82.1.143

Davies, A, Brown, A, Elder, C, Hill, K, Lumley, T, & McNamara, T. (1999). *Dictionary of language testing.* Cambridge: Cambridge University Press.

Elder, C. (2007). *OET-IELTS benchmarking study report*. Australia: Language Testing Research Centre, University of Melbourne.

Grand, J. A., Ryan, A. M., Schmitt, N., Hmurovic, J. (2010). How Far Does Stereotype Threat Reach? The Potential Detriment of Face Validity in Cognitive Ability Testing. *Human Performance,* 24(1), 1-28. http://doi.org/10.1080/08959285.2010.518184

He, L., & Shi, L. (2008). ESL test takers' perceptions and experiences of standardized English writing tests. *Assessing Writing*, *13*(2), 130-149.

Hill, K. (1998). The effect of test-taker characteristics on reactions to and performance on an oral English proficiency test. In A.J. Kunnan (Ed.) *Validation in language assessment* (pp. 299-323). Mahwah, NJ: Lawrence Erlbaum.

Holden, R. R. (2010). Face Validity. *The Corsini Encyclopedia of Psychology*: John Wiley & Sons, Inc.

Hughes, A. (2003). *Testing for language teachers.* Cambridge: Cambridge University Press.

Kafer, K. L., & Hunter, M. (1997). On testing the face validity of planning/problem-solving tasks in a normal population. *Journal of the International Neuropsychological Society*, 3(2), 108–119.

Kim, H., & Elder, C. (2015). Interrogating the construct of aviation English : Feedback from test takers in Korea. *Language Testing*, *32*(2), 129–149. http://doi.org/10.1177/0265532214544394

Kelecioğlu, H., & Şahin, S. G. (2014). Geçmişten Günümüze Geçerlik. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 5(2), 1–11.

Knoch, U. (2014). Using subject specialist to validate an ESP rating scale: The case of the International Civil Aviation Organization (ICAO) rating scale. *English for Specific Purposes*, 33, 77–86.

MacLellan, E. (2001). Assessment for learning: The differing perceptions of tutors and test takers. *Assessment & Evaluation in Higher Education*, *26*(4), 307–318.

Nevo, B. (1985). Face Validity Revisited. *Journal of Educational Measurement*, 22(4), 287–293. https://doi.org/10.1111/j.1745-3984.1985.tb01065.x

Sato, T., & Ikeda, N. (2015). Test-taker perception of what test items measure: a potential impact of face validity on student learning. *Language Testing in Asia*, *5*(10). http://doi.org/10.1186/s40468-015-0019-z

Scouller, K. M., & Prosser, M. (1994). Students' experiences in studying for multiple choice question examinations. *Studies in Higher Education*, 19(3), 267-279.

Schmitt, N. (2002). Do reactions to tests produce changes in the construct measured? *Multivariate Behavioral Research*, *37*(1), 105–126. http://doi.org/10.1207/S15327906MBR3701

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing*, *18*(4), 373-391.

Paker, T. (2013). The backwash effect of the test items in the achievement exams in preparatory classes. *Procedia-Social and Behavioral Sciences*, *70*, 1463–1471.http://doi.org/10.1016/j.sbspro.2013.01.212

Teemant, A. (2010). ESL student perspectives on university classroom testing practices. *Journal of the Scholarship of Teaching and Learning*, *10*(3), 89-105.

Tsaia, Y., & Tsou, C.-H. (2009). A standardised English language proficiency test as the graduation benchmark: Student perspectives in higher education. *Assessment in Education: Principles,Policy & Practice*, *16*(3), 319-330.

Watkins, D., Dahlin, B. O., & Ekholm, M. (2005). Awareness of the backwash effect of assessment : A phenomenographic study of the views of Hong Kong and Swedish lecturers. *Instructional Science*, (33), 283–309. http://doi.org/10.1007/s11251-005-3002-8

Xie, Q. (2011). Is test taker perception of assessment related to construct validity? *International Journal of Testing*, *11*, 324–348. http://doi.org/10.1080/15305058.2011.589018

SUMMARY

Giriş

Dil becerilerinin ölçülmesi ve değerlendirilmesi eğitim öğretim çalışmaları kapsamında önemli bir alandır ve eğitim öğretim süreçlerini de şekillendiren önemli sonuçlara sahiptir. Sınavlar farklı amaçlar için hazırlanıp uygulanabileceği için bu amaçlardan her birisi öğrenme ve öğretme süreçleri ile bu süreçlere dâhil olan bireyleri etkilemektedir. Bu sebeple, sınavların farklı özelliklerini ve etkilerini araştırmak üzere birçok araştırma gerçekleştirilmiştir (Kim, & Elder, 2015; Schmitt, 2002; Watkins, Dahlin, & Ekholm, 2005). Dil eğitimi özelinde bakılacak olursa, dil öğrenimi sınavların çok sık bir şekilde uygulandığı dinamik bir süreç olduğu için sınavların bu sürece etkileri de o oranda fazladır. Böylece sınavların özelliklerinin ve etkilerinin incelendiği çalışmalar, sınav hazırlayıcıları, öğretmenler ve öğrenciler açısından çok önemlidir.

Sınavların özellikleri ve etkilerinin yanı sıra sınava giren öğrencilerin sınav ile ilgili algıları da öğrenme sürecine direk etki eden ve öğrencinin sınav performansını etkileyen unsurlardan bir tanesidir. Uygulanan sınavlar genellikle bir dersin programını ve içeriğini etkiler. Böylece sınava girecek öğrenciler de genellikle sınavda ölçüleceğini düşündükleri içerik ve beceri üzerine daha fazla odaklanırlar. Bu duruma kaba tabir ile sınavın yankılanma (backwash) etkisi denir (Bachman, 1990). Hughes'ün (2003) ifadeleri ile öğretmenler ya da sınav hazırlayanlar sınıfta öğretilen şeyleri sınavlarında ölçmeliler çünkü sınava giren öğrenciler sınavda ölçülmeyeceğini bildikleri konulara genellikle çalışmazlar. Öte yandan, sınavda ölçülen konular ve sınavdaki maddeler eğer öğretim programı ve içerik ile uyumlu ise olumlu bir yankılanma etkisi gözlemlenir, aksi durumda ise yankılanma etkisi olumsuz olur (Brown & Hudson, 2002). Paker (2013) ölçme değerlendirme alanında yaptığı çalışmada İngilizce hazırlık programlarında uygulanan yeterlik sınavlarının yankılanma etkisinin olumlu ve yapıcı olacak şekilde tasarlanması gerektiğini ortaya koymuştur.

Sınava girecek öğrenciler için olumlu ve motive edici bir öğrenme ve sınav ortamı oluşturmak hem sınav güvenilirlik ve geçerlik kavramlarının temel unsurlarındandır hem de sınavı hazırlayanların amaçlarından bir tanesidir. Bu amaçla, bu çalışmada sınava giren öğrencilerin, sınavı hazırlayan ve derse giren öğretim elemanlarının sınavdaki maddelerin hazırlanma amaçları ile ilgili görüşleri araştırılmıştır.

Bu çalışmanın temel kavramını oluşturan görünüş geçerliği kavramı "bir testin ölçtüğünü iddia ettiği bilgi ve becerileri ne kadar ölçtüğü" şeklinde tanımlanabilir (Davies, Brown, Elder, Hill, Lumley, & McNamara., 1999, p.59). Görünüş geçerliği az araştırılmış bir konudur, bunun sebebi olarak da sınavı hazırlayan öğretmenlerin ya da uzmanların genellikle öğrencilerin değerlendirmelerinin güvenilirliğinden emin olmamalarıdır (Xie, 2011). Benzer bir şekilde konunun farklı bir tarafından araştıran çalışmalarda sınava giren öğrencilerin kötü performans göstermelerinin sebeplerinden birinin de sınavın düşük görünüş geçerliği olması olduğunu bulmuşlardır; yani, sınavın o sınavı hazırlayanlar tarafından hazırlanırken belirlenen ölçmesi gereken bilgi ve becerileri ölçmüyor gibi görünmesidir (Brown & Abeywickrama, 2010; Hughes, 2003; Bachman, 1990).

Sato ve Ikeda (2015) tarafından gerçekleştirilen görünüş geçerliği araştırıldığı güncel bir çalışmada, bu geçerliğin öğrencilerin öğrenme etkinlikleri üzerine etkisi incelenmiştir, öğrenci ve sınavı hazırlayan öğretmenin sınavı oluşturan soruların ölçtüğü bilgi ve becerileri belirlemesi istenmiştir. Sonuç olarak öğretmen ve öğrencilerin belirledikleri bilgi ve becerilerin birbirini tutmaması sonucunda görünüş geçerliği düşük olduğunu ve bunun da olumsuz yankılanma etkisine yol açtığını bulmuştur.

Tüm bunların ışığında bu çalışmanın amacı Gaziantep Üniversitesi Yabancı Diller Yüksekokulu'nda İngilizce hazırlık eğitimi alan öğrencilerin, öğretim elemanlarının ve sınav hazırlama komisyonu üyelerinin seçilen bir sınav için görünüş geçerliği çalışması yaparak bu üç grup arasında ne kadar uyuşma olduğunu belirlemektir. İkinci olarak da eğer bir uyuşmazlık durumu olması halinde görüşmeler ile bunların olası sebeplerini ortaya çıkartmaktır.

Metot

Yukarıda belirtilen amaçlar doğrultusunda bu çalışmada Gaziantep Üniversitesi YDYO İngilizce hazırlık sınıfında öğrenim gören 38 öğrenci, aynı birimden yedi öğretim elemanı ve dört sınav hazırlama komisyonu üyesi olmak üzere toplam 49 katılımcı yer almıştır. YDYO'da verilen İngilizce hazırlık eğitimi sekizer haftalık dört modülden oluşmaktadır. Her bir modülde altı haftalık sınav (Quiz), ödev değerlendirme ve çıkış sınavları farklı yüzdeliklerle modül notunu belirler. Çalışma için seçilen haftalık sınavların modül içerisindeki toplam ağırlığı % 10'dur. Belirlenen sınav sekiz bölüm ve 43 maddeden oluşmaktadır.

Cevap kodlama formu ve bire bir görüşme yöntemleri kullanılarak nicel ve nitel olmak üzere iki farklı veri toplanmıştır. İlgili kurumda sınav hazırlama komisyonu tarafından oluşturulan haftalık sınav belirlenerek sınav kapsamındaki her bir maddenin hangi bilgi ve becerileri ölçtüğü katılımcılara sorulmuştur. Katılımcılar kendilerine verilen cevap kodlama formu üzerine her bir maddenin ölçtüğünü düşündükleri bilgi ve becerileri not etmişlerdir. Daha sonra kodlama formu verileri incelenmiş ve elde edilen sonuçlara göre düşük seviyede uyuşmaya sahip olan katılımcılar ile görüşmeler yapılmıştır.

Sonuçlar ve Tartışma

Araştırma sonuçları iki kısım olarak verilmiştir. İlk olarak cevap kodlama formu sonuçlarına göre sınav hazırlayanlar ve derse giren öğretim elemanlarının sınav sorularının amaçlarını eşleştirmede orta düzeyde bir uyum olduğunu, öğrencilerde ise düşük seviyeli bir uyum olduğunu belirlemiştir. Daha detaylı ifade etmek gerekirse, üç farklı gruptaki katılımcılar arasındaki uygunluk hesaplandığında öğretmenler ve sınav hazırlama komisyonu arasında %61.10'luk bir uyum belirlenmiş ve öğrenciler ile komisyon üyeleri arasında ise bu oran %36.32 olarak hesaplanmıştır.

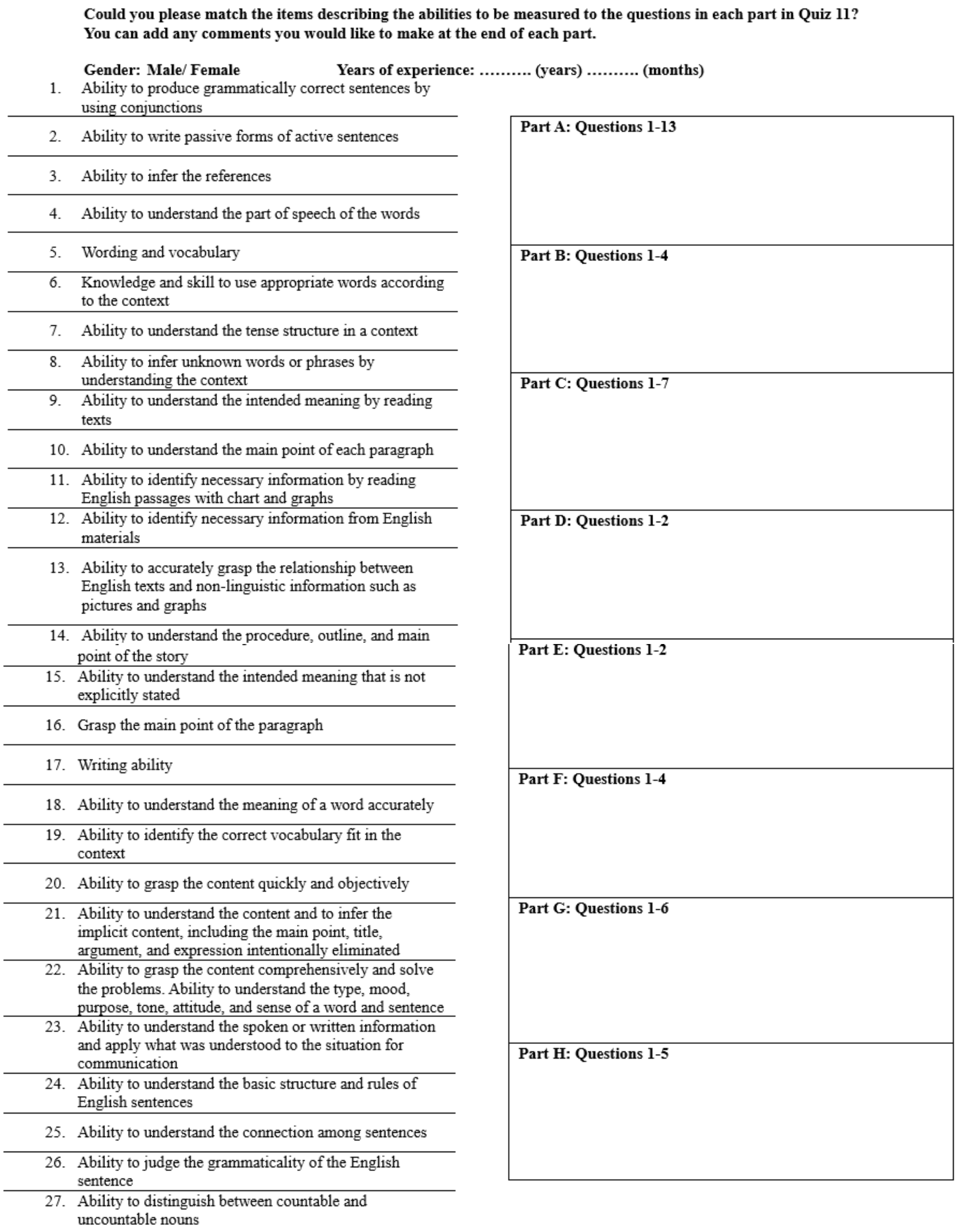
Bu uygunluk seviyelerini sınavın bölümleri arasında karşılaştırmak için her bölüm için öğretmen ve sınav hazırlama komisyonunun verdikleri cevapların birbirlerine uygunluğu her bir bölüm için hesaplanmıştır. Elde edilen sonuçlara göre, en yüksek uygunluk oranı, %82 ile öğrencilerden bir cümleyi verilen kelimeleri kullanarak tekrar yazmalarının istendiği B Bölümünde hesaplanmıştır.

İkinci kısımdaki bulgular yapılan bire bir görüşmelerden elde edilen verilerle oluşturulmuştur. Buna göre, görüşme yapılan iki öğretim elemanı da görünüş geçerliğini bir anlamda belirleyen faktörlerden bahsetmişlerdir. Her iki öğretim elemanına göre sınavın görünüş geçerliğinin olabilmesi için organizasyonu ve düzeni çok iyi ayarlanmalıdır. Sınavın farklı bölümleri kolaylıkla birbirinden ayrılmalı, yönergeler anlaşılır olmalı ve önemli noktaların altı çizilmeli ya da koyu yazılmalıdır. Bir diğer önemli nokta da her bir maddenin ölçtüğü tek bir bilgi ya da beceri olmalı ve kolaylıkla bu madde ilgili bilgi ya da beceri ile ilişkilendirilebilmelidir.

Görünüş geçerliği kuvvetlendirilmesinde önemli bir faktör de sınav maddelerinde verilen önergeler ve soru kökleri ile ilgilidir. Öğretim elemanlarından bir tanesi her bir madde ya da bölüm için ölçülen beceri ya da bilgiyi gösteren bir notun olmasını önermiştir.

Öğrencilerle yapılan görüşmelerde ise öğretim elemanları ile bazı ortak noktalara işaret edilmiştir. Örneğin bir öğrenci soruların ölçtüğü becerinin soru başlığı olarak verilmesi gerektiğini ve özellikle düşük seviye gruplarda bunun öğrenci başarısını artıracağını belirterek dil becerileri olarak ileri seviyedeki öğrencilerin ihtiyacının daha az olduğunu ifade etti. Özetle, yapılan görüşmelerde öğretim elemanları ve öğrenciler verilen sınav sorularının ölçtüğü beceri ve bilgileri belirlemede organizasyon, düzen ve soruların açık bir şekilde ifade edilmesi gibi bir takım faktörlerin önemli bir rol oynadığını belirtmişlerdir.

Appendix: Face Validity Response Sheet



Could you please match the items describing the abilities to be measured to the questions in each part in Quiz 11?
You can add any comments you would like to make at the end of each part.

Gender: Male/ Female **Years of experience: ………. (years) ………. (months)**

1. Ability to produce grammatically correct sentences by using conjunctions
2. Ability to write passive forms of active sentences
3. Ability to infer the references
4. Ability to understand the part of speech of the words
5. Wording and vocabulary
6. Knowledge and skill to use appropriate words according to the context
7. Ability to understand the tense structure in a context
8. Ability to infer unknown words or phrases by understanding the context
9. Ability to understand the intended meaning by reading texts
10. Ability to understand the main point of each paragraph
11. Ability to identify necessary information by reading English passages with chart and graphs
12. Ability to identify necessary information from English materials
13. Ability to accurately grasp the relationship between English texts and non-linguistic information such as pictures and graphs
14. Ability to understand the procedure, outline, and main point of the story
15. Ability to understand the intended meaning that is not explicitly stated
16. Grasp the main point of the paragraph
17. Writing ability
18. Ability to understand the meaning of a word accurately
19. Ability to identify the correct vocabulary fit in the context
20. Ability to grasp the content quickly and objectively
21. Ability to understand the content and to infer the implicit content, including the main point, title, argument, and expression intentionally eliminated
22. Ability to grasp the content comprehensively and solve the problems. Ability to understand the type, mood, purpose, tone, attitude, and sense of a word and sentence
23. Ability to understand the spoken or written information and apply what was understood to the situation for communication
24. Ability to understand the basic structure and rules of English sentences
25. Ability to understand the connection among sentences
26. Ability to judge the grammaticality of the English sentence
27. Ability to distinguish between countable and uncountable nouns

Part A: Questions 1-13

Part B: Questions 1-4

Part C: Questions 1-7

Part D: Questions 1-2

Part E: Questions 1-2

Part F: Questions 1-4

Part G: Questions 1-6

Part H: Questions 1-5