Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi (KEFAD)
Cilt 15, Sayı 3, Aralık 2014, Sayfa 1-13

Öğrenci ve Öğretim Elemanlarının Lisans Akademik Amaçlı İngilizce Sınavları ile İlgili Geçerlilik ve Güvenirlik Algıları

Nesrin ORUÇ ERTÜRK[1]

ÖZ

Bu çalışma özel bir ünivesitede lisans 1.sınıf öğrencilerinin almak zorunda oldukları, Akademik Amaçlı İngilizce dersi için hazırlanan sınavların geçerlilik ve güvenirliğinin araştırıldığı bir eserdir. Çalışmayı alandaki diğer eserlerden ayıran özelliği ise, bu sınavları alan öğrencilerin de veri sağlayıcısı olarak çalışmaya dâhil olmalarıdır. Çalışmanın ana araştırma sorusu hazırlanan testlerin ne derece geçerli ve güvenilir olduğudur. Çalışmaya katılan tüm öğrenci ve öğretim elemanları, çalışmanın bir diğer hedefi testleri hazırlayan ve cevaplandıran bu iki grubun görüşlerini kıyaslamak olduğu için, Likert tarzı bir anket doldurmuşlardır. Çalışmaya 19 öğretim elemanı ve 111 lisans öğrencisi katılmıştır. Analizler Kruskal-Wallis ve ortalamalar kullanılarak SPSS ile yapılmıştır. Sonuçlar testlerde bazı olası sorunları ortaya çıkartırken, anketteki birçok maddede de öğrenci ve öğretim elemanları arasında istatistiksel farklılıklar belirlemiştir.
Anahtar kelimeler: ölçme, değerlendirme, geçerlilik, güvenirlik.

Student and Teacher Perceptions on Reliability and Validity of Freshman EAP Tests

ABSTRACT

This study is an investigation of the reliability and validity of the tests written for freshman EAP course of a private university. The study is different from the empirical assessment of tests, in the sense that the students who took the tests participated in the study as data providers as well. The main research question of the study tested the extent to which the tests prepared; pose the face validity and reliability. Both instructors and students were asked to fill in a Likert Type scale since another focus of the study is to compare and contrast the opinions of both the test takers and test givers on the same test. 19 freshman instructors and 111 freshman students participated in the study. Data was analysed on SPSS using Kruskal-Wallis. Results reveal some potential weaknesses in the tests and statistically significant differences for most of the items between teachers and students.
Keywords: testing, assessment, validity, reliability.

INTRODUCTION

As Berlak, et. al. (1992:186) state tests and other forms of assessment are forms of schooling practice in the same sense as a school curriculum, teacher pedagogy, or school policies and procedures with respect to student discipline, grading, or staff development. The importance of tests for schooling practice has also been discussed by Linn (1986) with respect to five major aims of testing in

[1] Doç. Dr., İzmir Ekonomi Üniversitesi, e-posta: nesrin.oruc@ieu.edu.tr

education: placement in special education, certification of student achievement, teacher certification and recertification, educational assessment, and instructional diagnosis. Kellough et al (p. 418-419); on the other hand, characterize seven purposes of assessment and testing:

- To assist student learning.
- To identify students' strengths and weaknesses.
- To assess the effectiveness of a particular instructional strategy.
- To assess and improve the effectiveness of curriculum programs.
- To assess and improve teaching effectiveness.
- To provide data that assist in decision making.
- To communicate with and involve parents.

Any of the above mentioned points can be emphasized according to the purpose of testing and the context; however, no matter why and to whom, tests as a result present particular forms of discourse. For example, as Berlak et. al. (1992) state, standardized tests lead to discussion of talking about academic achievement in terms of numerical scores, norms and percentiles. For me, as a tester, the test and its results should also communicate with the tester, teacher, student, school administrators and the families with a record which accurately states the test taker's strengths, knowledge and areas of best performance. This information is needed for the tester to decide if the tests written serve the purpose or not. The teacher needs the same information to make any adjustments in his or her teaching depending on the test scores. School administrators and families need this information in order to evaluate their school program (curriculum, teaching, materials, etc.).

I use the term "test" in the sense of a test given for educational purposes with the aim of determining the presence, quality, or truth of something. In a test, a series of questions, problems, or physical responses are designed to determine the knowledge, intelligence, or the ability of the test taker on a specific topic, course, subject, etc. In order to determine this, different types of tasks, each of which can be called as an "item" is used. A test can be composed of many and varied numbers of items. Items can come in the form of multiple choices, in which the test takers are required to make a selection among a set of three or four possible responses, one of which is designated by the test-makers as the correct or the best possible answer. The test taker then fills a space provided, generally on a separate answer sheet which is subsequently machine scored, or in other cases, the students can be asked to answer on the test booklet within a specific predetermined duration of time.

Scores are usually computed by counting correct responses and subtracting this number from the number of incorrect responses. A variety of statistical operations are employed for summarizing test results, so that they may be used for comparing scores of individual or groups. Some tests may include open-ended test items, those which require a writing sample or solving a math problem. In scoring such items, responses are assigned a number by a person

trained in the use of a set of scoring conventions. The scores are then treated in the same way as those derived from multiple choice items (Koretz 2008). Having clarified the meaning of the term, let us continue with the study in hand.

THE STUDY

The study is conducted at a private English medium university in İzmir, Turkey, which provides one year intensive prep education (25-30 hours per week). Freshman students, after their one-year study in prep school, are required to take two courses, ENG 101 and ENG 102, apart from their department courses by the Higher Education Council. All state and private university freshman students in Turkey are supposed to take the same course. This course is taught with different materials and content at different universities. At the university where the study was conducted, ENG 101 and ENG 102 courses are designated as English for Academic Purposes (EAP). In this context, the main goal is to engage first year students in type of activities they are asked to carry out in their academic classes. It is believed that the EAP courses should address the curricula and syllabi of the students from different departments for every semester. The program takes into account the challenges non-native English speakers (NNES) face in their content classes, therefore to reduce this disadvantageous situation for the students, the academicians have decided to write the course books for each faculty. After a three year process, the material design group has produced six separate faculty-specific course books, each based on a needs analysis focusing on target requirements. These faculties were Faculty of Economics and Administrative Sciences, Faculty of Arts and Sciences, Faculty of Fine Arts and Design, Faculty of Communication, Faculty of Engineering and Computer Sciences, and School of Applied Management Sciences, Culinary Arts.

According to Fulcher (1999) testing and assessment in English for Academic Purposes (EAP) contexts has traditionally been carried out on the basis of a needs analysis of learners or a content analysis of courses. This is not surprising, given the dominance of needs analysis models in EAP, and a focus in test design that values adequacy of sampling as a major criterion in assessing the validity of an assessment procedure (Clapham 2000). During this process, teachers and the tester of Freshman English Department were well aware of the fact that any EAP course should first analyse the students' needs, develop a coherent course and sequence of learning, decide on the appropriate tasks and teaching method, apply reliable and valid tests, monitor learning progress and provide effective intervention.

At the time of the study, 1254 freshman students were enrolled in the program and the teaching was given by 20 freshman instructors. 2 tests (one midterm and one final exam) were administered for each department each semester by one tester only. Therefore, basically the tester was responsible for 24 tests throughout the year and 4 make-up exams for the students who were not able to take the midterm or final because of health issues.

It is very common in the literature for an institution to test the reliability or the validity of their tests. It is actually seen as a fundamental requirement for the development of the Testing Unit and the testing process. This study emerged in a different way, because in addition to the need for the institutional validation of the tests, the tester had a personal interest in analyzing the tests she herself has written.

The research questions posed for the study tested to what extent the tests prepared posses face validity in the eyes of the instructors and the students. Having both instructors and students as participants, another aim of the study is to compare and contrast the opinions of the test takers and test givers on the same tests.

Participants

The data was collected in the academic year of 2009-2010 fall semester. 111freshman students from all six departments participated in the study. The ages of the students ranged from 17 to 22. The gender ratio was 54 % female and 46 % male.

Apart from the students, the second group of participants was the instructors. All freshman instructors (n=19) except for the tester herself participated in the study. Teacher participants had between 12 to 28 years of experience and 5 were females. Among 19, 7 were non-native English speakers.

Instruments

In order to collect data, two different questionnaires were adapted from Küçük & Walters (2009). To serve the purposes of the study, separate questionnaires were designed for instructors and students. The first sections of both questionnaires contained the same questions, concerning participants' perceptions of face validity of the tests. There were items like "The content of the course book was represented in the exams sufficiently." and "There were a variety of tasks used in the exams." Instructors' second section aimed to collect data on scorer reliability. The instructors were asked questions such as "The questions included in the exams permitted objective scoring", "Testing Office provided a detailed answer key" and "The scorers who scored the exam papers were trained." The Teacher Questionnaire had 25 items, 9 of which were the same as the Student Questionnaire.

Students' second section was about issues related to test takers' performance. The test takers were presented with items about their perception of the tests, including items like "Sometimes two (or more) questions in the test seemed to be closely related, so that if I could not answer one question, I could not answer the other question either", "The exams included too many questions." and "The exams included an insufficient number of questions". The Student Questionnaire

had 21 items, again which had 9 same items. Both questionnaires were 5-item Likert scale.

Data Analysis

The scales in Table 1 were used in interpreting the means of the Likert scale items.

Table 1. *Scales Used*

Mean	Degree	Opinion
4,5-5	Very High	Strongly Agree
3,5-4,4	High	Agree
2,5-3,4	Moderate	Undecided
1,5-2,4	Low	Disagree
1,0-1,4	Very Low	Strongly Disagree

Since it was not possible to conduct t-test because of the imbalance in the number of students and teachers as participants (teachers=19, students=111), Kruskal Wallis -a non-parametric test (distribution-free) used to compare three or more independent groups of sampled data- was used. Unlike the parametric independent group ANOVA (one way ANOVA), this non-parametric test makes no assumptions about the distribution of the data (e.g., normality). This test is an alternative to the independent group ANOVA, when the assumption of normality or equality of variance is not met (Sall, Lehman & Creighton 2001).

It should be stated here that ANOVA compares the sample means. But it also assumes the populations to be normal with equal variances, so in fact, it tests whether these populations are identical. The Kruskal-Wallis test does not assume normality or equal variances, and instead of comparing sample means, it compares sample means of ranks. This similarity is the reason why the Kruskal-Wallis test is sometimes called "one-way ANOVA on ranks" (Büyüköztürk 2006).

RESULTS

Table-2 below represents the perceptions of teachers and students about the face-validity of the tests administered in the 2009-2010 educational year for ENG 101 and ENG 102 fall and spring midterm and final exams.

Table 2. *Teacher and Student Perception of Face Validity*

Questions	Instructors' Mean	Students' Mean
Q1. The content of the course book was sufficiently represented in the exams.	1.3	2.0
Q2. The content of the listening module was sufficiently represented in the exams.	1.5	2.5
Q3. The content of the reading module was sufficiently represented in the exams.	1.3	1.8
Q4. The content of the speaking module was sufficiently represented in the exams.	1.5	2.1
Q5. The content of the writing module was sufficiently represented in the exams.	1.4	1.7
Q6. Main objectives of each unit were sufficiently represented in the exams.	1.4	2.0
Q7. The vocabulary taught in the courses were sufficiently represented in the exams.	2.0	1.7
Q8. The task types made in the courses were sufficiently represented in the exams.	1.3	1.9
Q9. There were a variety of tasks used in the exams.	1.5	2.1

As can be seen in the table, except for item number 7, instructors' perception of face validity is lower than the students for all items. The biggest difference is observed in question 2, concerning whether the listening module was sufficiently represented in the exam. Instructors' mean for this item is low; however, students' mean is moderate. Among the four skills, both teachers and students identified the listening module content as being represented in the exams most sufficiently. Writing, however, was the skill which was represented the least in the exams according to the participants.

Instructors considered the face validity of the exams to be very low (m=1.0-1.4) for all items apart from the listening (m=1.5), speaking (m=1.5), vocabulary (m=2.0), and the variety of tasks used in the exam (m=1.5). Students, on the other hand, considered face validity to be low (m=1.5-2.4) to moderate (m=2.5-3.4) for all items on the questionnaire.

As was stated above, because of the incomparable number of the participants the mean scores of two groups could not be compared. Therefore, instead of ANOVA, Kruskal Wallis was used to compare the groups.

The results of Kruskal Wallis indicate no difference for the items 4 and 5 between teachers and students, which mean both teachers and students, considered reading and speaking modules were sufficiently represented in the

exams. However, for the remaining 7 items, the results indicate significant difference between the groups.

Table 3. *Kruskal Wallis Results*

Item #	Level of Significance
Q1. The content of the course book was sufficiently represented in the exams.	.012
Q2. The content of the listening module was sufficiently represented in the exams.	.002
Q3. The content of the reading module was sufficiently represented in the exams.	.005
Q4. The content of the speaking module was sufficiently represented in the exams.	.115
Q5. The content of the writing module was sufficiently represented in the exams.	.160
Q6. Main objectives of each unit were sufficiently represented in the exams.	.014
Q7. The vocabulary taught in the courses were sufficiently represented in the exams.	.063
Q8. The task types made in the courses were sufficiently represented in the exams.	.067
Q9. There were a variety of tasks used in the exams.	.025

The largest group of participants was the students and the table below represents their perceptions of test structure. In this part of the student questionnaire, the students were asked about some basic issues about the tests that they have taken. Two important things about the exam were time and points allotted. The students disagree with the statements about the distribution of the points for each section (m=1.9) and the adequacy of the time allowed for the exam (m=1.6). Another item which students disagreed with was the extent to which they have been informed about the writing criteria after the tests (m=1.9). The students disagreed with the item that they were shared the criteria after the tests were administered. The item which had the highest mean concerned the insufficient number of questions on the tests. The test takers with a mean score of 3.4 (moderate to high) stated that they found the number of questions on the tests insufficient.

Table 4. *Students' Perceptions of Test Structure*

Questions	Mean
Q1. Sometimes two (or more) questions in the test seemed to be closely related, so that if I could not answer one question, I could not answer the other question either.	2.5
Q2. The exams included too many questions.	2.7
Q3. The exams included an insufficient number of questions.	3.4
Q4. The instructions explaining what to do in each section in the exams were explicit and clear.	2.1
Q5. The points allotted for each section of the exam were always stated in the exam papers.	1.6
Q6. Time given to the students to complete the exam was always stated in the exam paper.	1.9
Q7. The questions in the exams had different difficulty levels.	1.9
Q8. The exam questions were explicit and clear.	2.4
Q9. The layout of the exam papers was clear.	2.0
Q10. The exam papers were legible.	1.8
Q11. In general, the structure of the tests helped me to display my best performance in the exams.	2.7
Q12. Information about how much the given tests would affect the final grade was always announced.	1.7
Q13. The instructors helped us to get used to the format of the exams.	1.7
Q14. The time given to complete the exams was enough.	2.9
Q15. The environmental conditions of the classrooms in which I took the tests were appropriate.	2.6
Q16. The criteria which the exam papers was graded was explained to me.	1.9

Teachers, as the second group of participants, were asked to give opinions on reliability. The table below shows the mean scores of 12 items on the questionnaire. The lowest mean score came for the item concerning whether they invigilated their own classes or not. A mean score of 1.1 (very low) showed that most invigilated different classes. The mean for Q12, the instructors' overall perception of scorers' reliability, falls into the range of 'disagree'. This indicates that, in the eyes of the instructors, scores have a low degree of reliability. This overall impression and the low mean scores of most of the items on the questionnaire indicate some potential problems in scorers' reliability. The instructors strongly disagree with the statement that their opinion was sought before the exam was administered, and they disagree with the statement that they believe their colleagues score the exam papers in a reliable manner.

Table 5. *Teachers' Perception of Reliability*

Questions	Mean
Q1. The questions included in the exams permitted objective scoring.	2.2
Q2. Testing Office provided a detailed answer key.	2.3
Q3. The scorers were trained.	2.3
Q4. The rating scales included on the criteria (for writing) helped scoring the exam papers.	1.6
Q5. My opinion was sought before the exam was administered.	1.2
Q6. I had the chance to discuss the answers after the exams with the testing unit.	1.5
Q7. The class which I instructed and the class which I invigilated during the exams were different.	1.1
Q8. I would like to be given the opportunity to grade papers of classes which I do not teach.	3.8
Q9. The deadline affected my scoring practices.	4.1
Q10. I score the exam papers in a reliable manner.	1.3
Q11. All my colleagues score the exam papers in a reliable manner.	2.8
Q12. In general, the scoring system was reliable.	2.0

CONCLUSION and DISCUSSION

The immediate pedagogical implications drawn from the study largely concern the researcher- the tester of the institution and the materials design office. The results derived from the study about face validity and reliability display some potential weaknesses in the tests, test administration and test scoring.

One of the weaknesses that arise is about the representation of goals and objectives of the course in the tests. A test writer should be given clear and well defined objectives for each unit and for the whole course book in order to be able to write questions and items which test the stated objectives. When the objectives are not clear, the tests may have the risk of having low validity and reliability since it is not clear which language points to give weight to on the test. This means the members of the group responsible for the curriculum may need to initiate the process of establishing clear goals and objectives for the course, explaining these to the teacher and making sure that both the teachers and the testers have understood these goals and objectives. It is especially vital for the test writers to grasp these since they have to determine the extent to which these objectives are tested and represented in the tests.

There are some interesting results obtained from the study as well. When it was decided to ask students about their perceptions of test structure, they were considered to be reliable sources for data gathering. However, their replies to some statements on the questionnaire made the researcher question their

reliability as data providers. For example; Q6 asked whether the time given to complete the tests was stated on the exam papers. The students had a mean score of 1.6 which is low in terms of degree, and contradicts the facts. As the test writer, the researcher can confirm that times were clearly stated on the cover page of the exams; however, this shows that students do not always read the cover page and therefore may overlook this information.

This study reveals the importance of giving test writers the required time and support to allow the production of better quality test and test items. In the present study, a single tester, who also had teaching duties, was responsible for carrying out all tests, with almost no time allowed for quality control. In order to sustain good test design and analysis in a school setting, it is recommended that more than one teacher should be involved in the assessment process (Coniam 2009). If it is possible, there should be a testing unit, including an appropriate number of trained, proficient test writers, with enough time and resources to accomplish the task.

There are some limitations to the study as well. It is a fact that teachers understand some of the basic principles of educational measurement. With this background they are capable of recognizing and/or addressing some of the pitfalls in the tests administered. In this study, the teachers were asked to make reliability judgments through intuitive methods. However, it should be stated here as a limitation to the study that the reliability should also be tested with more detailed empirical methods. Another limitation is about the data analysis. Kruskal Wallis, like many non-parametric tests, uses the ranks of the data rather than their raw values to calculate the results. Since this test does not make a distributional assumption, it may not be as powerful as the ANOVA.

To conclude, as Uysal (2010) points out "there is no perfect test that is valid for all purposes and uses" this does not mean, however, that test writers, test takers and school administrations should not make every effort to improve the quality of the tests they write.

REFERENCES

Anastasi, A. (1988). *Psychological testing*. New York, New York: MacMillan Publishing Company.

Berlak, H., Newmann, F. M., Adams, E., Archbald, D. A., Burgess, T. Raven, J. & Rumberg, T. A. (1992). Toward a new science of educational testing & assessment. State University of New York Press: New York.

Büyüköztürk, Ş. (2006). *Sosyal bilimler için veri analizi* (Data analyses for social sciences). Ankara: Pegem.

Clapham, C. (2000). Assessment for academic purposes: Where next? *System*. 28.511-521.

Coniam, D. (2009). Investigating the quality of teacher-produced tests for EFL students and the effects of training in test development principles and practices on improving test quality. *System*, 37. 226-242.

Frederiksen, J. R. & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, 18. 9. 27-32.

Fulcher, G. (1999). Assessment in English for academic purposes: Putting content validity in its place. *Applied Linguistics*. 20.2. 221-236.

Kellough, R.D. and Kellough, N.G. (1999). *Secondary school teaching: A guide to methods and resources: Planning for competence.* Prentice Hill: New Jersey.

Koretz, D. 2008. *Measuring up: What educational testing really tells us.* Harvard University Press: Harvard. USA.

Küçük, F. & Walters, J. (2009). How good is your test? *ELT Journal*. 63.4. 332-341.

Linn, L. R. (1986). Educational testing and assessment: Research needs and policy issues. *American Psychologist.* 41.10. 1153-1160.

Sall, J., Lehman, A., Creighton, L. (2001). *JMP start statistics*. Canada: Duxbury.

Stiggins, R.J. (1994). *Student-centred classroom assessment*. New York: Merrill.

Tomlinson, B. (2005). Testing to learn: A personal view of language testing. *ELT Journal*. 59/1. 39-46.

Uysal, H. H. (2010). A critical review of the IELTS writing test. *ELT Journal*. 64/3. 314-320.

GENİŞLETİLMİŞ ÖZET

Eğitimin hangi aşamasında olursa olsun, ölçme ve değerlendirme amacıyla yazılan sınavların kalitesi her zaman kurumların öncelikli hassasiyetlerinden biri olmuştur. Her ne kadar, sınavı veren kurumun eğitim algısına göre şekillense de, sınavların en temel görevlerinden biri öğrencilerin zayıf ve güçlü noktalarını ortaya çıkarmak, dolayısıyla da öğretim elemanına bu zayıf noktaları güçlendirmek adına yol göstermesidir. "Sınav" sözcüğü ile kastedilen, katılımcının belli bir konu, zekâ düzeyi ya da yeteneği ile ilgili soru, problem ya da fiziksel tepkilerinin ölçüldüğü ortamlardır. Bir sınav, farklı soru tiplerinden oluşabilir. Katılımcıdan beklenen, doğru cevabı bulması ve değerlendiricinin de bazen bir makine bazen de kendisinin doğru cevap sayısına göre katılımcıya belli bir puan vermesidir. Çalışmada kullanılmış olan ana temayı tanımladıktan sonra, çalışmanın detaylarına geçebilirz.

Bu çalışma, İngilizce eğitim veren özel bir üniversitenin lisans eğitiminin 1. sınıfında olan tüm öğrencilerinin 2 dönem boyunca almak zorunda oldukları Akademik Amaçlı İngilizce (ENG101-ENG102) dersi ve bu ders için hazırlanan sınavlarını incelemiştir. ENG101 ve ENG102 derslerinin amacı lisans eğitimine ikinci bir dilde devam edecek olan öğrencilerin bölümlerinde verilen dersleri takip edebilmeleri için gerekli akademik becerileri edinmelerini sağlamaktır. Dersin ana hedefleri dört dil becerisinin geliştirilmesi doğrultusunda şekillendirilmiştir ve dolayısıyla bu derste edinilen becerileri test etmek amacıyla oluşturulan sınavlar da, bu dört beceriyi ölçmek hedefli hazırlanmaktadır.

Her kurum, verdiği derslerin ve dolayısıyla kullanılan sınavların geçerlik ve güvenirliklerini test etmek ister. Bu çalışmada yöneltilen araştırma sorusu da öğrenci ve öğretim elemanlarının değerlendirmesine göre bu sınavların geçerlik ve güvenirliklerinin ne ölçüde olduğudur. Çalışmaya 2009–2010 güz döneminde üniversitedeki 6 fakültede okuyan yaşları 17 ve 22 arasında değişen 111 lisans öğrencisi katılmıştır. Katılımcıların %54'ü bayan, %46'sı ise erkektir. Bu öğrenciler halen eğitim aldıkları Mühendislik ve Bilgisayar, İşletme, Güzel Sanatlar ve Tasarım, Fen-Edebiyat, İletişim Fakültesi ve Uygulamalı Yönetim Bilimleri Yüksekokulu (Mutfak Sanatları) öğrencileridir. Her bir fakülteden öğrenci sayısının belli bir yüzdesi oranında katılımcı çalışmaya dahil edilmiştir.

Öğrenciler dışında bir de bu bölümde (Lisans İngilizce Bölümü) derse giren 19 öğretim elemanı çalışmaya dâhil olmuştur. Bu gruptaki katılımcılar 12 ila 28 yıl arasında deneyime sahiptir ve 7 tanesi yabancı uyrukludur. Çalışmaya katılan öğretim elemanlarının 9 tanesi kadın geri kalan 10 tanesi erkektir. Bu öğretim elemanlarının çeşitlilik göstermek koşuluyla en az lisans eğitimi en çok ise iki katılımcının doktora eğitimi bulunmaktadır. Yabancı uyruklu katılımcılar 7 ila 20 yıl arasında değişen rakamlar doğrultusunda ülkemizde bulunmaktadırlar.

Çalışmada öğrenci ve öğretim elemanları için iki ayrı anket kullanılmıştır fakat her iki ankette de ortak bir bölüm bulunmaktadır. Likert tipi anketlerin

analizinde Kruskal-Wallis kullanılmış ve analizler SPSS aracılığıyla değerlendirilmiştir.

Analiz sonuçları bize gösteriyor ki öğretim elemanlarının yüzeysel geçerliliğe ait algıları anketteki her soru için öğrencilerinkine oranla daha düşüktür. Öğretim elemanları sınavlardaki yüzeysel geçerliliği çok düşük bulurken bu öğrenciler için daha yüksektir. Öğretim elemanları ve öğrencilerin anketlerinde bulunan ortak bölüm sorularının her iki grup açısından ortalamaları incelendiğinde 2 soru hariç gruplar arasında farklılıklar gözlemlenmiştir. Her iki grupta okuma ve konuşma becerilerinin sınavlarda yeterince sorulduğunu düşünürken, geri kalan 7 maddede farklılıklar bulunmuştur. Öğrencilere verilen anket sonuçlarına göre, katılımcılar kendilerine verilen sınavları iki konuda eleştirmişlerdir. Bunlardan birincisi sınav için verilen süre ikincisi ise sınavda belli bölümlere verilen puanlardır. Katılımcılara göre sınavda kendilerine verilen süre yeterli değildir.

Öğretim elemanlarına ise daha çok güvenirlik algıları sorulmuştur. Sonuçlara göre, katılımcılar kendi sınıflarında gözetmenlik yapmamanın güvenirliği arttırdığını fakat değerlendiricilerin güvenirliği açısından bazı sorunlar olduğunu ve genel olarak verilen sınavların güvenirliğini düşük bulduklarını çünkü sınavların hazırlanması aşamasında kendilerinin fikirlerinin sorulmadığını belirtmişlerdir.

Bu çalışmanın sonuçlarından yapılacak çıkarsamaların ilk ve belki de en önemlisi kurumun test yazma birimi ve materyal geliştirme birimi ile ilgilidir. Çalışma sonucunda edinilen bulgulara göre sınavların yüzeysel geçerliği ve güvenirliğine dair bazı zayıf noktalar ortaya çıkmıştır. Bunlardan ilki ders hedeflerinin sınavlarda yeterince test edilmemesidir. Buna göre, bundan sonra sınav yazarlarına genel ders hedefleri dışında, her bir üniteye ait ders hedef çıktıları verilecek ve sınavlar buna göre yazılacaktır. Hedefler net olmadığında sınavların düşük geçerlilik ve güvenirliğe sahip olması kaçınılmazdır. Ayrıca, bu çalışma, sınav hazırlayan birimlerin ve bu birimlerde çalışan öğretim elemanlarının söz konusu sınavları hazırlamaları için kendilerine yeterince süre verilmesi gerektiğini de ortaya çıkarmıştır. Çalışmanın belirtilmesi gereken bazı eksiklikleri de bulunmaktadır. Çalışmaya katılan öğretim elemanlarına sınavların güvenirliğine dair görüşleri sorulmuş ve bu sadece fikir olarak alınmıştır. Fakat bu sınavların geçerlilik ve güvenirlik değerlendirmeleri daha bilimsel yöntemlerle de yapılmalı ve sonuçlar birbirleriyle kıyaslanmalıdır. Geçerlilik çalışmalarında yıllardır bu dersi veren ve deneyimli öğretim elemanlarının kişisel fikirlerinin alınması elbette çalışmanın önemli noktalarından biridir ancak bu görüşlerin nesnel ve tarafsız olmama ihtimali aynı nedenle çalışmanın geçerliliğini etkilemiştir.

Sonuç olarak hepimizin bildiği ve Uysal (2010)'un da belirttiği gibi, aslında her amaç ve kullanım için uygun mükemmel bir sınav yoktur. Her sınavın kendi içinde yetersiz olduğu durumlar olabilir önemli olan bu zayıflıkları ortaya çıkarmak ve gidermek adına çaba sarfetmektir.