GEFAD / GUJGEF 33(3): 531-548 (2013)

Standard Assessment and Alternative Assessment in English Language Teaching Program

İngilizce Öğretmenliği Programında Standart Değerlendirme ve Alternatif Değerlendirme

Cemal ÇAKIR

Gazi Üniversitesi, Yabancı Diller Eğitimi Bölümü, İngiliz Dili Eğitimi Anabilim Dalı.
e-posta: ccakir@gazi.edu.tr

ABSTRACT

Classroom-based assessment practices within English as Second/Foreign Language (ESL/EFL) contexts have started to appear in the literature. Despite the importance of assessment in FL teaching, studies on different assessment methods at college level FL have remained limited. In this study, scores of EFL trainee teachers from a multiple-choice test, an oral presentation, and a translation are analyzed, and success levels of ten highest multiple-choice test scorers and ten lowest multiple-choice test scorers in two alternative assessment tools are compared. Results reveal that assessing EFL learners only through a single tool may not be objective, and that assessment tools should be diversified.

Keywords: *Foreign language teaching, Language assessment, Standard tests, Alternative assessment.*

ÖZ

İkinci/yabancı dil olarak öğretildiği ortamlarda İngilizce'nin sınıf-temelli değerlendirilmesi, alanyazında son zamanlarda yer almaya başlamıştır. Yabancı dil öğretiminde değerlendirmenin önemine rağmen, üniversite düzeyinde yabancı dil öğretiminde farklı değerlendirme yöntemleri konusundaki araştırmalar sınırlı sayıda olmuştur. Bu çalışmada, İngilizce Öğretmenliği Programındaki öğrencilerin çoktan seçmeli bir sınavdan, bir sözlü sunumdan ve bir çeviri görevinden aldıkları puanlar incelenmekte; çoktan seçmeli sınavdan en yüksek notu alan on öğrenciyle, en düşük notu alan on öğrencinin iki alternatif değerlendirme aracındaki başarı düzeyleri karşılaştırılmaktadır. Bulgular, İngilizce öğrenenleri yalnızca bir araçla değerlendirmenin nesnel olmayabileceğini ve değerlendirme araçlarının çeşitlendirilmesinin faydalı olacağını ortaya koymaktadır.

Anahtar Sözcükler: *Yabancı dil eğitimi, Dil değerlendirmesi, Standart testler, Alternatif değerlendirme.*

INTRODUCTION

Assessment is of great importance in foreign language (FL) teaching and it should be approached from different angles in order for it to help develop learners' FL knowledge and skills. Assessment techniques were mostly discrete-point tests like the multiple-choice and true-false tests predominantly in the 1950s and 1960s, the integrative tests like cloze and dictation in the 1970s and early 1980s, and more communicative tests like task-based and other new assessments in the 1980s and 1990s (Brown & Hudson, 1998). Recently, studies investigating classroom-based assessment practices within the ESL/EFL school contexts have begun to appear (Cheng, Rogers, & Hu, 2004). Also, the past decade has witnessed such alternative forms of assessment as portfolios, work samples, and classroom-based teacher assessment (Leung & Lewkowicz, 2006). However, college level FL assessment is rarely studied by the researchers (Norris, 2006).

Assessment procedures range, on a continuum, from discrete-point tests to more open-ended performance assessments (Brown & Hudson, 1999). Various definitions of assessment include: (a) "The process of collecting information about a student to aid in decision making about the progress and language development of the student" (Cheng et al., 2004, p. 363); (b) "the systematic gathering of information about student learning in support of teaching and learning" (Norris, 2006, p. 579); and (c) "a general term that includes the full range of procedures used to gain information about student learning" (Linn & Gronlund, 2000, p. 31, cited in Sullivan, 2006, p. 591).

When it comes to evaluation, it is defined differently from assessment. Evaluation is often defined as (a) "using the evidence from assessment data to judge the worth or effectiveness of students or services" (Gottlieb, 2006, p. 186); (b) "the interpretation of assessment results that describes the worth or merit of a student's performance in relation to a set of learner expectations or standards of performance" (Cheng et al.,

2004, p. 363); or (c) that "evaluation brings evidence to bear on the problems of programs, but the nature of that evidence is not restricted to one particular methodology" (Norris, 2006, p. 579).

In the assessment literature, the term testing is used as well, which is used to mean a systematic procedure by which a sample of student behaviour at one point in time is collected (Gottlieb, 2006), or "one particular form of assessment" (Leung & Lewkowicz, 2006, p. 212). Measurement is also used as an alternative term to testing (Sullivan, 2006). Through assessment and evaluation, information on students' progress is obtained, feedback is provided to students as they progress through the course, strengths and weaknesses in students are diagnosed, final grades for students are determined, students are motivated to learn, and growth in learning of students is formally documented (Cheng et al., 2004).

In standard tests students are presented with language and required to pick the correct answer from among a limited set of options, no language is created by the students (Brown & Hudson, 1998). Typical standardised assessment techniques are true-false, matching, and multiple-choice. It is administered, scored, and interpreted in the identical manner without considering when it is given (Gottlieb, 2006). Standardised testing is often favoured because (a) they are useful for measuring a number of different kinds of precise learning points (Brown & Hudson); (b) they are objective (Brown & Hudson; Simkin & Kuechler, 2005); (c) they are convenient (Rowley, 1974); and (d) large numbers of test takers can take them, a large number of questions can be asked, they are to student advantage, student anxiety is reduced, inconsistent grading is avoided, and timely feedback is offered.

On the other hand, a great number of scholars report weaknesses of standard assessment. Standard tests are not able to "accurately and fairly measure student understanding of course concepts" (Simkin & Kuechler, 2005, p. 74); they cannot represent real-life language (Brown & Hudson, 1998); and they offer students success due to guessing (Henning et al., 1981); they cannot "adequately document learner strengths or capture actual progress" (Balliro, 1993, p. 558); and they "cannot on their

own tell teachers much about how learners are acquiring academic contents. Thus, as suggested by Barootchi and Keshavarz (2002), "these instruments, if used as the sole indicators of ability and/or growth, may generate faulty results" (p. 280).

What's more, they restrict what to be tested, their backwash effect may be harmful, and they may facilitate cheating (Hughes, 1990); no opportunity is offered to the learner to behave as an individual (Underhill, 1992); and their results do not often truly indicate learners' performance, and references made from them may not always be valid (Gottlieb, 2006). Moreover, student knowledge may be hidden rather than being revealed, students can be denied "the opportunity to organize, synthesize, or argue coherently, to express knowledge in personal terms, or to demonstrate creativity" (Simkin & Kuechler, 2005, p. 76). Finally, as Norris et al. (1998, p. 15) and Braun and Mislevy (2005, p. 495) say, they measure ability to recognize or recall only and cannot measure higher order thinking skills.

When it comes to alternative assessment, it has been approached in many different ways. McNamara (2001) sees it as a movement "away from the use of standardized multiple-choice tests in favour of more complex performance based assessments" (p. 329). Lynch (2003) notes that alternative assessment "views language ability and use as a reality (or realities) that do not exist independently of our attempts to know them" (p. 6). The purpose of alternative assessment is to collect information on and document the abilities, skills, progress, and attitudes of the students (Varela, 1997). In alternative assessment, (a) learners acquire problem solving and higher level thinking skills, (b) real-world contexts or simulations are utilised, and (c) both process and products are focused on (Norris et al., 1998).

Procedures of alternative assessment include checklists, journals, logs, videotapes and audiotapes, self-evaluation, teacher observations, portfolios, conferences, diaries, self-assessments, and peer assessments and so on (Brown & Hudson, 1998). Alternative assessment can have a lot contribution in FL teaching. It can provide valuable information about learners' performance in educational contexts (Barootchi & Keshavarz, 2002); it "connects students' experiences with the curriculum through active

involvement", and has "students produce original work around major themes, ideas, or issues", encouraging deep learning and supporting in-depth teaching (Gottlieb, 2006, p. 111, 123); and it can minimise the bad washback effect of standardized tests, helping align classroom assessment and classroom activities with authentic, real-life activities (Norris et al., 1998).

As for the limitations of alternative assessment, it can be observed that designing authentic performance is often extremely complex (Leung & Lewkowicz, 2006). Simkin and Kuechler (2005) maintain that in alternative assessment measures, teachers need a lot of time to grade them. Furthermore, Norris et al. (1998) raise the question if alternative assessment "adequately covers all skills, processes, and knowledge related to the task" (p. 18). It is also reported that studies focusing on how teachers assess their students' foreign language skills while teaching and learning are very few (Edelenbos & Kubanek-German, 2004). In addition, there is limited literature on what is happening at the classroom level of test development (Alderson & Banerjee, 2001, cited in Leung & Lewkowicz).

In one of the few studies, Rowley (1974) investigated the use of a multiple-choice test in measuring vocabulary and found that "that the use of multiple choice tests can produce scores which favour certain types of examinees and penalize others for reasons not explainable in terms of their knowledge of the material being tested" (p. 21). In another study, Barootchi and Keshavarz (2002) sought if there is any correlation between portfolio assessment scores and those of teacher-made tests and found a correlation between the scores of portfolio assessment and those of the tests made by teachers.

This study investigates the assessment in an elective course and aims to analyze the scores obtained through three different assessment tools: a multiple-choice test, an oral presentation, and a translation. It makes a comparison and contrast between the success levels of ten highest multiple-choice test scorers and ten lowest multiple-choice test scorers in two other tools.

METHOD

Participants

178 freshman trainee teachers at an ELT Program of a Faculty of Education in Ankara, Turkey participated in the three assessment tools. They constituted almost 75% of the freshman students at the ELT Program in question. Since they had all passed a national English proficiency test and a preparatory school exemption exam, in this study their English proficiency levels were considered to be almost homogeneous. They all followed the same curriculum and most of the activities and assignments they had had and were having at the time being were similar.

Procedure

A multiple-choice test and two assignments were given in an elective course and the test was given as the final exam; translation was assigned as an end-of-term work and oral presentations were made during the term. The multiple-choice test was composed of 15 fill-in-the-blanks items for prepositions, 15 items for matching collocations, 10 fill-in-the-blanks items for collocations and 10 fill-in-the-blanks items for clichés.

In the second tool, each student as a member of a team gave a 7-minute oral presentation to an audience of 25-30 classmates, in the presence of the course instructor (the present author), and compared and contrasted a pair of mass communication channels. The assessment criteria for the presentation were given to the students before they started. An oral presentation evaluation form was followed while the students were presenting. Each presentation was evaluated for language components, performance, and body language.

In the third tool, they translated news pieces from local Turkish newspapers out of class and submitted them at the end of the term. They were evaluated in terms of syntactic, semantic and pragmatic equivalence. Special attention was paid to the use of formulaic language in general and collocations in particular. News about events in very specific locations in Turkey were required in order to make students completely translate the

text by themselves because, in the case of national and international events, they would probably have found English versions of the events reported in the international media.

When it comes to data analysis, grade means for three assessment tools were calculated. Next, ten highest multiple-choice test scorers (Group A) and ten lowest multiple-choice test scorers (Group B) were chosen on purpose with a view to seeing differences far more clearly as these groups were expected to provide the most significant differences. Then, their scores in oral presentation and translation were compared to their multiple-choice test scores. Finally, for each student, the differences between (a) the score of multiple-choice test and that of oral presentation, (b) the score of multiple-choice test and that of translation, and (c) the score of multiple-choice test and average grade were found and tabulated.

Instruments

A standard assessment tool, namely a multiple-choice test, and two alternative assessment tools, namely a translation task and an oral presentation task, were given to the participants to investigate whether there are differences in students' levels of FL knowledge recognition, of FL knowledge transmission or pseudo-communication, and of FL knowledge application. (a) The multiple-choice test was given (out of 50 points) – to assess FL knowledge recognition**,** (b) the students were assigned to orally present a topic concerning the channels of mass communication (out of 50 points) – to assess FL knowledge transmission or pseudo-communication, and (c) they translated from Turkish into English a piece of news about something that is only concerned with the local people of a town or village in Turkey and that is of no international concern (out of 50 points) – to assess FL knowledge application.

Limitations

This study is limited to descriptive comparison of the scores of students obtained from three different assessment tools. Since all three tests are not parallel and do not assess the same content, a correlation among the test scores obtained from three tools is not sought. Since the multiple choice test and translation task assess written language, and

the oral presentation task assesses spoken language, they are not comparable and no case exists for examining correlations. Furthermore, since the third assessment item, the translation task, assesses syntactic, semantic, and pragmatic equivalence, it is quite different to the other two. Also, since the assessment tools were used as part of a course and there were a lot of groups and presentations, only one instructor, who is the author of this paper, assessed the tools. It was technically impossible for another rater to assess all of the oral presentations. Hence, the lack of interrater reliability is a weakness of the study.

RESULTS

When the grades of all three assessment tools are analysed, different grade intervals are observed. The mean for each assessment is as follows: (a) 38.17 (out of 50 points) for the multiple-choice test; (b) 36.63 (out of 50 points) for the oral presentation; and (c) 34.57 (out of 50 points) for the translation. The means of three instruments reveal that, of three, multiple-choice is the easiest, translation is the most difficult, and oral presentation is somewhat difficult. Possible reasons for difficulty levels of the three instruments will be handled in the discussion part to come.

When the presentation and translation grades of Group A and those of Group B were analyzed, it was observed that there are significant differences between them. As shown in Table 1, of the top 10 scorers of the objective test (45-50 interval), only Student 1 could take place in the 45-50 interval of the presentation assessment and Students 1, 2, 3, and 6 could keep their positions in the 45-50 interval of the translation.

Table 1. Presentation and Translation Grades of 10 Highest Multiple-Choice Test Scorers of Group A

Group A Student	Highest Multiple-Choice Test Score	Presentation Grade	Translation Grade	Average Grade
1	49	50	46	48
2	48	42	50	47
3	48	36	50	45
4	47	36	38	40

5	47	32	34	38
6	46	40	46	44
7	46	34	38	39
8	46	44	38	43
9	46	34	30	37
10	46	44	38	43

On the other hand, as Table 2 indicates, of ten students who got the lowest grades from the objective test (26-30 interval), Student 6 got 46 points from the presentation and Student 5 could get 42 points from the translation. Also, there are three students (Students 5, 8, and 10) who have an oral presentation grade, 38, in the 36-40 interval.

Table 2. Presentation and Translation Grades of 10 Lowest Multiple-Choice Test Scorers of Group B

Group B Student	Lowest Multiple-Choice Test Score	Presentation Grade	Translation Grade	Average Grade
1	30	32	30	31
2	29	34	34	32
3	29	30	26	28
4	29	34	30	31
5	29	38	42	36
6	29	46	34	36
7	28	30	30	29
8	27	38	26	30
9	26	30	30	29
10	26	38	30	31

Table 3 gives the ten highest multiple-choice test scores, presentation grade differences from multiple-choice test scores, and translation grade differences from multiple-choice test scores of Group A. While there is one person, Student 1, with a positive difference of 1 point in presentation grade, and two persons, Students 2 and 3 with positive difference of two points in translation grades, all the students have negative differences in presentation and translation grades. The most striking negative differences are observed in Students 4, 5, 7, and 9.

Table 3. Differences between Multiple-Choice Test Scores and Presentation/Translation Grades of Group A

Group A Student	Highest Multiple-Choice Test Score	Presentation Grade Difference from Multiple-Choice Test Score	Translation Grade Difference from Multiple-Choice Test Score
1	49	+1	-3
2	48	-6	+2
3	48	-12	+2
4	47	-11	-9
5	47	-15	-13
6	46	-6	0
7	46	-12	-8
8	46	-2	-8
9	46	-12	-16
10	46	-2	-8

When it comes to the scores and grades of Group B, as Table 4 shows, there are positive differences between multiple-choice test scores and presentation grades, and between multiple-choice test scores and presentation grades. While there is one person, Student 8, with a negative difference of 1 point in translation grade, all the students have positive differences in presentation and translation grades. The most striking positive differences are observed in Students 5, 6, and 10.

Table 4. Differences between Multiple-Choice Test Scores and Presentation/Translation Grades of Group B

Group B Student	Lowest Multiple-Choice Test Score	Presentation Grade Difference from Multiple-Choice Test Score	Translation Grade Difference from Multiple-Choice Test Score
1	30	+2	0
2	29	+5	+5
3	29	+1	-3
4	29	+5	+1
5	29	+9	+13
6	29	+17	+5

7	28	+2	+2
8	27	+11	-1
9	26	+4	+4
10	26	+12	+4

Finally, when the differences between multiple-choice test scores and average grades of Group A and Group B are analysed (Table 5), significant differences are noted. All average grades of Group A are lower than multiple-choice test scores whereas, except for Student 3, all average grades of Group B are higher than multiple-choice test scores. Great differences are observed: negative in Students 4, 5, 7, and 9 in Group A; and positive in Students 5, 6, and 10 in Group B.

Table 5. Differences between Multiple-Choice Test Scores and Average Grades of Group A and Group B

Group A Student	Highest Multiple-Choice Test Score	Average Grade Difference from Multiple-Choice Test Score	Group B Student	Lowest Multiple-Choice Test Score	Average Grade Difference from Multiple-Choice Test Score
1	49	-1	1	30	+1
2	48	-1	2	29	+3
3	48	-3	3	29	-1
4	47	-7	4	29	+2
5	47	-9	5	29	+7
6	46	-2	6	29	+7
7	46	-7	7	28	+1
8	46	-3	8	27	+3
9	46	-9	9	26	+3
10	46	-3	10	26	+5

DISCUSSION AND CONCLUSIONS

The results of the present study suggest that assessing ELT students through such standard tests as multiple-choice may have misleading results. The mean of the multiple-choice test was the highest, and many multiple-choice test-wise students might

have made of use of the advantages of multiple-choice test and have done better than many other students. However, this could be a weakness because standard tests may not be able to "accurately and fairly measure student understanding of course concepts" (Simkin & Kuechler, 2005, p. 74). For instance, the highest multiple-choice scores of some students (e.g. Students 4, 5, 7, and 9 in Group A) might be due to guessing (Henning et al., 1981) and ability to recognize or recall (Norris et al., 1998, p. 15; Braun and Mislevy, 2005, p. 495). The means of presentation grades and translation grades are lower than that of the multiple-choice test because they are more difficult as they require the student to "organize, synthesize, or argue coherently, to express knowledge in personal terms, or to demonstrate creativity" (Simkin & Kuechler, 2005, p. 76).

Although both test and translation contents were designed to test almost the same coverage of language, that is, linguistic competence elements of lexicon, formulaic language, collocation and grammar, translation task proved more difficult than the multiple-choice test, bringing about striking differences in some students' grades. It is evident that answering a multiple-choice test and translating an original text into the foreign language are quite different tasks since the latter involves a multitude of factors. A multiple-choice test assesses Bloom's knowledge level, i.e. "simple recall of facts", and comprehension level, i.e. "the ability to follow a set of problem-solving steps on test material that is similar to what students have seen in class or in textbooks" (Simkin & Kuechler, 2005, p. 82), whereas translation takes the learner to a step further, i.e. to the application level, i.e. "the ability to transfer the knowledge to new, but structurally similar, domains" (p. 83).

More specifically, translation is a reconstruction task (Bruton, 1999), which House (2006, p.243) describes as "an act of performance, of language use," and "a process of recontextualization". She defines translation as "the replacement of a text in a source language by a semantically and pragmatically equivalent text in a target language. An adequate translation is thus a pragmatically and semantically equivalent one" (House, p. 345). On the other hand, although Underhill (1992, p. 47) views making presentation "an authentic and communicative activity both for professional and academic

purposes", the oral presentation in this study is "pseudocommunicative task" (Upshur & Turner, 1999, p. 104), "informational talk" (Weir, 2005, p. 105), "monologic informational routine" (Weir, p. 160), "ready-made or pre-packaged … artificial …, not … authentic" (Norris et al., 1998, p. 50, 61), "information related talk" (Louma, 2004, p. 187), and transmission of "already organized material" (Louma, p. 187).

When the three tools are analysed in terms of the task difficulty variables, the test is the least difficult, translation is the most difficult and the presentation is in-between (Norris et al., 1998). If the assessment had been made only by means of the multiple-choice test, some test-wise students would have been favoured and some others would have been misassessed. It is likely that fairness and validity were almost realised by taking the average of the scores obtained through three tools, tapping different skills and knowledge areas of the students.

Now that assessment is an indispensable part of teaching/learning process, either high-stakes or low-stakes, multiple measures have to be tapped so that what the learner actually knows and what he can/cannot do could be assessed. As Brown and Hudson (1998) suggests, "virtually all of the various test types are useful for some purpose, somewhere, sometime. In other words, all of the different types of tests are important to keep, because all of them have distinct strengths and weaknesses" (p. 657). One might minimise subjective assessment, and thus increase validity and reliability by taking steps like the following (CEFR, 2001):

- Developing a specification for the content of the assessment, for example based upon a framework of reference common to the context involved.
- Using pooled judgements to select content and/or to rate performances.
- Adopting standard procedures governing how the assessments should be carried out.
- Providing definitive marking keys for indirect tests and basing judgements in direct tests on specific defined criteria.

- Requiring multiple judgements and/or weighting of different factors.
- Undertaking appropriate training in relation to assessment guidelines (p. 188).

To reach a balance between standard assessment and alternative assessment, "it is essential that teachers complete their assessments while they are instructing", instead of "'stop teaching' in order to assess their students" (Gottlieb, 2006, p. 7). The problems of the alternative assessment can be avoided by systematic rating procedures for learners' performances, by providing raters with example work samples, and with clear task descriptions and directions, and by using various tasks (Norris et al., 1998). As for very specific alternative ways, Tomlinson (2005) suggests the following:

- Presenting the learners with new knowledge during the test and then asking them to apply it (e.g. teaching elementary learners about the interrogative during a test and then asking them to design a questionnaire).
- Teaching the learners a new strategy during the test, and then asking them to apply it (e.g. teaching the learners ways of scanning a text for specific information during a test, and then giving them a short time to find information from a text).
- Teaching new language whilst testing something different (e.g. giving a comprehension test on a text teaching a feature of the language).
- Teaching new skills whilst testing something different (e.g. teaching half the class about the skill of visualization when reading or listening and testing them on their ability to give advice on visualization to the other half of the class in groups of four).
- Testing the learners' ability to use a skill in a novel context (e.g. testing learners who have given short oral presentations on their hobbies or on their ability to give presentations to a group of potential customers) (pp. 42-44).

Finally, although the findings from this study contribute to existing literature on FL teaching at the college level, the results in this study were drawn from a non-random

sample with a limited number of students. Therefore, the results should be interpreted and generalised cautiously.

REFERENCES

Balliro, L. (1993). What kind of alternative? Examining alternative assessment. *TESOL Quarterly*, *27* (3), 558-561.

Barootchi, N., & Keshavarz, M. H. (2002). Assessment of achievement through portfolios and teacher-made tests. *Educational Research*, *44* (3), 279–288.

Braun, H. I., & Mislevy, R. (2005). Intuitive test theory. *Phi Delta Kappan*, *86* (7), 489-497.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, *32*(4), 653-675.

Brown, J. D., & Hudson, T. (1999). The authors respond. *TESOL Quarterly*, *33* (4), 734-735.

Bruton, A. (1999). Comments on James D. Brown and Thom Hudson's "the alternatives in language assessment" a reader reacts. *TESOL Quarterly*, *33* (4), 729-734.

Cheng, L., Rogers, T., & Hu, H. (2004). ESL/EFL instructors' classroom assessment practices: Purposes, methods, and procedures. *Language Testing*, *21* (3), 360–389.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Edelenbos, P., & Kubanek-German, A. (2004). Teacher assessment: The concept of 'diagnostic competence'. *Language Testing*, *21* (3), 259–283.

Gottlieb, M. (2006). *Assessing English language learners: Bridges from language proficiency to academic achievement.* Thousand Oaks: Corwin Press.

Henning, G. H., Ghawaby, S. M., Saadalla, W. Z., El-Rifai, M. A., Hannallah, R. K., & Mattar, M. S.(1981). Comprehensive assessment of language proficiency and achievement among learners of English as a foreign language. *TESOL Quarterly*, *15* (4), 457-466.

House, J. (2006). Text and context in translation. *Journal of Pragmatics*, *38*, 338-358.

Hughes, A. (1990). *Testing for language teachers*. Cambridge: Cambridge University Press.

Leung, C., & Lewkowicz, J. (2006). Expanding horizons and unresolved conundrums: Language testing and assessment. *TESOL Quarterly*, *40* (1), 211-234.

Louma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

Lynch, B. K. (2003). *Language assessment and programme evaluation*. Edinburgh: Edinburgh University Press.

McNamara, T. (2001). Editorial: Rethinking alternative assessment. *Language Testing, 18* (4), 329–332.

Norris, J. M. (2006). The why (and how) of assessing student learning outcomes in college foreign language programs. *The Modern Language Journal, 90* (4), 574–601.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments.* USA: Second Language Teaching & Curriculum Center. University Of Hawai'i.

Rowley, G. L. (1974). Which examinees are most favoured by the use of multiple choice tests? *Journal of Educational Measurement, 11* (1), 15-23.

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education, 3* (1), 73-97.

Sullivan, J. H. (2006). The importance of program evaluation in collegiate foreign language programs. *The Modern Language Journal, 90* (4), 574– 601.

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

Underhill, N. (1992). *Testing spoken language: a handbook of oral testing techniques.* Cambridge: Cambridge University Press.

Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing, 16* (1), 82-111.

Varela, E. (1997). Review: Authentic assessment for English language learners: Practical approaches for teachers. *TESOL Quarterly, 31* (1), 188-189.

Weir, C. J. (2005). *Language testing and validation: an evidence-based approach*. Houndmills: Palgrave Macmillan.

GENİŞ ÖZET

İkinci/yabancı dil olarak öğretildiği ortamlarda İngilizce'nin sınıf-temelli değerlendirilmesi, alanla ilgili çalışmalarda son zamanlarda yer almaya başlamıştır. Yabancı dil öğretiminde değerlendirmenin önemine rağmen, üniversite düzeyinde yabancı dil öğretiminde farklı değerlendirme yöntemleri konusundaki araştırmalar sınırlı sayıda olmuştur. Hem standart değerlendirmenin hem de alternatif değerlendirmenin zayıf ve güçlü yanları vardır. Bu çalışmada, İngilizce Öğretmenliği Programında yürütülen seçmeli bir dersteki değerlendirme ele alınmakta ve

öğrencilerin bir standart değerlendirme aracı ve iki alternatif değerlendirme aracı yoluyla elde ettikleri puanlar incelenmektedir. Söz konusu araçlar şunlardır: çoktan seçmeli sınav, sözlü sunum ve çeviri. Özel olarak, çoktan seçmeli sınavdan en yüksek notu alan on öğrenciyle, en düşük notu alan on öğrencinin iki alternatif değerlendirme aracındaki başarı düzeyleri karşılaştırılmaktadır. Öğrencilere bir çoktan seçmeli test, bir çeviri görevi ve de bir sözlü sunum görevi verilerek, öğrencilerin şu noktalarda düzeyleri değerlendirilmiştir: yabancı dilde bilgi tanıma, bilgi aktarımı ve bilgi uygulaması. Çoktan seçmeli sınavdan en yüksek notu alan on öğrenciyle, en düşük notu alan on öğrenci seçilmiş olup, bu öğrencilerin sunum ve çeviriden aldıkları puanlar çoktan seçmeli testten aldıkları puanlarla karşılaştırılmıştır. Son olarak, her öğrenci için (a) çoktan seçmeli test notu ile sözlü sunum notu arasındaki fark, (b) çoktan seçmeli test notu ile çeviri notu arasındaki fark ve (c) çoktan seçmeli test notu ile ortalama notu arasındaki fark bulunmuş ve tablada gösterilmiştir. Çoktan seçmeli sınavdan en yüksek notu *(45-50 aralığında) alan on öğrenciden yalnızca bir öğrenci, sunum değerlendirmesinde 45-50 aralığında bir not almış olup, yine bu gruptan dört öğrenci çeviri değerlendirmesinde 45-50 aralığında bir performans göstermişlerdir. Öte yandan, çoktan seçmeli sınavdan en düşük notu (26-30 aralığında) alan on öğrenciden bir öğrenci, sunum değerlendirmesinde 46 puan almış olup, yine bu gruptan bir başka öğrenci çeviri değerlendirmesinde 42 puan almıştır. Bunlara ilaveten, çoktan seçmeli sınavdan en düşük notu alan on öğrenciden üç öğrenci sözlü sunumdan 38 puan almışlardır. Görev zorluğu bakımından değerlendirildiğinde, çoktan seçmeli test en kolay olup, çeviri en zor görev durumdayken, sözlü sunum orta düzey bir zorluğa sahiptir. Bulgular, İngilizce öğrenenleri yalnızca bir araçla değerlendirmenin nesnel olmayabileceğini ve değerlendirme araçlarının çeşitlendirilmesinin faydalı olacağını ortaya koymaktadır. Eğer değerlendirme yalnızca çoktan seçmeli test yoluyla yapılmış olsaydı, test tekniğine yatkın olan öğrenciler bundan daha fazla yararlanmış olacaktı ve test tekniğine yatkın olmayan öğrenciler yanlış değerlendirilmiş olacaktı. Muhtemeldir ki üç araçtan elde edilen puanların ortalaması alınarak adalet ve geçerlilik büyük oranda gerçekleştirilmiş olup, öğrencilerin farklı becerileri ve bilgi alanları değerlendirmede dikkate alınmıştır. Değerlendirme, öğretim/öğrenim süreçler*inin

vazgeçilmez bir parçasıdır ve öğrencinin ne bildiğinin/bilmediğinin yanında neyi yapabildiğinin/yapamadığının da değerlendirilmesi için birden fazla değerlendirme aracına başvurulmalıdır.