SOSYAL BİLİMLER Yıl : 2010 Cilt :8 Sayı :2 Celal Bayar Üniversitesi S.B.E.

Yabancı Dil Olarak İngilizce Öğretiminde Avrupa Birliği Ortak Dil Çerçevesine Göre B2 Düzeyindeki Kompozisyonları Değerlendirmek Üzere Analitik Bir Ölçme Aracının Geliştirilmesi

Sarp Erkır

Doç. Dr. Mehmet Ali Yavuz

Dokuz Eylül Üniversitesi, Eğitim Fakültesi, Yabancı Diller Bölümü

ÖZET

Araştırmacılar dil sınavlarının geçerliliği ve güvenilirliği konusuna her geçen gün daha fazla önem vermektedirler. Bu makale, ölçme ve değerlendirmede güvenilir bir ölçme aracı kullanmak yoluyla geçerliliğin arttırıldığı bir çalışmayı konu alır. Yeni geliştirilen bir ölçme aracı ile elde edilen yüksek düzeydeki güvenilirlik bütünsel bir ölçme aracı yerine analitik bir ölçme aracı seçilmesinden kaynakladığı düşünülmektedir. Ölçme işlemini gerçekleştiren okuyucuların da eğitime tabi tutulmasının, okuyucuların ölçme aracını doğru kullanmalarını sağladığı görülmüştür.

Anahtar Sözcükler*: Ölçme, değerlendirme, yazma eğitimi, dil seviyeleri, yabancı dil olarak İngilizce, analitik ölçüm, bütünsel ölçüm, geçerlilik, güvenilirlik*

The Development of an Analytic Scoring Rubric to be Used for Marking B2 Level ESL Compositions

ABSTRACT

Researchers are increasingly interested in validity and reliability of language exams. This article describes a study in which scoring validity is increased through the use of a reliable scoring rubric. High reliability scores achieved through the use of the newly developed rubric can be attributed to the choice of an analytic format over a holistic format. Rater training was also found to have a positive effect on increasing raters' familiarity with the scoring rubric.

Key Words: *Meausrement, evaluation, teaching writing, language proficiency levels, teaching English as a foreign language, analytical scoring, holistic scoring, validity, reliability*

I. INTRODUCTION: MARKING WRITING COMPOSITIONS OF EFL LEARNERS

One of the obvious shortcomings of any syllabus on earth is the fact that one can never cover everything that has to be taught. Just like the impossibility of teaching everything that exists in the world, exams can only cover a small part of what has been taught (learned). This commonly cited limitation of testing has a lot to do with time. A teacher has very limited time to teach and even less time to test what has been taught. In order to make the most of this limited resource, test developers try to strike a balance between receptive and productive skills of language learners so that they can elicit a more comprehensive sample representing the learners' linguistic development. By the same token, however, learners' productions– be it written or oral –cannot properly represent a learner's language capacity due to the fact that "it is unnatural for a learner to write a draft ... and submit it for a grade" (Cohen, 2001:534). This unnatural feel that is associated with marking writing brings to mind some construct validity considerations. Construct validity comprises context, cognitive and scoring validity and according to Shaw and Weir (2007:7), "there is a symbiotic relationship that exists between context, cognitive and scoring validity". Decisions regarding one of them will inevitably have an impact on the others and this paper deals with the decisions that are predominantly related with scoring validity of written papers.

Over the years, numerous studies have pointed out that written productions are scored differently by different graders. The poor reliability among graders is generally the result of a failure to pay adequate attention to scoring rubrics. As argued by McMillan (2001), when markers are not guided by clearly defined outcomes and scoring rubrics that reflect these outcomes, evaluations tend to be less stable. This study aims at describing the steps to be followed in order to develop a scoring rubric that reflects these outcomes. The study focuses on steps to be followed when developing an analytic scoring rubric. The stages that have been followed to design five analytic rubrics to be used to mark the written productions of preparatory school students with different levels of English proficiency ranging from beginner to upper-intermediate are mentioned with the hope that this overview may be helpful to other testing boards in understanding the activities involved in the creating of marking rubrics. Before that, however, a decision has to be made regarding the rubric type.

A. The Choice between Two Scoring Rubrics

A rubric is a device which clearly outlines the criteria for an assignment and articulates various levels of what an instructor is looking for on each graded dimension from weak to excellent (Goodrich, 1997). Recently, grading rubrics have gained increasing popularity as assessment tools promising to remove reliability concerns. There are two commonly-used rubric types; Holistic Scoring Rubrics and Analytic Scoring Rubrics.

B. Holistic Scoring Rubrics

As can be inferred from the name, holistic scoring rubrics "yield a single overall score taking into account the entire response" (McMillan, 2001:252). This type of scoring rubrics use scores that are described by statements about the features of response and "calls for the reader to rate overall writing proficiency on a single rating scale" (Stiggins & Bridgeford, 1983:26).

Holistic scoring has the highest construct validity when overall attained writing proficiency is the construct to be assessed (Perkins, 1983). In spite of the fact that holistic scoring has the highest construct validity, it also has many drawbacks: In scoring holistically, the grader reads the composition, forms a general impression, and assigns a mark to that composition based on some standard. Such an evaluation can, therefore, "be highly subjective due to bias, fatigue, internal lack of consistency, previous knowledge of the student, and/or shifting standards from one paper to the next" (Perkins, 1983:653). Holistic scoring is also criticised for not having theoretical underpinnings. Weigle (as cited in Shaw and Weir, 2007) claims that holistic scoring is problematic for second-language writers because different aspects of writing ability develop at different rates for different writers.

C. Analytic Scoring Rubrics

Analytic scoring rubrics "enable a teacher to focus on one characteristic of a response at a time" (McMillan, 2001:249). Analytical scoring breaks performance down into component parts (e.g., organization, wording, ideas) for rating on multiple scales (Stiggins & Bridgeford, 1983:26). In addition to showing students how their particular grades have been determined, analytic scoring has a further advantage:

> because an analytical scale focuses graders on scoring, the procedure is supposed to ensure sufficient agreement among graders to permit a reliable score to be derived from summed multiple ratings. (Perkins, 1983: 656)

As mentioned above, scoring holistically means that the grader reads the composition, forms a general impression, and assigns a mark to that composition based on some standard. Such an evaluation can be highly subjective and it lacks theoretical underpinnings. In addition, holistic scoring is problematic for second-language writers because different aspects of writing ability develop at different rates for different writers. For these reasons, analytic scoring has been chosen over holistic scoring.

To what extend an agreement among raters is ensured is at the focal point of this paper. Before that, however, the steps that were followed in the development of the analytic rubric have to be described.

II. THE STUDY

The study focuses on steps to be followed when developing an analytic scoring rubric. Due to the fact that analytic scoring rubrics, when compared with holistic scoring rubrics, provide a more useful feedback regarding student performance, and more importantly, that they are supposed to yield more reliable results, analytic marking rubrics are preferred in this study. The formidable procedure of developing an analytic scale, according to Bachman (2003), is composed of at least six steps. The following is a list of these steps and how these steps are transferred into life:

A. A large corpus of original student writing must be amassed; the features that make up the scale are derived from this corpus.

In line with Bachman's (2003) suggestion, a large number of student writings were gathered and it was noticed students have a tendency to use some memorised phrases in their introductions, transitions and conclusions. This is believed to be caused by previously-used rubrics that encouraged students to use some signalling phrases. From a Lexical-Approach perspective, this might appear to be a desirable situation and one might argue that, in teaching composition writing, such phrases are time-saving. However, in practice, these phrases are by no means samples of authentic language use. On the contrary, they are the indicators of memorised language production that lack meaning. In other words, the analysis of the previously amassed writing exams and student papers revealed that the writing prompts and the former scoring rubrics had a negative wash-back effect on the learners' performance. A closer look at the previous rubrics showed that the descriptors were worded in such a way that students were expected to use some certain phrases and chunks for certain essay types. For example, almost seventy percent of the exam takers started a 'compare & contrast' type essay with the same sentence: 'X and Y can be compared and contrasted of the bases of A, B and C. The first bases of contrast is A.' Such formulaic approaches to writing led teachers to get their students to memorise these chunks, and students who did so received higher graders for copying those phrases without any communicative attempt being exerted to convey a real and authentic message. On the other hand, those students who took risks trying to convey a real message instead of writing memorised phrases one after the other were penalised for their incorrect usages, which are in fact inevitable components of the process of language acquisition.

After it was noted that those papers were unsuitable for the purposes of this study, a decision was taken to analyse other rubrics that are currently used by international examination centres such as the Cambridge or the IELTS marking schemes for writing. In addition, features that make up the rubric were also derived from the Common European Framework of Reference. These were used to draw a list of pass-level descriptors from A2 to C2 levels. According to the CEF, the number of possible categories should be reduced to a feasible number – no more than five. The four categories that are used in this rubric are displayed in Figure 1:

	BANDS					
CONTENT						
COHERENCE AND COHESION						
LEXIS						
GRAMMATICAL STRUCTURES						

Figure 1. The four categories used in the analytic scoring rubric

Features were then reduced into a smaller set of assessment criteria appropriate to the requirements of the assessment task concerned. As argued in the CEF, the resultant criteria might be equally weighted, or alternatively certain factors considered more crucial to the task at hand might be more heavily weighted. As shown in Figure 2, due to concerns relating the practicality of standardisation and with a view of user-friendliness, a decision was taken to equally weight all the strands.

	BANDS					
CONTENT **Task achievement** **Length** **Effect on reader**						
COHERENCE AND COHESION **Organisation** **Fluency** **Linking devices**						
LEXIS **Appropriacy** **Range** **Accuracy**						
GRAMMATICAL STRUCTURES **Range** **Accuracy**						

Figure 2. The sub-categories of the four strands used in the analytic scoring rubric

B. The scale must be field-tested, after which the features may be modified.

Field testing, or piloting, is of paramount importance because of the difficulties that are associated with creating a continuous set of scales – one in which the fifth band of a lower level serves as the third band of a higher level. On top of this difficult task is another problem which stems from the range of intended uses of the scale. That is to say, the set of scales is intended to be used not only for different levels, but also for different tasks within the same level. The categories, the subcategories and the descriptors of those strands have to be worded in such a way that they exclude salient features imposed by different task types. The language and format one would expect to see in an e-mail task is far different from those one would expect in an essay. For this reason, consistent features of writing at different levels for different tasks have to be pinpointed by the descriptors.

During the field testing stage, particular attention was paid to whether the descriptors included elements that are observable. Being observable calls for maximum objectivity, and in order to avoid the influence the previous impressions of students might have on markers, it has been decided not to have boxes where students write their names. Alternatively, learners can be asked to write their student numbers. Almost a hundred students sat the pilot exam. In line with the feedback given by proctors and markers, some changes were made to the rubrics. While making these changes, expert advice was also sought.

C. Descriptors must be written to describe the high, mid, and low quality levels for each feature.

One way of drafting descriptors is to gather a corpus of candidate writing performances, and to analyse their distinguishing characteristics. Instead, the characteristics and categories that have been identified were rationalised into band descriptions for the five proficiency levels, starting with A to E.

According to the CEF, while norm-referencing is the placement of learners in rank order, their assessment and ranking in relation to their peers, in criterion-referencing, the learner is assessed purely in terms of his/her ability in the subject, irrespective of the ability of his/her peers. So as to be able to make this judgement, a mapping of the continuum of proficiency must be vertically designed. This involves the identification of the scores on a test deemed necessary to meet the proficiency standard set. The score which is deemed necessary to meet the Writing Proficiency Standard is 60 out a 100, and the descriptors given in the 3rd band below include expected behaviours from a student who deserves a pass grade of 3 out of 5 for each strand or 12 out of 20 in total, which is tantamount to 60 out of a 100.

	BANDS					
CONTENT **Task achievement** **Length** **Effect on reader**	No rateable language. Totally incomprehensible. Totally irrelevant.	Less than 50% of the given or 'original' content elements dealt with; and /or few content elements dealt with successfully: message not generally communicated and fully; message requires excessive effort by the reader. Less than 50% of specified length	Features of bands 1 and 3. Serious effort by reader is required.	All content elements addressed successfully and with appropriate expansion: message clearly and fully communicated to the reader. Length meets specifications. (+ 10% allowed)	Features of bands 3 and 5. Little effort by reader is required.	All content elements addressed successfully: message clearly and fully communicated to the reader using appropriate register in a natural way. Length meets specifications. (+ 10% allowed)
COHERENCE AND COHESION **Organisation** **Fluency** **Linking devices**		Response is at times incoherent . Insufficient control of organisational features beyond sentence level for task to be carried out successfully.		Good control of connected sentences and use of linking devices. Complex sentence forms attempted and achieved to a certain degree. Information is organised logically.		Good control of connected sentences and use of linking devices. Complex sentence forms attempted and generally achieved. The flow of ideas aids fluent reading.
LEXIS **Appropriacy** **Range** **Accuracy**		Insufficient range to carry out task successfully. Lack of control of spelling and sentence formation.		Appropriate and adequate range for the task. Only minor errors, which do not reduce communication and may be 'slips' or due to risk taking and ambition.		Appropriate and adequate range for the task. Only minor errors, which do not reduce communication and may be 'slips' or due to risk taking and ambition. Wide range of vocabulary. Use of formal or informal vocabulary appropriately where required.
GRAMMATICAL STRUCTURES **Range** **Accuracy**		Insufficient range of structures to carry out task successfully. Lack of general control of structures and /or punctuation		Sufficient range of structures for the task. Only minor errors, which do not reduce communication and may be 'slips' or due to risk taking and ambition		A broad range of grammar allowing the student to use expressions in a clear and appropriate style without having to restrict what they want to say. Consistently maintains a high degree of grammatical accuracy.

Figure 3. The four categories described in the analytic scoring rubric

A work to compare the newly emerging scale band descriptors with performance levels used in international exam scales was also carried out. These included mainly Cambridge examination rubrics. This helped increase criterion related validity of the process.

D. The points must be "anchored" on a scoring line.

The scoring rubrics described here consist of four analytic dimensions. These are (1) content, (2) coherence and cohesion, (3) lexis and (4) grammatical structures. Descriptions for score points of 1, 3 and 5 for each of the four dimensions are given. Intermediate scores of 2 and 4 are to be inferred as having qualities in between those specified for the points above and below. While writing the quality levels for each of the four features, it has been agreed to number high, mid and low performances as 5, 3 and 1 respectively. Therefore, the analytic marking rubric developed here assigns 20 points as the highest total that can be accomplished by learners. 0 is the mark to be given when there is not sufficient language sample to make judgements, or when the language production is totally irrelevant or when the language is unintelligible. It is worth explaining the rationale behind this irrelevancy issue. There are times when learners present the marker with a good piece of writing which is off topic or off task. Much as you might like the writing, it is never possible to tell whether the writing is the exam-takers' production, or it is a copy of a memorised text that was written in the classroom or even at home. Therefore, learners are strictly expected to write about the prompts they are given in the exam. Figure 4 illustrates how points are anchored on scoring lines in line with these considerations. A timely reminder would be to say that the rubric given in Figure 4 is designed to mark writings by D Level/Intermediate students.

	0	1	2	3	4	5
CONTENT ▪ **Task achievement** ▪ **Length** ▪ **Effect on reader**	No rateable language. Totally incomprehensible. Totally irrelevant.	Less than 50% of the given or 'original' content elements dealt with; and /or few content elements dealt with successfully: message not generally communicated and fully; message requires excessive effort by the reader. Less than 50% of specified length	Features of bands 1 and 3. Serious effort by reader is required.	All content elements addressed successfully and with appropriate expansion: message clearly and fully communicated to the reader. Length meets specifications. (+ 10% allowed)	Features of bands 3 and 5. Little effort by reader is required.	All content elements addressed successfully: message clearly and fully communicated to the reader using appropriate register in a natural way. Length meets specifications. (+ 10% allowed)
COHERENCE AND COHESION ▪ **Organisation** ▪ **Fluency** ▪ **Linking devices**		Response is at times incoherent . Insufficient control of organisational features beyond sentence level for task to be carried out successfully.		Good control of connected sentences and use of linking devices. Complex sentence forms attempted and achieved to a certain degree. Information is organised logically.		Good control of connected sentences and use of linking devices. Complex sentence forms attempted and generally achieved. The flow of ideas aids fluent reading.
LEXIS ▪ **Appropriacy** ▪ **Range** ▪ **Accuracy**		Insufficient range to carry out task successfully. Lack of control of spelling and sentence formation.		Appropriate and adequate range for the task. Only minor errors, which do not reduce communication and may be 'slips' or due to risk taking and ambition.		Appropriate and adequate range for the task. Only minor errors, which do not reduce communication and may be 'slips' or due to risk taking and ambition. Wide range of vocabulary. Use of formal or informal vocabulary appropriately where required.
GRAMMATICAL STRUCTURES ▪ **Range** ▪ **Accuracy**		Insufficient range of structures to carry out task successfully. Lack of general control of structures and /or punctuation		Sufficient range of structures for the task. Only minor errors, which do not reduce communication and may be 'slips' or due to risk taking and ambition		A broad range of grammar allowing the student to use expressions in a clear and appropriate style without having to restrict what they want to say. Consistently maintains a high degree of grammatical accuracy.

Figure 4. The analytic scoring rubric with points anchored

Using a marking rubric means that one has actually chosen criterion referencing (CR) over norm referencing and you have two choices ahead of you: Mastery CR and continuum CR. The former is an approach in which “a single ‘minimum competence standard’ or ‘cut-off point’ is set to divide learners into ‘masters’ and ‘non-masters’, with no degrees of quality in the achievement of the objective being recognised” (Common European Framework of Reference). In the latter case, individual ability is referenced to a defined continuum of all relevant degrees of ability in the area in question. As a continuum approach has been endorsed, levels according to which learners are grouped are matched to the Common Reference Levels. Accordingly, the rubrics that are developed to mark learner papers in different levels also reflect this match. With the aim of achieving a continuum of rubric descriptors, a bottom-up process has been followed. In other words, first the A/Beginner level analytic marking rubric is developed, then the 5th band of this rubric is copied to the 3rd band of the B/Elementary level and the same process is repeated until the E/Upper Intermediate level rubric is completed. For time and space concerns, the statistical results of the analysis carried out for D level analytic marking rubric is displayed.

E. The graders practice using the scale on a fresh set of compositions.

This stage requires that the raters are trained in the use of an analytic rubric. The key word in the 5th stage of the development of an analytic marking rubric is practice. However, training is required before practice. For this reason, a group of 40 markers/teachers were chosen. These teachers were exposed to a training session on how to use an analytic rubric. It was also during this training that markers were familiarised with the descriptors.

The training on the use of the rubric should not be limited to markers. Training was also carried out in the classroom. Students were expected to be familiar with the rubric. This made it possible for us to create a transparent exam by helping students understand what is expected from them in writing exams.

The issue of objectives is always a mystery to the students. Although students can understand the task types they have to master during the course of a lesson, they can barely understand the extent to which these objectives should be mastered. In this regard, an analytic marking rubric can prove to be of great use. It can serve not only as a tool for giving post-writing feedback; it can also be used as a pre-writing feedback tool, providing learners with detailed justifications of the prospective marks they might receive. Even the best rubrics are just not entirely self-explanatory to students. When an agreement between what the student sees and what the professor says is lacking, students will not perceive that they have been graded fairly.

The new rubrics were used operationally for the first time in October, 2010. At this stage, feedback from examiners was sought. Although general feedback from examiners was very positive, they complained that the new criteria did not help with the problem of marking potentially memorised scripts. Feedback from test developers was also consistent with the remarks made by the raters. They also argued that the new scale was not effective in dealing with the problem of candidates supplying off topic responses. For this reason, the sixth step, which is described under the title 'Results', was even more important.

III. RESULTS

When we want to know the extent to which two variables are related to each other, *Pearson's r* helps us to tell if these two are related. It is a statistical method describing the strength of a relationship between two variables. Put differently, when one variable goes up, does the other go down? As part of this study, the aim is to see whether or not the analytic scoring rubric allows a similar grading among two separate raters.

Table 1

Correlations Showing Inter-Rater Reliability Scores for Each Strand/Category

	Content	Coherence	Lexis	Grammar
Pearson correlation	66**	68**	62**	63**

*Note.*** *p*<.01.

If r is close to 1, there is a strong relationship between the two variables. There is a weak relationship between two variables when r is close to 0. As shown in Table 1, a Pearson product-moment correlation coefficient was computed to assess the relationship between grades given by 3 different pairs of raters using the newly developed analytic scoring rubric. There is a statistically significant correlation between the pairs (r = .80, n = 212, p = 0). This means that there is a strong relationship between the grades given by three pairs of raters. That is to say, changes in one rater's scores are strongly correlated with changes in the second rater's scores. Pearson's r is 0.80 and this number is very close to 1. As a result, we can conclude that there is a strong relationship between the three pairs of raters. When Pearson's r is positive, one variable's increase in value is followed by the second variable's increase in value. Similarly, as one variable decreases in value, the second variable also decreases in value. In the study, the *Pearson's r* value of 0.80 is positive. We can conclude that there is positive correlation between 3 pairs of raters.

Table 2

Correlations Showing Inter-Rater Reliability Scores of Each Pair

	1st pair	2nd pair	3rd pair
Pearson correlation	59**	85**	84**

*Note.***p*<.01.

As for the Sig (2-Tailed) value, it is less than .05, which means that there is a statistically significant correlation between the two variables. To be more specific, the value is .000, meaning increases or decreases in one rater's grades significantly relate to increases or decreases in the grades by the second rater.

IV. DISCUSSION AND POSSIBLE FUTURE RESEARCH

It is no secret that analytical scales have some serious disadvantages. The most cited disadvantage of analytic scoring has to do with our understanding of text and discourse. Discourse analysis tells us that a written text is more than the sum of its parts. Analysing bits and pieces and then reaching an overall evaluation may sometimes seem problematic. In other words, when using an analytic scoring rubric, the features to be analyzed are isolated from context and are scored separately. So as to overcome this pitfall, a subcategory titled "effect on reader" is included to cater for the absence of a focus on text as a whole. The test-taker can score high on this category provided that his or her text as a whole can achieve to leave the desired impression. A second related problem stems from a limited column space to provide a description of a pass grade or a fail grade that is detailed enough. Rater trainings that are mentioned in the fifth step above ~~is~~ are the right place to compensate for this shortcoming. Here, detailed information can be provided for raters regarding descriptors. There is another problem here about the wording of the descriptors. Because of the limited space in the columns, rubrics designers tend to resort to using meta-language. Consequently, a description can be difficult for a student to understand when meta-language is used in the descriptors. One way of solving this problem is to translate the rubric into the learners' native language. It is a known fact that the use of L1 contributes highly to the retention and recall of the learnt material (Simsek, 2009: 167). A rubric is not a reading task per-se, where teachers assess learners' comprehension and there is no reason to argue that L1 should not be used in learner training. This would also add to the transparency of the examination system that is implemented in an institution. Lastly, by nature, description of a particular graded category needs to be somewhat generic so that the assessment falls neatly into a particular category. If the performance by a student can fall into multiple categories, the rubric breaks down (North, 1991). The only way of overcoming this problem has to do with objectivity. Observations made by using the rubric should be objective, and this objectivity can be achieved by determining in advance the behaviours that are expected to be displayed by learners. Descriptors in a rubric should then be worded in a way that they lend themselves to objective observation without requiring extensive effort on the part of the rater.

As it is, this study is of descriptive nature, and further research is required to confirm its findings. For instance, it was mentioned earlier in the article that the holistic scoring rubric has a major drawback in that it fails to guide raters towards a reliable and justifiable grade. If it were possible to form two groups of raters with similar training and experience in marking writing compositions, then a comparative analysis could be carried out to compare the marks given to the same paper by the two groups of raters. This would enable researchers to make more informed decisions regarding their choice of rubric.

REFERENCES

BACHMAN, L. F. (2003). Fundamental considerations in language testing. Oxford: OUP.

COHEN, A. D. (2001). “Second language assessment”. In M. C. Murcia (Ed.), Teaching English as a second or foreign language (pp. 515-534). Boston, MA: Thomson Learning.

Common European framework of reference for languages: Learning, teaching, assessment. Cambridge University Press.

Goodrich, H. (1997). “Understanding Rubrics”. Educational Leadership, 54, 14-17.

MCMILLAN, J. H. (2001). Classroom assessment: Principles and practice for effective standards-based instruction. Boston: Allyn & Bacon.

NORTH, B. (1991). “Standardisation of continuous assessment grades”. In J. C. Alderson & B. North (Eds.), Language testing in the 1990’s: The communicative legacy (pp. 167-177). London: MacMillan.

PERKINS, K. (1983). “On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability”. TESOL Quarterly, 17, 651-671.

SHAW, S. D. & Weir, C. J. (2007). Research and practice in assessing second language writing. Cambridge: CUP.

SIMSEK, M. (2009). “Dilbilgisi kavramlarının öğretiminde ana dil kullanımının öğrenci başarısına etkileri” (Doktora Tezi, Dokuz Eylül Üniversitesi, 2009). Ulusal Tez Merkezi (No. 239327)

STIGGINS, R. J. & Bridgeford. N. J. (1983). “An analysis of published tests of writing proficiency”. Educational Measurement: Issues and Practices, 2, 6-19.