Available online at www.jlls.org

JOURNAL OF LANGUAGE AND LINGUISTIC STUDIES

ISSN: 1305-578X

Journal of Language and Linguistic Studies, 14(4), 342-351; 2018

Does indirect writing assessment have any relevance to direct writing assessment? Focus on validity and reliability

Faruk Kural[a*]

[a]*Hasan Kalyoncu University, Havalimanı Yolu, Şahinbey, Gaziantep, 27410, Turkey*

APA Citation:

Kural, F. (2018). Does indirect writing assessment have any relevance to direct writing assessment? Focus on validity and reliability. *Journal of Language and Linguistic Studies, 14*(4), 342-351.

Submission Date: 18/02/2018

Acceptance Date: 11/08/2018

Abstract

The present paper, which is a study based on midterm exam results of 53 University English prep-school students, examines correlation between a direct writing test, measured holistically by multiple-trait scoring, and two indirect writing tests used in a competence exam, one of which is a multiple-choice cloze test and the other a rewrite test measured on the sentence grammar scale. The purpose of the study is to compare and analyse the results obtained from the use of these distinct strategies in writing assessment and to see to what extent they are of value to account for the students' writing skills and in what areas they can offer validity and reliability in diagnostic terms as well as investigating whether any quantitative correlations exist between the choice of topics and the level of achievement in essay-writing, in which the choice of topic also implied the choice of a particular rhetoric to be used as part of the essay writing task.

Keywords: Indirect writing assessment; direct writing assessment; validity; reliability

1. Introduction

With the exclusion of portfolio writing assessment from its scope, the focus area of this study is concerned with two major practical approaches to writing assessment: direct writing assessment and indirect writing assessment. The former uses samples of student writing to judge writing proficiency, and the latter relies on students' own writing of various types (Grabe & Kaplan, 1988, p. 396). Although there are views suggesting that correlation between the two approaches reveals a consistent and relatively strong relationship at various educational levels, there are strong doubts about this relationship based particularly on certain areas of 'validity' and 'reliability'.

Although significantly distinct in form, both direct and indirect writing assessment methods are useful in the prep-school context. Each provides a slightly different kind of information regarding the student's ability to use or recognize standard written English. In direct assessment, the student is required to write in response to a given prompt; then, the results are evaluated according to pre-specified criteria. In indirect assessment, the student is asked to judge the appropriate use of language in a series of objective test items which often follow a multiple choice format. Each testing method requires the

* Corresponding author. Tel.: 0342-211-8080
E-mail address: farukkural@gmail.com

students to apply their previously acquired knowledge in language processing in one way or another. General knowledge and all the other pre-specified content awareness expected of the students do not really constitute much difference in essence. Each approach provides sufficient information that is useful to make a variety of decisions on the students' progress.

All the decisions pertaining to assessment and its management, which include diagnosing the students' strengths and weaknesses, determining the appropriate level of the prep-school program for the student, and predicting their future progress upon completion of the prep-school program, are always based on either of the same common measure. Direct measures of writing skills are valuable in this context as long as the results are analysed by using detailed and well prepared criteria and as long as the scoring criteria used to rate writing samples explicitly cover predefined essential skills. Likewise, indirect writing assessment methods also serve well if well and carefully prepared criterion based tests are used if these tests break the students' overall writing performance into component parts and if test items test those skills allowing for a detailed analysis of these skills. Both approaches can plan and use the scoring criteria or test items equally well, in a similar fashion, by basing them upon what has been taught and what outcomes are intended. In short, it is possible to use each of the method to test what is taught.

Since the validity and reliability criteria and issues relevant to these criteria make up some of the fundamental grounds for all account efforts that focus on correlation between the two methods, it is important to outline what these criteria refer to. Davidson & Lynch (2002) and Alderson, Clapham, and Wall (1995) provide a detailed explanation of these criteria used in the general context of assessment in education, which can be interpreted for writing assessments in the following terms:

'Validity' is a general term which is related to questions about what the test is actually assessing. It is to do whether or not the test provides the information for our intended questions or investigations, and whether or not it measures what we intend to measure. A test is not valid, for example, if it is intended to test a student's ability of using connectors but instead if it tests background knowledge or level of intelligence. What is valid is usually depends on the purpose of the researcher or the tester.

'Construct validity' refers to the overall construct or trait being measured. If a test is supposed to be testing the construct of writing, it has to consider items that are necessary to construct writing, such as connectors, topic sentences, development structures, etc. Construct validation also relates to the test method. If language teaching emphasizes a communicative approach, for example, a test that contains only out-of-context items which test only one linguistic point at a time (such as single-sentences and multiple-choice questions) is unlikely to be considered to have construct validity. However, some rewrite exercises that intends to test sentence grammar might be considered to have some construct validity in writing since they imply the correct and appropriate usage of a number of linguistic items at the same time in the construction of a sentence as the components of the writing production.

'Content validity' in testing refers to comparing test items with the test specifications to see whether the items are actually testing what they are supposed to be testing. In writing, the test specifications may include some relevant vocabulary, certain category of information relevant to the topic, etc. Choosing the appropriate words in multiple-choice cloze tests and synonym/antonym tests may give us some knowledge of what students have understood from the knowledge content area of the syllabus.

'Concurrent validity' is one kind of criterion-related validity. Students' test scores may, for example, be correlated with other measures of students' language ability or with scores on a similar test.

'Predictive validity' is another kind of criterion-related validity which helps to assess the future ability of students. For example, students with high proficiency exam scores could be expected to do better at graduate studies than those with low proficiency exam scores.

The ‘reliability’ of a test is an estimate of the consistency of its scores. For example, a reliable test is one where a student will get the same score if s/he takes the test, possibly with a different examiner or at a different time. Tests that contain multiple choice, true/false or short answer questions are much reliable than those with open-ended questions or essays.

1.1. Literature review

Three studies conducted on college freshmen students (Breland, Conlon & Rogosa, 1976; Breland&Gayner, 1979; Huntley, Schmeiser & Stiggins, 1973) and three studies on high school students at varying grade levels (Godshalk, Swineford & Coffman, 1966; Hogan&Mishler, 1980; Moss, Pamela, Cole, Nancy&Khampalikit,1981) suggest that the two methods assess at least some of the same performance factors, and at the same time each deals with some unique aspects of writing skills. On the other hand, there are claims indicating the problems that these arguments have in the areas of construct and content validities to which recent writing assessments give a lot more consideration and for which indirect writing assessment is unable to account. (Grabe&Kaplan 1988, p. 397).

The study conducted by Godshalk, Swineford & Coffman(1966) also proved that it was possible to attain predictive validity and reliability in the assessment of writing by using objective or indirect measures. They also stressed that validity and reliability could be slightly enhanced by the inclusion of a short essay writing assessed holistically. Thus, the consideration of the three tests used in this study together as a combined single unit of proficiency measures would be expected to provide much stronger reliability and validity.

On the other hand, Hout (1990) alleges that in current practice too much emphasis is put on the reliability of writing sample scoring leaving validity largely assumed. As holistic scoring practices have become widely accepted, much work has been done to demonstrate the reliability of such measures. He argues that there is more need now to direct some critical attention to the validity of holistic scoring practices. In parallel to this outlook, Ruth&Murphy (1989) stress the same need in a similar fashion claiming that little research has been done on what makes a good topic and the method and language with which it is introduced to the writer, the rhetorical aspects demanded by the task, and several other variables, all of which have major effects upon the quantity and quality of writing being produced. They consider these variables to be a "neglected variable" in writing assessment and research. To this end, this study also intends to investigate whether any quantitative correlations exist between the choice of topics and the level of achievement in the essay test, in which the choice of topic also implied the choice of a particular rhetoric to be used as part of the essay writing task in the test.

1.2. Research questions

1. This study intends to answer the following research questions:
2. Does indirect writing assessment have any relevance for direct writing assessment?
3. Is there any quantitative correlation between the choice of topics and the level of achievement in the essay test?

2. Methodology

2.1. Subjects

3 midterm exams of a total of 53 subjects were used in the studywho were allA Level students attending the English preparatory program of the Northern Preparatory School of Yeditepe University.

The program is offered at 3 gradual levels (C,B,A with A level being the highest) where students are placed in the appropriate-levels according to their performance in a placement exam given after an initial proficiency exam, the success of which is an exemption criterion from the prep-school's study program. Students placed at the C level are required to complete the programs of all 3 levels successfully to be eligible to enrol an undergraduate program of the Faculty of Fine Arts. Therefore, the subjects of this study consists of those who either were placed at A level after a placement exam or have completed both or one of the lower programs successfully. It is possible to say that the linguistic and educational backgrounds of the subjects, as far as their skills in English are concerned, were very similar perhaps with some minor variations related to their language training experience during their secondary education. They were all English L2 learners with L1 in Turkish and they all have been through the same program for a minim period of six weeks prior to the midterm exams.

2.2. Procedure and treatment

The midterm exams, consisting of 5 individual exams (reading, writing, grammar, listening and ESP) each representing a measure in one area of language skills, were given to the candidates to diagnose their progress at the end of the study period of 6 weeks. The exams were prepared to measure the predetermined instructional points in the syllabi reflecting in-class teaching contents.

Although the subjects had already had sufficient training in composition writing and the topics concerning. However, to avoid washback effect, the subjects, testers and the assessors were not informed of any aspect of this study, which is based entirely on some of the exam result data representing the candidates' achievements in numerical-measure terms. The data collection was accomplished two weeks after the assessment by random selection of three classes at A Level, and it does not contain any prejudicial research value to hinder objectivity in this study.

The writing exam results were based on the achievement grade out of 100; the multiple-choice cloze test out of 30; and the rewrite test out of 15. For the purpose of maintaining consistency in the data treatment in this study, the latter test results were also converted to the achievement grade out of 100.

This study did not involve in any stages of the preparation, conduct and assessment processes of the tests. The exam papers were obtained after all these stages were fully completed. As indicated earlier, the present study is based on a direct writing test, which is a composition written on a topic, a multiple-choice cloze test, and a rewrite test.

2.2.1. The composition writing exam

The compositions were written in a formal assessment period of one hour on one of the following topics chosen by the students and they were evaluated by an assessment team who had had some training on the scoring procedures.The topic choice also included the rhetoric type choice to be used in the composition as in the following:

Topic 1: *Write a* ***process essay*** *about* **how to get ready for a graduation party**. *Please follow the plan in your book.*

Topic 2: *Write a* **process essay** *about* **how to decorate your room**. *Please follow the plan in your book.*

Topic 3: *Write an* ***opinion essay*** *about the following statement. Please follow the plan in your book.* **"Children should not eat in fast food restaurants."**

Topic 4: *Write an essay* ***classifying*** **inventions according to their importance for human beings**. *Please follow the plan in your book*.

The papers were assessed by two raters separately who wrote their scores on separate answer sheets. The success grade was calculated by taking the average of the two scores, and if there was a big difference between the two scores it was sorted out by the team leader. The raters used a multiple-trait scoring rubric developed by the raters as a diagnostic measure scale on various points in essay writing, which considered the following essay writing points to be assessed within the distribution of a 100-point grading system:

Essay writing point	**Value**	**Skill criteria included**	**Achievement scales**
CONTENT	30 pts.	adherence to topic main idea presentation supporting ideas and evidence	Excellent: 26-30 pts. Good: 16-25 pts. Fair: 11-15 pts. Poor: 1-10 pts.
VOCABULARY	15 pts.	choice of words and idioms appropriate words for the topic clarity of word usage	Excellent: 12-15 pts. Good: 8-11 pts. Fair: 4-7 pts. Poor: 1-3 pts.
ORGANISATION	20 pts.	fluent expression of ideas logical sequencing cohesion linking ideas	Excellent: 1-20 pts. Good: 11-15 pts. Fair: 6-10 pts. Poor: 1-5 pts.
LANGUAGE USE	30 pts.	use of complex constructions tense, word order or function use of pronouns, prepositions, articles	Excellent: 26-30 pts. Good: 16-25 pts. Fair: 11-15 pts. Poor: 1-10 pts.
MECHANICS	5 pts.	spelling, punctuation, capitalization paragraphing	Excellent: 5 pts. Good: 4 pts. Fair: 3 pts. Poor: 1-2 pts.

2.2.2. The cloze test

The cloze test was developed by the raters was used as part of the midterm grammar exam covering the instruction content of the grammar syllabus, and it contained 20 multiple-choice questions representing 30 points of the grammar exam with 1.5 points for each question to be answered by picking up the correct choice of the three given ones.

2.2.3. The rewrite test

The rewrite test prepared by the raters was also used as part of the midterm grammar exam. It consisted of 5 items of sentences to be rewritten from sentence construction prompts representing 15 points of the grammar exam with 3 points for each item.

2.3. Data collection and analysis

The data used in this study came from the midterm exam results of 53 English prep-school students and they included the results of a multiple-choice cloze test, a rewrite test, and a holistically scored composition written on a topic chosen from four given ones. The quantity of each topic chosen by the students for their composition writing assessment is also included in the data.

The data collected in this study were used in two ways of descriptive statistics, one comparing the means of the three tests, and the other comparing the percentages of the topics chosen for composition writing and success grades. (See Tables 1 & 3)

3. Results and discussions

Table 1 presents the mean scores of the cloze test, the rewrite test and the composition exam. The cloze test mean score is much higher than the rewrite mean score. However, this is not surprising as the cloze test requires picking the correct one from the given choices while the rewrite test requiring some construction performance skills. Although the cloze test does not provide information on the students' construction performance, it offers a clear indication about the students' proficiency level, which is included as a small-proportion measure in the rubric representing only a small part of the cohesion in the multiple-trait holistic scoring within the *'Vocabulary'* and *'Language Use'* parts.

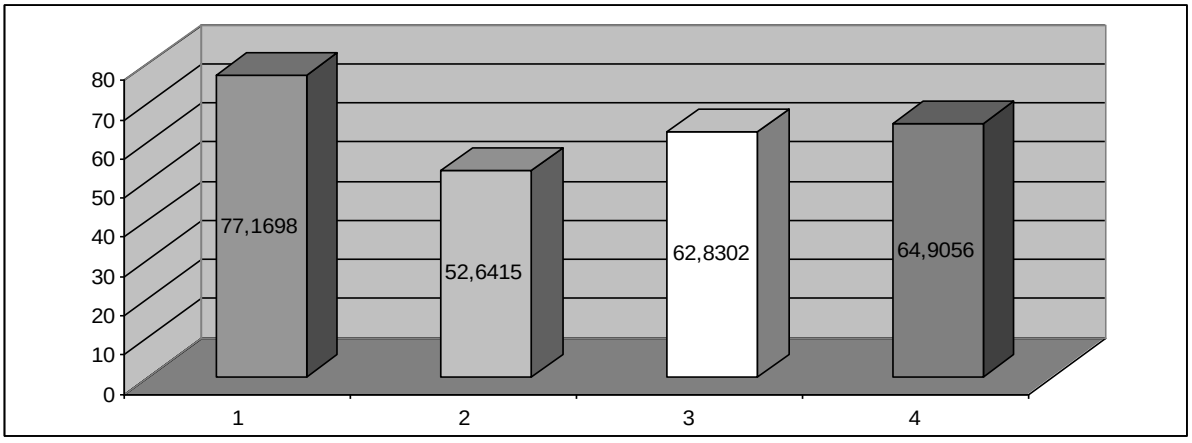
The 20 multiple-choice questionsin thecloze test include verb tenses, connectors, prepositions,adverbial clauses that are essential to 'cohesion' and 'clarity', which are included as a small-proportion measure within these parts. Although it is not possible to discern to what extent this measure was considered by the raters in their assessment judgements, it is possible to say that the students' correct choices of these items are closely related to their writing performance at the discourse level. In other words, the higher the cloze test score the clearer the expression in writing; the higher the proficiency the clearer the expression. Thus, in this regard it would not be wrong to say that the cloze test contains, if not completely, at least some degree of statistical reliability and concurrent validity.

Table 1. The mean scores of the test results

Test/Exam Type	Number	Minimum	Maximum	Mean	Std. Deviation
CLOZE TEST	53	35.00	95.00	77.1698	14.12545
REWRITE TEST	53	20.00	100.00	52.6415	18.62081
COMPOSITION EXAM	53	38.00	96.00	62.8302	12.70181
Valid Number	53				

The lower rewrite test mean score provides a more realistic clue about the students' writing performances as it represents a measure for a correct sentence construction skill by the correct use of the given prompts, which is more difficult a task to accomplish compared to the task required for the cloze-test. In this sense, it has a significant degree of construct validity for the students' composition writing performances as sentence construction capability is an indispensable skill required for a good composition writing performance. Furthermore, the consideration of the average mean scores of the two tests demonstrates much better statistical reliability and concurrent validity between the two indirect tests and the composition writing exam indicating the level of competence required for composition writing as it also represents 45% of the midterm grammar exam matching almost completely the composition exam mean score (See Table 2).

Table 2. Comparison of the mean scores of the direct and indirect assessment results.



1-Cloze test 2-Rewrite test 3-Composition exam 4-Average of Cloze & Rewrite tests

The comparison of the mean score per topic chosen for composition writing also constitutes a strong consistency between the mean scores of the topics and the mean score of the whole test (See Table 3), though a large proportion of the students, representing more than half of the total subjects (30 students with 56 percent), chose topic 3 with the rhetoric '*expressing opinion*' preferred over two '*process'* and one *'classification'* rhetoric choices. It is not possible to discern, from the data and materials used in this study, whether the students' choices were due to the preference of the rhetoric allowing them to express their opinion or whether they were due to the students' better familiarity with one topic compared to the other three.

Table 3. The mean scores and the frequency of the Composition Exam per topic

Topic No.	Mean	Number	Percent	Std. Deviation
1	64.8571	7	13.2	8.21439
2	62.4167	12	22.6	17.24929
3	62.3667	30	56.6	11.86848
4	64.0000	4	7.5	13.95230
Total	62.8302	53	100.0	12.70181

4. Conclusion

Indirect writing assessments can be used to diagnose direct writing performance skills provided that the questions are carefully prepared from proficiency measures essential for writing construction. Questions containing some sentence grammar construction tasks can enhance statistical reliability and concurrent validity of the indirect writing measure along with offering a limited degree of construct validity as a competence measure in cohesion.

On the other hand, indirect writing measures fall too short to offer any content validity, which is a fundamental element of direct writing assessment measures, as writing tasks cannot be accomplished without knowledge on a particular topic. The subjects used in this study have undergone the same prep-school training program, and their materials constitute this commonality in their indirect and direct

writing assessment performances and results, which were prepared in accordance with the syllabi they commonly followed and on which their assessment measures were based. The assessment results used in this study show that studies in one language skill area contributes to the improvement of other language skills, demonstrating the fact that each language skill is an essential and inseparable element of the target language acquisition.

Despite the disadvantages of the indirect assessment method compared to the direct writing assessment method - such as lack of fidelity to real world writing tasks, heavy reliance on examinees' reading rather than writing proficiency, and in many cases lack of face validity in the objective measure - the findings of this study posits the strong connection between the two approach as stated in the literature and indirect writing assessment results being strong indicator of direct writing performance(Breland, Conlon & Rogosa, 1976; Breland & Gayner, 1979; Huntley, Schmeiser & Stiggins, 1973; Godshalk, Swineford & Coffman, 1966; Hogan & Mishler, 1980; Moss, Pamela, Cole, Nancy & Khampalikit, 1981). The outcome of this study has also some significant pedagogical implications, underlining the usefulness of indirect writing assessment in exams given to a large population to be assessed within a limited time period, and the importance of direct writing assessment as part of formative assessment functioning as an important means of the identification of the target population needs and shortcomings in specific areas of writing performance and as part of the needs analysis of writing syllabus design.

4.1. Limitations

The data used in this study, which came from two tests used as part of a competence exam and a composition writing exam, constitute some heterogeneity in scale for comparison as indicated above (cf. 3.2). Although the two competence tests were also converted to the 100-point grading system, they represented different proportion and weight in the exam with different numbers of questions. The cloze test consisted of 20 questions with 1.5 points for each question and represented a 30-point proportion of the 100-point grammar assessment scale whereas the rewrite test, which can be said to contain a lot more assessment measure points in the composition writing assessment rubric, contained only 5 questions with 3 point for each question and represented only a 15-point proportion of the scale, which is half of the weight the cloze test carried in the grammar exam.

Another limitation relates to the assessment purpose for which these measures were used. The data used in this study were part of the midterm exam and were prepared to assess the students' achievement limited to the syllabus instruction, naturally being distinct from a purpose concentrating on the direct/indirect writing assessment comparison.

Related to the above limitations, the data used in this study fall short to offer a legitimate base for reliable frequency statistic analysis as the contents on which these data are based do not constitute enough frequency matching points for comparison. Thus, the data could not offer any validity for frequency analysis to contribute to the findings of the study.

There is a point related to the level of difficulty in the cloze test that might be considered to be an issue of weakness in the study. The cloze test questions offered only three choices for the candidates to pick the correct one. It is possible to allege that this provides some degree of convenience for guessing the correct choice instead of making a conscious judgement. Although this seems to undermine the reliability of the cloze-test, to what extent the candidates were tempted to guess rather than consciously chose and to what extend this temptation facilitated the correct choice coincidentally can be argued, not only for the cloze test used in this study but also for all cloze tests including the ones requiring the correct choice to be made from more options. Thus, this weakness is rather an inherent one, at least to

a certain degree, for all the studies concerning multiple-choice measures not peculiar to the present study only.

Another area which might be considered a weakness factor in this study is the limited number of the subjects used in it as a larger number would provide clearer indication with more reliability in data, with more acceptability and more validity of generalisation on the outcome of the study.

The scope of this study did not include any observation or data regarding the preparation of the rubrics and the assessment training the raters had been given prior to the scoring. It is not clear as to how the raters judged the students' direct writing performance, to what extent their judgements were consistent with what required in the rubric, and as to whether or not their judgements were influenced by other factors.

4.2. Suggestions

Indirect writing assessment measures should be planned more carefully with the inclusions to the broadest possible proficiency points relevant to writing activities if they are to be used as measures for direct writing assessment. This is particularly essential when indirect assessment measures are used for diagnostic purposes of large scale assessment where predictions on the candidates' present proficiency skills are judged as parameters of their present writing performance and their future academic progress. The proficiency measure areas should be carefully identified to form efficient data base for 'concurrent validity', 'statistical reliability', and to a certain extent, 'construct validity' required for direct writing assessment judgements.

As indirect assessment measures fall too short to provide efficient 'construct validity' and any 'content validity' direct writing assessment should be preferred for proper judgement of the candidates' writing performance, particularly in small-scale assessments. The criteria used in the rubrics should be well defined and carefully designed to reflect appropriate measure points to cater for all aspects of writing performance. The raters should be well trained on the content and the purpose of the rubrics in the assessment process, and if possible, the raters should be included in the development process of the rubrics.

References

Alderson, J. C., C. M. Clapham & D. Wall (1995). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

Breland, H.M., Conlon, G.C., and Rogosa, D. (1976). *A preliminary study of the test of standard written English*. Princeton, NJ: Educational Testing Service.

Breland, H.M. and Gaynor, J.L. (1979). A comparison of direct and indirect assessments of writing skill. *Journal of Educational Measurement, 16*, 119-128.

Davidson, F. & B. K. Lynch (2002). *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications.* London: Yale University Press.

Godshalk, F.I., Swineford, F., and Coffman, W.E. (1966). *The measurement of writing ability*. New York, NY: College Entrance Examination Board.

Grabe, W. and Kaplan, R.B. (1988). *Theory and Practice of Writing*. New York: Addison Wesley Longman.

Hogan, T.P. and Mishler, C. (1980). Relationships between essay tests and objective tests of language skills for elementary school students. *Journal of Educational Measurement, 17*, 219-227.

Huntley, R.M., Schmeiser, C.B., and Stiggins, R.J. (1979). *The assessment of rhetorical proficiency: The role of objective tests and writing samples.* Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA.

Huot, B. (1990). Reliability, Validity, and Holistic Scoring: What We Know and What We Need to Know. *College Composition and Communication, 41*(2), 201-213.

Moss, Pamela A., Cole, Nancy S., and Khampalikit, C. (1981). *A comparison of procedures to assess written language skills in grades 10, 7, and 4*: Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA.

Ruth, Leo and Sandra Murphy. (1988). *Designing Writing Tasks for the Assessment of Writing*. Norwood, New Jersey: Ablex Publishing Corp.

Yazma becerisini dolaylı olarak ölçmek doğrudan ölçmek bakımından yeterli midir?
Geçerlilik ve güvenilirlik konularının incelenmesi

Öz

Bu makale hazırlık okulunda öğrenim gören 53 üniversite öğrencisinin ara sınav sonuçlarını temel alarak gerçekleştirilmiştir. Çalışma, öğrencilerin dil becerilerinin değerlendirildiği sınavda, öğrencilerin kompozisyon yazılı sınavındaki performansları ile bir adet boşluk doldurma testinin ve bir adet gramer düzeyinde becerilerin değerlendirildiği yeniden yazma testinin sonuçlarının karşılaştırılması ile gerçekleştirilerek, yazı becerisinin doğrudan ölçme ile dolaylı olarak ölçme arasındaki korrelasyonu ortaya koymayı amaçlamaktadır. Çalışmanın amacı, öğrencilerin dil becerilerini kullanmada kullandıkları stratejilerin, ölçme ve değerlendirme bakımından kullandıkları diğer becerilerin, yazı yazma becerileri ile konu seçimi, kompozisyon yazımı ve retorik açısından ne oranda nicel ilişki olduğunu geçerlilik ve güvenilirlik açısından incelemektir.

Anahtar sözcükler: Yazı becerisinin dolaylı olarak ölçme; yazı becerisini doğrudan ölçme; geçerlilik; güvenilirlik

AUTHOR BIODATA

Dr. Faruk Kural is a lecturer at the English Language Teaching Department of Hasan Kalyoncu University. He has a BA degree from Deakin University, a MA in Applied Linguistics from Monash University, and a PhD in English Language Education from Yeditepe University. He is interested in ELF, ESP, syllabus design, multicultural education, functional grammar, and language planning and policy.