The Reliability of the Holistic Method When Grading Language Essays

Mª PILAR GARCÍA MAYO
Dpto. de Filología Inglesa y Alemana
Universidad del País Vasco
C/ Marques de Urquijo s/n
VITORIA - 01006

ABSTRACT

The evaluation of students' writing ability has become increasingly important in second language teaching. There are two main approaches to writing skill assessment, direct and indirect methods. This paper deals with direct methods of assessing writing ability, in particular with the method known as holistic evaluation. After providing a brief description of this method, the alleged sources of its lack of reliability (writers, readers and topics) are analyzed in turn and some possible ways of handling rhem are considered. Some suggestions for further research on the topic are offered before concluding that, from the author's point of view, the holistic method is an important measure to assess the underlying constructs of writing and that its use should be encouraged.

KEY WORDS: Writing, Holistic method, Foreign language.

RESUMEN

La evaluación de la producción escrita de los alumnos es uno de los temas de investigación más importantes dentro del campo de la enseñanza de segundas lenguas. Dicha evaluación se puede realizar utilizando mérodos directos o métodos indirectos. El presente trabajo se centra en los métodos directos y, específicamente, en el conocido como de evaluación global. Comenzamos por presentar una breve descripción de este método para analizar seguidamente las supuestas fuentes de su falta de fiabilidad (escritores, lectores y temas), así como posibles formas de solventar esos problemas. Ofrecemos algunas sugerencias para posteriores investigaciones antes de concluir que, desde nuestro punto de vista, el método de evaluación global es una forma importante de evaluar las ideas fundamentales que se reflejan en la producción escrita y que, por tanto, se debe fomentar su uso.

PALABRAS CLAVE: Escritura, Evaluación global, Lengua extranjera

Mª PILAR GARCÍA MAYO

I. INTRODUCTION

During the last few years the evaluation of students' writing ability has become increasingly important in second language teaching. A great deal of research has been done on the areas concerning the various methods available and the reliability of their results.[1] There are two main approaches to writing skill assessment. H. Breland and J.L. Gaynor (1979:119) define them as follows: «The direct assessment requires that actual essays be written and usually such essays are read and scored independently by two or more readers. The indirect assessment, sometimes called objective assessment, requires no writing at all - the examinee only responds to stimuli in a multiple-choice format.» Both methods have proved to be successful. However, according to H. Breland and J.L. Gaynor (1979:127) «indirect methods lack face validity and credibility among the members of the English profession and educators generally, and they tend to deliver the message to students that writing is not important.»

This paper will deal with one of the techniques the direct approach uses to grade writing skills: holistic evaluation.[2] Although holistic evaluation has much to offer. it has drawbacks as well. Therefore, the research questions to he answered in this paper are: (i) What are the sources of the alleged lack of reliability when grading foreign language essay tests through the holistic method? and (ii) How can optimum reliability be obtained using the same method? We first provide a brief description of the holistic method, examine the sources of its alleged lack of reliability and list some possible ways of handling them. Finally, we offer some personal comments and suggestions for future research. We hope

[1] Test reliability. The reliability coefficient for a set of scores from a group of examinees is the coefficient of correlation between that set of scores and another set of scores of an equivalent test obtained independenritly from tlie members of the same group. The more appropriate a test is to the level of abilities in the group, tlie higher the reliability of the scores it will yield.

[2] Analytical scoring and prirnary trait scoring are other so-called direct composirion scoring techriiques. Analytical scoring is an evaluation in which performance is brokeri down into cornponrnt parts (e.g. organization, granimar, vocabulary, meclianics, fluency), whereas in primary trait scoring a holistic score is assigned to a particular featiire of writing siich as structure. tone and vocabulary. In this paper I do not address issues such as different types of objective measures used (e.g measures of length, of subordinatiori and relativization, of sentence connectors or of syntactic complexity (T-units)) or the reliability - or lack of it - of those measures A thorough analysis would go well beyond ihe scope of this article. The reader rnay refer to Evola, Mamer and Lentz (1980), Flahive and Snow (1980), Gaies (1980), Hornburg (1980), Mullen (1980) or Perkin (1980).

that ESL professionals can make use of this discussion as they try to decide how best to handle questions related to writing ability within the holist approach.

II. HOLISTIC EVALUATION

Ch.R.Cooper (1977:3) provides a summary statement about holistic evaluation:

> Holistic evaluation is a guided procedure for sorting or ranking written pieces. The rater takes a piece of writing and either (i) matches it with another piece in a graded series of pieces or (ii) scores it for the prominence of certain features important to that kind of writing and (iii) assigns it a letter or number grade. The placing, scoring or grading occurs quickly, impressionistically, after the rater has practiced the procedure with other raters. The rater does not make corrections or revisions in the paper. Holistic evaluation is usually guided by a holistic scoring guide which describes each feature and identifies high, middle and low quality levels for each feature.

According to K. Perkins (1983:652) «Of all the composition evaluating schemes available today, holistic scoring has the highest construct validity[3] when overall attained writing proficiency is the construct to be assessed.» This method is a recommended tool for certification, placement. proficiency and research testing. However. as every method, it also has drawbacks.

In scoring holistically. the grader reads the composition, forms a general impression, and assigns a mark to that composition based on some standard. The standard may be either a model composition to which the reader has access, or a general impression the reader has based on previous experience in reading student compositions. Such evaluation can, therefore, be highly general and subjective due to bias, fatigue. previous knowledge of the student and shifting standards from one paper to the next. These drawbacks are referred to collectively as 'threats to reliability' and they constitute the major criticism levelled against the holistic evaluation of essay tests. Let us take a closer look at the sources of this lack of reliability and consider the ways they can be handled.

---

[3] The term validity. when applied to a test. refers to the precision with which the test rneasures some cognitive ability Therc are thus two aspects to validity: what is rneasured and how precisely it is measured

III. SOURCES OF VARIABILITY

The division established by W. McColly (1970:149) for the sources of variability in grading language essays by means of the holistic method will be basically adopted here. W. McColly divides the possible sources into three groups: writers, readers and topics. Each one of these will be dealt with in order and suggestions to improve reliability in each area will be offered.

III.1 WRITERS

W. McColly points out that there are no research findings dealing with reliability and the writer's role in essay testing. He agrees with R. Bradock et al. (1963) in the fact that, even if all the other sources of error were controlled, it would still be true that we cannot be sure that the students are fully using their ability. We may attribute this low performance to some of the factors pointed out by R.L. Thorndike (1951) (reprinted in R.L. Ebel and D.A. Frisbie (1986:74)): adherence to time constraints (but see T. Caudery 1990), the examinee's physical condition, external conditions of light, heat ... However. W. McColly thinks that this is not a real problem: if we want to measure a student's performance such distractions should not count because they are part of life.

Research on the writer's role in essay testing is certainly needed: there will always be differences among students in a classroom, differences that will be present not just in testing situations but in everyday classroom interaction. Students' physical condition or even their psychological condition is something over which we have no control. The best we can do to avoid the alleged lack of reliability of the holistic method in this area is to get the writers involved in their task. Writing quality has a direct connection with conveying meaning, with cornmunicating a message to an audience. A first step could be grouping within the classroom. Separated into small groups of four-five, students have a natural audience to write for. Such a procedure, promising interaction and feedback, is an interesting departure from the usual system involved when students write solely for the teacher.

II.2 READERS

The role played by readers in the holistic procedure has been often criticized. Both interrater and intrarater reliability should be considered. Different graders may assign the same composition to different categories affecting interrater reliability. The same composition grader may assign the same

composition to different grading categories at different times affecting intrarater reliability. This latter type of variability is what K. Perkins (1983:653) refers to as 'lack of consistency'. We will be concerned here with interrater reliability because different studies have shown that 'days don't matter much but readers do.' This last statement summarizes H.I. Braun and H. Waner's (1989) findings in one experiment carried out for an essay question in English literature. Over the course of a four-day experiment. each of the 12 readers available read each of 32 essays exactly once. Each reader read eight essays every day and, therefore. estimates were obtained of systematic differences between days. The result was that each reader was consistent over the four-day period but there was some variability among readers.

As we have previously mentioned, the main criticism levelled against holistic evaluation carried out by different judges is that they assign different evaluations to the same composition. In other words, the evaluation is highly subjective. Research on holistic scoring in terms of reliability has yielded contradictory findings. C.M. Kaczmarek (1980:151), for example, reports that «Subjective methods of evaluating essays work about as well as objective scoring techniques and are strongly correlated with other measures of ESL proficiency which have independent claims to validity.» J.C. Follman and J.A. Anderson (1967) reported interrater reliability coefficients as high as 0.90. Along the same lines, T.J. Homburg (1984:88) offers data from the study developed by the Testing Certification Division of the English Language Institute of Michigan. The reliability coefficients, based on correlations between the scores assigned to a certain composition read by two readers, ranged from a low of .721 to a high of .932 with a median of .880.

On the other hand, research exists showing that professional persons, including English teachers, vary in their assessment of attained writing proficiency. K. Perkins (1983:653) cites the work done by P. Diederich et al. (1961) in which «Sixty professionals were asked to grade 300 papers by college freshmen from different schools. The readers, who represent six occupational fields [...] were asked to sort the 300 papers into nine groups.» The result was that some essays ended up with every possible grade from 1 to 9. I do not think that this criticism can be used against holistic methods of grading essay tests, though. The readers in the Diederich study came from different backgrounds: some were coliege English teachers. social science teachers, natural science teachers, writers and educators, lawyers and business executives. One cannot expect these people to have the same standards when it comes to grading an essay test.

Searching for reasons to explain the lack of agreement among these readers, C.D. Hirsh (1977) suggested that different weights were attached to a few traits of writing and noted that reliable agreement in the scoring of writing samples is out of the question until agreement is obtained about what should be judged. It would be ideal, therefore, if all the raters had the same background but, if that is not the case, their training is the best way to increase reliability. Most researchers agree on this point. K. Perkins (1983:654) calls for the insistence on rater competence and expertise. T.C. Homburg (1984:103) states that «the training of readers can be important to the reliability and, hence, the validity of the holistic grading process.» McColly (1970:150) points out that «it is plain that readers must be given the proper training and orientation, regardless of how knowledgeable they are.» This training may be done either by providing the judges with predetermined standards and criteria for evaluation or by having the judges themselves arrive at a determination of their own standards. W. McColly considers this latter procedure to give better results. If we have a group of writing teachers, for example. each one will have certain aspects that s/he looks for in an essay. They will not need the same kind of training that people from different backgrounds will need. But they will have to agree on the aspects they are going to consider when grading.

The setting of common standards for judging quality of writing is another aspect to consider when trying to increase reliability. The importance of this common standard is emphasized in the following quote by W.F. Irmsher (1979) (in R.M. Terry (1989:51): «Evaluation obviously implies values. but many teachers evaluate without defining them or just feel frustrated because they can't quantify the value they hold. Without clearly defined values, it is impossible to make consistent judgments and discrirninations [...]. Not knowing what else to do, teachers proofread instead of reading critically.»

It is obvious that a specific set of values, common to all raters, has to be established. This set of common values will avoid the shifting of standards and will help to focus the rater's attention on significant aspects of the cornposition (see J.D.Brown 1991). A good idea is to monitor the readers periodically to check if they are consistent in applying the agreed upon criteria. W.E. Coffrnan (1972) demonstrates that both high reliabilities and validities for direct assessment can be obtained when multiple readings of each essay are made. Clearly this is a good piece of advice. but a difficult one to follow in a normal classroom setting because multiple readings are very time-consuming and require the availability of many people to collaborate with the grading process.

Another good method to increase reliability is to remove students' names from the essays and replace them by the last digits of their I.D. number.

Whenever possible we should have *students type their essays to avoid the 'handwriting factor', the most tangible* source *of unreliability and invalidity in essay tests. Research has shown that there* is *a significant interaction between qualities of handwriting and order of reading, which possibly indicates a tendency for the reader to progressively develop negative bias toward poor handwriting.*

As a teaching assistant in *the ESL program at the University of Iowa* (U.S.A), I *participated in the English proficiency testing of* all *new University of Iowa international graduate students.* Some *time before the beginning of each* semester (spring, *summer and* fall) *we graded writing samples holistically (using a general impression holistic* evaluation)[4]. *Before each grading session, we had a meeting to review the standards that had* been *agreed upon in previous semesters and were provided with sample essays with a discussion of the marking of each. A total of 10112 ESL teachers scored the handwritten essays during two to three six-hour sessions (this was a placement test so typewritten essays were out of the question). Each essay was read independently by two readers and scored on a scale from 1* (poor) *to 6 (good). In the few cases in which a significant difference (two scale points) was observed, a third reader was asked to adjust the score (only for admission purposes). The interrater reliability coefficients reported ranged from .712 to a high of .924. These coefficients were based on the correlation computed from the compositions written/rated in the* Fall *91 - Spring 93 period.*

If we now return to the drawbacks K. Perkins (1983) found in the holistic method of evaluation (the generality and the subject- tivity of the evaluation), we see that solutions have been *offered to overcome them. Thus, to avoid «*the *generality and* [...] *shifting of standards from one paper to the* next,» *we* have *our common standard for judging the* quality *of the essay. The subjectivity in grading is* also *avoided by removing the students' names from the essays (or by even having those essays typed whenever possible). And,* finally, *to* prevent *subjectivity we will* have *a group of trained professionals who know what they are looking for in the essay and will* provide *independent readings for each sample.*

111.3 TOPICS

The performance of writers varies from topic to topic and is another source *of variability when grading essays. W .* McColly *(1970:153) «*For *a writing topic to be* valid, *it should* have *the property of filtering out not only differences*

[4] General irnpression rnarkirig is the simplest of the procedures in holistic evaluation. It requires no detailed discussion of features and no summing of scores given to separate features.

ascribed to knowledge but also those arising from fluency in logical operations.» The kind of topic proposed as an alternative to the topic highly structured in content, in which all students are given something to say, is a topic in which all students are deprived of something to say. The following two are examples of the latter type:

a. You have heard the saying «The best things in life are free». Decide whether this is true or false, then write an essay in which you defend your opinion.

b. Consider these contradictory proverbs «Look before you leap» and «He who hesitates is lost.» Decide which of the two offers better advice, then write an essay in which you defend your choice.

K. Perkins (1983:654) encourages «the elicitation of multiple writing samples to control for the fact that attained writing ability may vary with topic and time of day.» H. Breland and J.L. Gaynor (1979:120) agree in that «both high reliabilities and validities for different assessment can be obtained when multiple samples of writing and multiple readings of each are made.« What we learn and apply to a classroom situation is that as many samples as possible should be obtained from each student during the semester/term. When an essay topic is given, choices should be avoided because if different examinees answer different questions, the basis for comparing their scores is weakened. According to R.L Ebel and D.A. Frisbie (1986:133), «when students choose the questions they can answer best, the range of test scores is likely to be narrowed - hence the reliability of the scores would be expected. The essay topic should be carefully phrased so that students know what they are expected to write about. And, finally, as it was mentioned earlier, another factor one needs to consider to increase reliability is the communicative aspect of any writing task. Providing the students with meaningful topics plays a very important role in achieving greater reliability through the holistic method.

IV. FINAL COMMENTS

The sources of the lack of reliability in the grading of essays using the holistic method have been identified in order of importance as: readers, topics and writers. Some possible solutions to obtain optimum reliability through the holistic method have also been proposed. To increase reader reliability, proper training of those involved in the grading process should be considered and the setting of common standards and possible methods of concealing students' identities should

also be taken into account. As for topics, it is suggested that, to achieve a greater reliability, rnultiple sarnples of writing should be obtained frorn each student. Also, it is irnportant that everyone write on the sarne subject, and that the chosen topic emphasize the communicative aspect of writing. Finally, it is concluded that the writer's variable does not play as great a role as the other two variables insofar as being a source in the lack of reliability. However, it is suggested that students should be provided with meaningful writing tasks in order to motivate them and to obtain the best results. Definitely. rnuch more research should be done on the areas of the role played by topics and writers, especially in the latter where no research findings exist. As far as topics are concerned, it would be interesting to know not only to what extend topic difficulty influences writing performance but also the possibility, related to the reader's variables, that readers may reward the choice of a more difficult topic.

In surnrnary, what can be deduced from this brief survey of the literature is that rneasuring writing ability with the holistic rnethod is not as easy as it seems. However, apparent difficulties should not lead us to the use of objective rneasures instead of the holistic rnethod. In rny opinion, objective rnethods can be used as a cornplernentary source of information about the writing ability of students but no surnrnative evaluation should be rnade based on results frorn objective tests alone. As K. Perkins (1983:662) says: «While objective rneasures rnay be of interest to researchers, they. seerningly, are of little value in assessing the underlying constructs of writing because the intent to cornrnunicate is neither assessed nor rneasured by them.« I think that the cornmunicative and rneaningful part of the writing task is the one that should be ernphasized and that is the part that indirect rnethods lack.

Perhaps a fitting coda to this paper is the following frorn Ch.R. Cooper (1977:3): «When there is cornrnitrnent and time to do the work required to achieve reliability of judgrnent, holistic evaluation remains the rnost valid and direct means of rank-ordering students by writing ability.»

Fecha de recepción: 14 - 3 - 1994

WORKS CITED

Braddock. R., R. Lloyd-Jones & L. Shoer (1963) ***Research in Written Composition***. Champaign. Illinois: National Council of Teachers of English.

Braun, H.I. & H. Waner. (1989) «Making Essay Test Scores Fairer with Statistics», in J. Tanur et al. (eds.) *Statistics: **a Guide to rhe Unknown.*** Pacific Grove, CA: Wordsworth and Brooks: 178-187.

Breland, H. & J.L. Gayiior (1979) «A Comparison of Direct and Indirect Assessments of Writing Skill», Journal of Educational *Measurement*, 16(2): 119-128.

Brown. J.D. (1991) «Do English and ESL Faculties Rate Writing Samples Differently?», *TESOL* Quarterlv. 25(4): 587-603.

Caudery, T. (1990) «The Validity of Tiined Essay Tests in the Assessment of Writing Skills», ELT Journal, 44(2): 222-231.

Coffman. W.E. (1972) «The Validity of Essay Tests» en G.H. Bracht, K.D. Hopkins & J.C. Stanley (eds.) *Perspectives in Educational* and *Psychological Measurements*. New Jersey: Prentice Hall.

Cooper, Ch.R. «Holistic Evaluation of Writing» (1977) in Ch.R. Cooper & L. Odell (eds.) *Evaluating* Writing: Describing, Measuring, Judging. Urbana, Illinois: National Council of Teachcrs of English: 3-31.

Diederich, P.. J. Frcnch & S. Carlton (1961) *Factors in Judgments* of *Writing Ability*. Educational Testing Service Research Bulletin RB-61-15. Princcton, New Jerscy: Educational Testing Service.

Ebel, R.L. & D. A. Frisbie (1986) *Essentials* of Educational *Measurement*. Englewood Cliffs: Prentice Hall.

Evola, J.. E. Mamen & B. Lentz (1980) «Discrete Point versus Global Scoring for Coheaive Devices» in J.W. Oller & K. Perkins (eds.) Research *in Language* Testing. Rowley, MA: Newbury House: 177-181.

Flahive, D. & B. Snow (1980) «Measures of Syntaciic Complexity in Evaluating ESL Compositions», in J.W. Oller & K. Perkins (eds.) Researrh *in Language* Testing. Rowlcy. MA Newbiin House: 171-176.

Follman, J.C. & J.A. Anderron (1967)«An Investigation of the Reliability of Five Procedures for Grading English Themes», Research in *the Teaching* of *English*, 1(2): 190-200.

Gaies, S.J. 11980) «T-unit Analysis in Second Language Research: Applications, Problems and Limitations», *TESOL* Quarterly, 14(1): 53-60.

Hirsch, E.D. (1977) *The Philosophy* of *Composition*. Chicago: The University of Chicago Press

Hombug. T. J. (1980) A *Syntactic Complexity Measure of attained ESL Writing Proficiency*. M.A Thesis. Carbondale: Southern Illinois University

Hombug. T.J. (1984) «Holistic Evaluation of ESL Composition: Can It Be Validated Objectively?», *TESOL Quarterly*, 18(1): 87-107.

Irrnscher, W.F. (1979) *Teaching Expository Writing* New York: Holt. Rinehart and Winston

Kaczmarek. C M. (1980) «Scoring and Rating Essay Taskr- in J W. Oller & K. Perkins (eds.) Research *in Language* Testing. Rowley, MA: Newbury Hourc Publishers: 151-159.

McColly, W. (1970). «What does Educational Research Say about the Judging of Writing Ability?», *The Journal of Educarional Research,* 64(4): 148-156.

Mullen, K. (1980) «Evaluating Writing Proficiency in ESL» in J. W. Oller & K. Perkins (eds.) *Research in Language Testing*. Rowley. MA: Newbury House Publishers: 160-170.

Perkins. K. (1980) «Using Objective Mcthods of Attained Writing Proficiency to Discriminate Among Holistic Evaluation», *TESOL Quarterly*, 14(1): 61-69.

Perkins, K. (1983) «On the Use of Composition Score Techniques, Objective Measures and Objective Tests to Evaluatc ESL Writing Ability», *TESOL Quarterly*, 17(4): 651-671

Terry, R. M. (1989) «Teaching and Evaluating Writing as a Communicative Skill», *Foreign Language Arinals,* 22(11. 43-53.

Thorndike, E.L. (1951) «Reliability» in E.T. Lindquist (ed.) *Educarional Measuremetir.* Washington D.C.: American Council on Education.

APPENDIX

HOLISTIC CRITERIA

1. Easily recognizablc. Barely cohcrent with many structural errors, misspellings, diction problems. Nothing noteworthy in the way of ideas and comments.
2. Errors in diction, spelling. agrccinent and structure. Sonie fluency and order to commonplace thoughts or ideas
3. Errors in struciure and grammar. Some coherence and more fluency, but little pointcd organization or unity, though there is evidence of some thought. Repetitive syntactical patterns.
4. Some sense of unity and effective cohcrence. Usually only one commonplace illustration or one line of argument. Weak paragraphing: paper has beginning, and shows some effort at a coiiclusion. Still a fcw errors in grammar and mechanics. Varied patterns.
5. Generally skillfully written with effective sentence sense and good control of mechanics. Usually effective paragraphing, cniploying more than a single illustration. Occasional errors in spelling, punctuation or agreement. Unity and coherence are evident.
6. Skillfully written. as in rating 5, but with the addition of sonie sense of style, or an argument or multiple illustrations that are more than merely commonplace. Perhaps effective use of personal expcrience as well as hunior or irony though these last are rare. Generally free of errors. (H. Breland & J.L. Gaynor 1979: 121-122)