A REASSESSMENT OF ERROR-COUNT

Jonathan D. Picken

Abstract

Research in the field of error-count evaluation of EFL writing has not had a history of results consistently successful enough to establish the method as a full-fledged alternative to analytic or general impression marking. In this article it is suggested that this record is not so much due to an inherent weakness of error-count evaluation as such, but rather to the lack of a theoretical foundation in the methods used to date.

In order to make this point, extensive reference is made to the Dutch CITO writing proficiency test (CITO, 1984) and to a related CITO study, Melse and Verstralen (1986). Data from the latter are marshalled by the present author in making his case in support of error-count evaluation. The described evaluation procedure is used to determine the English language writing skills of test subjects, and not to correct student essays.

Error-Count Evaluation of EFL Writing and the CITO Writing Proficiency Test

Over the years, error-count (EC) methods of writing proficiency assessment — variously known as frequency-count or objective methods — have not had an easy time in establishing themselves as viable alternatives to analytic or to general-impression approaches. While coming under fire for a lack of theoretical underpinnings, EC has been hard put to come up with results consistent enough to silence its critics. The problem has been compounded by the fact that EC, concentrating as it does on errors, has found itself out of step with mainstream writing pedagogy where work on mistakes has had to take something of a back seat, and only errors that interfere with communication have been deemed worthy of attention.

Jonathan Picken has obtained Master's degrees from Groningen State University in the Netherlands, and the University of London Institute of Education. He is a full-time lecturer at Tokai University in Kanagawa prefecture, and has lectured on writing proficiency testing at two JALT conferences.

The key issue that has not been addressed so far is the question of whether EC as a method is in principle misguided. Criticism so far has concerned itself with the weaknesses of EC studies to date, without considering the broader issue of whether EC in whatever form is bound to fail. The purpose of this paper is to show that the case is far from closed, and that EC with a broader theoretical orientation has a very considerable potential that deserves to be recognized.

The Background

Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey (1981) define EC methods as ones that

> tally or enumerate certain elements in the composition such as: the number or type of words, clauses, T-units [see Table 1], cohesive devices, misspelled words, misplaced commas, or sentence errors. (p. 29)

The key word is *tally*, for this is what distinguishes EC from general impression or analytic marking. The EC marker counts; markers using the latter methods do not.

As stated in the introduction, the validity of all this counting has been called into question, mainly because it tends to ignore what are considered major aspects of student compositions. Perkins (1983), a former EC researcher, writes one of the strongest formulations of this view, claiming that EC methods are

> of little value in assessing the underlying constructs of writing [and that] currently used objective methods do not quantify cohesion, coherence, organization, . . . idiom, diction, tone, relevance or focus — all factors which contribute to good writing. (p. 662)

Against the background of this criticism, it is unfortunate, though perhaps not surprising, that even some of the most popular EC methods have failed to distinguish consistently between students of different ability levels. A comparison of the performance of four of these measures over three studies (see Table 1) shows that not one of them succeeds in distinguishing *both* between students of roughly the same level of ability, and between students of different ability levels. Thus the same measure may differentiate well between, for example, good and poor beginners, but not between, for example, beginners and intermediate students. This is clearly unsatisfactory.

Table 1
Ability of Four EC Methods to Evaluate Performance
within levels and Distinguish *between levels*

Basis of Evaluation	Perkins (1980)	Larsen-Freeman (1983)	Homburg (1984)
length error-free T-unit	within level	only between 2 of her 4 levels	
errors per T-unit	within level		not between levels
total errors per composition	within level		not between levels
error-free T-units per composition	within level	not between levels	between levels

(Note: T-units or terminable units are defined as the shortest possible units of a passage that are grammatically allowable to be punctuated as sentences.)

Further criticism could be adduced, but we need not labor the point; with neither a claim to theoretical respectability, nor a solid record of good experimental results, the case for EC methods seems seriously flawed.

Limitations of Research to Date

The main weakness of research carried out so far has been that the theoretical justification for the EC methods used has often been questionable. Given this situation, it would have been purely coincidental if any of the measures had performed more satisfactorily than they did. There is no obvious reason, after all, to expect that a measure of writing ability based on, for example, a mechanical tallying of the number of errors in a composition is likely to perform in the same way as an analytic rating scale that takes, for example, content and organization into consideration. Research so far seems to bear out this point.

The question that has gone unanswered, however, is whether an EC method that does start from an informed perception of what matters in composition would fare as badly as its predecessors. In the course of the following discussion of the CITO writing proficiency test, I hope to provide at least a partial answer to this question.

The CITO Writing Proficiency Tests

The writing tests that will be discussed in this section were developed by the CITO, the Dutch Central Institute for Test Research, with the aim of assessing the writing proficiency of students in their final, examination year at three different types of secondary school: VWO, HAVO, and MAVO. The VWO has a six-year curriculum, and the final examination at this school admits successful students to university education. Graduates of HAVO and MAVO — the former has a curriculum of five and the latter of four years — are admitted to various other kinds of further education, but not to universities. In their examination years, VWO, HAVO, and MAVO students would typically be 18, 17, and 16 years old respectively.

The CITO tasks are highly controlled and require students to write formal or informal letters to a more or less clearly defined audience and with a purpose and content specified in detail by means of a number of sub-assignments formulated in Dutch. The tasks vary in degree of difficulty according to the kind of school for which they were designed.

The marking protocol used (see Table 2 below) is the same for all three types of school. A notable contribution to its final form came from a group of schoolteachers who cooperated with CITO in carrying out trials and providing feedback. It was essential to take these future users' views into account, because within the Dutch educational system, secondary-school teachers were under no obligation to use these particular tests for examination-year writing proficiency assessment. Consequently widespread acceptance of the tests could only be achieved by gaining a broad consensus of support among teachers. The teachers insisted on having a much more detailed and specific marking protocol[1] than CITO had originally intended (Melse, 1984, p. 358). Further support for such changes came from the test trials (Melse, 1984, p. 358).

Given the test's history, it will come as no surprise to find that the marking protocol is not the product of one single perception of writing proficiency. It does, however, take a very broad range of errors into consideration. Beyond punctuation, spelling, and grammatical errors, it looks for errors of style (excessive repetition), discourse (illogical connection; lack of clarity; too sudden change of subject) and of appropriateness (use of word or expression inappropriate to context) (see Appendix I for a clarification of these error types). Content errors are manifested as "incompleteness" or "absolutely incomplete" in cases where students have failed to carry out a sub-assignment partially or completely. The bonus/malus or penalty system allows for a limited number of points to be added or subtracted where unquantified strengths or weaknesses of compositions require this.[2]

Table 2
Summary of the CITO Marking Protocol (after CITO, 1984, p. 20)

Sign	Meaning	Characteristics	Points
O (O O etc.)	incompleteness	element of the sub-assignment missing	–2 (–4 etc.)
?	absolutely incomplete	assignment not carried out	–8
+ (+ +)	bonus	style or content of item observably above average	+1 (2)
– (– –)	malus	style or content of item observably below average	–1 (2)
═══	wordgroup error	network of errors such that it is difficult to decide how many errors have been made	–2
———	primary error	— removal or replacement of word required — grammatical error — word-order error — excessive repetition — illogical connection — lack of clarity — use of word or expression inappropriate to context	–1
—V—	primary error	— date partially missing — word needs to be added — too sudden change of subject	–1
—/—	secondary error	— spelling error — secondary preposition error	–1/2
X	punctuation	— punctuation, apostrophes, capital and lower-case letter errors	–1/4
	dependent error, repetition of error, punctuation error of type not to be counted		

All students start with a basic score of 40 points. After correction, points are added to and subtracted from this total. The resulting total is translated into a score on a scale of 10, using CITO score conversion tables. (See Appendix A for definition of Characteristics. See Appendix B for a sample letter marked along these lines.)

In broadness of orientation, the CITO test is indisputably superior to its predecessors, and this is reflected in the test statistics, for the reliability figures are, if not ideal, at least sufficiently strong for our present purposes. Thus if the tests are administered according to CITO instructions (students write one formal and one informal letter on separate occasions, with each letter being marked independently by two teachers), test-retest reliability scores of around .80 are achieved. (A score of 1.0 would mean complete agreement among markers.)

Test-retest reliability is a measure of the extent to which the same test administered on a different occasion produces the same result. Table 3 below shows how CITO's reliability scores improve when either the number of compositions or the number of markers or both are increased. The scores of roughly .80 referred to in the previous paragraph are displayed in the fifth column (2 compositions, 2 markers). In all cases, the reliability has been checked by giving a separate test and comparing the point result with that of the earlier test(s). Thus, the figures in column one were reached when results of a second test were used to determine the accuracy of the first evaluation.

Table 3
Test-Retest Reliability of the CITO Writing Proficiency Test

	Number of compositions					
	1			2		
	number of markers			number of markers		
	1	2	3	1	2	3
HAVO English	.51	.68	.76	.68	.81	.86
VWO English	.49	.66	.74	.66	.79	.85

(Melse & Verstralen, 1986, p. 122)

Interrater reliability, that is the extent to which different raters assign the same scores to compositions, consistently hovers around the .65 mark (Melse & Verstralen, 1986, p. 111). Intrarater reliability, which shows the degree to which a rater is consistent in assigning grades, was not calculated.

To validate the test, CITO carried out a rank-order correlation of 25 letters (12 formal, 13 informal) rated by, on the one hand, a group of 10 HAVO/VWO teachers, and on the other a group of 16 NS raters (teachers and non-teachers) who ranked the letters using general impression. The resulting Spearman's rank-correlation figures[3] are r=.92 for the informal letters, and r=.86 for the formal ones. These results show that native speakers and Dutch foreign language teachers rank writing samples similarly when the latter use the CITO method.[4]

The reliability figures compare favorably with the popular Jacobs et al. (1981) ESL Composition Profile ones. Jacobs et al. have a test-retest reliability score of .72 (p. 73) and an interrater reliability coefficient of .65 (p. 69).[5]

The significance of all these figures, within the context of this article, is not so much that they tell us how good a test the CITO's is, but rather, what they by extension suggest concerning the potential of similar EC methods in general. Earlier we saw that uni-dimensional EC measures failed to distinguish consistently between levels of proficiency. However, once we start using a measure that is multi-dimensional and to take a broad range of errors into account, the picture changes We get a test that, in terms of conventional reliability and validity falls broadly within the range of currently available analytic and general impression tests of writing proficiency. The test, in other words, that the case for EC methods of writing proficiency is still very much an one.

A Drawback of EC Marking

One important drawback of the EC method, and especially one along CITO lines, is that it is more time-consuming than either analytic or general marking, and it is an open question as to whether future improvements of EC in this respect will serve to reduce the difference significantly. At present EC would appear to be inappropriate for large-scale testing where cost-efficiency imposes stringent limitations on the time allowed for marking. EC would seem to be much more suitable for contexts such as the Dutch one where teachers, being markers of their students' own compositions, need a method that comes up with consistent and valid scores and at the same time provides students with detailed information on how their scores have been determined.

Conclusion

In the course of this article, it has been argued that up to the present the case for EC methods of grading has been poorly made. Such measures as "errors per T-unit" or "error-free T-units per composition" are not only too narrow from the theoretical point of view, but they also fall far short of a convincing performance in practice. Unfortunately it has also been assumed that, by extension, EC in general has no merit.

This conclusion is premature. The CITO writing proficiency tests show that if errors are weighted and categorized so that many relevant aspects of compositions are included in their assessment, EC can operate with the same success as general impression or analytic methods of marking. In addition it seems likely that EC marking could be significantly improved by means of further research into methods of delineating error-categories even more clearly and by looking into ways of relating these categories to aspects of writing proficiency drawn from models of communicative competence.

Notes

1. In fact CITO calls the method analytic, even though it clearly belongs to the error-count category, that is, it requires markers to count errors and deduct the resulting, weighted, total from a basic score of 40.
2. The contribution of the bonus/malus system to score variance is very small: .4% (Melse & Verstralen, 1986, p. 111).
3. CITO uses a rather more complicated statistical method to calculate the correlations. In the main text I have reported Spearman's rank correlation figures — based on my own re-analysis of the data — as readers are more likely to be familiar with this method.
4. Strictly speaking one should say here that the hypothesis that there is no interdependence between the two sets of rankings has not been confirmed. As Woods, Fletcher and Hughes (1986) point out, it is usually "... difficult to interpret [Spearman's rank correlation] as a measure of 'degree of interdependence'..." (p. 174).
5. To calculate their test-retest reliability score, Jacobs et al. (1981) assessed the performance of two groups of students (size not reported) who took writing tests with two different tasks; the authors found a correlation coefficient of .72 between the mean scores for each task, a figure that they characterize as being "... in effect a test-retest reliability coefficient." (p. 73). Without specifying exactly how many readers took part in their experiment, Jacobs et al. (1981) explain that the "... ranges of reliability coefficients over subsamples of sets of readers who read at least 30 papers each were, for two readers, .59 to .96; for three readers, .89 to .94; and for four readers, .92 to .94" (p. 69). The figure cited in the main text is therefore an *average* interrater reliability score.

References

CITO (1984). *Schrijftoetsen Engels Havo/VWO.* Arnhem, Holland: CITO.

Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly 18*, 87-107.

Jacobs, H. L., Zinkgraf, S.A., Wormuth, D.R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach.* Rowley, Mass.: Newbury House.

Larsen-Freeman, D. (1983). Assessing global language proficiency. In H. W. Seliger and M. Long (Eds.), *Classroom oriented research in 2nd language acquisition* (pp. 286-305). Rowley, Mass.: Newbury House.

Melse, L. (1984). Toetsing en beoordeling van schrijfvaardigheid in de moderne vreemde talen. *Levende Talen*, 353-360.

Melse, L., & Verstralen, H. H. F. M., (1986). *De ontwikkeling van schrijftoetsen voor de moderne vreemde talen.* Specialistisch bulletin nr. 49. Arnhem, Holland: CITO.

Perkins, K. (1980). Using objective measures of attained writing proficiency to discriminate among holistic evaluations. *TESOL Quarterly 14*, 61-70.

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.

Woods, A., Fletcher, P., & Hughes, A. (1986). *Statistics in language studies.* Cambridge: Cambridge University Press.

Appendix A
Characterization of Error Types (based on CITO, 1984)

Incompleteness: All of the letters consist of a number of sub-assignments specifying, in Dutch, the contents of the letter. If parts of the sub-assignment have been omitted or carried out improperly, students lose points.

Absolutely Incomplete: Students lose points for not having carried out a sub-assignment at all.

Wordgroup Error: This category refers to wordgroups that contain a complex of errors such that it is difficult to establish exactly how many errors have been made. CITO (1984) gives the following example:
I wanted to be that you are writing about yourself. (p. 15)

Primary Errors: This category contains a considerable variety of errors, most of which are self-evident. Therefore only the less obvious ones will be discussed here.

Excessive repetition: In the example below, the writer is penalized for using the "it was nice" construction with excessive frequency:

It was nice to see you. It was also nice to see your parents and your little brother. And it was nice to be in England again. (CITO, 1984, p. 15)

Illogical connection: This is a discourse error and results from inadequate textual cohesion:

We like English t.v. programmes. Apart from that we often watch them.

Lack of clarity:

When you arrive at the station, I'll be waiting for you at the door. (CITO, 1984, p. 16)

Stations tend to have many doors — and exits, which is presumably what the original author meant — and by not specifying which one is being referred to, the writer is being insufficiently clear about where the meeting is to take place.

Use of word or expression inappropriate to context: These are errors of register, such as:

Further to our telephone call. . . . (in an informal letter)
All the best, greetings from. . . . (in a formal letter)
(CITO, 1984, p. 16)

Too sudden change of subject: When changes of topic are inadequately signposted, they are penalized. "V" indicates something is missing.

I have two pets, a cat and a dog, and I like them very much. They always sleep in my bedroom. V My grandfather always snores when he's sleeping. (CITO, 1984, p. 16)

Secondary Errors:

Spelling errors: self-evident.

Punctuation errors: self-evident.

Secondary preposition error: If a student has used an incorrect preposition, this normally means adding, changing, or removing a word in the correction process, and consequently one would have to call it a primary error. CITO, however, makes an exception here:

They haven't looked <u>for</u> the child very well.

John was sitting <u>at the back of</u> Mary.

CITO (1984) points out that in the letters from which these examples came, it was clear that students had intended to writer "after" and "behind" respectively. The first error, however, is counted as a primary one as "to look for" and "to look after" have completely different meanings. This affects comprehension much more than the second error, which does not have a meaning of its own that could confuse readers.

Appendix B
Letter Marked According to CITO Method

Brugstraat 2
Amsterdam
The Netherlands

January 25, 1987

Dear David,

How have you been since my last letter? I'm writing again
because I'm wondering if you have got my pictures 3 months ago.

I really want to know if you liked them. What did you think of
5 our house and my family? I am not very good at taking pictures, but
I hope you got an impression of what our place looks V.

Maybe it would be a good idea for me to pay a visit ~~at~~ the
O photography club in our neighbourhood to learn how to take better
pictures, that would also be nice for you.

10 Last week there has been took pictures of our class. Can you
see me standing in the middle row? The tall boy next to me is Daan
Jansen, my best friend.

I hope you will send me a letter soon. Could you please send
me some pictures too? I'm looking forward to hearing from you.

15 Yours sincerely,

Richard

kind of error	correction
line 3 grammatical error	Change to "got"
line 6 word needs to be added	Add "like"
line 7 secondary preposition	Change to "to"
line 9 punctuation	Change to full stop + capital letter or to semicolon.
line 10 wordgroup	It is difficult to see exactly how many corrections would be required here to get, for example, the more acceptable "some pictures were taken."
line 15 use of word or expression inappropriate to context	This being an informal letter, one has to change the expression used to, for example, "best wishes."
paragraph 3	This paragraph is incomplete because the sub-assignment here specified that the student should explain how he had heard of the photography club. The writer of the letter above gives no such explanation, and therefore the paragraph has to be treated as incomplete.

Point value of errors: –7-3/4 Final score: 32-1/4
(letter and some of the comments adapted from CITO, 1984, p. 27)