Holistic Scoring Based on Analytic Features: Can Holistic Scoring Be a Constant Winner?[1]

Ching Kang Liu[2]
College Entrance Examination Center

This paper introduces the unique approach of holistic scoring CEEC has been using for years. It combines the strength of analytic scoring but provides efficiency and practicality of holistic scoring for a large scale of test. This paper also reports the results of a study examining whether analytic scoring equals holistic scoring with the unique approach. Three hundred written products were evaluated by 12 experienced raters with these two methods respectively. The results show that analytic scores and holistic scores differed slightly. But no difference was found when the analytic scores were compared with the holistic scores elicited from the database at CEEC. This paper reassures that holistic scoring, with some modification, can still be an efficient and reliable evaluation tool.

Introduction

In assessment of a large number of writing, holistic evaluation has been considered reliable, practical, and economical (Davies et al, 1999). But the disadvantage of holistic scoring is that it is not clear which feature each rater focuses on, and therefore, scores might vary if the training is not solid enough (Yuji Nakamura, 2004). On the other hand, analytic evaluation has been considered more reliable and more beneficial to EFL students (Weigle, 2002; Table 1). For language instructors or language training programs, there seem to be more room for evaluators to adopt either approach according to the need at a certain time. However, for a national test center, a most desirable and least controversial approach is the only choice. After many years' exploration and modification, the College Entrance Exam Center (CEEC) has developed an approach that requires raters to "read holistically and adjust analytic scores to match holistic impressions" (Weigle, 2002). Since the evaluation at CEEC is a national task that requires high confidentiality and impartiality, the rationale is to employ an approach that combines the strength of both holistic scoring and analytic scoring. And this study is to examine whether our modified approach works as we have expected.

The research question for this study is very straightforward:

> Will the scores based on the unique holistic scoring differ from the scores based on five analytic features under the specific EFL environment in Taiwan?

Meanwhile, this study is also responsible for clarifying the doubts from many

[1] This paper is based on a project sponsored by the College Entrance Examination Center in Taiwan. I also here extend my warmest acknowledgement for Professor Zhaoming Gao, National Taiwan University, who has contributed a great deal of original ideas and technical supports to this project. I would also like to thank Miss Chunchi Yu and Miss Connie Lin for their assistance in collecting and analyzing the data.

[2] Ching Kang Liu is an associate professor of the department of foreign languages and applied linguistics, National Taipei University. E-mail: ckliu@mail.ntpu.edu.tw

parents or English instructors who have been assuming that holistic scoring could involve unfairness in the process of assessing the written products. They have conjectured that analytic scoring may bring in higher scores than the scores based on holistic scoring. Many also doubt that slight grammar errors may affect raters' judgment in determining the final holistic scores if grammar is not excluded in the process. Therefore, in addition to reporting the results of this study, the following sections will also include discussions on how raters are trained at CEEC, what analytic features are adopted, and whether the doubts mentioned above can be clarified.

Table 1. A comparison of holistic and analytic scales in terms of six qualities of test usefulness (Weigle, 2002, p.121)

Quality	**Holistic Scales**	**Analytic Scales**
Reliability	lower than analytic, but still acceptable	higher than holistic
Construct Validity	assume that all relevant aspects of writing ability develop at the same rate and can thus be captured in a single score; correlate with superficial aspects such as length and handwriting	more appropriate for L2 writers as different aspects of writing ability develop at different rates
Practicality	relatively fast and easy	time-consuming; expensive
Impact	single score may mask an uneven writing profile and may lead to misleading placements	more scales provide useful diagnostic information for placement and/or instruction; more useful for rater training
Authenticity	White(1995) argues that reading holistically is a more natural process than reading analyticly	Raters may read holistically and adjust analytic scores to match holistic impressions
Interactiveness	n/a	n/a

The rater-training process at CEEC:

In the past five years, the rater-training process has been stable and the inter-rater agreement has been consistently higher than 0.75 (Pearson *r*)[3]. The process has been consistent and well controlled. Below are the descriptions of how the raters are selected and trained, what rubrics for each analytic feature have been adopted, and how the holistic scoring has been merged with the analytic scoring.

For each evaluation task, two directors are recommended[4] after the registration process is completed. The directors need to determine how many evaluators are needed on the basis of the number of students who register for the exam. Since the number of high school graduates are quite stable, the number of the raters has remained stable, too. Once the number of the raters is determined, 12 to 14 group leaders will be recommended by the directors and each leader is responsible for the selection of qualified and stable raters (also university faculty members) for his/her specific group. The group leader is also responsible for the pace and any factors that might intervene with the quality of evaluation. Below are the specific steps of how the

[3] Data provided by the College Entrance Examination Center.
[4] Directors are university faculty members with rich experience and high credit in the evaluation task.

whole process is gone through after the exam is taken and before the written products are evaluated:

Step 1: The raters (directors and the group leaders) refresh the rubrics (Table 2) of each feature that have been used for years and make sure that the basic principles are still appropriate for this specific task.

Step 2: To verify the rubrics, the directors and the group leaders randomly select 1,500 copies of the written products produced by the examinees.

Step 3: Each rater tries to read about 100 copies of the selected written products and choose 10 pieces of written product to represent each of the different levels (with a tentative holistic score for each piece).

Table 2: The rubrics and the score range of different analytic features

Features	Scores	Rubrics
Content	5-4	**Excellent to very good:** well-stated thesis related to the assigned topic with relevant, substantive, and detailed supports
	3	**Good to average:** limitedly-developed or vague thesis with irrelevant statements
	2-1	**Fair to poor:** poorly-developed or obscured thesis; too much repetition of limited relevant sentences
	0	**Very poor:** not pertinent; or no written products (if this stands, all the other features are counted as "0")
Organization	5-4	**Excellent to very good:** well-organized structure with beginning, development, and ending; effective transition with logical sequencing and coherence
	3	**Good to average:** loosely-organized structure with imbalanced beginning, development, and ending; less effective transition that obvious affects logical sequencing and coherence
	2-1	**Fair to poor:** choppy ideas scattering without logical sequencing and coherence
	0	**Very poor:** no organization, no sequencing and coherence; or not pertinent
Grammar & rhetoric	4	**Excellent to very good:** well-structured sentences with variety; appropriate rhetoric; few grammatical errors
	3	**Good to average:** less well-structured sentence with some errors of tense, agreement, etc.; but meaning seldom obscured
	2-1	**Fair to poor:** major errors of conjunctions, fragments, or ill-structured sentences that make meaning confused or obscured
	0	**Very poor:** being dominated by errors that blocks communication
Vocabulary	4	**Excellent to very good:** specific and effective wording; idiomatic and no spelling error
	3	**Good to average:** dull and repeated wording; occasional errors of word/idiom form, choice, usage but meaning not obscured
	2-1	**Fair to poor:** inappropriate wording; frequent spelling errors; meaning confused or obscured
	0	**Very poor:** some relevant words found, but meaning incomprehensible
Mechanics	2	**Excellent to very good:** no errors of format, punctuation, or capitalization
	1	**Fair to poor:** limited errors of format, punctuation, or capitalization, but meaning not obscured
	0	**Very poor:** too many errors of format, punctuation, or capitalization; violating basic conventions of writing

Step 4: Then the directors and the group leaders sit together and discuss on samples for each level with the chosen written products and determine whether the analytic rubrics fit in the specific topic assigned for this exam, and whether the raters have the agreeable criteria.

Step 5: After the discussion, 20 most agreeable samples will be chosen with 2-3 samples reflecting each level of the holistic scores. Among the 20 samples, 10 will serve as the model papers for the training of the raters in each group and the other 10 will serve as "experimental" papers for the raters to "try" and "readjust themselves to" the new rating job[5].

Step 6: Each group of raters will spend about one hour getting familiar with the

[5] "Readjust" is used here because these raters are mostly experienced and with good credit, based on the records at CEEC.

rubrics, the assigned topic, and the 10 model samples. As soon as they are ready, they will have to evaluate another 10 samples analytically according to the rubrics in Table 2 and the model samples they have just read. When they finish, their analytic scores will be compared with the scores rated by the directors and the group leaders. If the scores they produced are not agreeable to the scores rated by the directors and the group leaders, reasons should be given. By the end of the first hour, all the comments or questions will be collected by each group leader and the directors and the group leaders will have another meeting, trying to understand the comments and questions and work out with good solutions.

Step 7: Then the raters in each group will evaluate the first 50 written products (bound as a book) as an experimental package and give only one holistic score (Table 3) for each written product. The group leader will monitor the first 50 evaluation of each rater to make sure that the raters in that group are ready to score holistically on the basis of analytic features, and then to continue and complete the evaluation task in a valid and reliable way.

The rubrics in Table 2 look similar to Jacobs et al.'s (1981) scoring profile (see appendix 1). The scoring system is slightly different. In Table 2, we can see that the total scores for each feature are different: it is 5 for content, 5 for organization, 4 for grammar and rhetoric, 4 for vocabulary, and 2 for mechanics. In each feature, rubrics for different levels of quality are also slightly different from Jacobs et al.'s profile. For instance, in Jacobs' profile, "knowledge" is a concern for the feature of content while "detailed illustrations" are important in Table 2. As for the scoring, in Jacobs' profile, minor grammatical errors like agreement, tense, article, number, etc. may cost nothing more than 3 out of 25 points while in table 2, the same type of error will cost 1 out of 4 point. In other words, in our system, grammar is treated with more weight than in Jacobs et al.'s rubrics. Besides, while the scores on the feature of mechanics in Jacob's table sum up to 5, which is only one sixth of the total scores of the content, they sum up to 2, which is almost a half of the scores of the content. In a word, people who developed the first version of the rubrics for CEEC seemed to pay much attention to the feature of grammar.

After the raters are familiar with the scoring based on the analytic features, they will have to give each written product a holistic score as shown in Table 3, where each level of scores reflects the total of the scores added up from the five features listed in Table 2. The raters are expected to "read holistically and adjust analytic scores to match holistic impressions" (Weigle, 2002) at a very limited time. If the raters have difficulties determining the holistic scores, they can always go back to the rubrics for each analytic feature. This back-and-forth process will soon familiarize all the raters with the rubrics and hence will enable them to assess the written products holistically.

Table 3: The levels of the holistic scores

Scores	Quality of the written products
19-20	excellent
15-18	excellent-very good
10-14	good-average
5-9	fair-poor
0-4	very poor

Raters for this study

Twelve raters (5 males and 7 females) were selected from the name list of the raters with at least 5 years of experience working for the CEEC in evaluating written products. They were asked to go through "step 6" mentioned above and were well prepared for the rating task before the experiment.

In order to avoid the ordering effect, we used a balanced incomplete block design (BIBD) (See Table 4) in which the raters were divided into two groups with one group (3 males and 3 females) doing the analytic evaluation first and then the holistic scoring; and the other group (4 females and 2 males), doing holistic scoring first and then analytic scoring.

Materials

Three hundred copies of written products randomly selected from different districts in Taiwan were used for this study. They were divided into six books, each containing 50 copies. In the balanced incomplete block design, each book was rated four times by four different raters: two based on holistic scoring with only one holistic score; the other two, on analytic features with 5 separate scores (Table 2). We planned to imitate the real situation of evaluation where there was no chance that the same composition would be evaluated by the same person twice.

Table 4: The arrangement of the materials and the raters

Group 1			Group 2		
	Period 1	Period 2		Period 1	Period 2
Evaluation / Raters	Holistic	Analytic	Evaluation / Raters	Analytic	Holistic
1	A	B	7	A	B
2	B	A	8	B	A
3	C	D	9	C	D
4	D	C	10	D	C
5	E	F	11	E	F
6	F	E	12	F	E

Raters: 1-12; materials to be rated: books A to F

The results

The correlations

First, we looked at the correlations between the holistic scores and analytic scores rated by the 12 raters. Table 5 shows three sets of inter-rater correlations: between the two sets of holistic scores, two sets of analytic scores, and two sets of holistic scores retrieved from the data at CEEC. The correlation (r = .811) between the two sets of holistic scores in this study is obviously higher than the correlation between the two sets of scores retrieved from CEEC (r = .71). It is also higher than the correlation between the two sets of analytic scores (r = .754).

Table 5: Pearson correlation between different groups of raters

Groups of scores	Correlations	N
holistic (1)-holistic (2) in this study	.811**	299
holistic (1)-holistic (2) from CEEC	.710**	300
analytic (1)-analytic (2) in this study	.754**	299

**Correlation is significant at the 0.01 level (2-tailed).

When looking at the correlations between the total scores and the scores of the

five analytic features, Table 6 shows the two groups of raters had the same perspective in assessment: the lowest correlation fall on the feature of mechanics. It implies that this feature is over-weighed so that it is easy for raters to either overrate or underrate this feature while assessing this feature. Other than this, almost very feature reflects surprisingly high correlations (with *r* from .893 to .938), indicating each feature can serve as a significant predictor to the total score.

Two more things are worth mentioning here. First, the correlation (r = .814) between the holistic scores and the analytic scores is slightly higher than the correlation between the two sets of holistic scores (r = .811), the correlation between the two sets of analytic scores (r = .754), and the correlation between the two rated scores elicited from the data at CEEC (r = .710). This implies that the 12 raters in this study agreed quite well in both holistic scoring and analytic scoring. Second, Table 7 shows that when we looked at the average scores of each groups, we noticed that the more raters involved in the rating, the more agreement could be expected because the correlations coefficients of the mean scores are higher than those listed in Table 5.

Table 6: Pearson correlation between the total score and the score of each feature

Groups		Content	Organization	Grammar	Vocabulary	Mechanics
Group 1	Total	.932**	.893**	.938**	.927**	.825**
	N	300	300	300	300	300
Group 2	Total	.964**	.971**	.943**	.951**	.852**
	N	299	299	299	299	299

**Correlation is significant at the 0.01 level (2-tailed).

Table 7: Pearson correlation between the mean scores

Means of scores	Correlations	N
holistic-analytic (in this study)	.908**	298
holistic (this study)-holistic (CEEC)	.845**	299
analytic (this study)-holistic (CEEC)	.861**	299

**Correlation is significant at the 0.01 level (2-tailed).

The reliability

Since the analytic features are in fact multiple-item additive scales, the reliability analysis might help us determine how accurate, on the average, the estimate of the true score is in the written products we measured. The reliability coefficient (α) of the five features was calculated and the results (Table 8) show that all reliability coefficients of the six books (300 copies of written products) are high enough to support that the scores of these features can truly reflects the quality of writing.

Table 8: The reliability coefficients of the five features of the analytic scoring

	All	Book1	Book2	Book3	Book4	Book5	Book6
Reliability coefficient (α)	0.9414	0.9487	0.9555	0.9482	0.9546	0.9518	0.8979
N	599	100	100	100	100	99	100

Is holistic scoring differs from analytic scoring?

The idealized result for this study is that the holistic scores are statistically equal to the analytic scores, meaning that the two groups of scores are not significantly different. Since the final scores of the written products at CEEC are presented as the mean of the scores from two evaluators, the comparison here were also based on the data of the averaged scores produced by two raters. First, we compared the two

averaged holistic scores and analytic scores collected in this study. The results were not exciting but quite encouraging: the difference was barely significant ($t = 2.007$, $p = 0.045$). If we take a strict criterion (e.g., $p < 0.01$), we can say that the holistic scores and the analytic scores do not differ. On the other hand, when we compared the analytic scores collected in this study and the holistic scores elicited from the data bank at CEEC, we noticed that these two sets of scores do not differ ($t = 1.413$, $p > .05/ = 0.158$).

Table 9: Independent Samples *t* Test on the averaged holistic scores and the analytic scores for each book

	t	*df*	Sig. (2-tailed)	Mean difference	Std. Error Difference
Book 1	-.408	98	.684	-.3450	.8454
Book 2	.841	98	.402	.7400	.8799
Book 3	.542	98	.589	.3900	.7196
Book 4	1.657	98	.101	1.4500	.8753
Book 5	2.636**	97	.010	2.1433	.8130
Book 6	-.339	97	.735	-.2749	.8104

**Significance at the 0.01 level (2-tailed).

When we did the separate *t* test based on the six books (each containing 50 copies of written samples), we noticed that only one of the six books showed significant difference ($t = 2.636$, $p = 0.01$, $df = 97$) in comparing the averaged holistic scores and the averaged analytic scores (Table 9). When we checked the descriptive data, we noticed that there was one rater who tended to give much higher scores (2 points more than the average scores). In reality, the committee in charge of the evaluation could simply "eliminate" this type of raters who provide extremely high or low scores in average to maintain the quality and reliability of the rating task (however, the high correlation coefficients also show that these raters are also good raters. They might have some personal preferences in giving too high or too low a score in assessing a certain feature of the written products).

Conclusion

Though this is only a small study with 12 raters and 300 written products, several impressive conclusions still can be drawn here.

1. For a large scale of test as the task at CEEC, holistic assessment is still a more practical and economical choice. Data showed that the 12 raters spent, in average, 78 minutes in assessing the 50 written samples analytically while they spent only 54 minutes assessing the same number of copies holistically.
2. To avoid uncontrollable idiosyncratic deviation of ratings, a training process on the basis of analytic features, as suggested the previous sections, can be adopted.
3. This study suggests that the more ratings each written product receives, the higher the agreement among the scores can be expected (compare the correlations in Table 5 and Table 7).
4. The high reliability of the analytic features suggests that analytic scoring is undoubtedly a most desirable type of assessment for written products. However, since there is no strong significant difference found between the holistic scoring and the analytic scoring, it would not be necessary to spend more money and efforts in assessing a large scale of written products with analytic scoring.
5. When we asked the 12 raters whether they would think of holistic scores when they were doing the analytic scoring, most of them (9 out of 12) said that they had a holistic impression on the written products before they gave scores for each

analytic feature. Since the cognitive process of assessment is too complicated for the raters to figure whether the holistic impression determines the score or the analytic scoring contribute more to the total score, this issue will be remained unanswered here.

6. There is one thing for sure that we can tell the parents and the English instructors in Taiwan that the scores based on analytic features are not higher than the holistic scores. And the holistic assessment turns out to be one of the best choices in assessing a large scale of written products.
7. Since the correlation among the 12 raters regarding the feature of mechanics is lower than other features (Table 6), A new rubric for this category requires modification. The College Entrance Examination Center might want to consider lower the weight of this category to see if the correlation could rise in other related studies to come.

Limitations

We know that if we had wanted to make this study more reliable and more valid, we should have asked the raters to evaluate the same work, at different periods of time with at least an interval of one month, with each of the two approaches. And each rater should evaluate hundreds of writing samples so that we can have more information regarding the discussed issue. However, reality did not allow us to do this because most raters were terribly busy and could not have such spare time for the study. On the other hand, there is not much evidence showing that, for experienced raters, rating more copies will make the rater's judgment or rating quality more stable. On the contrary, fatigue effect might arise and new issues will be threats to the study. With our current design, it may still be a good way to provide us enough information to justify the choice of holistic scoring for the assessment of writing.

References

Davies, A., Brown, A., et al. (Eds.). (1999). *Dictionary of language testing.* Cambridge: Cambridge University Press.

Hughes, A. (2003). *Testing for language teachers.* Second Edition. Cambridge: Cambridge University Press.

Jacobs. H. L., Zingraf, S. A., Wormuth, D. R., Hartfield, V. F., and Hughey, J. B. (1981). *Testing ESL Compositions: a practical approach*. Rowley, Mass: Newbury House.

Nakamura, Y. (2004). A Comparison of Holistic and Analytic Scoring Methods in the Assessment of Wrting. JALT Pan-SIG Proceedings. Downloaded from http://www.jalt.org/pansig/2004/HTML/Nakamura.htm

Weigle, S. C. (2002). *Assessing writing.* Cambridge: Cambridge University Press.

Appendix 1: Jacobs et al.'s (1981) scoring profile (cited from Hughes, 2002, p.104)

ESL COMPOSITION PROFILE

STUDENT DATE TOPIC

SCORE	LEVEL	CRITERIA	COMMENTS
CONTENT	30-27	EXCELLENT TO VERY GOOD: knowledgeable • substantive • thorough development of thesis • relevant to assigned topic	
	26-22	GOOD TO AVERAGE: some knowledge of subject • adequate range • limited development of thesis • mostly relevant to topic, but lacks detail	
	21-17	FAIR TO POOR: limited knowledge of subject • little substance • inadequate development of topic	
	16-13	VERY POOR: does not show knowledge of subject • non-substantive • not pertinent • OR not enough to evaluate	
ORGANIZATION	20-18	EXCELLENT TO VERY GOOD: fluent expression • ideas clearly stated/supported • succinct • well-organized • logical sequencing • cohesive	
	17-14	GOOD TO AVERAGE: somewhat choppy • loosely organized but main ideas stand out • limited support • logical but incomplete sequencing	
	13-10	FAIR TO POOR: non-fluent • ideas confused or disconnected • lacks logical sequencing and development	
	9-7	VERY POOR: does not communicate • no organization • OR not enough to evaluate	
VOCABULARY	20-18	EXCELLENT TO VERY GOOD: sophisticated range • effective word/idiom choice and usage • word form mastery • appropriate register	
	17-14	GOOD TO AVERAGE: adequate range • occasional errors of word/idiom form, choice, usage *but meaning not obscured*	
	13-10	FAIR TO POOR: limited range • frequent errors of word/idiom form, choice, usage • *meaning confused or obscured*	
	9-7	VERY POOR: essentially translation • little knowledge of English vocabulary, idioms, word form • OR not enough to evaluate	
LANGUAGE USE	25-22	EXCELLENT TO VERY GOOD: effective complex constructions • few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions	
	21-18	GOOD TO AVERAGE: effective but simple constructions • minor problems in complex constructions • several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions *but meaning seldom obscured*	
	17-11	FAIR TO POOR: major problems in simple/complex constructions • frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions • *meaning confused or obscured*	
	10-5	VERY POOR: virtually no mastery of sentence construction rules • dominated by errors • does not communicate • OR not enough to evaluate	
MECHANICS	5	EXCELLENT TO VERY GOOD: demonstrates mastery of conventions • few errors of spelling, punctuation, capitalization, paragraphing	
	4	GOOD TO AVERAGE: occasional errors of spelling, punctuation, capitalization, paragraphing *but meaning not obscured*	
	3	FAIR TO POOR: frequent errors of spelling, punctuation, capitalization, paragraphing • poor handwriting • *meaning confused or obscured*	
	2	VERY POOR: no mastery of conventions • dominated by errors of spelling, punctuation, capitalization, paragraphing • handwriting illegible • OR not enough to evaluate	

TOTAL SCORE READER COMMENTS