DOCUMENT RESUME

ED 305 647 CS 211 759

AUTHOR Froese, Victor
TITLE Aspects of Holistic Scoring Validity.
PUB DATE Apr 89
NOTE 18p.; Paper presented at the Annual Meeting of the National Testing Network on Writing Assessment (7th, Montreal, Canada, April 9-11, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Secondary Education; Foreign Countries; Grade 3; Grade 6; Grade 9; *Holistic Evaluation; Statistical Analysis; *Validity; *Writing Evaluation
IDENTIFIERS Canada; General Linear Model; *Writers Workbench (Computer Software)

ABSTRACT

A study investigated two questions regarding the validity of holistic writing evaluation: (1) How well do holistic scores predict sentence length, passage length, and spelling errors in grades 3, 6, and 9 narrative and explanatory writing? and (2) Which model best predicts these variables (across grades within type, or within grade within type)? Forty randomly selected narrative and 40 explanatory passages originally scored holistically as part of a Canadian provincial writing assessment program were reanalyzed with the Writer's Workbench software. Findings indicated that holistic scoring is apparently sensitive to sometimes irrelevant factors such as composition length, sentence length, and spelling errors, but that these operate differentially at different grade levels and for different types of writing. (Two tables of data are included and 13 references are attached.) (SR)

## ABSTRACT

Title: Aspects of Holistic Scoring Validity
Author: Dr. V. Froese, University of British Columbia

There is insufficient research into whether those trained in holistic scoring base their judgements on substantive rather than on superficial characteristics of writing. In this study holistic scores were compared with composition length, sentence length, and spelling errors. For each of grades 3, 6, and 9 forty randomly selected narrative and forty explanatory passages originally scored holistically as part of a provincial writing assessment program were reanalysed with the Writer's Workbench software.

Statistical analysis were performed using GLM procedures. It was concluded that holistic scoring is apparently sensitive to sometimes irrelevant factors such as composition length, sentence length, and spelling errors, but that these operate somewhat differentially at different grade levels and for different types of writing.

ASPECTS OF HOLISTIC SCORING VALIDITY

Paper presented at the Seventh Annual Conference on Writing Assessment, Montreal, April 9-11, 1989

INTRODUCTION

Three related observation prompted this study. First, as noted by Charney (1984) "insufficient research has been done into the question of whether readers trained in holistic rating base their judgements on substantive criteria or on superficial characteristics of the writing sample". Second, since qualitative scores correlate with quantitative scores, can the scores from computer analyzed compositions adequately predict those based on holistic scores? And third, a comparison of machine marked essays with human-scored essays could say something about the validity of such marking. Since previous work (Stewart & Grobe, 1979) has indicated that teacher-markers are more influenced by such quantitative measures as length of composition and freedom from mechanical errors (spelling, etc.) rather than by the control of syntactic resources of their language, it seemed probable that text analysis programs such as the Writers' Workbench could predict holistic scores.

The major questions asked in this study are: 1)How well do holistic scores predict sentence length, passage length, and spelling errors (or a combination of these variables) in Grade 3, 6, and 9 narrative and explanatory writing? 2)Which model best predicts these variables (across grades within type, or within grade within type)? Subsidiary questions were: 3)Are there differences between

types of writing within grades? 4)Are there differences between grades within types of writing?

## BACKGROUND

### Holistic Scoring

Hillocks (1986) in his comprehensive review on the literature on written composition concluded that holistic scoring appears to be a valid and reliable measure when adequate safeguards are taken. However, in a critical review of the validity of holistic scoring Charney (1984) states that insufficient research has been done into the question as to whether markers base their judgments on substantive or superficial characteristics of the writing sample and as to whether writing samples representing different aims of discourse should be compared. He concludes that "Holistic ratings should not be ruled out as a method of evaluating writing ability, but those who use such ratings must seriously consider the questions of the validity of scores that result." These are, of course, some of the concerns of interest to the present study since holistic scoring is used so widely today.

### Writer' Workbench Research

The current decade has seen a heightened interest in the impact of computers on the writing process not only because of the ubiquitous

word processor, but also because text analyzers are gradually creeping into the market place for microcomputers and minicomputers. The Writers' Workbench is one of these. It was developed originally by Bell Laboratories and its Collegiate Version is now marketed by AT&T for UNIX operating systems. While it can provide feedback on a variety of concepts related to reading and writing--style, diction, organization, vagueness, spelling, usage (split infinitives, nominalizations, etc.), a variety of counts (words, sentences, sentence types, word usage, etc.), and readability measures--in this study only three a counts are used. These are sentence length, passage length, and number of spelling errors.

In the development of the Writers' Workbench (WWB) attention was given to the reliability and validity of the program in relationship to expert opinion on which aspects of writing to include. Additionally, computer and human judgments were compared. There was a 90% agreement on parts of speech and 70% agreement on judgments of abstractness (Frase & Diel, 1986). Several studies have also investigated whether use of the WWB results in improved writing. Sterkel, Johnson & Sjogren (1986) found mixed results when comparing regular grading with WWB feedback, but found gains for the WWB group in the third semester of a writing course in a business communication class. Another study by Needles (1988) found that the writing of business letters was not improved by WWB feedback over the instructor's marking of errors. Kiefer and

Smith's study (1983) investigated whether use of the WWB would improve editing skills over a control group which spent the same amount of time on its writing. The experimental groups identified more errors in editing but based on holistic scoring no differences between the groups was found. These authors concluded that holistic scoring did not lend itself to measuring changes in the more specific aspects of writing about which the DICTION, STYLE, and SUGGEST sections provides information. However, students attitude toward the WWB feedback was positive.

A more recent study by Carlson (1989) investigating the relationships of reasoning and writing skills to GRE Analytic Ability Scores also investigated WWB analyses and found that they provide " information about the characteristics of text as well as measures that identify features of written discourse that are relatively independent." This was based on her finding that a factor analysis isolated three factors reflecting WWB fluency, content, and sentence variety.

In summary, even though improvements in writing have not been convincingly demonstrated through use of the WWB, its analysis agrees reasonably with human ratings and certain aspects--fluency, content, and sentence variety--are relatively independent factors related to writing. One might conclude that evidence for assessment and evaluation is stronger than for teaching.

## Sentence Length, Passage Length & Spelling

Charney (1984) pointed out in his review on the research behind holistic scoring that such ratings should not be "unduly influenced by superficial features if they are to be considered valid." Three aspects of writing--sentence length, passage length, and spelling errors--have been documented as being such "superficial" features. Hillocks (1986) review of research on written composition discussed these aspects in several ways. Average sentence length was found to be significantly different among grades 4, 8, and 12 by Hunt (1965) as was average length of T-unit (main clause with all its appended modifiers, including subordinate clauses). Composition length has been found to be related to holistic scores (Stewart and Grobe, 1979) as well, but Hillocks nevertheless contended that "length could be a productive area for research as an indication of the elaboration of structure" as did Langer (1986). Finally, it is commonly observed (Hillocks, 1986,28) that "weaker writers have a tendency to be preoccupied with mechanics, particulary spelling." Since all three measures may be accurately counted by the WWB, it was decided to use them as independent variables.

## Manitoba Writing Assessment Program

The passages used in this study were a small sample of provincial writing assessment program conducted in May, 1982 (Manitoba Writing Assessment Program, 1982). The writing tests were designed to

describe student writing performance at grade levels 3,6,9, and 12. The tests were administered to a sample of 8-9 percent of the students at each grade level (i.e. 1333 at Grade 3, 1430 at Gr. 6, 1197 at grade 9, and 9980 at Gr. 12). Their were four components--a dictated spelling list, sentence combining, a composition of a story, and a composition of description or explanation. For purposes of this study only the compositions were used and since Gr. 12 did only the explanatory essay, some analysis could only be performed for the remaining grades. Some 80 teachers were trained in grade level groups to score compositions according to a set procedure. The writing samples were evaluated in a number of ways--General Impression Marking, an Analytic Scale, an Attribute Scale, and by a Descriptive Scale. For purposes of this study only the General Impression Marking results were used. Each writing sample was read by two raters who assigned it a score of 1 - 6 and if the score did not differ more than one point the scores were added together. If the difference was greater the paper was read by an arbitrator and the score consisted of an average of the first two scores plus the arbitrator's score. These procedures were patterned after those described in Cooper & Odell (1977) and allowed a score of 2 - 12 to be achieved by any individual paper. A randomly drawn set of 40 composition of each type--narrative and descriptive--were obtained for this study. A test of means indicated that the this sample was not significantly different from the original larger sample.

DESCRIPTION OF STUDY

As indicated earlier, the purpose of this study was to compare holistically scored composition scores with commonly associated measures--sentence length, passage length, and spelling errors (within the compositions)--identified by a text analysis program, the Writers' Workbench. For each of grades 3,6,9 forty randomly selected narrative and forty randomly selected explanatory passages formed the basis of this study. A graduate research assistant experienced in using the Writers' Workbench entered the same compositions for computer analysis being careful to encode spellings and punctuation exactly as in the originals. The WWB analysis were run and coded but only the three measures mentioned above were used for this study. Statistical analysis were performed using General Linear Model (GLM) procedures and the results are presented and discussed below.

FINDINGS

Each of the questions previously posed will be discussed separately below.

1. Which model best predicts sentence length (SL), passage length (PL), or spelling errors (SE) from holistic scores?

Model 2 (grade within type) when compare with Model 1 (grades

pooled)turns out to be the better model ($F(8,228)=1.94^*$) for sentence length, ($F(8,228)=16.18^*$) for passage length, but not for spelling errors $F(8,228)=1.33$. Model 2 is also significantly better for predicting a combination of SL, PL, and SE ($F=14.28^*$). If Model 1 is compared with Model 0 (slope=0) for spelling errros, the resulting $F(2,236)=1.74$ and is not significant at the .05 level.

2. How well do holistic scores predict sentence length, passage length, and spelling errors (or combinations thereof)?

Using Model 2 for predicting results in sentence length, significant differences for Gr.3 Narrative and Gr. 3 Explanatory passages were found, but no significant differences were found at any of the other grade levels (See Table 1).

---

Insert Table 1 About Here

---

Again, using Model 2, to consider passage length resulted in significant differences at all three grade levels for Narrative, and for Grades 3 and 9 for the Explanatory passages.

For spelling errors no significant results were found at any grade level within writing type.

When SL, PL, and SE were considered together, significant slopes were found for Grades 6 and 9 for Narrative but only for Gr. 9 in Expository writing.

3. Are there differences between types of writing (narrative or explanatory) within grade levels?

Because grade x type interactions were found, for Holistic measures narrative scores exceed explanatory scores at the Gr.3 level but explanatory scores were approximately equal to narrative scores at the Gr. 6 and 9 levels.

For Sentence Length again interactions were found and hence explanatory scores exceeded narrative scores for Gr.3 and 9 but they were approximately equal for Gr. 6.

Passage Length measures were also accompanied by grade x type interactions and hence narrative scores exceeded expository scores for all three grades (and differences increased with grade).

For Spelling Errors there were no significant interactions and hence narrative scores approximately equalled explanatory scores (plus a constant of 1.54 (0.36)) across all three grades.

4. Are there differences between grades within type of writing (narrative and explanatory)?

For narrative writing significant differences were found only for passage length ($F(2,117)=28.52^{**}$). There were no significant differences among grades for holistic scores ($F(2,117)=2.46$), for sentence length ($F(2,117)=1.60$), or for spelling errors ($F(2,117)=1.64$). Table 2 contains the means as well as F-ratios.

Insert Table 2 About Here

For explanatory writing in addition to Gr. 3,6, and 9, Gr. 12 scores were available. While there were no significant differences for the holistic measure ($F(3,156)=0.46$), differences in the remaining measures were all significant. For sentence length $F(3,156)=2.74^{*}$, for passage length $F(3,156)=226.22^{**}$, and for spelling errors $F(3,156)=5.15^{**}$.

CONCLUSIONS

The general purpose of this study was to consider the validity of holistic scoring when compared to the distracters--sentence length, passage length, and number of spelling errors in narrative and explanatory text. It was concluded from the findings above that indeed holistic scores could predict sentence length and passage length, with the qualifications that this could be done at certain

grade levels only since grade x type interactions occurred in all but the spelling category. Combining SL, PS, and SE did not result in any meaningful improvements in prediction.

A second aim was to explore whether machine scoreable measures such as sentence length, passage length, and spelling errors could be predicted from human-scored holistic marks. This worked best for passage length for both types of writing--narrative and explanatory (except for Gr. 6 expository). Holistic scores at the Gr.3 level also predict sentence length for both types of writing.

A third purpose was to establish whether types of writing within grade were different. Because of grade x type interactions the answer varied depending on which measure was considered. For holistic scores narrative scores exceeded expository at the Gr. 3 level only. For the sentence length measure, Gr.3 and 9 expository writing exceeded narrative but for Gr. 6 they are about the same. For passage length narrative exceeded expository at all levels. And for spelling errors there were no statistically significant differences. One might expect "narrative" type of writing to be superior initially since that is the type of writing frequently used but why the different measures result in different patterns is unclear.

Examinations of the means shed light on the fourth question. It was interesting but expected that there were no significant

differences between holistic scores among grades for narrative and expository texts. It was somewhat surprising, however, to find no difference in sentence length across the grades for narrative writing. At every grade level expository writing exceeded narrative in sentence length, and differences among grades 3 - 12 were also significant. Passage length was the best differentiating measure and it appeared that expository text generally was shorter (about 1/2 the length of narrative) than narrative, but it contained longer sentences. More elaboration and qualification is used in explanatory writing but why it also resulted in shorter writing is left unexplained.

## DISCUSSION

Clearly this study cannot resolve the issue as to whether machine-scored attributes of writing can substitute for human-based judgements. However, prediction is statistically possible, but it is a matter of degree since only a reasonably small amount of the variance (about 20%) is explained by these three measures. Prediction is also not a simple matter since statistical significance is not necessarily practical significance. From these data it is clear that one cannot simply talk about narrative or expository writing; one must put the type of writing within the context of a grade (or age) since interactions occurred.

Since "passage length" is related to quality ratings, it needs to

be explored further. Trained markers probably do not react to simple length but a number of other variables which relate to it. We need to find out what these contributing factors are.

Because of its reliabilty technology can be used to supplement human judgement and perhaps that is the best use of it since it has the possibility of allowing time for more high-level interaction between student and teacher.

REFERENCES

Carlson, S.B. (1988). Relationship of reasoning and writing skills to GRE analytical ability scores. GRE No. 84-23 (45 pp.+Appendices).

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. Research in the Teaching of English, 18, 1, 65-81.

Cooper, C.R. and Odell, L. (Eds.).(1977). Evaluating Writing. Urbana, Illinois: NCTE.

Frase, L.T. and Diel, M. (1986). UNIX Writers' Workbench: Software for streamlining communication. Technical Horizons in Education Journal, 14,3, 74-78.

Hillocks, G. (1986). Research on Written Composition:New Directions for Teaching. Urbana, Illinois.

Hunt, K.W. (1977). Early blooming and late blooming syntactic structures. In Cooper, C.R. & Odell, L. (Eds.) Evaluating Writing, 91-104. Urbana, Illinois:NCTE.

Kiefer, K.E. and Smith, C.R. (1983). Textual analysis with computers: Tests of Bell Laboratories' computer software. Research in the Teaching of English, 17, 3, 201-214.

Langer, J.A. (1986). Children Reading and Writing· Structures and Strategies. Norwood, N.J.:Ablex.

Manitoba Writing Assessment Program 1982. A report of the Curriculum Development and Implementation Branch, Department of Education, Province of Manitoba, Winnipeg, Manitoba.

Needles, M. (1988). Instructor-graded versus computer-graded business letters. Journal of Education for Business, 63, 269-272.

Sterkel, K.S., Johnson, M.I. and Sjogren, D.D. (1986). Textual analysis with computers to improve the writing skills of business communication students. The Journal of Business Communication, 23, 1, 43-61.

Stewart,M.F. and Grobe, C.H. (1979). Syntactic maturity, mechanics of writing, and teachers' quality ratings. Research in the Teaching of English, 13, 3, 207-215.

Writer's Workbench, UNIX Collegiate Edition (Running on UNIX System V, Release 3.1), AT & T, 1984.

TABLE 1
Grade 3, 6, and 9 intercepts and slopes for sentence length, passage length, spelling errors, and SL+PL+SE

| MODEL | INTERCEPT | SLOPE |
|---|---|---|
| Sentence Length - Narrative | | |
| Gr. 3 | 30.32 | -1.92* |
| Gr. 6 | 21.24 | -0.76 |
| Gr. 9 | 13.13 | -0.12 |
| Sentence Length - Explanatory | | |
| Gr. 3 | 34.89 | -1.79* |
| Gr. 6 | 23.56 | -0.97 |
| Gr. 9 | 17.62 | -0.26 |
| Passage Length - Narrative | | |
| Gr. 3 | 74.29 | 12.72* |
| Gr. 6 | 131.60 | 23.17* |
| Gr. 9 | 211.20 | 18.09* |
| Passage Length - Explanatory | | |
| Gr. 3 | 12.97 | 11.27* |
| Gr. 6 | 141.10 | 4.62 |
| Gr. 9 | 42.57 | 16.71* |
| Spelling Errors - Narrative | | |
| Gr. 3 | 6.42 | -0.22 |
| Gr. 6 | 7.47 | -0.30 |
| Gr. 9 | 8.17 | -0.67 |
| Spelling Errors - Explanatory | | |
| Gr. 3 | 3.59 | -0.12 |
| Gr. 6 | 5.55 | -0.28 |
| Gr. 9 | 1.73 | 0.62 |
| SL + PL + SE - Narrative | | |
| Gr. 3 | 111.0 | 10.58 |
| Gr. 6 | 160.3 | 22.11* |
| Gr. 9 | 232.5 | 17.30* |
| SL + PL + SE - Explanatory | | |
| Gr. 3 | 51.46 | 9.37 |
| Gr. 6 | 170.20 | 3.37 |
| Gr. 9 | 61.91 | 16.51* |

* $p<.05$

TABLE 2
Means and F-scores for Narrative and Explanatory Text for Grades 3, 6, 9, and 12

| | | NARRATIVE | EXPLANATORY |
|---|---|---|---|
| Holistic | | | |
| | Gr. 3 | 8.125 | 7.250 |
| | Gr. 6 | 8.225 | 7.600 |
| | Gr. 9 | 7.275 | 7.600 |
| | Gr.12 | n/a | 7.175 |
| | | F(2,117)=2.46 | F(3.156)=0.46 |
| Sentence Length | | | |
| | Gr. 3 | 14.68 | 21.93 |
| | Gr. 6 | 14.97 | 16.22 |
| | Gr. 9 | 12.24 | 15.64 |
| | Gr.12 | n/a | 16.70 |
| | | F(2,117)=1.60 | F(3.156)=2.74* |
| Passage Length | | | |
| | Gr. 3 | 177.5 | 94.68 |
| | Gr. 6 | 322.2 | 176.20 |
| | Gr. 9 | 342.8 | 169.60 |
| | Gr.12 | n/a | 575.80 |
| | | F(2,117)=28.52** | F(3,156)=226.22** |
| Spelling Errors | | | |
| | Gr. 3 | 4.625 (2.6/100) | 2.750 (2.9/100) |
| | Gr. 6 | 5.025 (1.6/100) | 3.400 (1.9/100) |
| | Gr. 9 | 3.325 (1.0/100) | 2.200 (1.3/100) |
| | Gr.12 | n/a | 5.180 (1.3/100) |
| | | F(2,117)=1.64 | F(3,156)=5.15** |

* p<.05
** p<.001