R E P O R T R E S U M E S

ED 016 679 TE 500 039

AN INVESTIGATION OF THE RELIABILITY OF FIVE PROCEDURES FOR GRADING ENGLISH THEMES.

BY- FOLLMAN, JOHN C. ANDERSON, JAMES A.
NATIONAL COUNCIL OF TEACHERS OF ENG.,CHAMPAIGN,ILL

PUB DATE 67

EDRS PRICE MF-$0.25 HC-$0.56 12P.

DESCRIPTORS- *ENGLISH, *COMPOSITION (LITERARY), *COMPARATIVE ANALYSIS, *GRADING, *RATING SCALES, EVALUATION TECHNIQUES, TEST RELIABILITY, ACHIEVEMENT RATING, GRADE EQUIVALENT SCALES, TESTING PROBLEMS, TEST INTERPRETATION, MEASUREMENT TECHNIQUES, STATISTICAL DATA, CALIFORNIA ESSAY SCALE, CLEVELAND COMPOSITION RATING SCALE, FOLLMAN ENGLISH MECHANICS GUIDE, DIEDERICH RATING SCALE,

A STUDY OF FIVE DIFFERENT ESSAY EVALUATION PROCEDURES WAS CONDUCTED TO DETERMINE THE INTRARELIABILITY OF EACH AND TO USE THESE RELIABILITY SCORES AS A BASIS FOR COMPARING THE FIVE PROCEDURES. TEN THEMES WERE ASSIGNED TO FIVE RATING GROUPS, EACH USING A DIFFERENT EVALUATION PROCEDURE--(1) THE CALIFORNIA ESSAY SCALE, (2) THE CLEVELAND COMPOSITION RATING SCALE, (3) THE DIEDERICH RATING SCALE, (4) THE FOLLMAN ENGLISH MECHANICS GUIDE, AND (5) THE "EVERYMAN'S SCALE." THE RATERS WHO GRADED THE THEMES WERE CONSIDERED HOMOGENEOUS--ALL WERE UPPER DIVISION COLLEGE ENGLISH MAJORS WITH THE SAME COURSE IN ENGLISH METHODS. RESULTS INDICATE THAT (1) THE DIFFERENCES AMONG RATING GROUPS DID NOT CHANGE WITH THE SUBJECT MATTER OF THE ESSAYS, (2) THE ESSAYS RECEIVED SUBSTANTIALLY THE SAME SCORES FROM ALL FIVE RATING GROUPS, (3) THERE WERE HIGH INTERCORRELATIONS AMONG SYSTEMS, EXCEPT FOR THE DIEDERICH SCALE, AND (4) THE HIGHEST RELIABILITY SCORES WERE EVIDENCED BY THE ENGLISH MECHANICS SCALE, THE LOWEST BY THE CLEVELAND COMPOSITION RATING SCALE. THE HIGH RELIABILITY OBTAINED ACROSS DIFFERENT EVALUATION PROCEDURES MAY BE DUE TO THE HOMOGENEOUS NATURE OF THE RATERS RATHER THAN TO THE RATING SYSTEM. HYPOTHESES SUGGEST THAT A RATING SYSTEM WOULD HAVE ITS GREATEST EFFECT IN RAISING THE RELIABILITY OF GRADING WHEN USED BY A GROUP WITH HETEROGENEOUS TRAINING AND BACKGROUNDS, AND THAT RATERS OF HOMOGENEOUS TRAINING WILL BE EQUALLY CONSISTENT WITH OR WITHOUT A RATING SYSTEM. THIS ARTICLE APPEARED IN "RESEARCH IN THE TEACHING OF ENGLISH," VOLUME 1, NUMBER 2, FALL 1967, PAGES 190-200. (BN)

In this comparison of five procedures for grading themes, the authors found a remarkable reliability of rating, which they attribute in part to the homogeneity of the raters—upper division college students with the same course in English methods. The findings lend support to the kind of intradepartmental rating espoused and described by Paul Diederich in the April, 1967, **English Journal.**

An investigation of the reliability of five procedures for grading English themes

JOHN C. FOLLMAN
University of South Florida
JAMES A. ANDERSON
Wisconsin State University at Oshkosh

It is common knowledge to student and teacher alike that the grading of essay materials can be highly inconsistent. The grade given to an English theme may vary considerably among different raters and even with the same rater at different times. In recent years a great deal of emphasis has been placed on developing procedures which improve the consistency of grades given to English themes. While reliability scores have been obtained for each procedure, little comparison of the different procedures has been made. This study compared five different essay evaluation procedures.

RELATED LITERATURE

In her review of the literature, Huddleston[1] points out that essay grading unreliability was recognized as far back as the

[1] Edith Huddleston, "Measurement of writing ability at the college entrance level: objective vs. subjective techniques," *Journal of Experimental Education*, 1954, 22, 165-213.

RESEARCH IN THE TEACHING OF ENGLISH

Volume 1, Number 2, Fall 1967

Richard Braddock, *Editor*
Rhetoric Program
The University of Iowa
Iowa City, Iowa 52240

Nathan S. Blount, *Associate Editor*
Research and Development Center for
Cognitive Learning
University of Wisconsin
1404 Regent Street
Madison, Wisconsin 53706

Contributors should follow the format of *Research in the Teaching of English*. Articles should include captions at appropriate places. Articles should be submitted (with a self-addressed return envelope to which stamps are clipped) to Richard Braddock. Abstracts and bibliographic suggestions should be submitted to Nathan S. Blount. Books and research reports for "Roundtable review" and entries for "Notes and comment" should be sent to Richard Braddock. Deadlines for manuscripts: November 1 for spring issue, May 1 for fall issue.

Published each spring and fall by the National Council of Teachers of English. Subscription price, $6.00 for three years; single copies, $1.50 each. Send subscription and business communications to NCTE, 508 So. Sixth St., Champaign, Illinois 61820. Entered as second-class matter January 31, 1967, at the Post Office at Champaign, Illinois 61820, under the Act of March 3, 1879.

1880's. Her review, as well as extensive reviews by Meckel[2] and by Ebel and Damrin,[3] have documented the unreliability of typical grading procedures.

Diederich and others[4] examined the ratings of 300 short essays, 150 on each of two topics, made by 53 readers from six different professional areas. The raters were instructed to sort the papers into nine grading categories. More than one-third of the 300 papers received all nine grades; 94% of the papers received seven or more of the nine possible grades; and no paper received less than five different grades. The median correlation between readers was .31.

It can be suggested that the Diederich findings accurately represent the unreliability of essay grading in general. However, as Ebel and Damrin point out, if experienced, trained raters follow clearly articulated criteria, reliability can be increased, especially if rater teams are used to evaluate subject matter essays. The assumptions in the preceding statement indicate the conditional nature of the achievement of high reliability. That improvement in reliability can be accomplished has been demonstrated by Stuckless,[5] Stalnaker and Stalnaker,[6] Stalnaker,[7] and Diederich.[8]

Review of Grading Systems

A number of formalized systems have been used as criteria, standards, guides, models, etc., in the grading of essays. These may be categorized into several different types of systems.

A number of format-type systems consist of the following facets: content, style, organization, mechanics, wording, but

[2] H. C. Meckel, "Research on teaching composition and literature," *Handbook of research on teaching*, edited by N. L. Gage (Chicago: Rand McNally, 1963), pp. 966-1006.

[3] R. L. Ebel and D. E. Damrin, "Tests and examinations," *Encyclopedia of educational research*, edited by C. Harris (New York: The Macmillan Co., 1960), pp. 1502-1517.

[4] P. B. Diederich and others, *Factors in judgments of writing ability* (Research Bulletin RB-61-15. Princeton, N.J.: ETS, 1961).

[5] E. Stuckless, *Assessment of written language for deaf children* (Progress Report, U.S. Office of Education Cooperative Research Project. Univer. of Pittsburgh, 1965).

[6] J. Stalnaker and R. C. Stalnaker, "Reliable reading of essay tests," *School Review*, 1934, *42*, 599-605.

[7] J. Stalnaker, "Essay examinations reliably read," *School and Society*, 1937, *46*, 671-672.

[8] P. B. Diederich, "Problems and possibilities of research in the teaching of English," *Research design and the teaching of English*, edited by D. H. Russell and others (Champaign, Ill.: NCTE, 1964), pp. 52-73.

with differing emphases. *The California Essay Scale*[9] is representative of this type of system. (See Figure 1.)

A similar system is the *Cleveland Composition Rating Scale*.[10] This features a format similar to that of the *California Essay Scale* but in addition provides a percentage weighting for each major category to assist the grader in making his evaluation. (See Figure 2.)

Another approach is to employ point-scale ratings. In gathering the data, Diederich[11] simply advised the graders to sort the essays into nine piles. This is similar to what most graders do, although they generally use five categories or points (A, B, C, D and F) rather than nine.

Some systems feature a combination of the format- and point-scale ratings. An example is the *Diederich rating scale*,[12] composed of eight facets, each to be evaluated on a 5-unit rating scale. (See Figure 3.)

Figure 1
California Essay Scale

I. Content: Is the conception clear, accurate, and complete?
   A. Does the student discuss the subject intelligently?
      1. Does he seem to have an adequate knowledge of his subject?
      2. Does he avoid errors in logic?
   B. Does the essay offer evidence in support of generalization?
II. Organization: Is the method of presentation clear, effective, and interesting?
   A. Is it possible to state clearly the central idea of the essay?
   B. Is the central idea of the paper as a whole sufficiently developed through the use of details and examples?
   C. Are the individual paragraphs sufficiently developed?
   D. Are all the ideas of the essay relevant?
   E. Are the ideas developed in logical order?
      1. Are the paragraphs placed in natural and logical sequence within the whole?
      2. Are the sentences placed in natural and logical sequence within the paragraph?
   F. Are the transitions adequate?
   G. Are ideas given the emphasis required by their importance?
   H. Is the point of view consistent and appropriate?

(Figure 1 continued on next page)

[9] P. Nail and others, *A scale for evaluation of high school student essays* (Sponsored by the California Association of Teachers of English. Champaign, Ill.: NCTE, 1960).

[10] L. Fryman, *Composition rating scale* (Cleveland Heights, Ohio: Cleveland Heights-University Heights City School District, n.d.)

[11] Diederich and others, *op. cit.*

[12] Diederich, *op. cit.*

III. Style and Mechanics: Does the essay observe standards of style and mechanics generally accepted by educated writers?
   A. Are the sentences clear, idiomatic, and grammatically correct? (For example, are they reasonably free of fragments, run-on sentences, comma splices, faulty parallel structure, mixed constructions, dangling modifiers, and errors of agreement, case, and verb forms?)
   B. Is the sentence structure effective?
      1. Is there appropriate variety in sentence structure?
      2. Are uses of subordination and coordination appropriate?
   C. Is conventional punctuation followed?
   D. Is the spelling generally correct?
   E. Is the vocabulary accurate, judicious, and sufficiently varied?

Figure 2
Cleveland Composition Rating Scale

ASSIGNMENT STUDENT DATE
PURPOSE

A. Content—50%	
Convincing persuasive, sincere, enthusiastic, certain	Unconvincing
Organized logical, planned, orderly, systematic	Jumbled
Thoughtful reflective, perceptive, probing, inquiring	Superficial
Broad comprehensive, complete, extensive range of data, inclusive	Limited
Specific concrete, definite, detailed, exact	Vague
B. Style—30%	
Fluent expressive, colorful, descriptive, smooth	Restricted
Cultivated varied, mature, appropriate	Awkward
Strong effective, striking, forceful, idioms, fresh, stimulating	Weak
C. Conventions—20%	
Correct Writing Form paragraphing, heading, punctuation, spelling	Incorrect Form
Conventional Grammar sentence structure, agreement, references, etc.	Substandard

A fourth system consists of a specific check list of errors the grader uses as a guide to evaluate themes. Such a system (*The Follman English Mechanics Guide*)[13] was developed by one of the experimenters. (See Figure 4.)

The final means of evaluation used in this study was the "Everyman's Scale," in which a rater individually judges essays by whatever criteria he chooses. (See Figure 5.)

OBJECTIVES

The purpose of this study was to determine the reliability of the five grading systems and the differences among them.

The specific questions were these:

1. What is the mean reliability of each rating system?
2. What are the differences in mean reliabilities among the five systems?
3. What are the relationships among the scores from the five systems?

Figure 3
Diederich Rating Scale

TOPIC..... READER..... PAPER.....

	Low		Middle		High	
Ideas	2	4	6	8	10	
Organization	2	4	6	8	10	
Wording	1	2	3	4	5	
Flavor	1	2	3	4	5	.....
Usage	1	2	3	4	5	
Punctuation	1	2	3	4	5	
Spelling	1	2	3	4	5	
Handwriting	1	2	3	4	5	.....
					Sum	.....

Figure 4
Follman English Mechanics Guide

Reference
- accuracy
- completeness
- appropriate style

Spelling

Punctuation
- period, exclamation mark, question mark, comma, semicolon, colon, dash, hyphen, word division (syllabification), apostrophe, italics, quotation marks, parentheses, capitals, abbreviations

Sentence structure
- frags, runons, comma splices (cs)
- faulty parallel structure
- mixed constructions, dangling modifiers

(Figure 4 continued on next page)

[13] J. C. Follman, "An investigation of the differential effects of different methods of teaching critical thinking" (Unpublished study, 1966).

number, case, verb form, verb tense
agreement and referent, noun and antecedent, pronoun and antecedent
word order
comparison fault
adjectives and adverbs

Paragraphs
sentence sequence within paragraphs and paragraph sequence; naturalness and logic
unity, consistency
length and propriateness
logical paragraph division

Diction
word meaning
correctness, preciseness, adequacy, specificity
waffling, vagueness, ambiguity

Word usage
repetition, redundancy, wordiness
slang, triteness, idiom, colloquialism
relevancy, appropriateness
word omission or coinage

Figure 5
"Everyman's Scale"

The purpose of this study is to determine how undergraduate English majors grade essays. Please evaluate the ten essays you have been given.

Grade each essay independently, in other words, grade the first essay, then grade the second essay, etc.

There is no particluar grade that each essay should receive. You evaluate each essay according to your own judgment as to what constitutes writing ability. Use your own judgment about the writing ability as indicated by each essay. Don't use any system other than your own judgment.

Sort the papers into five piles in order of merit with at least one paper in each pile.

Write comments on each essay indicating what you liked or disliked about it.

When you have completed sorting the essays write on each essay which pile you assigned it to. Then return the entire handout in the self-addressed return envelope. Please do this by June 3, 1966. We will then pay you.

PROCEDURE

Ten themes averaging 370 words in length were used. Five themes were essentially expository and five were essentially argumentative. Five of the themes were from Nail,[14] and the other five were from the basic English composition course at

[14] Nail and others, *op. cit.*

the Wisconsin State University at Oshkosh. The themes were chosen to represent the conventional A to F grading continuum insofar as the grades actually assigned to the themes by the respective instructors did accurately represent these different gradations of quality.

The themes were graded by five groups of five raters. Each rater was randomly assigned to his respective rater group. Each rater group used a different rating procedure. Each rater judged the ten themes independently of the other four raters using the same rating system, as well as independently of the other twenty raters. The ratings were all made within ten days.

With two exceptions, the raters were upper division students in the School of Education majoring in English and enrolled in an English methods course at the Wisconsin State University at Oshkosh. The two other raters were upper division education students minoring in English who had completed the English methods course.

Rater Group 1 used *The California Essay Scale,* Group 2 *The Cleveland Composition Rating Scale,* Group 3 *The Diederich Rating Scale,* Group 4 *The Follman English Mechanics Guide,* and Group 5 the "*Everyman's Scale,*" grading as they saw fit with the exception that each of the five possible grades had to be used at least once. All raters completed the *English Expression, Cooperative English Tests, Form 1B.*

Each rating group was met individually. At this meeting the *Cooperative English Tests* were administered, the instructions for the rating system to be used were read and discussed, and written instructions on the system and the essays themselves were distributed. Each rater was assured that his responses were confidential and that he would be paid $5 for his work. This study was primarily supported by the Board of Regents of the Wisconsin State Universities through the Wisconsin State University at Oshkosh.

RESULTS

Analysis of the *Cooperative English Test* scores showed no significant differences in raw scores among the five groups of raters ($F = 1.33$; $df = 4/20$). The five groups were considered equal in basic English skills.

Raw scores for the essays from all the rating systems were analyzed according to Lindquist's Type I analysis of variance

with the between subjects factor (B effect) being the rater groups and the within factor (A effect) the essays.[15]

The interaction of essays and rating groups was not significant at the .05 level of confidence ($F = 1.06$; $df = 36/180$). This result indicates that the differences among rating groups were not changing with the subject matter of the essays.

As was expected, the A effect (essays) was significant ($F = 67.07$; $df = 9/180$; $P<.001$). An inspection of the mean scores for each theme showed a relatively regular distribution, with no ties, spreading from 1.44 to 4.52 in a possible 1 - 5 range. The themes, then, apparently represented varying degrees of achievement.

Table 1

Matrix of Correlations of Rating Group Scores

	Eng. Mech.	*Dieder.*	*Calif.*	*Cleve.*	*Every.*
Eng. Mech.	—				
Dieder.	.59	—			
Calif.	.975	.60	—		
Cleve.	.935	.51*	.95	—	
Every.	.955	.61	.99	.955	—

*All figures but this one are significant at 5% level of confidence.

The B effect (rating groups) was not significant ($F<1.00$; $df = 4/20$). The essays received substantially the same scores from all five rating groups.

The average rating group score for each essay was correlated (Pearson product-moment) with the same score from each of the other rating groups. Table 1 presents the matrix of these correlations. All correlations are significant with the starred exception ($P = .05$; one-tailed test). All of the *Diederich Scale* correlations were significantly lower than all the other correlations ($t = 1.84$; $df = \infty$). All other systems, then,

Table 2

The Average Reliability Scores for One and Five Raters for Each Rater Group

	One Rater	*Five Raters*
Calif.	.769	.943
Cleve.	.460	.810
Dieder.	.727	.930
Eng. Mech.	.813	.953
Every.	.788	.949

[15] E. F. Lindquist, *Design and analysis of experiments* (Boston: Houghton Mifflin Company, 1953).

were significantly better predictors of the common score received by an essay from the five groups.

Reliability measures were obtained for each of the five groups using Ebel's intraclass correlation method.[16] The average reliability scores for one rater and five raters are presented in Table 2. The highest reliability scores were evidenced by the *English Mechanics Scale*, the lowest by the *Cleveland Composition Rating Scale*.

DISCUSSION

The findings of consistently high intercorrelations among systems, with the exception of the *Diederich Scale*, and consistently and markedly high reliability for each rating group, lead to a number of possible inferences. First, four of the five systems intercorrelated highly. Other evidence can be cited to support the notion that the different systems in actuality measure a substantial number of elements in common. Inspection of the various systems except *Everyman's* indicates a number of common elements. Secondly, the instructions for all procedures required the raters to use five categories on scoring, i.e., A, B, C, D, F. Thirdly, related studies using factor analysis have found correspondence in composition annotation. This evidence coupled with the high intercorrelations found here may be interpreted to support the notion that evaluation systems do in fact measure a substantial number of elements in common.

Second, what might be characterized as the abnormally high within-group rater consistency which we obtained was not expected—particularly the *Everyman's*, for which the second highest reliability coefficient, r = .949, was found. The instructions to the raters using this system were explicitly open: "You evaluate each essay according to your own judgment as to what constitutes writing ability. Use your own judgment about the writing ability as indicated by each essay. Don't use any system other than your own judgment." A reasonable expectation is that such instructions would permit great individual difference and inconsistency among the raters using the *Everyman's Scale*. This did not occur; in fact, the opposite did.

It may now be suggested that the unreliability usually obtained in the evaluation of essays occurs primarily because raters are to a considerable degree heterogeneous in academic background and have had different experiential backgrounds

[16] R. L. Ebel, "Estimation of the reliability of ratings," *Psychometrika*, 1951, *16*, 407-424.

which are likely to produce different attitudes and values which operate significantly in their evaluations of essays. The function of a theme evaluation procedure, then, becomes that of a sensitizer or organizer of the rater's perception and gives direction to his attitudes and values; in other words, it points out what he should look for and guides his judgment.

The authors of this paper propose the following framework for dealing with reliable and unreliable ratings. This framework is that of homogeneous *versus* heterogeneous raters. As all of the raters with two exceptions were upper division School of Education English majors with recent experience in an English methods course (the two exceptions were upper division School of Education English minors who had completed their English methods course earlier), and since there were no significant differences in their tested basic English skills, it may be suggested that these raters represented a homogeneous group with respect to the factors that determine their theme evaluating. Therefore, the high reliability which was obtained across different evaluation procedures may be due primarily to the homogeneous nature of the raters rather than to a rating system. The finding of high mean reliability of the *Everyman's* procedure may be interpreted to support the authors' rationale in that the second highest reliability score was obtained with, in effect, a nonsystem, suggesting that the raters were reliable to begin with, perhaps substantially because of their common English methods course experience.

The situation with heterogeneous raters would be somewhat different. When a group of heterogeneous raters uses an evaluation system and the mean reliability is higher than what it would be without the system, it appears that the system provides a sensitizing to certain elements of a theme and to certain values used in theme evaluation.

A number of studies seem to be suggested by this research. Initially it would seem desirable to replicate this study with homogeneous and heterogeneous raters.

We would suggest the following hypotheses.

We would hypothesize that a rating system would have its greatest effect in raising the reliability of grading when used by a group with heterogeneous training backgrounds. In these cases the procedure serves two purposes: First, it provides an observation routine. That is, each procedure spells out what

items are to be evaluated in a theme. Second, the rating procedure provides a set of values to be applied to each element. As a result of the grading procedure, raters are directed to the same elements and these elements receive the same weightings in the evaluation process. Without the rating system, raters of heterogeneous training backgrounds apparently observe different elements or evaluate elements differently or both.

We further hypothesize that raters of homogeneous training, such as that provided by a common English methods course, will be equally consistent with or without a rating system. Without the system, the raters would use the most readily applicable experience for rating, which would be their similar college training. As a result, raters of homogeneous backgrounds would continue to observe similar elements in the theme and to evaluate them in similar ways.

SUMMARY

The unreliability of grading English essays has been widely documented in experience and research. Various evaluation systems and procedures have been developed to improve the reliability of grading English themes. This study was an attempt to determine the intrareliability of each of five different kinds of evaluation procedures and to use those reliability scores as a basis for comparison of the five procedures. Mean group reliability scores were generally high within each system, less high for the individual rater. The intercorrelations of mean group scores for evaluation systems were high for all systems except the Diederich system. Hypotheses were extended concerning homogeneity and heterogeneity of rating groups in training backgrounds.