DOCUMENT RESUME

ED 194 616 TM 800 747

AUTHOR Brossell, Gordon
TITLE Validation of Topics and Comparisons of Three Presentation Modes for the Writing Subtest of the Florida Teacher Certification Examination. Volume Five of Five.
INSTITUTION Florida State Univ., Tallahassee. Coll. of Education.
SPONS AGENCY Florida State Dept. of Education, Tallahassee.
PUB DATE Mar 80
NOTE 29p.; For related documents, see TM 800 743-46.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Competency Based Teacher Education; Education Majors; Elementary Secondary Education; *Essay Tests; Higher Education; *Minimum Competency Testing; Student Evaluation; Teacher Certification; *Test Validity; Writing (Composition); *Writing Skills
IDENTIFIERS *Essay Topics; Florida; Florida Teacher Competency Examination; Interrater Reliability; *Presentation Mode; Writing Evaluation

ABSTRACT

The hypothesis that role-playing scenarios specifying full rhetorical contexts were a superior means of eliciting valid writing samples for the purpose of assessing compositional skills of prospective Florida teachers was tested. In addition, topics for use on the initial administrations of the Florida Teacher Competency Examination's writing subtest were developed and validated. Six essay topics selected by a panel of validators were cast in three different presentation modes according to degree of specification of rhetorical context, or "informational load (IL)." A low IL presented the topic in a brief phrase leaving the writer free to make decisions about audience, purpose, form, and tone without guidance. A moderate IL gave the writer an orientation to the task without including specifications. A high IL gave a full rhetorical context. The six topics in the three modes, 18 combinations in all, were administered randomly to a sample of college education majors. Evidence gathered in statistical and rhetorical analyses points clearly at the moderate IL as the presentation mode most likely to stimulate the best writing of large number of examinees. (RL)

VALIDATION OF TOPICS AND COMPARISON OF THREE PRESENTATION MODES FOR THE WRITING SUBTEST OF THE FLORIDA TEACHER CERTIFICATION EXAMINATION

Gordon Brossell

College of Education
Florida State University

VOLUME FIVE OF FIVE

Under Contract to the Department of Teacher Education
Florida State Department of Education

#080-064

Department of Education

March, 1980

TABLE OF CONTENTS

Purpose of the Study

Despite a considerable body of professional opinion, there is little empirical evidence that shows the effects of essay topics and their mode of presentation on writers. It is a widely held notion among composition teachers that specification of rhetorical context--that is, the identification for a writer of the purpose of a piece of writing, its intended audience, its subject, its form, and its "voice" --will enable him or her to understand more fully the demands of a particular writing task and so to produce a more fully realized, coherent piece of work. Some researchers have passed beyond speculation about the necessity for rhetorical specification, preferring instead to ask questions about the impact of certain elements within a full writing context rather than about the effects of the context itself. Yet when it comes to demonstrating these effects, to showing how changes in the mode of presentation of an essay topic result in variations in the quality of an essay written on it, there is but scant data available upon which to decide how best to design a valid test of writing competence.

That is the question. How to devise an essay examination whose topics and whose presentation modes will offer a fair test of compositional skill to a population of thousands of examinees. In an attempt to answer this question, and to cut a path through the jungle of research reports, testimonials, reviews, commentaries, and self-styled theories which comprise the relevant literature on the subject, James Hoetker undertook Volume IV of the Writing Subtest Handbooks, "On Writing Essay Topics for a Test of the Composition Skills of Prospective Teachers." In it he put forth the

recommendation that topics for the Florida Teacher Certification Examination should indeed present full rhetorical contexts to the writers and that furthermore they should be based on situations that prospective teachers would be apt to face in the real world. The particular form he advocated was the hypothetical situation, or scenario, that is predicated on a variety of roles typically played by practicing teachers. Despite careful, extensive research, logically sound reasoning, and a convincing argument, Hoetker's recommendation had to be qualified by a lack of empirical data:

> In the absence of the needed experimental evidence, and in the presence of the possibility that partial specification of context might make a difference, the safest course seems to be not to take chances, and to produce topics . . . that give full specification of the class of discourse that has been demonstrated to have direct pertinence to a teacher's job.[1]

It remained to test the hypothesis that role-playing scenarios specifying full rhetorical contexts were a superior means of eliciting valid writing samples for the purpose of assessing the compositional skills of prospective Florida teachers. The current study did just that.

In addition, the study sought to develop and validate a minimum of four topics for use on the initial administrations of the writing subtest. The validation of topics for writing examinations is neither a standardized nor a widely practiced procedure and has been used consistently only by professional testing organizations that administer writing exams regularly, such as the Educational Testing Service and College Entrance Examination Board. Essentially,

---

1. James Hoetker, "On Writing Essay Topics for a Test of the Composition Skills of Prospective Teachers," Florida State Department of Education, 1979, p. 76.

validation is a process whereby a verbally expressed writing stimulus--a topic--is certified by a number of experienced reviewers to be free from the kinds of rhetorical, structural, and psychological biases which might otherwise affect a writer. The process normally includes, as it did in this study, a series of critical reviews, emendations, and editions of topics, as well as a trial of them under conditions similar to those that will exist in the actual examination. It is--and this process was--aimed at attaining a high level of agreement among expert consultants about the possible impact of writing stimuli and is in keeping with the subjective but rigorous and criteria-oriented procedures that mark good programmatic writing assessment.

Design of the Study

Overview. Six essay topics were selected from a list generated by a panel of validators specially chosen by the investigator. Each topic was cast in three different presentation modes according to degree of specification of rhetorical context, or "informational load." The first mode presented the topic in a brief phrase only, leaving the writer free to make decisions about audience, purpose, form, and tone without guidance of any kind. This mode, Mode 1, was said to contain low informational load. Mode 2, characterized by moderate informational load, presented a general introductory statement about the topic and then asked the writer to state his or her own views on it, giving the writer an orientation to the task but leaving specifications aside. Mode 3 described a hypothetical situation placing the writer in the position of having to state personal views on a subject as in Mode 2, but this time in a full rhetorical context--that is, with an identified audience, a specific

form, and a stated or clearly implied purpose. The writer was thus given a point of view, or rhetorical stance, from which to write, though the substance of the piece--the actual views expressed--was always left to his or her personal disposition. This mode was said to be marked by high informational load.

The six topics in the three modes, eighteen combinations in all, were then administered randomly in an essay examination trial--one to each writer--to a sample population of undergraduate education majors at two state universities, Florida State University in Tallahassee and the University of South Florida in Tampa, under test conditions closely resembling those anticipated for the first administrations of the actual writing subtest. The resulting papers were collected and screened by the investigator for degree of commitment to the task (the writers were asked to take the test seriously though they knew, of course, that it would have no real consequences for them), and those few found to exhibit clearly insufficient effort were removed. Twenty essays were retained in each mode of each topic, or "cell," leaving a total of 360 in the sample population.

The essays were read and rated by a panel of three raters who had undergone training in holistic evaluation of writing samples conducted by the investigator using the Training Manual (Volume II of the Writing Subtest Handbooks) developed for this purpose. Essays receiving discrepant scores were read and rated by a referee whose ratings replaced the discrepant ones according to the procedures established in Volume III of the Handbooks. The resulting final

scores of the essays were then entered, together with their estimated lengths, into the FSU computer for statistical analysis.

Topic Selection. The investigator recruited three experienced teachers of written composition to serve as validators: Barbara Ash, a second-year doctoral candidate in English Education at FSU and a former high school English teacher, who also served as the study's chief administrative assistant; Linda Clarke, an English teacher at Lincoln High School in Tallahassee who holds a master's degree in English Education from FSU; and Pamela Laws, a composition instructor at Tallahassee Community College who also holds an FSU master's degree in English Education.

The panel met several times in the fall of 1979, completing its work in mid-October. Initially, the validators were asked to generate a working list of possible topics using a set of criteria adapted in part from Volume IV of the Handbooks. Topics meeting these criteria were thought by the validators to be:

1) self-explanatory (i.e., clearly and explicitly phrased);
2) defined and limited;
3) familiar to every examinee;
4) stimulating;
5) fresh;
6) of middle-emotional ground (i.e., neither too pedestrian nor too sensational);
7) nonbiased and nonbiasing.

From an initial list of more than twenty possible topics, the panel, after deliberating the potential effects of each, selected eight it felt met the criteria in each of the three presentation modes. These eight topics in each mode were then sent for review and comment to the consultants to the Writing Subtest Handbooks, Dr. Nancy McGee of the Department of Secondary Education at the University of Central Florida in Orlando and Professor Dan Kelly of the English

Department at the University of Florida in Gainesville. Their responses to the topics together with further deliberation by the panel resulted in the elimination of two of the eight topics, leaving six judged valid for the essay trial.

Essay Field Trial. During November, 1979, the topics were administered to a sample population of undergraduate education majors in an essay field trial. Students in professional education classes at two universities, FSU and USF, took the tests, which were administered by the investigator and his assistant at FSU and by Dr. Annie Ward, technical consultant to this study, at USF. Test conditions were similar to those anticipated for the first administrations of the actual examination: directions were printed on the cover sheet of each test packet, which contained blank lined paper for the examinees' use; and a period of 45 minutes was allowed in which to complete the exam. No choice of topic or mode was offered, to insure that the proper number of essays in each cell could be obtained. The test administrators announced the purpose of the field trial at the beginning of each class period in which the testing was conducted. The examinees thus had no foreknowledge of the test, a condition which helped to insure their attendance in numbers adequate to collect sufficient essay samples. All three test administrators reported that this condition had no apparent effect on the attitudes of the writers toward the test, and the essays themselves made no mention of it.

A total of 360 essays comprised the final sample. Of these, 190 were written by students at FSU and 170 by students at USF.

There were 294 females in the sample and 63 males; three additional writers failed to identify themselves. The students represented a large variety of majors, about 30 in all, though the largest numbers of them came from programs in elementary education (including early childhood education and child development) and physical education--159 in the former case and 55 in the latter. The sample was thus roughly proportional to the distribution of academic majors currently being prepared in Florida teacher education programs. A breakdown of the sample population by academic major follows.

FSU

Major	# in Sample
Physical Ed.	55
Special Ed.	21
Music Ed.	17
Elementary Ed.	17
English Ed.	13
Speech Pathology	10
Social Studies Ed.	7
Home Economics Ed.	6
Art Ed.	5
Mathematics Ed.	5
Social Work	5
Early Childhood Ed.	4
Child Development	4
Science Ed.	3
Visual Disabilities	3
Library Science	3
Vocational/Business Ed.	2
Industrial Arts Ed.	2
Foreign Language Ed.	2
Theater Ed.	1
Political Science	1
ESL	1
Psychology	1
Art Therapy	1
Career Ed.	1
	190

USF

Major	# in Sample
Elementary Ed.	98
Early Childhood Ed.	36
Learning Disabilities	16
EMR	8
EMH	6
English Ed.	2
Gifted Ed.	2
Foreign Language Ed.	1
Deaf Ed.	1
	170

The Rating Team. Two of the validators, Linda Clarke and Pamela Laws, also served as raters of the essays; a third rater, Carol Gray, an experienced teacher of composition and a member of the English department at Leon High School in Tallahassee, joined the rating team in December. Dr. Dwight Burton, Professor of English Education and chairman of the Department of Curriculum & Instruction in the College of Education at FSU, agreed to serve as referee, whose task is to read and rate essays receiving discrepant ratings. (See Volume III of the Handbooks, p. 27-36, for a full treatment of this procedure.) These people comprised a first-rate holistic scoring team, meeting the requirements of professional experience and technical knowledge imposed by the study and, in the investigator's opinion, surpassing the degree of competence that might be expected of a typical team rating essays written in the actual subtest.

Rater Training. In early December, the rating team (except for Dr. Burton, who had served in a similar capacity before and who was thoroughly familiar with his referee's role) underwent initial training in holistic scoring of essays. The training session, which occurred on a Saturday on the FSU campus, was conducted by the investigator, aided by Barbara Ash, the administrative assistant, using Volume II of the Handbooks, the Training Manual. The raters spent roughly the first half of the session on the materials and procedures called for in the manual--practicing holistic rating with the appropriate criteria and rating guides, and attempting to reach a high level of consistency in their rating of the same essays.

When the formal training was completed, a check was made to determine the level of interrater reliability, or consistency,

achieved in the training session. Thirty of the essays written on the field trial were read and rated independently by each rater in three packets of ten. Their ratings were then analyzed in terms of the four indexes described in Volume I of the Handbooks--percentage of complete agreement among raters, average percentage of two of three raters agreeing, average percentage of agreement by pairs of raters as to whether an essay passes or fails, and percentage of complete agreement about whether an essay passes or fails. The following table shows how the reliability levels achieved by the rating team compared with the target levels established in Volume I.

	Raters' Level	Target Level
Index 1--% Complete Agreement	40	30-40
Index 2--Average % Two of Three Raters Agreeing	96.7	80-90
Index 3--Average % Agreement by Pairs as to Pass/Fail	82.2	80-90
Index 4--% Complete Agreement about Pass/Fail	73.3	70-80

In indexes 1, 3, and 4, the raters' reliability levels fell within the desired ranges; in index 2, their level exceeded that of the target range. (Only one of the thirty essays read in the reliability check received a discrepant set of ratings and needed subsequently to be read by the referee.) The investigator thus had convincing evidence that the training session had been successful and that the rating team had achieved a level of reliability sufficient to sustain a high degree of confidence in their ratings of field trial essays.

Essay Ratings: The Data

After initial training, the raters were given the task of

rating essays written on the field trial. All the essays were read and rated independently by each rater under conditions (the work was done for the most part in the raters' homes) as similar as is possible to obtain when raters are not gathered in one place, as they were for the initial training session. When the ratings were submitted, they were reviewed and those essays receiving discrepant ratings--72 in all, or 20% of the total number--were given to the referee. His ratings were substituted for the discrepant ones, and the scores of all 360 essays were finalized. The following table summarizes the results of the essay field trial.

Score	N	% of N
3	37	10.3
*4	0	0
5	100	27.8
6	71	19.7
7	76	21.1
8	47	13.1
9	13	3.6
10	13	3.6
11	2	.6
12	1	.3
Total	360	100

Mean Score --6.19
Median Score--6
Modal Score--5

*Initially there were 23 scores of 4. Of these, 10 became 3's and 13 became 5's--the result of the substitution of the referee's ratings, in each case, for a 1 or a 2 in the ratings distribution comprising a score of 4 (1 1 2).

In early January, the investigator, having secured the assistance of the Office of User Services at the FSU Computing Center, entered the ratings of the essays, together with their estimated lengths,[1] into the computer, the resulting data file becoming the basis of the statistical analysis that followed.

Interrater Reliability

The level of reliability achieved by the rating team in rating the field trial essays was measured, using the four indexes described on page 9. The resulting figures reflect the referee's ratings, unlike those of the training session where a referee was not involved; thus they represent one final measure of the team's reliability as essay raters. The following table shows how the team's rating effectiveness compared with the target ranges in the four indexes of consistency. The figures in parentheses are those the team attained in the initial training session and are supplied for the sake of comparison with the levels of reliability it achieved in the entire essay trial.

---

1. Lengths were estimated according to the following procedure: every tenth line of each essay was given a word count; the sum of these counts was divided by the number of lines counted to get an average number of words per line, which was then multiplied by the number of lines in the whole essay. The resulting product was the estimated length of the essay. To insure word counts accurate enough to be meaningful, a check was run against the actual word count in 60 essays. In one set of 30 essays, estimated and actual word counts differed by 8.5% on the average with twelve cases in which differences exceeded 10%, a tolerable margin of error in this kind of estimate. The average number of words per essay differed by only ten from estimated to actual count, however. In another set of 30 essays, the average difference between the estimated and actual word counts was 7.35% with only six cases in which the 10% margin of error was exceeded, and the average word count per paper differed by only ten words from estimated to actual count. This check indicated that the estimated word count was accurate enough for inclusion in the statistical analysis. (Dr. Tom Denmark, Professor of Mathematics Education at FSU, rendered advisory assistance on the word-count procedure.)

	Raters' Level		Target Level
Index 1--% Complete Agreement	32.2	(40)	30-40
Index 2--Average % Two of Three Raters Agreeing	98.3	(96.7)	80-90
Index 3--Average % Agreement by Pairs as to Pass/Fail	81.3	(82.2)	80-90
Index 4--% Complete Agreement about Pass/Fail	71.7	(73.3)	70-80

On three of the four indexes the rating team's level of consistency fell within the target ranges; in one case, Index 2, the team's level exceeded not only the target range but also the level it had achieved in the training session. In indexes 1, 3, and 4, the small dropoff from the training session levels to the field trial levels is in all likelihood a result of the tenfold increase in the number of essays read, and was hardly unexpected.

In addition to the four indexes, a coefficient of interrater reliability was obtained for pairs of raters and for the rating team both before and after the substitution of the referee's ratings. Known as the Alpha coefficient, it is in simplest terms a statistical indication of the expected correlation between the ratings of the team on this task and those of a hypothetical team of similarly comprised and similarly trained raters doing the same task. The following table shows the Alpha coefficients for the rating team.

	Without Referee's Ratings	With Referee's Ratings
Raters 1 & 2	.619	.640
Raters 1 & 3	.720	.799
Raters 2 & 3	.686	.815
Raters 1, 2, and 3	.759	.828

The figures reflect the effect of the referee's ratings on the team's between-rater consistency, increasing the level of reliability in every instance and increasing it substantially in some. The most important coefficient--that of raters 1, 2, and 3 (i.e., the whole team) with the referee's ratings--is, as would be expected, the highest, since the reliability of a group of trained raters generally increases as its number increases and since the substitution of a referee's ratings is, in and of itself, a deliberate upward adjustment in interrater reliability. The level of reliability achieved by the rating team is, in the judgment of the investigator, sufficiently high to justify firm reliance on the data yielded in the essay field trial.

Analysis of the Data

The purpose of the statistical analysis was to determine the extent to which the final scores of essays written on the field trial depended on three factors--topic, mode, and length. Toward this end, two statistical operations were undertaken: analysis of variance (ANOVA) and, later, a multiple regression analysis (including a scattergram) of the effect of length on score.

Three separate analyses of variance were run on the SPSS program at the FSU Computing Center. The first ANOVA tested for the effects and interactions of topic and mode only, ignoring length. The second ANOVA processed length with the main effects of topic and mode, in effect treating all three factors equally. In the third ANOVA, length was treated as a covariate, and the effects of topic and mode were corrected for the effects of length.

The following table summarizes the statistical data for the first ANOVA.

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Main Effects	18.183	7	2.598	.853	.544
Topic	6.167	5	1.233	.405	.845
Mode	12.017	2	6.088	1.974	.140
Two-Way Interactions	36.817	10	3.682	1.210	.283
Topic Mode	36.817	10	3.682	1.210	.283
Explained	55.000	17	3.235	1.063	.389

There were no statistically significant effects or interactions of topic and mode on score, but the effect of mode was clearly much stronger than the effect of topic, ~~achieving statistical significance at the .14 level~~ with a probability of .14.

In the second ANOVA, length was entered into the equation along with topic and mode; its statistical summary follows.

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Main Effects	295.983	8	36.998	16.042	.001
Topic	11.817	5	2.363	1.025	.403
Mode	3.235	2	1.617	.701	.497
Length	277.800	1	277.800	120.454	.001
Two-Way Interactions	13.475	10	1.347	.584	.827
Topic Mode	13.475	10	1.347	.584	.827
Explained	309.458	18	17.192	7.454	.001

Once again the main effects and the interaction of topic and mode were insignificant, but the effect of length on score was significant at the .001 level.

In the third ANOVA, length was held constant in assessing the effects of topic and mode. Under these conditions--with length treated as a covariate--the ANOVA produced the following data.

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Main Effects	18.183	7	2.598	1.126	.346
Topic	6.167	5	1.233	.535	.750
Mode	12.017	2	6.008	2.605	.075
Covariate					
Length	277.800	1	277.800	120.454	.001
Two-Way Interactions					
Topic X Mode	13.475	10	1.347	.584	.827
Explained	309.458	10	17.192	7.454	.001

This time length was again significant at the .001 level, and the effects of topic and mode were again insignificant; but the effect of mode approached significance at the .05 level. That is, when the effects of topic and mode were adjusted for the effects of length, mode was significant at the .075 level.

To help identify the nature of the significant correlation between length and score, a multiple regression analysis, together with a scattergram, was run. This analysis, the statistical summary of which follows below, revealed a correlation of moderate statistical significance between them, and the scattergram (not reproduced here) showed the correlation to be curvilinear in nature and of the following order: up to and including a score of 9, the mean length of the essays gradually increased; after 9, mean length varied widely (with the number of writers scoring higher than 9 greatly diminishing), and the correlation broke down. A table showing mean lengths (including standard deviations and variance coefficients) by score follows the multiple regression analysis summary below.

Multiple Regression Analysis Summary

F to Enter or Remove	Signifi-cance	Multiple R	R Square	R Square Change	Simple R	Overall F	Signifi-cance
122.97136	.000	.50564	.25567	.25567	.50564	122.97136	.000

Mean Lengths by Score

Score	Mean Length (Words)	Standard Deviation (Words)	Variance	Coefficient	N
3	208	112	13049.5088		37
4	---	---	---		0
5	260	89	7905.0961		100
6	315	88	7688.5968		71
7	342	83	6864.4370		76
8	394	138	18972.3784		47
9	418	92	8436.0897		13
10	374	102	10409.3333		13
11	405	9	84.5000		2
12	493	0	0		1
Total	312	114	13049.5088		360

The correlation between length and scores 3 through 9 suggests that essays of certain lengths were more likely to get certain scores. Indeed in the table above, the mean length of essays varies directly with increasing scores up through 9--scores which account for more than 95% of the essays in the sample. It can be said then that the longer one's paper was--or, more specifically, the greater its length in the range of roughly 200-400 words--the more apt it was to get a higher score, at least up to a score of 9.

The level of significance of mode, particularly when adjusted for the effects of length, suggests that mode might too have affected the scores of essays, though that conclusion appears to be considerably more tenuous than that adducing length as a decisive factor. Mode,

after all, was not a statistically significant factor on essay scores, unless one wishes to argue that the .075 level attained in the third analysis of variance is, in this study, significant. There is no basis for doing so, however. What does appear to be a sensible inference is, rather, that mode is correlated with score by way of length; that is, a certain mode produced a tendency toward higher scores by virtue of having stimulated examinees to write essays of greater length. A look at mean scores and mean lengths by modes in the following table reveals the beginnings of a case for this line of reasoning.

	N	Mean Score	SD	Mean Length	SD
Mode 1	120	6.14	1.74	320	123
Mode 2	120	6.43	1.77	326	101
Mode 3	120	5.98	1.70	288	115
Total	360	6.19	1.74	312	114

Mode 2 essays scored higher on the average and were of greater average length than essays written in the other two modes. In addition, its standard deviation for length was a bit lower than those for the other two modes and for the entire sample, and its standard deviation for score was virtually the same as that of the others and of the entire sample's. To be sure, the differences are not dramatic, but one hardly expects dramatic differences in a study whose sources of variation are themselves notably subtle. The magnitude of the variation among the three modes is relatively small to begin with, especially given the careful attention to selection of topics, the controlled conditions of the essay trial, and the relatively large size of the sample population. So a difference in mean score of nearly a half point, for example--the difference between the mean scores of Mode 2 and Mode 3 essays--is

not a negligible one, particularly if it can reasonably be attributed to specific, deliberate changes in the wording of topics. And while the differences in the mean lengths of the modes are quite small, the fact remains that Mode 2 writers exhibited a tendency to write slightly longer essays and, in so doing, to achieve somewhat higher scores.

A similar comparison of mean scores and mean lengths by topics and by topics within modes lends greater weight to the argument that mode is correlated to score by way of length. The table below and that which follows present these data.

	N	Mean Score	SD	Mean Length	SD
Topic 1	60	6.35	1.96	310	118
" 2	60	6.25	1.59	335	127
" 3	60	6.22	1.54	293	133
" 4	60	6.28	1.68	299	104
" 5	60	5.97	2.08	311	97
" 6	60	6.05	1.58	321	102
Total	360	6.19	1.74	312	114

The figures here are inconclusive. Mean scores for all topics except #5 [redacted] cluster within a range of .3 of a rating point; mean lengths do not vary proportionately with increasing mean scores. There is no discernible pattern in the standard deviations of either measure. A view across topics, in other words, seems to bear out the insignificant effect of topic on score revealed in the three analyses of variance. The view across topics within modes, however, is more instructive.

	N	Mean Score	SD	Mean Length	SD
Mode 1					
Topic 1	20	6.45	2.37	332	152
" 2	20	6.30	1.42	347	119
" 3	20	6.30	1.49	327	151
" 4	20	6.20	1.82	309	117
" 5	20	5.95	1.73	301	95
" 6	20	5.65	1.53	303	99
Mode 2					
Topic 1	20	6.30	1.59	289	95
" 2	20	6.20	1.77	335	103
" 3	20	6.80	1.64	316	117
" 4	20	6.15	1.50	301	111
" 5	20	6.10	2.38	337	80
" 6	20	7.05	1.61	381	82
Mode 3					
Topic 1	20	6.30	1.92	308	101
" 2	20	6.25	1.65	324	158
" 3	20	5.55	1.28	236	115
" 4	20	6.50	1.76	289	87
" 5	20	5.85	2.18	295	111
" 6	20	5.45	1.10	278	99
Total	360	6.19	1.74	3.12	114

Here the figures for Topics 6 and 3 [redacted] in Mode 2 stand out: the mean scores are the highest in the entire breakdown, as is the mean length for Topic 6. Clearly the presentation of these two topics in this mode had an influence on writers such that they wrote essays of greater mean length and thus of higher mean score--to an extent that these are the only two categories of essays in the statistical breakdown that can unqualifiably be referred to as above average in quality. Topic 6 essays in Mode 2 also had a low standard deviation for mean length, suggesting that the effects of this version of this topic were more consistent from writer to writer than were, with one

exception, all other combinations. The tendency exhibited by Topics 6 and 3 in Mode 2 is particularly interesting when a comparison is made of their mean scores and mean lengths with those of the other two modes of those two topics. In Modes 1 and 3 of Topic 6, mean scores were only 5.65 and 5.45, and mean lengths were 303 and 278, respectively; and in Modes 1 and 3 of Topic 3, mean scores were 6.30 and 5.55, while mean lengths were 327 and 236, respectively. The differences, while once again not dramatic, are systematic and even sizable given only the change in presentation mode to account for them.

It must be remembered that, given the degrees of discrimination on the rating scale and the rendering of at least three and sometimes four ratings for each essay, the difference between scores of 5 and 6 a 6 and 7 is not, qualitatively speaking, a minuscule one. An essay scoring 5 is only a marginally passing essay and had to have been considered incompetent by at least one rater. An essay scoring 6 is, by definition, an essay of average competence and most likely was judged average by all three raters. An essay scoring 7 on the other hand is an essay of better than average competence and had to have been judged that way by at least one rater.

The same can be said of the difference between essays scoring below and at the cutoff point on the scale, 5. An essay scoring 3 is an imcompetent essay and had to have been judged that way by all three raters or by two raters and the referee. An essay scoring 4 is discrepant by definition because it was rated average by one of three raters and incompetent by the other two. The referee's rating in such a case determines in all likelihood whether the final score

will be 3 or 5 (since there is little chance the referee will rate such an essay above 2). Either way there is a high level of consensus about the issue of competence (or incompetence) in any individual essay with an initial score of 4. An essay scoring 5 on the other hand, though only marginally competent, still earned its score by virtue of convincing two raters that it is of average competence (except in the highly unlikely event that the referee confirms a rating of 3 in a ratings distribution of 1 1 3). The point to remember, then, is that adjacent scores on the rating scale are the result of separate ratings and so <u>are</u> qualitatively different.

With respect to the influence of modes, topics, and combinations thereof on writers, it is useful to look at the numbers of essays scoring at and below the cutoff point. The following table presents this information.

	3's	5's
Mode 1		
Topic 1	5	1
" 2	1	4
" 3	1	5
" 4	3	3
" 5	2	8
" 6	2	10
Mode 2		
Topic 1	1	6
" 2	2	6
" 3	1	4
" 4	1	7
" 5	3	7
" 6	1	2
Mode 3		
Topic 1	2	6
" 2	2	5
" 3	2	9
" 4	2	3
" 5	4	6
" 6	2	8
Mode 1	14	31
" 2	9	32
" 3	14	37
Topic 1	8	13
" 2	5	15
" 3	4	18
" 4	6	13
" 5	9	21
" 6	5	20
Total Exam	37	100

Again, Mode 2 and Topics 6 and 3 in Mode 2 stand out as the categories in which the fewest writers, comparatively speaking, were judged incompetent or marginally competent. (Topics 2 and 3 in Mode 1 are notable in this regard as well, as are Topics 1 and 4 in Mode 2.) To the extent that these figures represent a reasonable approximation of the percentages of examinees who will attain similar scores on the actual subtest, they are important. Essays scoring 3 and 5 comprise fully 38% (137 of 360) of the sample population. If one assumes a proportionate distribution of such scores among modes, topics, and cells, one would expect one-third in each mode, one-sixth in each topic, and one-eighteenth in each cell. A glance at the table reveals a disproportionately low number of 3's in Mode 2 scores (as well as a similar condition for 3's and 5's in Topic 6 of Mode 2). A conclusion justifiable in terms of these data, then, is that Mode 2 essays were less likely than others to be judged incompetent; or, to put it another way, the influence of Mode 2 versions of topics produced a stronger tendency in writers to compose competent essays than did other modes. Certainly this is a tendency to consider in selecting a presentation mode for topics in actual administrations of the writing subtest of the Florida Teacher Certification Examination.

A Rhetorical Perspective on the Essays

A sample (about 20%) of the essays was drawn randomly from the eighteen cells and read by the investigator and the administrative assistant with an eye toward discovering any rhetorical characteristics or patterns that might be attributable to factors in the study.

The result was a series of impressions based not on specific criteria--as were the ratings--but rather on a sense of how the essays matched the expectations of the reviewers (who are, after all, experienced teachers of writing) for college educated young adults writing under these particular test conditions. Interestingly, the two readers in their independent reviews agreed substantially on what they felt were the dominant characteristics of essays and of certain groups of essays. These impressions, which follow, are meant to provide a supplementary gloss on the statistical analysis reported above.

As a group, the essays written on the field trial were a desultory lot, distinguished chiefly, though not uniformly, by blandness of expression, a tendency toward overgenerality, and an uncertain command of the rhetorical, structural, and mechanical conventions of written English. Many essays struck the readers as the linguistic equivalents of photographs taken by an unsteady hand, the contours hazy and uncertain and the entire subject somewhat out of focus. The writers either failed to find what they wanted to say soon enough--many rambled as if doing unstructured thinking exercises on paper--or they were uncommitted to flushing out their real feelings on a particular topic. In a number of instances, writers simply had--or chose to have--little or nothing of significance to say. While it is hazardous to guess what impact a lack of motivation might have had on these writers, their papers did not reflect anything like genuine involvement in the business of writing on these topics. Whether this circumstance mirrors the lack of real concern and effort which often attends simulated versions of experience, and whether a different set of characteristics will be manifested in the actual

subtest are, at this time, strictly moot questions. For this sample of essays, mediocre is an apt though possibly overgenerous description of them.

There were notable exceptions, however, the most compelling being those essays written in Mode 2 of Topic 6, [redacted]. Despite some variations, these essays were as a group better organized, more sharply focussed, and more interesting, lively, forthright, and personal than any other category of papers in the sample. As a rule, they addressed the topic more quickly than others, had more specific points to make about it, and were stylistically superior to their counterparts in the rest of the sample.

A Perspective on Modes and Topics

By design, Modes 1, 2, and 3 vary according to the amount of information each supplies about a topic; or, to put it another way, according to the degree to which each approaches a full rhetorical context. Mode 1 provides little information and no context whatsoever; Mode 2 supplies some information and an orientation to the topic; Mode 3 provides a good deal of information and contains all the elements of a full rhetorical context--audience, purpose, form, and subject. All affect writers in particular ways.

Mode 1, by virtue of its low degree of specification, challenges a writer first to define the topic at hand and then to say something about it. Such a task throws a writer on his or her own resources early, forcing quick decisions--or at least accelerated thinking--about what a topic means and what a writer feels about it. No method of organization or procedure is suggested or implied, and if a writer cannot bring some organizational principle directly to bear

on an essay, it will likely founder aimlessly into waters as muddy as those treaded by the excerpts of essays quoted earlier. When a writer meets the challenge successfully, however, the result is quite like a good essay in Mode 2, with the exception that it takes a bit longer to get in focus.

Mode 2, unlike Mode 1, gives a writer a definite place to begin. It makes a statement about a topic and then asks for a personal expression of a writer's own views on it. Many writers in the sample [redacted] used the statement in one way or another as a means of introducing their own positions. In fact, this seems to be the major difference--perhaps the one critical difference--between Mode 2 and the other presentation modes: it offers a ready method of organizing an essay by providing a kind of pre-established path along which writers may channel their thoughts on a topic. In short, it supplies enough structure for writers to begin writing quickly and purposefully.

Mode 3, the mode establishing a full rhetorical context, apparently wasn't very helpful to examinees in organizing and focussing their writing. Quite a few got more caught up in the format required, especially when a personal letter was called for, than in the development of their ideas on the topic. Many failed to do more than rehash the information given in the scenario; perhaps the information given acted as a boundary rather than as the stimulus it was intended to be. Perhaps too the hypothetical situations were, because of their locus in fixed, real-world events, inadequate introductions to the task of writing personal statements on issues conceived of originally as large-scale. That is, asking for a statement based on a particular event or situation may have elicited shallower responses than asking for a position on a general issue.

Recommendations

Mode. The evidence gathered in the statistical and the rhetorical analyses points clearly at Mode 2 as the presentation mode likeliest to stimulate the best writing of large numbers of examinees. The Mode 2 format is thus the preferred format for the writing subtest of the Florida Teacher Certification Examination.

Topic. Of the six topics generated and validated in this study, two of them--produced essays of higher quality in the preferred mode than the others. It is recommended that these topics definitely be among those used in the first administrations of the examination.

Rhetorical Modifications. Because of the apparent effect of Mode 2, Topic 6 on the quality of essays, it is recommended that all the topics used in the first administrations of the examination be worded as closely as possible like that of Mode 2, Topic 6. Such modifications will bring into line the particular charge of each topic and will offer a fairer writer-to-writer test of compositional skill.