DOCUMENT RESUME

ED 213 728 TM 820 056

AUTHOR Baker, Eva L.; Quellmalz, Edys
TITLE Results of Pilot Studies: Effects of Variations in Writing Task Stimuli on the Analysis of Student Writing Performance. Studies in Measurement and Methodology. Work Unit 1: Design and Use of Tests.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Nov 79
GRANT OB-NIE-G-78-0213
NOTE 99p.

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Essay Tests; Pictorial Stimuli; *Scores; *Scoring; Secondary Education; Testing Problems; Visual Stimuli; Writing (Composition); *Writing Evaluation
IDENTIFIERS Curriculum Test Overlap; Inter Rater Reliability

ABSTRACT

Three pilot studies--effects of writing prompt modality on writing performance; effect of topic familiarity; and effect of topic, sample and rater group membership on the stability of scoring criteria application--are used to identify the variability in student writing performance. A writing competency test aims to assess writing-specific skills and, in preparation for a study of the effects of writing prompt modality on essays written in two modes of discourse, this investigation identifies essay scoring criteria sensitive to probable effects of the different modalities. The following scoring criteria are identified: (1) Do the different writing modalities affect essay writing on Analytic Expository Scale subscales? Do differences occur on the General Impression, Organization, Support, and Total Score ratings? (2) Does prompt modality affect essay length? (3) Does prompt modality influence the number and types of facts included in the essay? and (4) Does prompt modality affect syntactic fluency? Since the final study examines eighth-grade writing performance, the pilot study examines the appropriateness of an analytic scoring rubric previously employed with high school and college level writing. Sample essay topics and rating formulas are also included. (CE)

DELIVERABLE - November 1979

STUDIES IN MEASUREMENT AND METHODOLOGY

Work Unit 1: Design and Use of Tests

Eva L. Baker and Edys Quellmalz
Project Directors

Results of Pilot Studies: Effects of Variations in Writing Task Stimuli on the Analysis of Student Writing Performance

Pilot Study #1 - Effects of Writing Prompt Modality on Writing Performance

Pilot Study #2 - Effect of Topic Familiarity

Pilot Study #3 - Effects of Topic, Sample and Rater Group Membership on the Stability of Scoring Criteria Application

GRANT # - OB-NIE-G-78-0213

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California - Los Angeles

PILOT STUDY #1

EFFECTS OF WRITING PROMPT MODALITY ON WRITING PERFORMANCE

Researchers have directed little systematic study to the effects of writing prompt modality on written production. Given the body of research identifying the variability in student writing performance from topic to topic, (F. Godshalk, Swineford and Coffman, 1973) CSE will conduct a series of studies aimed at reducing variability of student writing performance attributable to lack of information about a topic. A writing competency test aims, after all, to assess writing-specific skills. Availability of sufficient content to formulate a written response is a necessary resource, but distinct from the writing production skills of interest.

Test writers' primary strategy for reducing error variance due to accessibility of topic-related information has been to offer topics "familiar and interesting to students". The most common modality of writing prompt has been the written form which instructs examinees to write about a topic or, in some cases, presents limited information about the content and the aspects of it required for the exam. Less frequently, the visual modality has been used to present one or more pictures to stimulate writing.

From the perspective of cognitive information processing, the modality of the prompt may serve not simply to satisfy the preference or whim of the test writer, but to provide to the examinee quite different amounts and forms of input about required essay content. Hoping a topic is "interesting and familiar" to students risks permitting irrelevant variability among students. Research certainly provides empirical evidence of

large error variance due to subject-by-topic interactions (Godshalk, et al., 1966; Spooner-Smith, 1978; Pitts, 1978; Winters, 1978; Quellmalz, 1979). If multiple writing topics are intended to represent parallel, homogeneous items for sampling a domain of writing skill, then tests should attend to means of presenting the writing topic which will minimize individual differences attributable to topic familiarity.

A related issue in the use of written prompts is the level of reading comprehension required. Students with low comprehension skills may perform at a low level because of their poor comprehension of the topic requirements rather than because of their actual writing ability. A visual prompt might avoid some of the limitations of written prompts by providing the learner with all or most topic-relevant information and perhaps even assisting in the organization of that information.

In preparation for a study of the effects of writing prompt modality on essays written in two modes of discourse, the present investigation attempted to identify dimensions of essay scoring criteria that would be sensitive to probable effects of the different modalities. The potential research questions suggest a variety of dependent measures and, hence, scoring criteria:

1) Do the different writing modalities affect essay rating on CSE Analytic Expository Scale subscales? In particular, do differences occur on the General Impression, Organization, Support, and Total Score ratings?

2) Does prompt modality affect essay length?

3) Does prompt modality influence the number and types of facts included in the essay?

4) Does prompt modality affect syntactic fluency?

Since the final study will examine eighth-grade writing performance, the pilot study also examined the appropriateness of an analytic scoring rubric previously employed with high school and college level writing.

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Table of Contents

PILOT STUDIES

EFFECTS OF VARIATIONS IN WRITING TASK STIMULI ON THE ANALYSIS OF STUDENT WRITING PERFORMANCE

Technical reports on assessment instruments seldom include specifications that clearly describe the stimulus and response attributes of the test's design. It is not clear why these design features are absent; perhaps they were never considered by the test designers, or perhaps they are simply not documented in the technical reports. In an attempt to move test design from craft to science, CSE continues to investigate dimensions of test design that must be considered in producing valid and reliable measures. This report describes three pilot studies which examine influences on writing scores of writing prompt modality, level of topic information, and topic and subject variability.

The first pilot study explores the nature and structure of dependent measures sensitive to characteristics of student writing elicited by pictorial and written stimuli.

The second pilot study, in preparation for research attempting to control for level of topic knowledge, investigates techniques for analyzing the role of content in essay quality.

The third study recognizes that written samples, presumed to be direct measures of writing skill, are in fact mediated by rater judgement. The study examines techniques for detecting the influences of topic and sample variability on raters' application of scoring criteria.

Specifically, the pilot study asked:

1) Do eighth-grade essays vary on scores for essay and paragraph organization?

2) Do eighth-grade essays vary on elements within a mechanics subscale rating?

Method

Subjects. Essays were gathered from eighth-grade students attending schools in the planning or implementation stage of the California School Improvement Program. Within each school, which was representative of California schools on such dimensions as socioeconomic level, the principal was asked to identify four non-tracked, heterogeneously grouped classes and to select randomly two classes to complete writing tasks.

Writing tasks. Students wrote either an expository or narrative essay in response to an entirely written description of the topic or in response to a topic description referenced to a picture. For the purposes of this pilot of scoring techniques, 15 expository essays stimulated by a written prompt (EWP) and 15 expository essays stimulated by a picture prompt (EPP) were selected for exploratory analyses.

The expository written prompt (EWP) asked junior high school students to explain to elementary students how junior high differs from elementary school. The pictorial prompt (EPP) asked students to write a report for the city building department describing the condition of an old house portrayed in a picture. The picture displayed four panels, one of the outside of an old house in disrepair, and three panels of separate rooms in the house. Copies of the two expositiory writing tasks appear in Appendix A.

Test Administration. The teacher distributied the exams and read

each set of directions to the class. Forty minutes were alloted for completion of the essay. Student names were not written on the essays; each essay received a numerical code identifying the school, class, and student.

Dependent Measures

Introduction. The primary purpose of the pilot study was to explore the configuration of dependent measures which would capture any differences due to prompt modality and would reflect characteristics of the essays representative of eighth-grade writing. This section describes the amended CSE Analytic Expository scale restructured for the study and the other measures designed to detect the influence of the prompt modalities.

CSE Analytic Expository Scale. Previous CSE writing studies have employed domain-referenced scoring criteria related to explicit essay elements. Spooner-Smith developed two expository scales for rating high school essays. One scale called for a single general impression (GI) judgment of the essay according to the qualities of expository writing it demonstrated. The second scale required analytic rating of essay focus, organization, development, paragraphing, support, and mechanics (Spooner-Smith, 1978). Winters also employed the GI and Analytic scales to compare the effects of alternative scoring systems in competence decisions. She found that the GI rating contributed information quite different from the analytic ratings (Winters, 1979). In a subsequent study, Quellmalz, therefore, added a GI rating to the analytic scales and collapsed Spooner-Smith's Organization, Development, and Paragraphing subscales into one Organization subscale (Quellmalz, 1979).

Preliminary readings of the eighth-grade essays collected for the present pilot study suggested that the essays written by these younger writers differed along several dimensions from essays written by high school and college students.

1) The skill with which the eighth graders organized paragraphs differed from the overall essay cohesiveness.

2) The eighth-grade papers differed from each other and from upper-grade essays in their application of the skills subsumed in the previous rubrics' "Mechanics" subscale. Within and between eighth-grade papers there appeared to be variability in sentence construction, usage, spelling, punctuation, and capitalization.

A third issue in establishing criteria which would capture salient features of eighth-grade writing related to the standard of competence against which the General Impression rating would be awarded. Inherent in the definition of a criterion/domain-referenced test is the existence of an operational referent. Thus, a General Impression quality rating should refer to explicated skill dimensions in the domain of competent expository writing. Competence judgments should not be susceptible to relative interpretations of quality "for eighth-graders," "for high school students", or "for highest quality professional writing". For example, a criterion may be that "a main idea is clearly stated or implied", or that "generalizations are supported by specific concrete details". Junior high and college level writing may vary by how support is achieved, by the type, number, or variety of techniques employed, but the presence of main idea or support can still be uniformly judged. In this study, consequently, the GI rating was revised to reflect four levels of competency judgment implying levels of instructional intervention, ranging from "4-very competent, requires little or no additional instruction on basic skills" to "1-not competent, requires extensive remediation". (See Appendix B.)

Use of prompted and unprompted facts. One hypothesized influence of prompt modality was the amount of topic information available for the student to weave into the essay. In this pilot study, the written prompt provided minimum cues about the nature of information desired (e.g., differences between junior high and elementary school) and no information about the actual differences students had to retrieve from memory (i.e., the relevant categories of information or specific details related to the categories.). The picture prompt, however, portrayed most information necessary for the report requested on the condition of the house. Students with all relevant information readily accessible might incorporate more information into their essays (essay length) and marshal more support for their generalizations (support ratings).

To determine the extent to which students used information from the prompts in their essays, two information categories-- prompted and unprompted facts--were devised. Prompted facts were defined as those bits of information provided by the written or pictorial prompt. Raters constructed a "fact list" for each prompt, and a rating sheet for recording the number of prompted and unprompted details. Materials constructed for the fact counts appear in Appendix C.

T - Unit Analysis. A T-unit analysis was also performed on each essay. This analysis addressed the hypothesis that students with more readily accessible information (picture prompt) would struggle less with content and perhaps produces more fluent, facile sentences. Procedures for the T-unit analysis appear in Appendix D.

Procedures

CSE Analytic Expository Scale. Essay scoring according to the CSE rubric followed methods employed in previous CSE writing studies (Spooner-

Smith, 1978; Winters, 1978; Quellmalz, 1979). Two raters practiced applying scoring criteria on approximately 30 papers for approximately four hours. Raters then scored 15 pilot papers to assess interrater reliability.

Based on these results, raters refined decision rules for Usage ratings and checked subsequent agreement on eight additional papers. Final rater agreements included Alpha coefficients ranging from .43 to .75 on the subscales, and .71 on the Total score. One subscale, Sentence Construction, had an extremely low alpha of .11. Variance estimates attributable to raters ranged from .00 to .60 for the subscales. Appendix E presents the relevant tables.

Prompted and unprompted facts. Two raters referred to the Prompted Facts list to practice counting facts on approximately 10 essays. The raters then independently counted facts in the 30 essays previously scored according to the CSE Analytic Expository rating scale. Rater agreement as indicated by the Alpha coefficient was .96 and the estimated variance attributable to raters was 4.78. Generalizability statistics also appear in Appendix E.

T - Unit Analyses. Following procedures similar to those for the fact count, the two raters practiced T-Unit analyses together, then rated the 30 essays independently. Alpha coefficients ranged from .91 to .98, and rater variance estimated ranged from .08 to 2.1. Statistics for the generalizability coefficient appear in Appendix E.

Findings

Findings of the pilot study relate to hypothesized treatment effects and type and structure of dependent measures.

Effects of Prompt Modality. Analyses of preliminary hypotheses regarding effects of writing prompt modality contrast CSE Analytic Expository Scale scores, prompted and unprompted fact counts and T-unit data. A comparison of the means and standard deviations of CSE ratings for the two prompt modality groups appears in Table 1.

(Insert Table 1 about here)

Essay scores for students writing in response to the pictorial prompt (EPP) did not differ significantly from the EWP scores on the General Impression or Total scores. The EPP scores were significantly higher than EWP scores, however, for Essay Organization ($p < .006$), and Support ($p < .034$).

Table 2 presents the means and standard deviations of each group's use of facts. The EPP group used significantly more prompted facts ($p < .00$) than the EWP group and more total facts ($p < .04$). Conversely, the EWP group used significantly more unprompted facts ($p < .01$).

(Insert Table 2 about here)

Table 3 presents the means and standard deviations of the T-unit analyses. The groups did not differ significantly from each other on the total number of words in their essays nor in the number or length of T-units.

(Insert Table 3 about here)

TABLE 1

Comparison of
Means and Standard Deviations of CSE Essay Scores
for Written and Picture Prompt Groups

		Picture Prompt	Written Prompt	t Value
General Impression (4)	$\bar{X}$	2.1	2.0	.44
	s.d.	.573	.683	
Focus (4)	$\bar{X}$	2.17	1.94	1.08
	s.d.	.53	.66	
Organization Essay (4)	$\bar{X}$	2.63	1.97	2.95**
	s.d.	.61	.65	
Organization Paragraph (4)	$\bar{X}$	2.27	2.03	.67
	s.d.	.96	.99	
Support (4)	$\bar{X}$	3.13	2.31	2.23*
	s.d.	.99	1.06	
Mechanics: Sentence Construction (4)	$\bar{X}$	2.20	2.41	-.81
	s.d.	.73	.69	
Usage (4)	$\bar{X}$	2.90	2.72	.63
	s.d.	.76	.84	
Spelling (4)	$\bar{X}$	3.13	3.16	-.07
	s.d.	.92	.96	
Punctuation and Capitalization (4)	$\bar{X}$	2.63	2.19	1.60
	s.d.	.66	.87	
TOTAL	$\bar{X}$	23.17	20.72	1.48
(36)	s.d.	4.14	5.03	

* $p < .05$
** $p < .001$

Table 2

Comparison of
Means and Standard Deviations of Facts
Used in Essays by Written and Picture
Prompt Groups

		Picture Prompt	Written Prompt
Number of Prompted facts	X	17.8	1.94
	s.d.	8.12	.93
	t	7.52**	
Number of Unprompted facts	X	9.67	18.44
	s.d.	4.43	7.65
	t	-3.94**	
Total number of Facts	X	27.13	20.38
	s.d.	9.53	7.51
	t	2.18*	

* $p < .05$
** $p < .01$

Table 3

Comparison of
Means and Standard Deviations
of T-Unit Analyses

		PICTURE PROMPT	WRITTEN PROMPT	t-VALUE
Number of Words	$\bar{X}$	193.23	173.81	.78
	s.d.	75.96	60.62	
Number of T-Units	$\bar{X}$	20.10	16.22	1.54
	s.d.	8.43	5.08	
Number Words per T-Unit	$\bar{X}$	9.97	10.84	-1.20
	s.d.	1.89	2.18	

Analyses of dependent measures

As a major purpose of the pilot study was to identify and refine appropriate means for characterizing eighth-grade writing and for detecting possible treatment effects, the relationships of variables within and between measures were examined. Tabel 4 presents the intercorrelations of CSE essay scores within each treatment group. The General Impression rating is moderately related to the structural scales of Essay and Paragraph Organization and Support. The Focus scores for the Picture Prompt group are weakly related to the GI rating (.30); but strongly associated with the Paragraph Organization (.73). Essay Organization for the Picture Prompt group is strongly associated with the Support (.85) rating and moderately related to Paragraph Organization scores (.66). Essay Organization scores for the Written Prompt group are also strongly related to Support scores (.77), but considerably less so than in the Picture Prompt group (.85). Support scores, presumably likely to reflect Picture Prompt effects, are related more strongly to GI and Essay Organization (.76, .85) than are Written Prompt group scores (.57, .77). Support is highly related to Total scores (.68, .70) of both groups. Mechanics ratings are variably associated with each other and the structural subscales within and between treatments.

Table 5 displays correlations among all the dependent measures. Contrary to hypothesized effects, the General Impression scores are not associated with the number of prompted facts used for either treatment group. For the Picture Prompt group, the number of unprompted facts seems to be strongly associated with the GI (.71) and Total (.73) scores.

Table 4

Intercorrelations Among CSE Analytic

Expository Scale Subscales

for Picture and Written Prompt Groups

	G I	Focus		Organization				Support		Mechanics								Total	
				Essay		Paragraph				Sent. Constr.		Usage		Spelling		Punct.			
Treatment Group		1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
CSE Scale																			
General Impression		.30	.74	.72	.75	.73	.54	.76	.57	.08	.50	.35	.35	.25	.15	.20	.56	.80	.81
Focus				.48	.71	.95	.67	.51	.51	.36	.32	.49	.06	.21	.12	.07	.63	.46	.75
Organization Essay						.66	.65	.85	.77	.31	.37	.34	.11	.22	.04	.22	.51	.69	.78
Paragraph								.65	.73	.12	.56	.28	.25	.34	.02	.19	.44	.79	.79
Support										.41	.36	.30	.11	.14	.01	.14	.20	.68	.70
Mechanics																			
Sentence Construction												.14	.56	.49	.12	.53	.45	.24	.68
Usage														.64	.43	.10	.42	.66	.54
Spelling																.55	.46	.72	.40
Punctuation & Capitalization																		.39	.75
TOTAL																			

Group 1 = EPP - Expository Picture Prompt
Group 2 = EWP - Expository Written Prompt

TABLE 5

Correlations Between CSE Essay Scores

and Fact and T-Unit Analyses

by Treatment Group

	Prompted Facts		Unprompted Facts		Total Facts		Total Words		#T-Units		#Words/T-Units	
	Picture	Written	Picture	Written	Picture	Written	Picture	Written	Picture	Written	Picture	Written
G.I.	.16	.18	.71	.17	.41	.19	.28	.09	.33	-.02	-.07	.20
Focus	-.18	.21	.23	.01	-.03	.04	-.12	.00	-.30	-.10	.46	.07
Organization --Essay	.11	.41	.45	.04	.25	.09	.24	.04	.19	.01	.11	.01
--Paragraph	.21	-.11	.61	.03	.42	.02	.35	-.09	.34	-.06	.03	-.16
Support	.10	.17	.45	.31	.26	.34	.27	.14	.29	.15	-.02	-.09
Mechanics --Sentence Construction	.03	.02	.27	.03	.14	.04	-.02	.08	.06	-.10	-.16	.21
--Usage	-.34	-.09	.58	.13	-.05	.12	.08	-.02	-.10	-.22	.40	.25
--Spelling	-.34	.27	.40	-.19	-.14	-.16	-.15	-.32	-.29	-.42	.30	.10
--Punctuation	-.37	.16	.29	-.25	-.20	-.23	-.31	-.21	-.42	-.28	.27	-.05
TOTAL	-.10	.19	.73	.05	.20	.07	.13	-.06	.04	-.17	.22	.07

Notably, the number of prompted facts used have little or no association with Support or other essay ratings on the CSE scale, with the exception of Essay Organization scores in the Written Prompt group. For the Picture Prompt group, total number of facts and unprompted facts are weakly related to all structural subscales but Focus. CSE total scores have no relationship to any of the other dependent variables. T-Unit analyses also do not relate to any of the other measures.

Discussion

Effects of prompt modality. The pilot data suggest that the prompt modalities are influencing the form and content of student essays. Hypothesized differences on overall essay quality as indicated by GI and Total scores did not materialize in the pilot scoring. It may be that the modality will not affect all students. The final study will further examine modality effects on students who differ in verbal and writing ability indicated by other independent measures. Less verbal students may be the ones who would benefit most from having topic-related information available so that they can concentrate their processing strategies on structuring the essay. Also, a larger sample size will permit analyses more sensitive to the factors contributing to score variances. Hypothesized mean differences in Essay Organization and Support scores favoring the Picture Prompt group did occur. However, the low correlation in the Picture Prompt group between the use of Prompted Facts and Support rating (.10) or either Organization rating (.11, .21) cloud the role of the visually presented information. The stronger correlations in the Picture Prompt group between use of unprompted facts and GI (.71), Essay Organization (.45), Paragraph Organization (.61), Support (.45), and Total Score (.73) sug-

gest that information other than the concrete details provided in the picture contributed to these scores. Recommendations for further analyses of the nature of prompted and unprompted facts will be presented in the section on Implications for Dependent Measures. Information tallied as unprompted may have been generalizations or inferences drawn from the picture. It may also be, of course, that sheer number of facts used are not as important as how skillfully they are woven into the essay.

While Picture Prompted essays were not significantly longer than essays elicited by Written Prompts, the visually prompted essays did have more facts. The particular topic may have idiosyncratically caused this phenomenon as it elicited many strings of facts about what was wrong with the house. Further studies might reveal that other visual prompts and topics need not stimulate "laundry lists".

The T-Unit analyses did reveal for the Picture Prompt group a slight association between number (but not length) of T-units and scores on GI (.33), Essay Organization (.19), Paragraph Organization (.34) and Support (.29). Whether availability of topic-related information affects fluency is still somewhat unclear.

Implications for Dependent Measures

CSE Analytic Expository Scale. The new General Impression criteria emphasizing instructionally-referenced competency judgments correlated strongly with the Total score and the other structural criteria. Additional data would be necessary to test corroboration of the GI's utility as a global competency judgment. The two organization subscales correlate moderately well with each other (.65) and with the Total score (.78, .79). They do seem to capture different aspects of the essay and warrant further application to eighth grade writing. The Mechanics subscales do seem to differ from each other but do not seem sensitive to the treatment interventions. It may be that the detailed Mechanics scores, which add additional rating time, are not useful for detecting treatment effects and should be dropped for the final study ratings. Their diagnostic utility could be further examined in a differently oriented study. The Sentence Construction subscale did not relate to any of the T-Unit measures, suggesting that the sentence construction criteria require refinement.

Prompted and Unprompted Facts. Raters reported a number of problems in operational definitions of "facts". In the larger study, clearer distinctions will be drawn between facts and inferences and between related and directly translated information. A "sensory description" category might be a subscale to consider for capturing visually induced writing.

T-Unit analyses. It is not clear if the very weak correlations of number of T-Units with CSE structural scores for the Picture Prompt group indicates any treatment effects. In general, T-Unit analysis seems insensitive to other variations. T-Unit analysis of good and poor papers might be considered for a small portion of essays in the final study.

Conclusions

In general, the measures seem appropriate for detecting possible treatment effects. Fact check procedures will need refinement. Increased sample size and topic generalizability should allow further clarification of the contributions of the treatments and scoring schemes.

References

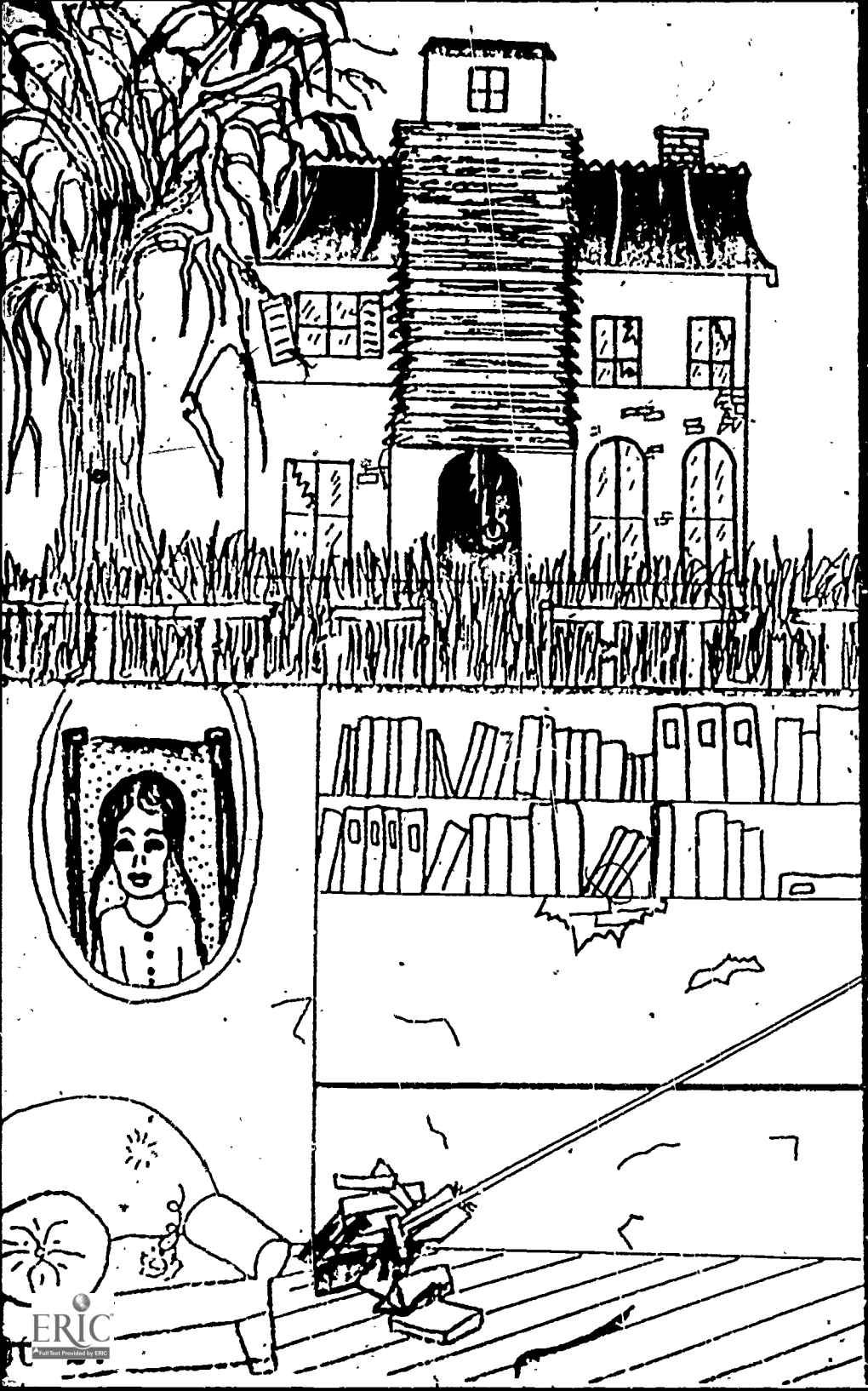
Godshalk, F. I., Swineford, F. and Coffman, W. E. The measurement of writing ability. New York: College Entrance Examination Board, 1966.

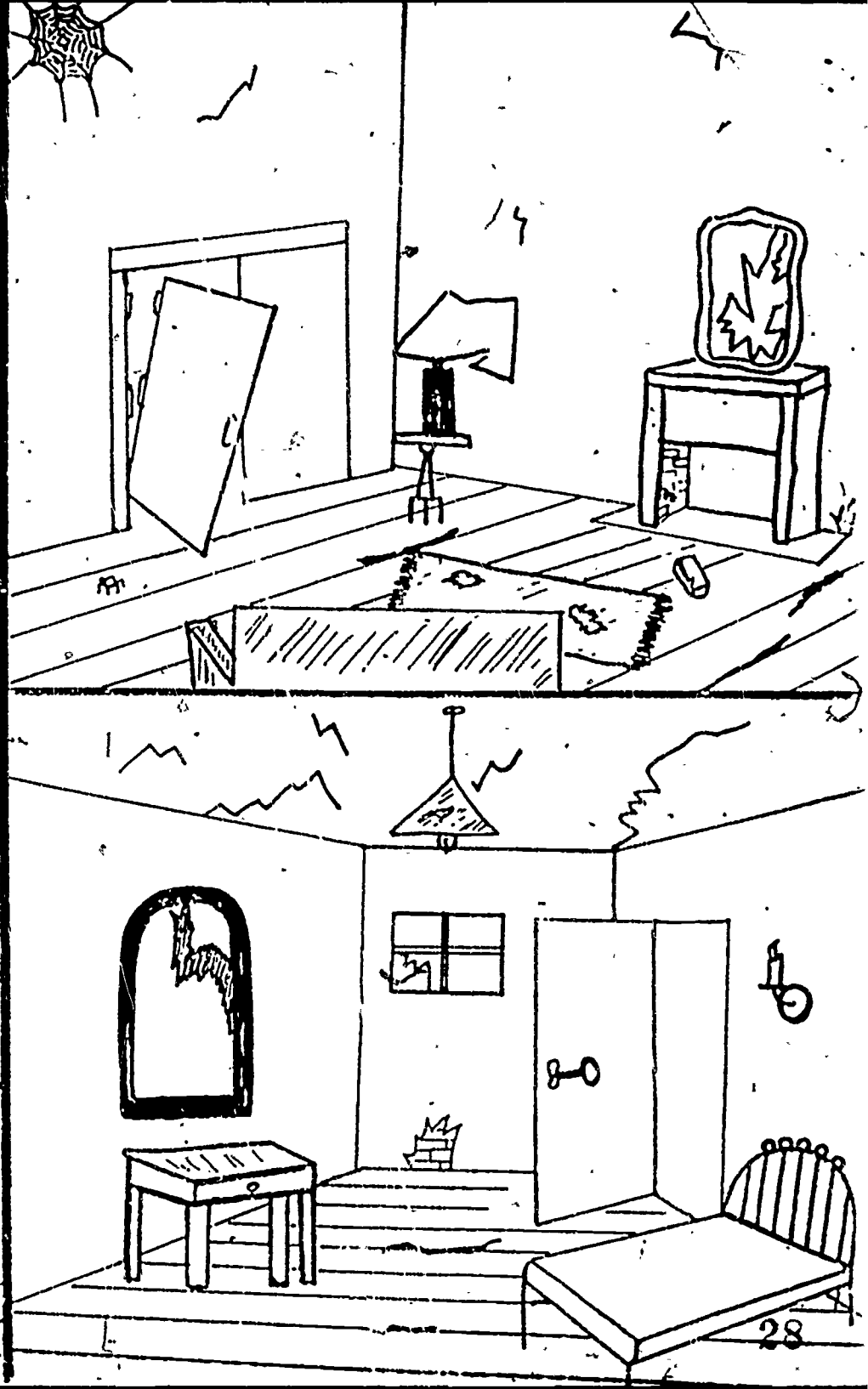
Pitts, M. The relationship of classroom instructional characteristics and writing in the descriptive/narrative mode. Unpublished project report prepared under grant number OB-NIE-G-78-0213. Los Angeles, California: Center for the Study of Evaluation, U.C.L.A., 1978.

Quellmalz, E. Defining writing domains: Effect of discourse and response mode. Unpublished project report prepared under grant number OB-NIE-G-78-0213. Los Angeles, California: Center for the Study of Evaluation, U.C.L.A., 1979.

Spooner-Smith, L. Investigation of writing assessment strategies. Unpublished project report prepared under grant number OB-NIE-G-78-0213. Los Angeles, California: Center for the Study of Evaluation, U.C.L.A., 1978.

Winters, L. The effects of differing response criteria on the assessment of writing competence. Unpublished project report prepared under grant number OB-NIE-G-78-0213. Los Angeles, California: Center for the Study of Evaluation, U.C.L.A., 1978.





APPENDIX A

READING KEY

1.	A	B	C	D	20.	A	B	C	D
2.	A	B	C	D	21.	A	B	C	D
3.	A	B	C	D	22.	A	B	C	D
4.	A	B	C	D	23.	A	B	C	D
5.	A	B	C	D	24.	A	B	C	D
6.	A	B	C	D	25.	A	B	C	D
7.	A	B	C	D	26.	A	B	C	D
8.	A	B	C	D	27.	A	B	C	D
9.	A	B	C	D	28.	A	B	C	D
10.	A	B	C	D	29.	A	B	C	D
11.	A	B	C	D	30.	A	B	C	D
12.	A	B	C	D	31.	A	B	C	D
13.	A	B	C	D	32.	A	B	C	D
14.	A	B	C	D	33.	A	B	C	D
15.	A	B	C	D	34.	A	B	C	D
16.	A	B	C	D	35.	A	B	C	D
17.	A	B	C	D	36.	A	B	C	D
18.	A	B	C	D	37.	A	B	C	D
19.	A	B	C	D	38.	A	B	C	D

NAME _____ AGE _____

SCHOOL _____ ROOM _____ SEX: M F

SUBJECT _____ PERIOD _____ TEST FORM: A B

WRITING SAMPLE DIRECTIONS

The city is planning to fix up this old house in the picture. Pretend you have been asked by the city Building Department to write a report explaining the condition of the house. Use the picture on the page of this page to help you.

- The purpose of your report is to help the city Building Department understand what shape the old house is in. Give them information about the house, not just your own opinions.
- Use details and facts to support your ideas.
- Be sure your report has a main idea, sticks to the topic and is well-organized.

READING KEY

1.	A	B	C	D		20.	A	B	C	D
2.	A	B	C	D		21.	A	B	C	D
3.	A	B	C	D		22.	A	B	C	D
4.	A	B	C	D		23.	A	B	C	D
5.	A	B	C	D		24.	A	B	C	D
6.	A	B	C	D		25.	A	B	C	D
7.	A	B	C	D		26.	A	B	C	D
8.	A	B	C	D		27.	A	B	C	D
9.	A	B	C	D		28.	A	B	C	D
10.	A	B	C	D		29.	A	B	C	D
11.	A	B	C	D		30.	A	B	C	D
12.	A	B	C	D		31.	A	B	C	D
13.	A	B	C	D		32.	A	B	C	D
14.	A	B	C	D		33.	A	B	C	D
15.	A	B	C	D		34.	A	B	C	D
16.	A	B	C	D		35.	A	B	C	D
17.	A	B	C	D		36.	A	B	C	D
18.	A	B	C	D		37.	A	B	C	D
19.	A	B	C	D		38.	A	B	C	D

NAME _____ AGE _____ SEX: M F

SCHOOL _____ ROOM _____ TEST FORM: A B

SUBJECT _____ PERIOD _____

APPENDIX A (cont.)

WRITING SAMPLE

Students sometimes feel a little nervous when they graduate from elementary school and move up to junior high. They must get used to a new situation, new friends and teachers, and different rules. Most students graduating from elementary school would like to have someone tell them what to expect and how to fit in in the upper grades.

Pretend that you have been asked to write about this topic for the students who will be coming to your school next year. WRITE AN ESSAY IN WHICH YOU EXPLAIN SOME OF THE WAYS YOUR SCHOOL IS DIFFERENT AND HOW THE NEW STUDENTS SHOULD DEAL WITH THESE NEW AND DIFFERENT THINGS IN THE UPPER GRADES.

- The purpose of your essay is to help others understand by giving them information, not just your own opinions.
- Use details and facts to support your ideas.
- Be sure your essay has a main idea, sticks to the topic and is well-organized.

APPENDIX B

Expository Scale

ELEMENT 1

Impressionistic Rating Procedures

The purpose of Impressionistic Rating is to form a single impression of a piece of writing as to how well it communicates a whole message to the reader. Impressionistic scoring assumes that each characteristic that makes up an essay -- organization of ideas, content, mechanics and so on -- is related to all other characteristics. Impressionistic scoring further assumes that some qualities of an essay cannot easily be separated from each other. In short, the procedure views a piece of writing as a total work, the whole of which is greater than the sum of its parts.

Discerning readers naturally will attend to, or be influenced by, some essay characteristics more than others. In the Impressionistic scoring, however, readers should arrive at a judgment regarding the overall quality of the essay.

For this element, you are being asked to form an overall impression concerning the effectiveness of the essays as examples of expository writing.

The Topics

You will be reading essays on two different topics. Both topics, though, were designed to elicit writing in the expository mode. Some views on exposition are given below:

- Exposition is the kind of discourse that explains or clarifies a subject.
- Exposition seeks to explain or inform through such methods as giving reasons or examples, comparing and contrasting, defining, ennumerating or through a combination of methods.
- Exposition explains why or how.
- Exposition promotes reader understanding of a subject.

APPENDIX B (cont.)
Expository Scale

General Impression

You are to read each essay quickly; first in order to form an overall impression of its quality. To assign the essay a score, consider the following question: To what extent is the essay an example of effective exposition?

Assign each paper a mark of 1 - 4 using the scale below:

4 = Very competent	- Few or no flaws. Demonstrates command of basic narrative skills, Requires little or no instruction on basic skills.
3 = Adequately competent	- A few obvious flaws imply a need for limited instructional attention.
2 = Not competent	- Many obvious flaws. Needs remediation in many skill areas.
1 = Not competent	- Requires extensive remediation in basic writing skills.

APPENDIX B (cont.)

Expository Scale

ELEMENT 2

Essay Focus/Main Idea

The introduction (if deductively structured) or conclusion (if inductively structured) of the essay clearly indicates the subject and main idea of the whole essay.

4. The introduction (and/or conclusion) of this paper clearly conveys the main idea of the whole essay. It also limits the topic by alerting the reader to the key points covered in the body of the essay.

   Specifically, in the introduction (and/or conclusion):

   a. The subject of the essay is clearly identified.

   b. The main idea of the whole essay is clearly stated or implied.

   c. The topic is clearly limited. That is, key points (e.g., reasons, ideas) or major line(s) of reasoning treated in the essay are identified or summarized.

3. The introduction (and/or conclusion) of this paper conveys the main idea of the whole essay. It sets limits on the topic, but does not clearly suggest how the main idea is developed.

   Specifically, in the introduction (and/or conclusion):

   a. The subject of the essay is clearly identified.

   b. The main idea of the whole essay is clearly stated or implied.

   c. An attempt is made to limit the topic. That is, the number -- or type -- of key points is specified, but there is not clear reference to the substantive issues treated in the body of the essay.

2. The introduction (and/or conclusion) of this paper gives the reader a fairly clear sense of the main idea of the whole essay. However, neither the introduction nor the conclusion help focus -- or bring direction to -- the body of the paper.

   Specifically, in the introduction (and/or conclusion):

   a. The subject of the essay is identified.

   b. The main idea of the whole essay is stated or implied.

   c. No attempt is made to limit the topic.

1. Neither the introduction nor the conclusion is helpful to the reader in obtaining any sense of the main idea of the essay.

   Specifically, in the introduction (and/or conclusion):

   a. The subject of the essay is not clearly identified or there is no reference to the subject

   AND/OR

   b. The main idea of the whole essay is not clearly stated or implied or no reference is made to the main idea, or the reference is confusing.

ELEMENT 3

Essay Organization

The main idea is developed according to a clearly discernible method of organization.

4. The plan by which the essay is structured is readily apparent and consistently applied throughout the essay. Some logical scheme serves as the underlying basis for the overall essay structure. That is, there is an easily discernible logical flow from one idea (subtopic) to the next. Development of ideas is not interrupted by digressions of thought or extraneous material.

   Specifically:

   a. The plan for organizing the essay is clearly established in the beginning, e.g., through reference to the subtopics or line(s) of reasoning to be developed

      AND/OR

      The plan for organizing the essay is readily evident to the reader due to the effective use of paragraphs, transitions and linking expressions.

   b. The plan by which the essay is organized is consistently applied throughout the essay All major subtopics (main points) clearly related to the main idea of the whole essay. There are no digressions of thought or extraneous material.

   c. The overall structure of the essay reflects some logical, underlying organizing rubric. That is, subtopics are presented according to some logical scheme, or order, such as: order of importance, seriousness; comparison and contrast; from specific to general; cause to effect; through classification.

   d. All statements related to the same idea (subtopics) are contiguous, i.e., are presented together as logically related units or blocks of thought.

3. The essay is organized according to some recognizable plan which is employed throughout the essay. Some organizing rubric is used to divide the content/ideas of the essay into units of thought. For example, each subtopic might represent a different reason or category of content. Subtopics, however, are not presented in any particular logical order.

   Specifically:

   a. The plan for organizing the essay is established in the beginning, e.g., through reference to the subtopic(s) or line(s) of reasoning to be developed

      AND/OR

      The plan for organizing the essay is evident to the reader due to the use of paragraphs, transitions and linking expressions.

b. The plan by which the essay is organized is employed throughout the essay. All subtopics relate to the one main idea. There are no major digressions; perhaps some minor digressions.

c. Major subtopics are not presented according to any underlying logical order; they could be "reshuffled" without dramatically altering the overall impact or clarity of the essay.

2. The essay is not structured according to any readily discernible plan. However, the reader is able to infer some association between major ideas. The logical flow from one idea to the next is so tenuous that the reader must supply his/her own links or transitions to form a clear relationship among ideas.

Specifically:

a. A plan for ordering ideas/details is attempted; the plan is not readily apparent to the reader. The essay does not make effective use of transitions or linking expressions to help the reader follow the trend of thought

AND/OR

b. The plan for ordering ideas/details is not consistently applied throughout the essay.

c. Some subtopics relate to the main idea. There may be major digressions of thought.

1. No plan for organizing ideas and details is apparent. Subtopics are not arranged in any discernible order. The reader has difficulty inferring any relationship among ideas or any pattern of thought.

Specifically:

a. No plan for organizing ideas and details is evident. Paragraphs, transitions and linking expressions, if present, are not helpful in following the writer's train of thought.

b. Few or no subtopics relate to the main idea and/or most or all of the essay is off the topic.

APPENDIX B (cont.)

Expository Scale

ELEMENT 4

Organization - Paragraph

4 - All major units of thought are set off by paragraphs.

3 - Most Subtopics are developed in paragraphs containing logically related support.

2 - A few paragraphs appear in which a subtopic is developed by related sets of supporting statements.

1 - Essay contains no differentiated units of thought, or, conventions of paragraphing are absent or consistently incorrect.

ELEMENT 5

Support

Generalizations and assertions are supported by specific, clear supporting statements.

4. All generalizations and assertions in this paper have logically related support. Supporting statements clearly are at a greater level of specificity than the generalizations they are intended to support. The writer makes effective use of concrete detail to support ideas.

   Specifically:

   a. Generalizations and assertions are supported by more specific statements.

   b. Supporting statements provide specific detail, such as illustrations, examples, facts, anecdotes, and/or employ concrete language to convey ideas.

3. Most generalizations and assertions are supported by more specific, clear and logically related statements.

   Specifically:

   a. Generalizations and assertions are supported by more specific statements.

   b. Many supporting statements provide specific detail, such as examples, facts, ancecdotes, and/or employ concrete language to convey ideas.

2. Support for many generalizations and assertions in this paper is not especially convincing. Supporting statements lack precision, either in language or in the use of specific detail.

   Specifically:

   a. An attempt is made to support generalizations and assertions through the use of statements at a greater level of specificity than the generalization they are intended to support.

   b. Supporting statements lack precision and clarity. The use of specific detail, such as examples, facts, or anecdotes, may be lacking and/or the language may be vague.

1. Support is provided for few or none of the generalizations or assertions. Few (or none) of the supporting statements are at a greater level of specificity than the generalizations themselves.

   Specifically:

   a. Supporting statements are provided for few or none of the generalizations or assertions.

   b. Supporting details, if present, are vague, confusing, or not logically related to the generalizations they apparently are intended to support, and/or the language of supporting statements is imprecise.

APPENDIX B (cont.)

Expository Scale

ELEMENT 6

Mechanics

Sentence Construction

4 - There are few or no major errors in syntax.

3 - There may be a few errors in syntax.

2 - There are numerous, serious syntactic errors.

1 - There are many syntactic errors which interfere seriously with communication.

Vocabulary

4 - There are few or no errors in standard usage.

3 - There may be a few minor errors in usage.

2 - There are numerous major usage errors.

1 - There are many serious errors which interfere with communication.

Spelling

4 - There are few or no spelling errors.

3 - There are a few errors.

2 - There are numerous errors.

1 - There are numerous errors which interfere seriously with communication.

Punctuation and Capitalization

4 - There are few or no errors.

3 - There are a few major errors.

2 - There are numerous errors.

1 - Errors are numerous and interfere with communication.

APPENDIX C

Written Prompt
Junior High School Writing Sample
Prompted Facts List

1. Students sometimes feel a little nervous when they graduate from elementary school and move up to junior high school.
2. They must get used to a new situation.
3. They must get used to new friends.
4. They must get used to new teachers.
5. They must get used to different rules.
6. Most students graduating from elementary school would like to have someone tell them what to expect.
7. Most students graduating from elementary school would like to have someone tell them how to fit in in the upper grades.

APPENDIX C

Picture Prompt
Prompted Facts List

1. City is planning to fix up a house and asked you to report on it.
2. The house is old
3. Two-story house.
4. Chimney on the roof.
5. Roof is old.
6. Grass is tall.
7. Large/old tree.
8. Limbs falling off tree.
9. Windows broken.
10. Shutters falling down.
11. Cracked/broken door.
12. Cracked paint/wallpaper/walls.
13. Broken fence.
14. Old/broken furniture.
15. Living room.
16. Cobweb.
17. Spider on floor.
18. Door off hinge.
19. Torn rug.
20. Lamp on 3-legged table.
21. Torn shade on lamp.
22. Mirror on table or fireplace.

23. Mirror is cracked.
24. Sofa/couch/bench.
25. Cracked/broken floor panels.
26. Study room.
27. Picture of girl with braids.
28. Chair with tear.
29. Spring popping out of chair.
30. Pillow on chair.
31. Bookshelves with books.
32. Empty bookshelf.
33. Fallen bookshelf.
34. Books (or garbage) on floor.
35. Bedroom.
36. Brass bed.
37. 4-legged table.
38. Cracked wall mirror.
39. Open door.
40. Large key or large doorknob on door.
41. Candle on wall stand.
42. Hanging lamp.
43. Torn shade on hanging lamp.
44. Cracked ceiling.
45. Brick on floor.

APPENDIX D

T-Unit Analysis

I. Procedure

A. Count the number of words in the composition and record on the rating sheet.

B. With a pen or pencil, section off all T-units.

C. Count the number of T-units in the composition.

D. Divide the number of words by the number of T-units to obtain the words-per-T-unit score.

II. Segmenting Rules

A. Rules for counting words in the composition

1. Compound nouns written as one word count as one word.
2. Compound nouns written as two words and hyphenated words count as two words.
3. Phrasal proper names count as one word.
4. Dates, such as June 21, July 2, count as two words.
5. Contractions such as he'd; shouldn't, count as two words.
6. Garbles (unintellible graphemes) should not be counted.

B. Rules for identifying T-units

1. T-units are minimally terminable communication units: i.e. a single independent predication together with any aubordinate clauses grammatically related to it.
2. One main clause plus any subordinate or non-clausal structure attached to or embedded in it counts as one T-unit.
3. Simple or complex sentences count as one T-unit.
4. Compound sentences count as two T-units.
5. When two independent clauses are linked by a coordinating conjunction, count the coordinating conjunction as the first element of the second clause.
6. If a fragment can be made into a clause with the addition of one word, add the word and count the fragment as one T-unit.
7. Discard fragments that cannot be made into independent clauses with the addition of only one word.
8. Discard garbles, i.e., unintelligible strings of words.
9. Ignore punctuation.
10. For direct discourse, the first expression after "he said" is counted as the direct object of 'he' and is considered as one T-unit. Marsha said, "I really like you John. However, Clarance's father is a millionaire and I like the idea of Palm Beach (3 units).

APPENDIX E

Pilot Test #1

Data For Estimating Generalizability Coefficients

General Impression

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	.166	30	.798	.423
Rater	-.015	1	.000	
Subject/ Rater	.467	30	.467	

α .67

Focus

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	.067	30	.731	.192
Rater	-.017	1	.645	
Subject/ Rater	.598	30	.598	

α .45

Organization - Essay

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	.385	30	1.195	.647
Rater	-.005	1	.258	
Subject/ Rater	.425	30	.425	

α .73

Organization - Paragraph

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	.667	30	2.047	.650
Rater	.004	1	.581	
Subject/ Rater	.714	30	.714	

α.66

Support

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	.937	30	2.606	.731
Rater	-.024	1	.000	
Subject/ Rater	.733	30	.733	

α .75

Mechanics - Sentence Construction

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	.516	30	1.096	.400
Rater	.559	1	2.726	
Subject/ Rater	.992	30	.992	

α .11

Mechanics - Usage

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	.388	30	1.528	.405
Rater	.387	1	1.952	
Subject/ Rater	.752	30	.752	

α.63

Mechanics - Spelling

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	.605	30	1.891	.641
Rater	-.003	1	.581	
Subject/ Rater	.681	30	.681	

α.69

Mechanics - Punctuation

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	.091	30	1.223	.798
Rater	.009	1	1.306	
Subject/ Rater	1.040	30	1.040	

α.43

Words in Essay

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	4621.648	30	9254.320	.999
Rater	2.122	1	76.790	
Subject/ Rater	11.022	30	11.022	

α .99

T - Units

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	49.131	30	99.913	.982
Rater	.155	1	6.452	
Subject/ Rater	1.652	30	1.652	

α .98

Words Per T - Unit

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	3.871	30	8.470	.906
Rater	.079	1	3.161	
Subject/ Rater	.728	30	.728	

α .91

Prompt Facts

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	93.978	30	192.322	.978
Rater	-.141	1	0.000	
Subject/ Rater	4.365	30	4.365	

α .98

Non - Prompt Facts

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	51.375	30	116.555	.855
Rater	3.584	1	124.903	
Subject/ Rater	13.804	30	13.804	

α .88

Total Facts

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	78.951	30	164.805	.931
Rater	4.774	1	154.903	
Subject/ Rater	6.903	30	6.903	

α .96

PILOT STUDY #2

EFFECT OF TOPIC FAMILIARITY

Introduction

A major project scheduled for the next funding period will investigate the effect of students' knowledge about an essay topic upon the rated quality of student writing (CSE Three Year Plan, 1979). As planned, the proposed study would control the amount of student knowledge on a particular writing topic, varying it by groups. Essays produced by subjects would then be compared both across and within (baseline) groups for effects due to differences in amount of relevant knowledge.

Before embarking upon the research effort, a pilot study was planned and carried out to (a) determine the feasibility of controlling topic knowledge through specific curricular treatment distinctions; and (b) investigate the likely areas of impact of knowledge upon the essay. This latter (b) information would inform the use and possible revision of the essay rating scale planned for use in the later study. In particular, we are considering the need for expanding content-sensitive categories, (e.g., supporting detail), with subcategories of greater specificity.

This report describes the pilot study, presents a brief review of findings and longer discussion of design and analysis implications for the larger planned study.

Study Overview

The pilot study presented a discrete, three week curriculum package on energy to one group of eighth grade students while a comparable group received no such treatment. Both groups received the curriculum-based

posttest and were asked to formulate an essay response to a curriculum-based question on energy sources. Essays were scored for average t-unit length, total number of words, and number of discrete energy facts included. Analyses compared the two groups on posttest scores and each of the three essay content measures.

Focus of Inquiry

The pilot study pursues two lines of inquiry, procedural and theoretic. In the first case, a primary concern was with the quality of the curricular material developed for the study. In particular, two questions on the feasibility and adequacy of the planned study design tested here were:

- Does the curriculum present knowledge not available elsewhere to the sample population?
- Is the curriculum presented at an appropriate level of difficulty for learning to occur?

The second focus of this pilot study was an exploration of the measurement implications and theoretical basis of our planned studies on effects of level of knowledge on writing quality. In particular, we needed information on the kinds of essay features that would be most affected by variations in student writers' level of knowledge about the topic examined. The earlier developed CSE rating scale (Quellmalz, 1979 and Spooner-Smith, 1979) might not be sensitive to those very changes we wished to measure, although, we anticipate the need to expand the obviously related subscale categories of support and focus. Also, the need for adding categories more closely aligned with essay structure was a possibility to be explored in this study. The following questions represent this study's measurement and theory concerns with essay rating.

- Does the content of student essays reflect student level of knowledge as measured by an objective test?
- Are differential levels of topic-relevant knowledge reflected in the syntactic structure of student writing?
- Are differential levels of topic-relevant knowledge reflected in the length and factual density of student writing?

Method

Subjects

The study sampled four heterogeneously grouped history classrooms of eighth grade students. Students were randomly assigned by classroom to either instructional treatment or control groups. The total number of students for whom complete data sets exist is forty; twenty in each group.

Independent Variable

Knowledge of essay topic was the variable under experimental control in this study. Two values of the variable (informed/not informed) were allowed by having one-group of subjects exposed to a topic-specific three week curriculum while a comparable student group did not receive this instruction. Unfortunately, no pretest or other baseline measure were available to us.

The curriculum, developed by Oliver and Niedermeyer (1979) is entitled Energy in American History and consists of pre- and post-test, source-book (test), filmstrips and cassettes, teacher's script and program record sheet. These materials provide fifteen, 50-minute lessons focussing on five student objectives. These objectives cover: (1) energy consumption; (2) energy sources; (3) the energy crunch; (4) the great discoveries; (5) energy in daily life. Table 1 offers the instructional objectives for each of these content areas.

TABLE 1

Instructional Objectives for the

Energy in American History Curricular Modules

I. Energy Consumption--Pupils Identify:

a. trends in the growth of energy consumption in the United States over the last 20 years;

b. the major causal factors for this growth;

c. changes in the proportion of energy used by the four sectors of society (residential, commercial, industrial and transportational);

d. the major causes for changes in the proportion of energy used by each sector.

II. Energy Sources--Pupils Identify:

a. the major changes that have taken place in America's fuel mix (i.e., sources of energy) over the last 200 years;

b. the rise and fall of each of the three major fuel epochs (i.e., wood, coal and oil).

III. The Energy Crunch--Pupils Identify:

a. characteristics of the present day energy crunch, including dwindling supply of fossil fuels, reliance upon foreign oil and high cost of providing energy;

b. major means by which the energy crunch can be remedied, including need for conservation practices and importance of developing alternative sources of energy;

IV. The Great Discoveries--Pupils Identify:

a. landmark energy--related inventions, discoveries and inventors;

b. significant impact of these energy-related discoveries upon the growth of America and upon our patterns of energy use.

V. Energy in Daily Life--Pupils Identify:

a. lifestyle changes in the way people work, run households, transport and entertain themselves, brought about by the changing pattern of energy use in America.

Dependent Variables

Four dependent variables were assessed with two measures. Three of these were drawn from students' essay responses to a question on energy. The fourth measure was students' scores on the curriculum-embedded objective posttest. Both essay and objective test measures were given to all subjects of both groups.

The essay measure allowed students 30 minutes to respond to the following prompt and directions:

> TOPIC: Many changes have taken place in America's energy history. Our energy consumption has risen. Our sources of energy have changed. Our daily lives have changed as a result of energy-related inventions and discoveries.
>
> Pick the one energy-related change you think is most important. Write an essay in which you describe this change and tell why you think it is important. Give as many specific details and facts as you can.
>
> Imagine that you are writing for someone who knows nothing about the history of energy in America. Tell them everything they need to know about the energy change you have selected.

Essays produced by each group were rated on three variables: length of t-units; number of essay facts; total number of words. A t-unit, or "minimally terminable syntactic unit" (Hunt, 1965), is an independent clause and its embedded or grammatically related components. T-unit length measures the number of words per t-unit. Considerable research supports a positive relationship between t-unit length and measures of writing quality (O'Hare, 1973; Combs, 1976; Howerton et al., 1977). Additionally, t-unit length has been tied to language development and is considered a measure of "syntactic maturity." (See Appendix A.)

We have selected t-unit length as a dependent variable for two reasons. First, small sample size (n=20 group) and our use of intact classrooms

rather than random assignment, created the desire to validate the comparability in writing skills between both groups. T-unit length analysis provides this comparison. That is, we would expect no difference between groups on this measure. Number of essay facts is also a dependent variable extracted from student essays. For this study we have defined facts to include all energy-related statements, generalizations or factual details regardless of accuracy, appropriateness and source.

The third essay variable, number of words, reflects the effect of knowledge upon the essay length. All students had 30 minutes to read and respond to the essay prompt. Assuming similarity between groups in writing ability and other individual traits, group differences in essay length should be due to level of knowledge. That is, students who are more familiar with and better informed on the essay topic will be able to respond more readily to the task. This may be due to their more quickly accessing information or structuring their responses, or it may be that they really do have more to say.

The fourth dependent variable is the curriculum-based posttest given to both the experimental and the control group. It consists of 25 objective test items drawn from the five content areas described in Table 1. These tests were scored for total number correct, using the answer key provided by the curriculum developers. (A copy of the posttest is in Appendix B of this report.) Obviously we expected the instructed gr to outperform the uninstructed group, confirming the possibility of controlling student knowledge in a topic area.

Procedures

In Spring, 1970, the pilot study was run using eighth grade students in the San Diego (California) City schools. Students in four heterogeneously grouped American history classes were randomly assigned, by class, to either the group receiving the instructional treatment or to the control, no treatment group. Treatment group students' own regular teachers were the instructors, interjecting this curricular package into the regular course work. As planned and carried out, students received three weeks (fifteen lessons) of instruction on Energy in American History in 50-minute lesson segments. Following completion of the course as outlined, students were given a 25-question objective test and a 30-minute essay question on the material covered. Students in the control group also received the same test and essay question.

Essays and objective tests were collected, scored, rated and coded here at CSE. Analyses performed and recommendations drawn up for this report will feed into the design and analysis of the "knowledge of essay topic" study planned and described in the CSE scope of work statement for 1979-1982.

Results

Analyses Performed

T-tests were run between means of the two groups on each of the four dependent variables. Table 2 presents the t-test results, along with means and standard deviations for both groups on the four dependent variables: posttest, essay length, essay t-unit length; number of essay facts.

Description of Findings

As would be expected, the two groups were significantly different in their posttest means. The instructed group mean was 20.8 out of 25

TABLE 2

Means, Standard Deviations, T-Values for Both Groups on the Dependent Variables

Variable		Mean (S.D.)	Standard Error	t-Value* (α)	DF's
Posttest:	Instructed	20.8 (2.1)	0.5	4.6[2]	25
	Uninstructed	15.1 (5.2)	1.2	($\alpha \leq .001$)	25
Essay--Number of Words:					
	Instructed	172.1 (55.1)	12.3	1.93[1]	38
	Uninstructed	135.2 (65.3)	14.6	($\alpha \leq .05$)	38
Essay--t-Unit Length:					
	Instructed	10.73 (1.6)	0.3	1.08[1]	38
	Uninstructed	10.08 (2.2)	0.5	(nsd)	38
Essay--Number of Facts:					
	Instructed	20.1 (6.0)	1.3	3.61[1]	38
	Uninstructed	13.1 (6.4)	1.4	($\alpha \leq .001$)	38

*Note: [1]Pooled variance estimate used

[2]Separate variance estimate used (i.e., F-value significant for tailed probability in variances.

possible (S.D.=2.1), or 83.2 percent correct. The mean for the uninstructed group was 15.1 (S.D. 5.2), or 60.4 percent correct. Note the variance of scores within each group was quite dissimilar. Based upon a significant F=value for the difference, separate variance estimates (df=25) were used to calculate the t-value.

Other significant differences between group means exist for essay length and number of essay facts. The instructed group averaged 172.1 words (S.D.=55.1) per essay; the uninstructed group 135.2 words (S.D.=65.3); the one-tailed probability for the t-value exceeded α=.05. Between group difference on number of essay facts was significant beyond .001. The instructed group mean was 20.1 (S.D.=6.0); the uninstructed group mean was 13.1 (S.D.=6.4).

The fourth variable, t-unit length, was very similar between groups. The means for instructed and uninstructed groups were 10.73 (S.D.=1.6) and 10.08 (S.D.=2.2), respectively. As an indicator of syntactic maturity level and given our assumed comparability of study groups, we expected no difference on this variable.

Discussion

This pilot study was a feasibility probe into the planned procedures for our later study. Results, discussed below, will influence our procedures and measures (previously described in the CSE Three Year Institutional Plan, 1979). In particular, alternative ways to administer or measure the level of treatment should be considered. Further, additional dependent variables are suggested for inclusion in the planned study. We expect to consider these recommendations along with input from our external consultants, to revise and perhaps pilot again the procedures and analyses for the level of knowledge study.

The discussion of results is divided into the two major foci of the pilot effort, procedural concerns and construct/measurement concerns. Each section includes recommendations for the larger, planned study.

Procedural Concerns

This study was designed in part to check whether or not essay topic knowledge could be manipulated, specifically, using the prepared curriculum, Energy in American History. As stated earlier, we wished to know:

(A) Feasibility of controlling topic knowledge:

- Is the curriculum presented at an appropriate level of difficulty for learning to occur?
- Does the curriculum present knowledge unavailable elsewhere to the general sample?

The findings suggested that topic knowledge was in fact manipulated in the study as indicated by the difference in performance between treatment and control groups (See Table 2). The mastery effect on the instructed group yielded, as expected, a restricted range of students' scores on the knowledge ppsttest, inhibiting our use of correlational analyses. It may, therefore, be more appropriate for us to conduct a preliminary dependent variable study, focussing on the relationship among various dependent measures of writing ability and topic knowledge in naturalistic settings before we move to treatment variations, which will undoubtedly result in variance restrictions. Such a study would involve the measurement of general and specific knowledge, potentially, general ability, and provide "baseline" information on writing ability. A second essay would be more specific in its expository prompting of information.

This essay topic might be a specific version of the "baseline" general essay, or on a different topic altogether. In this way, we might expect, through random sampling, to maintain sufficient variation in both writing ability and level of topic knowledge to allow exploration of the relationships among essay quality and knowledge variables.

Such a study would be followed by one in which either criterion levels (through a trials-to-criterion approach) or treatment intensity were specifically administered. We would assess the extent to which such treatments resulted in differential performance on both general and specific essay dependent measures. Of potential interest might also be a corollary study relating the amount of specific knowledge raters possess on their ratings of essay quality.

The manipulation of knowledge levels would also force us to confront analysis issues which inhere in assessments (or evaluations) of instruction where very different distributions occur as a function of a mastery or criterion-referenced instructional approach. Specifically, we would attempt to explore the utility of ranking-and-selection analyses in addition to more traditional correlational approaches for providing insight into the effects of variables of interest. Such procedures have not been well developed for use with multivariate studies, but may be most appropriate to the kind of distribution we encountered in this pilot effort.

Construct/Measurement Concerns

The questions posed originally to explore means for increasing the sensitivity of dependent measures (see page 3), reflected our expectations for collecting data to use in revising currently proposed measures for the 1980 study. T-test differences between the two study groups do provide some basis for recommendations.

Significant differences between group means occurred for essay length (measured in number of words) and for number of "facts." No difference could be found for syntactic maturity (measured by t-unit length).

As a developmental measure, syntactic maturity would not be expected to show the effects of topic knowledge, nor to vary between groups of comparable students at the same grade level. In this study, we viewed the syntactic maturity measure as a second check on the comparability of our student groups' general level of sophistication in writing. With greater sample size, or the incorporation of the baseline essay procedures described above, this measure does not seem necessary or appropriate for the planned study. We recommend it not be used.

The two variables, essay length and number of facts, were selected as our best guesses of where student knowledge might affect student writing on a particular topic. If essay content is mediated by students' knowledge, it may be that selection and synthesis of information are the processes facilitated for students with the greater knowledge base. This facilitating effect may be to increase the ease and/or speed of selecting and organizing/structuring the essay. The overt result of differences in topic knowledge then may be in essay length or in content quality itself. For our pilot, essay length did turn out significantly different between groups (α=.05). As expected, the instructed group wrote longer essays than the control group despite equal time allotment for responding. Although we cannot conclude that this is due to differences between groups in the amount of topic relevant knowlege, we do recommend that essay length be measured in the later study. The measurement of content quality in this pilot study was number of essay facts. Although our

t-test suggests significant differences between groups in the number of facts used, the operational definition of "facts" used to assess content is confounded with the number of sentences. This is due to the lack of distinction between relevant and irrelevant facts, veracity of "facts" and level of abstraction of "facts." We recommend a measure of essay content that distinguishes among these qualities, perhaps even accounting for the use of generalizations and opinions. In addition, we should explore the differential impact of knowledge of topic for expository analytic scales, which emphasize supporting detail, for instance, and more holistic or general impression procedures. We propose such measures as listed above in order to provide cues for developing assessments of writing which may be sensitive to the instructional efforts of writing programs.

References

Howerton, M. C., Jacobson, M., and Elden, R. The relationship between quantitative and qualitative measures of writing skill. Paper presented at the Annual Meeting of the American Education Research Association, New York, April, 1977.

Hunt, K. W. Grammatical structures written at three grade levels. NCTE Research Report No. 3, Urbana, Illinois, 1965.

O'Hare, F. Sentence combining: Improving student writing without formal grammar instruction. NCTE Research Report No. 15, Urbana, Illinois, 1973.

Oliver, L. and Niedermeyer, F.C. Energy in American History Sourcebook, cassettes and filmstrips, teacher's manual, posttest. Developed, 1979.

Spooner-Smith, L. Investigation of writing assessment strategies. Unpublished CSE Report under NIE Contract OB-NIEG-78-0213. Los Angeles: Center for the Study of Evaluation, November, 1978.

PILOT STUDY #3

EFFECTS OF TOPIC, SAMPLE AND RATER GROUP MEMBERSHIP ON THE STABILITY OF SCORING CRITERIA APPLICATION

Researchers, measurement specialists, and practitioners recognize the increasing need to assess samples of complex performance. In the area of language production, writing, oral language, and oral reading represent important developing skills which cannot be tested adequately by selected response formats. Performance sampling of complex human competencies provides direct measures that are presumably more valid than selected response formats. Yet these direct samples of highly variable responses are also affected by the highly variable judgments of raters who must score or characterize the performance samples along some dimensions.

A critical goal for performance measurement in general, and writing assessment in particular, is to refine methodologies for assuring uniform, reliable, valid application of scoring criteria. Just as student writing performance can fluctuate across occasion and topics, rater judgment can vary because of influences irrelevant to the scoring task. Procedures such as rater training and operationalization of scoring rubrics can reduce measurement error attributable to raters (Follman, Anderson, 1977; Winters, 1979). This pilot study examines the influence of three other potential sources of error: composition of rating group membership, variations in the quality of student writing sample, and changes in topic.

Most documented research on the validity and reliability of scoring written production has occurred within a norm-referenced testing framework.

Essays are usually scored holistically, on generally described criteria and involve "norming table" training procedures where raters rank essay by sorting them into piles anchored by the range of quality of that particular sample (Conlan, 1977). Thus, a particular paper's ranking could change from sample to sample if the range of quality of the compositions varied from one rating occasion to the next.

Analogous to the tenet that defendants must receive equal treatment under the law and that rules of evidence must be consistently observed, new competency testing systems must demonstrate equal application of criteria across testing occasions, sets of raters, and topics. Uniform application of performance criteria is a required technical quality of domain-referenced tests; it will also be a legal requirement when decisions based on these tests result in life-altering decisions about students.

Thus, criterion-referenced or domain-referenced tests of writing should reflect judgments referenced to absolute, concrete skill descriptions which are intended to provide a uniform standard. Regardless of how other essays demonstrate skill level, any particular essay score must be referenced to the competency description, e.g., "main idea is clearly stated or implied," "generalizations are supported by specific concrete details."

How, then, can rater judgments be operationalized and standardized? Even with the provision of systematic training and explicit scoring criteria, are rater applications of criteria stable across topics, student samples, and rating group membership? This pilot study attempted to explore these questions by investigating effects of stimulus variation not only in the writing task, but also within the rating context.

The study was preliminary to one examining the effects of writing stimulus student and rating variables on essay scoring. Potential research questions for the planned study are:

1. How stable are ratings of essays across topics?
2. How stable are ratings across student samples?
3. How stable are ratings by raters more and less experienced in scale application?
4. How stable are scores given by raters trained in different rating groups?

METHOD

To examine influences of stimulus variability in the assessment and rating contexts on subsequent scoring, the pilot study compared a small sample of essay scores given by raters in two different rating sessions. Variables of interest for the assessment stimuli were the topic and quality range of student papers. Rating context variables were level of rater experience with the scale and training group membership.

Scores were compared on essays written on three different narrative topics. Two narrative topics were given to high school students in a study (Study A) described by Quellmalz, 1979. The third topic was given to high school students as part of a college admissions exam study (Study B) investigated in a report by Baker and Quellmalz, et al in process. Two rater pairs scored essays in each study. One rater pair participated in both Study A and Study B. Of interest were comparisons of mean scores and levels inter-rater agreement within and between rater pairs and rating occasions.

Figure 1 presents the study design and numbers of essays forming the data base. Note that all essays rated in Study A were included in Study B, and were rated by all four raters. Appendix A presents the three topics.

Figure 1
Rating Occasion

	Study A		Study B		
	Raters x, y	3,4	Raters 1, 2	3,4	
Topics 1	6	8	16	16	16
Topics 2	7	8	16	16	16
Topics 3	-	-	11	11	11

Subjects

Raters. In both studies raters were drawn from the teaching staff of the UCLA English and Subject A Departments. All had extensive experience rating high school essays. Two of the raters, designated Raters 3 and 4, scored essays in Studies A and B. Raters X and Y scored only in Study A. Raters 1 and 2 scored only in Study B.

Writers. Examinees in Study A and Study B varied in their academic achievement and presumably in their writing ability. High school seniors in Study A were drawn from average and above average high school English classes. Mean Differential Aptitude Verbal Test score was 61%; mean PSAT and SAT scores available on only a subset of the examinees, were 445 and 516, respectively. Students in Study B were also high school students, but were from the set admitted to the University of California, the upper 12% of state high school seniors. These students all were required to take a writing placement examination which determines if students should take freshman or reredial English. Students are required to take the exam

if their College Entrance Examination Board scores falls between 450 and 600. Mean CEEB score for the Study B student sample was 491. Mean Scholastic Aptitude Test-Verbal Score was 510.

Raters used the same rating scale to evaluate essays in each study. Designed as a domain-referenced scale for factual narrative prose, the rubric included response criteria conventionally cited as important structural narrative elements by theoretical rhetoricians and educators (Kinneavy, 1971, Brooks and Warren, 1961). The Factual Narrative Scale consisted of five subscales: General Impression, Focus, Organization, Support, and Mechanics. Appendix B presents a copy of the scale.

Rating Procedures

Uniform training methods occurred in both studies.

Study A. In Study A, rater X, Y, 3, and 4 had about four hours of practice in applying criteria on approximately 30 papers written on Topics 1 and 2 (Quellmalz, 1979). A pilot test of inter-rater reliability was then conducted on 13 papers and yielded generalizability coefficients ranging from .63 to .74 and alpha coefficients ranging from .84 to .95. Rating of experimental essays on Topics 1 and 2 then proceeded. Final rating generalizability coefficients ranged from .67 to .85. Appendix C presents tables used for calculating the generalizability coefficients.

Study B. In Study B, raters 1, 2, 3, and 4 also practiced for 4 hours applying scoring criteria to approximately 20 papers on Topic 3 and 15 papers on Topics 1 and 2. A pilot test of rater agreement was then conducted on 20 Topic 3 papers. Generalizability coefficients ranged from .67 to .84.

Seven Topic 2 and 3 papers (3 of Topic 2; 1 of Topic 3) were also rated during the pilot testing, but the sample size was too small to allow calculation of agreement indices. Ratings on these seven papers were pooled with subsequent final ratings of additional Topic 2 and 3 papers. During rating of the experimental Topic 3 papers in the Study B "common check" papers were periodically read after approximately every 10 papers. Each rater pair compared scores and discussed descrepancies. Each "common check" included a Topic 3 paper and either a Topic 1 or Topic 2 paper. The eleven "common check" sets of essays read during Study B and the 7 aforementioned Topic 1 and Topic 2 papers formed the data base for Study B. Appendix D presents statistics used to calculate rater reliabilities on the 5 subscales. Generalizability coefficients ranged from .67 to .89. Alpha coefficients ranged from .72 to .85. To allow comparison between Study B scores on Topic 1 and 2 papers with Study A on those same papers, Study B raters then scored an additional five Topic 1 and 2 papers.

RESULTS

Introduction

To examine effects of writing task variables, topic and sample quality effects were analyzed first within Study B, then between scores on topics common to Study A and Study B.

Rating context variables were rater experience with the scale and training group membership. Scores and reliabilities of novice Raters 1 and 2 were compared to scores and reliabilities of scale-experienced Raters 3 and 4. Effect of rating group membership on criteria stability was examined by inspecting scores on Topics 1 and 2 in Study A and scores for Topics 1 and 2 in Study B. Analyses for this pilot study were exploratory, and clearly the sample size limits generalization. The results were intended to stimulate hypotheses rather than to provide definitive evidence.

Effects of Exam Variablity

Tables 1 and 2 present data for comparing scores within Study B given by the four individual raters on the three topics. Table 1 presents means and standard deviations of topic scores. Table 2 presents results of analyses of variance to detect differences due to topic and rater. Topic variability will be discussed in this section, reference to rater variability will be deferred to the section on rating context variation.

Insert Tables 1 and 2

Across all subscales, scores on Topic 3 are highest, scores on Topic 2 are lowest. The ANOVAs indicate that topic score differences are significant for ratings on General Impression, Organization, and Total Score.

TABLE 1

Means and Standard Deviations of CSE Essay Scores given in Study B by Raters 1-4 on Topics 1-3

General Impression

Rater		Topic 1	2	3	
1	$\bar{X}$	1.06	1.00	1.82	1.23
	SD	1.06	1.03	.40	
2	$\bar{X}$	1.06	.88	1.73	1.62
	SD	1.0	.89	.47	
3	$\bar{X}$	1.13	.75	1.64	1.12
	SD	1.20	1.06	1.02	
4	$\bar{X}$	.88	.44	2.00	1.00
	SD	1.15	.63	.63	
		1.03	.77	1.80	1.13
	n =	16	16	11	43

Focus

Rater		Topic 1	2	3	
1	$\bar{X}$	1.56	1.44	2.00	1.63
	SD	.63	.63	.63	
2	$\bar{X}$	1.63	1.63	2.00	1.72
	SD	.50	.62	.63	
3	$\bar{X}$	1.69	1.81	1.82	1.77
	SD	.48	.75	.60	
4	$\bar{X}$	1.69	1.44	2.09	1.70
	SD	.60	.51	.70	
		1.64	1.58	1.98	1.70
	n =	16	16	11	43

Organization

Rater		Topic 1	2	3	
1	$\bar{X}$	1.38	1.44	1.82	1.51
	SD	.50	.63	.40	
2	$\bar{X}$	1.44	1.31	1.82	1.49
	SD	.63	.48	.40	
3	$\bar{X}$	1.81	1.31	2.09	1.70
	SD	.91	.60	.54	
4	$\bar{X}$	1.62	1.25	2.36	1.67
	SD	.96	.45	.81	
		1.56	1.33	2.02	1.59
	n =	16	16	11	43

Support

Rater		Topic 1	2	3	
1	$\bar{X}$	1.56	1.44	2.00	1.63
	SD	.73	.51	.63	
2	$\bar{X}$	1.87	1.81	1.73	1.81
	SD	.62	.65	.65	
3	$\bar{X}$	2.31	2.00	2.45	2.23
	SD	.95	.73	.52	
4	$\bar{X}$	1.69	1.44	2.18	1.72
	SD	.79	.51	.75	
		1.86	1.67	2.09	1.85
	n =	16	16	11	43

Mechanics

Rater		Topic 1	2	3	
1	$\bar{X}$	1.94	1.69	1.91	1.84
	SD	.57	.70	.31	
2	$\bar{X}$	1.63	1.81	2.00	1.79
	SD	.62	.65	.63	
3	$\bar{X}$	1.87	1.94	2.27	2.00
	SD	.81	.68	.79	
4	$\bar{X}$	1.69	1.56	2.09	1.74
	SD	.87	.73	.70	
		1.78	1.75	2.07	1.84
	n=	16	16	11	43

Total Score

Rater		Topic 1	2	3	
1	$\bar{X}$	7.50	7.00	9.54	7.84
	SD	2.99	2.80	1.51	
2	$\bar{X}$	7.62	7.44	9.27	7.98
	SD	2.33	2.25	1.74	
3	$\bar{X}$	8.81	7.81	10.27	8.81
	SD	3.60	2.79	2.90	
4	$\bar{X}$	7.56	6.12	10.73	7.84
	SD	3.63	1.89	2.97	
		7.87	7.09	9.95	8.12
	n =	16	16	11	43

TABLE 2

Analysis of Variance of Rater (1-4) and Topic (1-3) on Study B Essay Scores

General Impression

Source	SS	DF	MS	F
Mean	239.08	1	239.08	91.44
Topic	28.61	2	14.30	5.47**
Error	104.58	40	2.61	
Rater	.80	3	.27	.35
Rater X Topic	2.92	6	.49	1.54
Error		120		

Focus

Source	SS	DF	MS	F
Mean	500.19	1	500.19	595.97
Topic	4.56	2	2.27	2.71
Error	33.57	40	.84	
Rater	.26	3	.08	.41
Rater X Topic	1.71	6	.29	1.34
Error	25.60	120	.21	

Organization

Source	SS	DF	MS	F
Mean	447.24	1	447.24	368.08
Topic	12.67	2	6.34	5.47**
Error	46.34	40	1.16	
Rater	1.84	3	0.61	3.66*
Rater X Topic	2.91	6	0.49	2.90*
Error	20.08	120	0.17	

Support

Source	SS	DF	MS	F
Mean	585.59	1	585.59	585.88
Topic	4.59	2	2.29	2.30
Error	39.98	40	1.00	
Rater	8.52	3	2.84	9.62**
Rater X Topic	2.87	6	0.48	1.62
Error	35.44	120	0.29	

Mechanics

Source	SS	DF	MS	F
Mean	580.87	1	580.87	479.23
Topic	3.03	2	1.51	1.25
Error	48.48	40	1.21	
Rater	1.54	3	0.51	2.19
Rater X Topic	1.51	6	0.25	1.07
Error	28.14	120	0.23	

Total Score

Source	SS	DF	MS	F
Mean	11508.00	1	11508.00	508.45
Topic	219.33	2	109.66	4.85*
Rater	24.42	3	8.14	3.34*
Rater X Topic	30.36	6	5.06	2.08
Error	292.18	120	2.43	

* p<.05
** p<.01

Planned comparisons contrasting scores on Topics 1 and 2 and Topics 1 and 2 together with Topic 3 revealed no significant difference between Topics 1 and 2, but a significant difference between Topics 1 and 2 vs. 3 on General Impression, ($t = -4.54$; d.f.37,3; $p<.01$), Focus (= -2.19, d.f. 16,0; $p<.05$), Organization ($t = -3.69$, d.f. 25,0; $p<.01$) and Support ($t = -2.16$; d.f. 23,3, $p<.05$). These differences might be attributed to the different populations who contributed writing samples on Topics 1 and 2 vs.Topic 3, however, aptitude scores from both groups appeared quite similar.

Table 3 presents the means and standard deviations of essay scores given by rater pairs in Studies A and B. The Table also presents the alpha coefficients for pair reliabilities in Study B. Sample sizes in Study A did not permit stable calculation of alphas; Study B n's are also quite small.

Insert Table 3

Looking first at the levels of agreement by topic in Study B, it appears that rater pairs do differ substantially in their scoring consistency from topic to topic. Interestingly, Raters 3 and 4, the more practiced with the scale, fluctuate more in their agreement, particularly on Mechanics and Support. These 2 subscales were fairly stable across topics in the original Study A. Again, the small number of essays involved in the calculations require caution in interpreting these magnitudes.

Inspection of means in Table 3 does reveal a trend toward lower ratings for Topics 1 and 2 in Study B. This trend may imply that the restricted quality range of primarily college bound student essays in Study B made the fewer high school essays seem worse. Thus, a subtle "norming" may

TABLE 3

Comparison of Rater Pair Scores Across Studies

		Study A				Study B					
Rater Pair		X & Y		3 & 4		3 & 4			1 & 2		
Topics		1	2	1	2	1	2	3	1	2	3
CSE Subscale											
General Impression	$\bar{X}$	1.92	1.36	1.28	1.0	1.00	.59	1.82	1.06	.94	1.77
	SD	1.32	1.28	1.37	1.3	1.13	.76	.72	.94	.91	.41
	n =	6	7	8	8	16	16	11	16	16	11
	α					.91	.67	.58	.81	.88	.86
Focus	$\bar{X}$	2.08	1.50	1.71	2.2	1.69	1.63	1.95	1.59	1.53	2.00
	SD	.38	.29	.9	.8	.48	.56	.61	.49	.50	.50
	n =	6	7	8	8	16	16	11	16	16	11
	α					.71	.70	.85	.67	.44	.40
Organization	$\bar{X}$	2.33	1.64	1.65	1.75	1.72	1.28	2.23	1.41	1.38	1.82
	SD	.98	.75	.6	.9	.86	.48	.61	.52	.50	.34
	n =	6	7	8	8	16	16	11	16	16	11
	α					.81	.79	.72	.82	.75	.56
Support	$\bar{X}$	2.42	2.36	2.76	2.45	2.00	1.72	2.32	1.72	1.63	1.86
	SD	.92	.56	1.15	.85	.68	.52	.60	.63	.50	.45
	n =	6	7	8	8	16	16	11	16	16	11
	α					.37	.50	.85	.86	.62	.00
Mechanics	$\bar{X}$	2.50	2.57	2.2	2.05	1.78	1.75	2.18	1.78	1.75	1.95
	SD	.84	.45	.7	.80	.77	.55	.68	.48	.58	.42
	n =	6	7	8	8	16	16	11	16	16	11
	α					.82	.35	.80	.47	.61	.58
Total	$\bar{X}$	11.25	10.43			8.19	6.97	10.50	7.56	7.21	9.41
	SD	3.71	2.67			3.40	2.10	2.73	2.52	2.35	1.42
	n =	6	7			16	16	11	16	16	11
	α					.86	.72	.84	.86	.83	.70

have occurred, suggesting that the composition of the student sample, i.e., the range of quality, may affect ratings.

Rating Context Variation

One aspect of rating hypothesized to affect application of rating criteria was level of experience of raters with the scale. It might be expected that raters less familiar with using the scale might be less consistent in their ratings. The ANOVAs presented in Table 2 indicated statistically significant variation due to raters on Organization, Support and Total essay scores. Table 4 presents results of planned comparisons between Rater pair 1 and 2 (inexperienced) and Rater pair 3 and 4 (experienced). The pairs do give significantly different scores on the Organization and Support criteria, suggesting differential application of criteria by rater pair. However, the substantive magnitude of the mean differences was relatively low, .20, or less than 5% variation on a 4-point scale.

Insert Table 4

Table 5 presents the interrater reliabilities for each rater pair in Study B and for all 4 raters. While mean scores given by each rater pair differed statistically significantly on Organization and Support, pair reliabilities differ most on Focus and Mechanics; the more experienced raters were more consistent on these subscales. These data may be considered as weak evidence that the rater pairs were perhaps becoming socialized during the common checks within pair.

Insert Table 5

TABLE 4

Pair Comparisons of Study B Essay Ratings
by Raters 1 & 2 vs. 3 & 4 Across Topics 1-3

CSE Scale		Raters 1 & 2	Raters 3 & 4	t
General Impression	$\bar{X}$	1.20	1.06	1.43
	SD	.89	1.01	
Focus	$\bar{X}$	1.67	1.73	- .78
	SD	.52	.55	
Organization	$\bar{X}$	1.50	1.69	-2.50*
	SD	.50	.76	
Support	$\bar{X}$	1.72	1.98	-2.94**
	SD	.53	.64	
Mechanics	$\bar{X}$	1.81	1.87	- .84
	SD	.50	.68	
TOTAL	$\bar{X}$	7.91	8.33	-1.55
	SD	2.35	3.07	

TABLE 5

Interrater Reliability Across Topics 1-3
Within and Between Rater Pairs

	Raters 1 & 2	Raters 3 & 4	Raters 1,2,3,4
CSE Scale	α	α	α
General Impression	.87	.82	.90
Focus	.57	.71	.76
Organization	.79	.82	.87
Support	.58	.58	.71
Mechanics	.53	.72	.81
TOTAL	.85	.85	.90

In addition to experience with the scale, training group membership was hypothesized as a possible influence on application of scoring criteria. In her study contrasting different scoring systems, Winters (1979) observed differences in criteria interpretation during training and practice. While no explicit, salient alterations in decision rules for any subscale criteria were noted in Studies A and B, more subtle unexplicated shifts could have occurred. In an attempt to detect criteria shifts, Topics 1 and 2 essay scores in Study A and B can be compared. Table 6 presents differences between Study A and Study B mean scores given by rater pair.

Insert Table 6

In general, differences between scores given by Raters 3 and 4 to papers they read in both Study A and Study B were smaller than mean score differences between scores given by Raters X and Y in Study A and scores given by Raters 1 and 2 in Study B. The small differences between Raters 3 and 4 first and second rating scores, particularly on the GI (.34) and Total Scores (1.46) can be viewed as evidence for scoring stability across occasions, analogous to test-retest comparisons. An exception is the Support subscale where Study A and B mean score differences of Raters 3 and 4 are higher.

Another comparison relating to scale stability is between scores given by Raters X and Y in Study A with scores given on the same essays read by Raters 1 and 2 in Study B. Comparisons between scores given by Raters X and Y in Study A and by Raters 1 and 2 in Study B are between different people, of course, and also between two sets of "first timers," i.e., rater pairs who had used the scale only once. Mean subscale score differences range from .57 to .78, higher generally than Rater 3 and 4 score differences.

TABLE 6

Differences between Mean Scores given during Study A and Study B

Rater	$\bar{X}_{(3\&4_A)} - \bar{X}_{(3\&4_B)}$			$\bar{X}_{(X\&Y_A)} - \bar{X}_{(1\&2_B)}$		
	Topic 1	Topic 2	$\bar{X}_D$	Topic 1	Topic 2	$\bar{X}_D$
GI	.28	.40	.34	.84	.42	.63
F	.02	.57	.29	.48	.97	.78
O	.07	.47	.27	.95	.26	.60
S	.93	.73	.83	.70	.73	.72
M	.42	.30	.36	.32	.82	.57
$\bar{X}_D$						
T	.44	2.48	1.46	3.69	3.22	3.45

Thus criteria application by novice raters do not appear to be as stable across studies. Further research employing more raters and essays are required to disentangle the influence of within-training group interpretation of criteria.

Discussion

The purpose of the pilot study was to explore effects of exam and rating context variation on essay scores. While the small number of raters and essays did not permit strong statistical comparisons, the descriptive statistics did provide bases for pursuing the tentative hypotheses. Topic assignment and quality range of examinees do seem to affect scoring. Also rater's scale experience and training group memebership do seem to yield different essay scores. Future research employing more raters, essays, and topics should approach the issues posed in the research questions both with a strong sampling design and more rigorous statistical analyses.

Implications for Study Design

In the pilot study, topic and examinee quality were confounded. The final study will gather multiple topics from a population varying widely in writing and verbal abilities indicated by multiple independent measures. Rating context variables might include both the scale experience of raters and the number of training groups, depending on logistical and economic constraints. Tapes of training group interaction might provide valuable information on both criteria interpretation and the sociological dynamics of different groups' interactions. Other dependent measures, in addition to other writing and verbal test scores, might include annotated "expert" scoring of essays and independent classifications of students into mastery/non-mastery writing groups. The differential classification "accuracy" related to different topics, sample quality, rater pair and training group membership conditions could then be

examined for the "real world" contingencies of the treatment effects. In sum, rater behavior seems to be affected by a variety of variables which bear further examination and control.

References

Brooks, C., & Warren R.P. Modern rhetoric: Shorter edition. New York: Harcourt, Brace & World, 1961.

Conlan, G. How the essay in the CEEB English test is scored. Princeton, New Jersey: Educational Testing Service, 1976.

Follman, J.C. & Anderson, J.A. An investigation of the reliability of five procedures for grading English themes. Research in the teaching of English, 1967, 190-200.

Kinneavy, J. L. A theory of discourse. The aims of discourse. Englewood Cliffs, N.J.: Prentice Hall, Inc., 1971.

Quellmalz, E. Defining writing domains: Effects of discourse and response mode. Los Angeles, California: Center for the Study of Evaluation, May, 1979.

Winters, L. The effects of differing response criteria on the assessment of writing competence. Los Angeles, California: Center for the Study of Evaluation, November, 1978.

APPENDIX A

Study A

Topic 1: Narrative Drugs

Many young adults use drugs. In fact, every year these substances hurt the lives of thousands of American teenagers by making them emotionally and physically ill.

Assume you have been asked by your school board to tell how drugs affect students in your district. Describe what happened over a period of time to one or more people you know who became involved with drugs.

- Your narrative should help others experience how drugs affect students.
- Use specific details and facts to make your account realistic.
- Be sure your narrative has a clear main idea and is well-organized.

APPENDIX A

Study A

Topic 2: Narrative Violence

Many news reports and books assert that today's students do not respect adults. The reports describe instances of discipline problems, even violence in schools, neighborhoods and in homes.

Assume you have been asked by your school board to tell how students in your district show their lack of respect for adults. Describe what happened over a period of time to one or more people you know who did not respect adults.

- Your narrative should help others experience how these students felt and acted.
- Use specific details and facts to make your account realistic.
- Be sure your narrative has a clear main idea and is well organized.

APPENDIX A

Study B

Topic 3: Narrative Change

Directions: You will have 45 minutes to plan and write the essay assigned below. Before you begin writing, consider the topic carefully and plan what you will say. Your essay should be as well organized and as carefully written as you can make it. Be sure to use specific examples to support your ideas.

Some people really change whereas others only appear to change. Describe either a real or an apparent change in a person you have known well for several years. If you are describing a real change, explain how it came about and what effect it had upon the person's life. If you consider the change only apparent, explain why you think so and how that judgment has affected your attitude toward the person.

APPENDIX B

Factual Narrative Scale

ELEMENT 1

Impressionistic Rating Procedures

The purpose of Impressionistic Rating is to form a single impression of a piece of writing as to how well it communicates a whole message to the reader. Impressionistic scoring assumes that each characteristic that makes up an essay -- organization of ideas, content, mechanics and so on -- is related to all other characteristics. Impressionistic scoring further assumes that some qualities of an essay cannot easily be separated from each other. In short, the procedure views a piece of writing as a total work, the whole of which is greater than the sum of its parts.

Discerning readers naturally will attend to, or be influenced by, some essay characteristics more than others. In the Impressionistic scoring, however, readers should arrive at a judgment regarding the overall quality of the essay.

For this element, you are being asked to form an overall impression concerning the effectiveness of the essays as examples of narrative writing.

The Topics

You will be reading essays on two different topics. Both topics were designed to elicit writing in the narrative mode. Some views on narration are given below:

- Narration is the kind of discourse that tells what happened, how it happened.
- Narration gives an impression of movement in time.
- Narrative action includes time, logic, and meaning.

General Impression

Read each essay quickly in order to form an overall impression of its quality. To assign the essay a score, consider the following question: To what extent is the essay an example of effective narration?

Assign each paper a mark of 1 - 4 using the scale below.

4 = An excellent example of narration

3 = A good or adequate example of narration

2 = A minimally adequate example of narration

1 = A poor example of narration, barely readable and/or off the topic

0 = Is not a narrative

APPENDIX B

Factual Narrative Scale

ELEMENT 2

Essay Focus

The beginning or end of the essay clearly identifies the subject and main point of the whole essay.

4. The beginning (and/or conclusion) of this paper clearly implies or states the main points of the whole essay. It also limits the topic by alerting the reader to the key events or stages covered in the body of the essay.

   Specifically, at the beginning (and/or end):

   a. The subject of the essay is clearly identified.

   b. The main point of the whole essay is clearly stated or implied.

   c. The topic is clearly limited. That is, key points (e.g., major events) or major line(s) of action treated in the essay are identified, summarized or implied.

3. The beginning (and/or end) of this paper implies or states the main point of the whole essay. It sets limits on the topic, but does not clearly suggest how the main point is developed.

   Specifically, in the beginning (and/or end):

   a. The subject of the essay is clearly identified.

   b. The main point of the whole essay is clearly stated or implied.

   c. An attempt is made to limit the topic. That is, the number -- or general type -- of key points is indicated, but there is not clear reference to the important stages, or events treated in the body of the essay.

2. The beginning (and/or end) of this paper gives the reader a fairly clear sense of the main point of the whole essay. However, neither the introduction nor the conclusion help focus -- or bring direction to -- the body of the paper.

   Specifically, in the beginning (and/or end):

   a. The subject of the essay is identified.

   b. The main point of the whole essay is stated or implied.

   c. No attempt is made to limit the topic.

(continued)

ELEMENT 2 (continued)

1. Neither the beginning nor the end is helpful to the reader in obtaining any sense of the main point of the essay.

   Specifically, in the beginning (and/or end):

   a. The subject of the essay is not clearly identified or there is no reference to the subject

   AND/OR

   b. The main point of the whole essay is not clearly stated or implied or no reference is made to the main point or the reference is confusing.

APPENDIX B

Factual Narrative Scale

ELEMENT 3

Organization

4. The temporal (chronological) order of events is clear.

   a. Transitions and linking expressions enhance a sense of movement. Events and ideas are ordered according to a readily discernible logical progression.

   b. All events and ideas clearly relate to the main point of the action.

   c. There are no digressions or extraneous material.

3. The temporal order of events is clear.

   a. Transitions and linking help establish a sense of movement.

   b. Events and ideas are ordered according to a discernible logical progression.

   c. Most or all events and ideas relate to the main point of the essay, there are minor digressions or extraneous material.

2. The temporal order of events is not consistently clear although the essay does have a discernible beginning and end.

   a. Events and ideas are not ordered according to a discernible logical progression or sequence. Events may be merely listed.

   b. The relationship of many events and ideas to the main point is not consistently evident. There are major digressions and/or extraneous material.

1. There is no clear temporal order.

   a. Transitions are not present or are ineffective.

   b. Events and ideas are in a confusing order. The reader has a difficult time inferring the relationship among events and ideas. The paper may be expository or descriptive.

APPENDIX B

Factual Narrative Scale

ELEMENT 4

Support

Generalizations and assertions are supported by specific, clear supporting statements.

4. All general points and assertions in this paper have logically related details. Supporting statements clearly are at a greater level of specificity than the generalizations they are intended to support. The writer makes effective use of concrete detail to support ideas.

   Specifically:

   a. Generalizations and assertions are supported by more specific statements.

   b. Supporting statements provide specific detail, such as examples, facts, anecdotes, and/or employ concrete language to convey ideas about events, actions, and/or characters. A character is described by reference to appearance, feelings, thoughts and actions.

3. Most generalization points and assertions are supported by more specific, clear and logically related details.

   Specifically:

   a. General points and assertions are supported by more specific statements.

   b. Many supporting statements provide specific detail, such as examples, facts, anecdotes, and/or employ concrete language to convey ideas about events, actions and/or character(s)' appearance, feelings, thoughts and actions.

2. Support for many general points and assertions in this paper is not especially convincing. Supporting statements lack precision, either in language or in the use of specific detail.

   Specifically:

   a. An attempt is made to support general points and assertions through the use of statements at a greater level of specificity than the generalization they are intended to support.

   b. Supporting statements lack precision and clarity. The use of specific detail, such as examples, facts, anecdotes, character description may be lacking and/or the language may be vague.

(continued)

ELEMENT 4 (continued)

1. Support is provided for few or none of the general points or assertions. Few (or none) of the supporting statements are at a greater level of specificity than the generalizations themselves.

   Specificially:

   a. Supporting statements are provided for few or none of the general points or assertions.

   b. Supporting details, if present, are vague, confusing, or not logically related to the generalizations they apparently are intended to support, and/or the language of supporting statements is imprecise. There is little or no attempt to delineate a character's appearance, thoughts, feelings or actions.

Factual Narrative Scale

ELEMENT 5

Mechanics

The essay is free of intrusive usage and mechanical errors.

4. The writer appears to have control over the usage and mechanical aspects of this essay.

   Specifically:

   a. There are only a few minor errors in usage and mechanics or no errors at all.

   b. Errors, if present, do not interfere with the clarity of communication.

3. Usage and mechanics are not a problem in this paper. However, the writer appears to be occasionally careless and makes a few minor errors which are readily evident to the discerning reader. The paper also may contain a few common errors in usage and mechanics.

   Specifically:

   a. There are only a few obvious usage or mechanical errors.

   b. Errors do not interfere with the clarity of communication, e.g., occasional confusion of subject-object pronouns ("between you and I..."), spelling errors on difficult words, misuse of colons and semicolons.

   c. There are no major errors, such as run-on sentences, inappropriate fragments, lack of subject-verb agreement in simple sentences.

2. This essay is flawed by errors in mechanics, although the paper does not strike the discerning reader as being illiterate.

   Specifically:

   a. There are obvious errors in mechanics throughout much of the essay.

   b. Errors occasionally detract from the clarity of the communication, such as confusing antecedents, consistent omission of key words.

   c. Many of the errors are major, e.g., many run-on sentences or inappropriate fragments, lack of subject-verb agreement in simple sentences.

1. Mechanical errors make this paper very difficult to read and understand.

   Specifically:

   a. There are excessive and obvious errors in mechanics throughout the essay; nearly every sentence contains some type of error.

   b. Errors consistently interfere with the writer's attempt to communicate.

   c. Errors are not restricted to one type of problem, such as run-on sentences. Errors are diverse in nature and major.

APPENDIX C

Data for Estimating Generalizability Coefficients: Study A

Raters 3 & 4

Expository Scale: General Impression

Source	$E\sigma^2$	DF	MS	G
Subjects	.312	25	1.606	
Topics	-.002	1	.038	
Raters	.105	1	5.538	.83
Subjects x Topics	.060	25	.298	
Subjects x Raters	.029	25	.238	
Topics x Raters	-.007	1	0	
Subjects x Topics x Raters	.180	25	.180	

Expository Scale: Focus

Source	$E\sigma^2$	DF	MS	G
Subjects	.245	25	1.463	
Topics	-.01	1	.962	
Raters	.032	1	2.163	.70
Subjects x Topics	.029	25	.430	
Subjects x Raters	.026	25	.423	
Topics x Raters	.004	1	.471	
Subjects x Topics x Raters	.371	25	.371	

Expository Scale: Organization

Source	$E\sigma^2$	DF	MS	G
Subjects	.217	25	1.493	
Topics	-.004	1	.154	
Raters	.017	1	3.846	.67
Subjects x Topics	.100	25	.454	
Subjects x Raters	.086	25	.426	
Topics x Raters	-.004	1	.154	
Subjects x Topics x Raters	.253	25	.254	

(Continued)

Expository Scale: Support

Source	$E\sigma^2$	DF	MS	G
Subjects	.538	25	.727	
Topics	.231	1	.471	
Raters	.039	1	2.163	.85
Subjects x Topics	.171	25	.591	
Subjects x Raters	.057	25	.363	
Topics x Raters	-.009	1	.961	
Subjects x Topics x Raters	.250	25	.250	

Expository Scale: Mechanics

Source	$E\sigma^2$	DF	MS	G
Subjects	.268	25	1.446	
Topics	-.002	1	0	
Raters	.166	1	8.653	.77
Subjects x Topics	.041	25	.400	
Subjects x Raters	-.012	25	.294	
Topics x Raters	-.011	1	.038	
Subjects x Topics x Rater	.318	25	.318	

APPENDIX C

Data for Estimating Generalizability Coefficients: Study A

Raters X & Y

Expository Scale: General Impression

Source	$E\sigma^2$	DF	MS	G
Subjects	.033	13	.815	
Topics	.010	1	3.02	
Raters	.021	1	.875	.22
Subjects x Topics	.115	13	.440	
Subjects x Raters	.120	13	.452	
Topics x Raters	-.014	1	-.018	
Subjects x Topics x Raters	.210	13	.210	

Expository Scale: Focus

Source	$E\sigma^2$	DF	MS	G
Subjects	.110	13	.946	
Topics	.099	1	3.02	
Raters	.005	1	.161	.60
Subjects x Topics	.115	13	.518	
Subjects x Raters	-.005	13	.276	
Topics x Raters	-.019	1	.018	
Subjects x Topics x Raters	.287	13	.287	

Expository Scale: Organization

Source	$E\sigma^2$	DF	MS	G
Subjects	.129	13	1.374	
Topics	.055	1	2.160	
Raters	.049	1	1.446	.61
Subjects x Topics	.302	13	.815	
Subjects x Raters	.022	13	.254	
Topics x Raters	-.014	-	-.018	
Subjects x Topics x Raters	.210			


ERIC
Full Text Provided by ERIC

(Continued)

Raters X & Y

Expository Scale: Support

Source	$E\sigma^2$	DF	MS	G
Subjects	.159	13	1.153	
Topics	.000	1	.286	
Raters	-.549	1	.071	.68
Subjects x Topics	.107	13	.362	
Subjects x Raters	.007	13	.302	
Topics x Raters	-.005	1	.071	
Subjects x Topics x Raters	.148	13	.148	

Expository Scale: Mechanics

Source	$E\sigma^2$	DF	MS	G
Subjects	.164	13	1.133	
Topics	-.049	1	1.446	
Raters	-.003	1	.161	.57
Subjects x Topics	.022	13	.254	
Subjects x Raters	.110	13	.430	
Topics x Raters	-.014	1	.018	
Subjects x Topics x Rater	.210	13	.210	

APPENDIX D

Study B

Statistics for Calculating Generalizability Coefficients:

Raters 1, 2, 3, 4

General Impression

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	.712	42	3.171	.897
Rater	.002	3	.411	
Subject/ Rater	.324	126	.324	

α.85

Focus

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	.173	42	.908	.763
Rater	-.002	3	.145	
Subject/ Rater	.217	126	.217	

α.74

Organization

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	.306	42	1.405	.866
Rater	.008	3	.504	
Subject/ Rater	.182	126	.182	

α.85

Support

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	.189	42	1.061	.673
Rater	.064	3	3.062	
Subject/ Rater	.304	126	.304	

α .72

Mechanics

Source	Estimates of Variance Components	Degrees of Freedom	Mean Square	G
Subject	.248	42	1.226	.804
Rater	.007	3	.533	
Subject/ Rater	.235	126	.235	

α .81

Snidman, N. S., & Quellmalz, E. S. Issues in criterion-referenced test construction. Paper presented at the annual meeting of the California Educational Research Association, San Diego, California, November, 1976.

Stalnaker, J. The construction and results of a twelve-hour test in English composition. School and Society, 39, 1934.

Stanley, J. C. Analysis of variance principles applied to the grading essay tests. Journal of Experimental Education. 1962, 30 (3).

Starch, D., & Elliott, E. C. Reliability of the grading of high school work in English. School Review, 1912, 20.

Veal, L. R. & Tillman, M. Mode of discourse variation in the evaluation of children's writing. Research in the Teaching of English, 1971, 5: 37-45.

Warriner, J. & Griffith, F. English grammar and composition, complete course. New York: Harcourt, Brace & Jovanovich, 1973.

West, W. Developing writing skills, second edition. New Jersey: Prentice-Hall, 1973.

Yeasmen, N., & Barker, D. A. Half a century of research on essay testing. Improving College and University Teaching, 1973, XXI(1).

VALIDITY

The concept of validity has long been one of the most fundamental concepts in educational and psychological measurement. Much has been written about the various types of validity and the process of validation. Yet, within the context of criterion-referenced measurement, validity has received relatively little attention. Certainly it has not received the amount of attention that has been given to various other topics such as reliability, determination of test length, and questions of the need for variability.

The relative lack of attention to issues of validity for criterion-referenced tests may be largely attributed to a view that of the three types of validity discussed in the Standards for Educational and Psychological Measurement (APA, 1974) only concent validity is particularly relevant to criterion-referenced measures. The key steps in content validation are first the definition of domain and, second, the demonstration that the measure adequately represents the domain (Linn, 1977). If the preceding is correct, then it is hardly surprising that validity has not been the topic of hot debate in a criterion-referenced context that it has been in other measurement contexts.

The great emphasis that has been placed on domain definition does much to solve the problem of content validation. Where the domain is sufficiently well defined, random sampling of items can provide a means of achieving representativeness. This is certainly a much stronger basis for making inferences about a content domain than is provided by traditional test construction procedures. Thus, issues of validity may have seemed less salient for criterion-referenced measures because, when based on adequate definition and sampling, the type of validity of primary concern

is assured.

Of course, the ideals of a completely defined domain and construction of measures by random sampling are rarely even approximated in practice. Even content validity will often be problematic. Furthermore, it is our contention that, like other measures, criterion-referenced measures are subject to many interpretations and uses. The various interpretations and uses require the consideration of a much wider range of evidence than would ordinarily be considered under the heading of content validity. Thus, we shall argue that validity is a topic worthy of much more consideration than it has received in most discussions of criterion-referenced measurement.

Types of Validity

The validity of a test is often defined as the degree to which it measures what it is supposed to measure. A broader and more useful conception, however, is that questions of validity are questions of the soundness of interpretations of a measure (Cronbach, 1971; APA, 1974). Validation is best conceived of as a process rather than an end product. It is the process of marshalling evidence to support interpretations. Measurement results can have many interpretations which differ in their validity and in the type of evidence required for the validation process.

There are a variety of types of questions of validity which require different types of validation, e.g. criterion-related, content and construct. It is convenient to distinguish different types of validation. Certainly, many labels and category systems have been offered by various authors who have written about validity. It is worth noting, however, that "these aspects of validity can be discussed independently, but only for convenience. They are interrelated operationally and logically; only rarely is one of them alone important in a particular situation" (APA, 1974).

The Variety of Validation Needs for Criterion-Referenced Measures

Millman (1974) described validity in terms of the "...accuracy of inferences made from test scores" (p. 360). This position is generally consonent with the one stated above. While recognizing a variety of types of inferences requiring different validation processes, Millman argued that the most appropriate inference for a domain-referenced test concerns the status of an individual relative to a well defined domain. For this inference, the logical analysis of the domain definition and procedures for generating items are of primary concern. That is, content validation is the primary issue. Hambleton, Swaminathan, Algina and Coulson (1975) stated a similar position. They acknowledged that other types of validity are appropriate to study, but claimed that content validity is the "center of validation concerns" for a domain-referenced [here, criterion-referenced] test.

We do not quarrel with the emphasis placed on content validity by Millman and by Hambleton, et al. Indeed, we think that one of the main advantages offered by criterion-referenced measurement is that it provides a much firmer basis for content validation than has been available with traditional psychometric approaches to constructing achievement tests.

Content validity is a necessary consideration for criterion-referenced measures. It is our contention, however, that it is rarely sufficient. This is so because rarely, if ever, are inferences restricted to the type that are adequately supported by content validity alone.

With a well defined domain and random sampling of items from that domain there is a firm basis for supporting an inference that a person can satisfactorily perform at least x percent of the tasks in that domain. This is the type of inference that is supported by content validity. It is a natural consequence of the adequacy of the domain definition and of item

generation rules. Seldom are our inferences limited to such a degree, however. Labels such as "master" or "competent" are often attached to persons whose performance is estimated to be above some specified level. Scores on the test are apt to be used as the basis of differential treatment. For example, some students may be given remedial instruction as a result of their scores. These additional interpretations and uses require more than content validity (Linn, 1976).

When a person is labeled a "master" certain expectations usually accompany that label. The student who has mastered certain types of addition problems may be expected to be "ready" to learn subtraction. Mastery of one skill may be expected to be a prerequisite for mastering another skill. Evidence of the soundness of expectations such as these is needed. The process of marshalling the evidence for the soundness of the expectations is construct validation (Cronbach, 1971; Messick, 1975).

Even where constructs are not invoked more than content validity will usually be needed. The decision to have students with scores on a test below some specified level repeat an instructional segment rather than go on to a new one is based on an implicit prediction that the student will, at least, learn the content of the segment to be repeated better by redoing it than by moving on. It may also be based on an implicit prediction that he or she would do more poorly on the next segment of instruction if allowed to move on to it before repeating the earlier segment than would otherwise be the case. Evidence of criterion-related validity is needed to support these predictions.

We have argued that all three of the traditionally recognized types of validity are relevant to criterion-referenced measurement just as they are to other types of measurement. It does not follow from this position, however, that the traditional validation procedures will be the most appropriate

ones for criterion-referenced measures. For instance, there are a variety of reasons that the traditional reliance on correlation coefficients may be inappropriate for criterion-referenced measures. One of these is the issue of variability which is discussed at length in the reliability section of this paper. Another reason is that, as suggested by Hambleton and Novick (1973), correlations are based on an inappropriate loss function for situations commonly encountered with criterion-referenced measures. Hambleton and Novick suggested that a threshold loss is more appropriate in a criterion-referenced context than is the traditional squared-error loss.

The need for a different type of statistics or for nontraditional procedures for collecting the needed data do not require different concepts of validation. They merely represent different details of implimentation within a common conceptual framework. A few comments regarding the special implimentation requirements for validation of criterion-referenced measurement may be useful, however. These are offered below under the three traditional headings of content, criterion-related, and construct validity.

Content Validity

Domain definition

Although content validity is formally recognized as one of the three main types of validity (APA, 1974), few would consider it to stand on an equal footing with the other two types of validity in terms of the rigor of the evidence that is usually provided to support a claim of validity. Indeed, some well known test theorists have argued that what traditionally goes under the heading of content validity shouldn't even be called validity. For example, Messick (1975) suggested that

we should "call it 'content relevance', ... or 'content representativeness', but don't call it content validity" (quoted by Guion, 1977, p. 2). Ebel (1975a) suggested that "perhaps... we should call it content reliability ..." (as quoted by Guion, 1977, p. 2).

Guion (1977) clearly stated a number of reasons for having reservations about content validity. Primary among these reasons is Guion's conclusion that "judgements of content validity have been too swiftly, glibly and easily reached in accepting tests that otherwise would never be deemed acceptable" (1977, p. 2). Despite his reservations, Guion argued that the ideas contained under the notion of content validity are extremely important.

For the acceptance of a measure on the basis of content validity Guion proposed a set of five minimal conditions. These conditions are:

1. that the content domain must involve "behavior with a generally accepted meaning" (p. 6),
2. that the definition of the domain must be unambiguous,
3. that the domain must be relevant to the purposes of the measurement,
4. that "qualified judges must agree that the domain has been adequately sampled" (p. 7), and
5. that the measure must have reliability.

The above list is useful. But, it involves considerations that go beyond what we would include under the heading of content validity. We will argue that content validity is derived from on two considerations: domain definition and representativeness. Our disagreement with Guion may be more one of semantics than substance however. Issues on "meaning" and "relevance" imply inferences that go beyond those justi-

fied on the basis of content validity alone. They involve constructs or external criteria and require other forms of validity evidence. As we have already argued, content validity is necessary but not sufficient to support "the acceptance of a measure" for any but the most limited types of used and inferences.

The key to content validation is the definition of the domain of interest. Ideally, the definition should include a description of every detail of the measurement procedure that might have an influence on the results. The ideal can never be fully achieved in practice, but improvements over common practice could certainly be made.

Consider, for example, the observations of student teachers in a classroom setting. There is an almost limitless list of potentially important details that might have an influence on the results of the observation procedures. There are the physical conditions of the situation(e.g., building, lighting, room size, furniture arrangement, etc.) There are considerations

of timing (e.g., time of day, day of week, part of the semester). There are the students and other persons in the room to be considered (e.g., number, age, sex, etc.). The topic of the lesson, the amount of previous contact of the student teacher with the students, the obtrusiveness of the observers, and the characteristics of the observers all require attention. The list could go on and on. The description can never be complete. Judgment is always required in determining what is most deserving of attention, i.e., those aspects of the measurement procedure that are most crucial to the results of the measurement.

Certain aspects of the procedure will be controlled. For example, in the observation of student teachers, the content of the lesson, the number of students and the length of time available for the lesson, among a number of other conditions, might be fixed for everyone. All the possible variations of other conditions not fixed by the definition, e.g., observers, physical location, etc, constitute what Cronbach, et al. (1972) refer to as "the universe of admissible observations."

It is probably more common in discussions of content validity to speak in terms of a "performance domain" (e.g., APA, 1974) rather than a universe of admissible observations. The latter label, though longer and more clumsy to use all the time, has the advantage of calling attention to the idea that any characteristic of the measurement procedure which is not controlled in the definition, could be varied and the observation would still be admissible. For example, the domain of all addition problems involving two single digit positive numbers appears fairly clear and specific. Such a definition, however, would include as admissible observations oral as well as written presentation, base eight as well as base ten numbers, speeded as well as unspeeded presentations, multiple-choice as well as constructed response formats, and many other variations. The universe of admissible

observations calls attention to the limits and possible vagueness of the definition of the measurement procedure that may go unattended in the description of a performance domain.

Domain Representativeness

As noted above, definition of the universe of admissible observations is the first of two crucial steps in content validation. The second step requires a demonstration that the observations yield results that are representative of that universe. Where the universe is sufficiently well defined, representativeness may be pursued by a formal sampling process. Random or stratified random sampling has great advantages of objectivity and enables one to calculate unbiased estimates of population parameters. For example, if test items are randomly sampled from a universe of items, the proportion of items correct in the sample of items administered to an individual is an unbiased, maximum likelihood estimate of the number of items in the universe that the individual could answer correctly (Harris, Pearlman & Wilcox, in press). Furthermore, it does not depend on who else was tested, either as an estimate or for interpretation.

Defining universes such that random samples of items or other types of observations can be selected has great advantages but is seldom accomplished. Some have argued that it doesn't even approximate reality and that it is misleading to use a sampling model (e.g., Loevinger, 1965). At least for some limited examples, however, domains have been defined such that sampling is possible. The work of Bormuth (1968, 1970) and of Hively and his associates (e.g., Hively, Patterson & Page, 1968) provide some important examples. Some of the potential advantages of this type of approach in the area of achievement testing have been discussed by Shoemaker (1975).

Lip service is often given to the requirements of clear definition. It is fundamental to any approach that is concerned with "representative" or

accepted as desirable (see for example p. 28 of the APA Standards for Educational and Psychological Tests, 1974). Unfortunately, satisfactory definitions are the exception rather than the rule.

There are a lot of reasons that our measures lack the sort of rigorous definitions and the sampling processes that are required. For many content areas, the problem is a very difficult one, certainly more difficult than the area of simple arithmetic problems which has become the overused example. It is also the case that test publishers have relied too heavily on a totally different approach to achieving meaning. Norms rather than content definition are used to provide meaning. Note that it is not, as some have implied, the existance of norms as a basis of comparison that is the problem. Any test can be normed. Rather it is the reliance on norms as the sole basis of interpretation that is the culprit.

The need for content interpretations of achievement tests has been recognized for a long time. Ebel (1962) argued forcefully for content interpretations and provided an example where the universe and sampling procedures were defined with such rigor that it made no difference whether he or his secretary constructed the test. Ebel concluded that "a test produced by objectively defined processes may be less efficient, or lack some kinds of excellence which a creatively artistic test construction might achieve, but the increase in objective meaningfulness and reproducibility could more than offset the cost" (1962, p. 22).

Applying the principles of universe definition and sampling to measures such as observations of teacher behavior, or ratings of student art projects is substantially more complicated than applications with spelling or arithmetic tests. The necessity of doing so, however, is just as great if we are to achieve the sort of "objective meaningfulness and reproducibility" described by Ebel. For measures, such as teacher observations, the universe of admissible

such a measure could be assessed by duplicate-construction experiment. Two independen. teams would use the same definition of the universe of admissible observations to construct the measures. The results of applications of the two independently constructed measures would then be compared. The ideal outcome would be that the results of the two measures were always equivalent except for sampling errors. The content validity of the measure would then be assessed in terms of the degree to which this ideal is approximated.

Why Call It Validity?

After describing what we consider to be the two essential aspects of content validity we must return to a nagging concern raised in the beginning of this section. What we have argued for might, as Ebel (1975a) suggested, better be called "content reliability". There is a close correspondence between the notions of generalizability theory and those of content validity as we have espoused them. We have retained the label of content validity for two reasons.

First, the conotation of reliability is generally narrower than what we think should be included under considerations of content validity. As commonly thought of, reliability need not involve the careful attention to unambiguous definition and sample representativeness. The common procedures for estimating various types of reliability, (e.g. internal consistency, parallel forms reliability, or test-retest reliability) can be applied with little concern for the clear definition of the limits of the domain or for questions of representativeness.

Our second reason for preferring to retain the content validity label is that it makes the tie among the three types of validity more apparent. Content validity can seldom stand alone. On the other hand, it is a

observations may be defined in terms of several independent characteristics, which Cronbach, et al. (1972) refer to as facets. For example, one facet might be observers. A population of potential observers would need to be defined from which random samples could be selected. Other facets might be occasions, subject matter of lesson, and items on which observers make observations. In practice, it may seldom be possible to actually draw random samples of observations representing each facet. For example, location and scheduling may make some compromise necessary in the assignment of observers. Does this mean that the notions of universe definition and sampling have no place in the discussion? We think not. As noted by Cronbach:

> "...it is customary to apply statistics derived from a random sampling model to groups of persons who were not truly drawn at random and are only loosely representative of the population of persons to which the investigation generalizes. Sometimes the mismatch between model and reality is serious, sometimes not, but the problem is not peculiar to content sampling" (1971, p. 455).

Cornfield and Tukey's (1956) famous analogy between a bridge with two spans and statistical and substantive aspects of data interpretation is apropos here. The statistical span of the bridge goes from one shore to an island and the substantive span from the island to the other shore. They argue that where the substantive span is weak, it may be better to have that span of the bridge short while stretching the statistical span.

The absence of the ability to randomly sample from a domain does not preclude the possibility of studying generalizability. Cronbach (1971) suggested a procedure that would provide a check of the content validity of a measure where components of the measure were selected judgmentally rather than by a set of sampling rules. He suggested that the content validity of

necessary accompaniment to the other types of validity. For example, content validity of a test may be given little consideration in a traditional criterion-related validity study but may the central basis for the acceptance of the criterion measure. To do otherwise could lead to the need for a criterion against which the first criterion ~~against which the first criterion~~ could be validated and the continued application of that logic only leads to an infinite regress (Guion, 1977, p. 8).

Criterion-Related Validity

Criterion-referenced tests are often used to make short term instructional decisions. Commonly a dichotomy is formed and people scoring below some specified level are given some form of remediation while those above the cut off move on to a new set of instructional materials. As previously noted, use of scores for decisions such as this involves implicit predictions. For example, it is assumed that those selected to move on will do better on the new material than those assigned to some form of remediation. Other implicit predictions are that students given remedial work will perform better on the test the second time around and will do better on the next segment of instruction after remediation than they would have done without it.

The implicit predictions such as the above can sometimes be supported by empirical evidence. The performance of students before and after remediation can be compared. Experiments could be conducted where some students below the cutoff score on a test are randomly selected to receive remediation while the others move on to new material. The performance of the two groups at a subsequent time could then be compared. The results of such studies would provide evidence of criterion-related validity, albeit not necessarily summarized in terms of a traditional "validity coefficient."

Hambleton and Novick (1973) suggested that the proportion of times

persons above the "qualifying score" on a criterion-referenced test also are above the qualifying score on a "new test" provides an indication of validity. The new test "...might well be derived from performance on the next unit of instruction, or it could be a job-related performance criterion" (Hambleton & Novick, 1972, p. 168). This position is in close harmony with traditional psychometric notions of criterion-related validity. The only differences are in the metric of the scores and the type of summary statistic used. As argued by Hambleton and Novick, a dichotomous score metric is more consistent with some uses of criterion-referenced tests than is the more finely gradated number right score and the proportion of agreement is a more useful statistic for the resulting binary variables than is a product-moment correlation.

Construct Validity

Construct validity has been a controversial topic even in the realm of psychological measurement where it was originated. Its role in educational measurement has been even more dubious. "Construct validity is not usually sought for educational tests, because they are typically thought to be valid on other grounds, namely, on the grounds of content validity" (Messick, 1975, p. 959). This is particularily true of criterion-referenced measures.

Operational definitions are thought of as the key for criterion-referenced tests and no need is seen to invoke constructs in the interpretation of performance (e.g. Osburn, 1968; Harris, Pearlman, & Wilson, in press). But, interpretations of scores on criterion-referenced tests that go beyond operational definitions are commonly made. Competency-based educational programs, for example, rely heavily on criterion-referenced measurement. Yet the very word "competency" implies a construct. Even in very simple contexts the claim of competence or incompetence involves an inference. The claim may be based on a set of well-defined procedures to measure the performance of an individual,

but poor performance does not necessarily imply incompetence.

> "The inference of inability or incompetence from the absence of correct performance requires the elimination of a number of plausible rival hypotheses dealing with motivation, attention, deafness, and so forth. Thus, a report of failure to perform would be valid, but one of inability to perform would not necessarily be valid. The very use of the term inability invokes constructs of attributes and process, whereas a content-valid interpretation would stick to the outcomes" (Messick, 1975, p. 960).

Inferences about competencies need to be supported by the process of construct validation. The need for construct validation is avoided only by avoiding the inferences. In some circumstances such inferences may be avoided altogether, but we think more often than not some inferences will be desired and construct validation will be called for.

It is not possible to provide a simple prescription for establishing construct validity. It is better thought of as a continuing process of marshalling evidence to support or refute inferences and interpretations of test results (Cronbach, 1971; Messick, 1975). The process involves logical analysis as well as a wide variety of possible empirical studies, including for example studies of the effects of experimental interventions.

The process of construct validation has been described most thoroughly in Cronbach's 1971 chapter in Educational Measurement. He illustrated many procedures but noted that "...the procedures cannot be cateloged exhaustively and no guide can tell just how to meet the requirement of hard-headed reasoning from data to conclusions" (p. 484). The emphasis, however, is on a purposeful approach starting "...with a reasonably definite statement of the proposed interpretation" (p. 483). Data which can support or refute the interpretation are collected and the results are used in the refinement of the interpretation.

Instructional theory is far from the stage of providing the type of elaborated nomological net that Cronbach has described. Construct validity cannot be "established" by merely plugging a test into the network and checking on the predictions. The process is both more difficult and less clearly defined than that. On the other hand, it is only through the doing that a network can be developed and refined. In this sense the process of construct validation is just as central to the development of theory as the theory is to the validation of the measures. It is a long term iterative process.

SETTING STANDARDS

Performance standards play a very important and pervasive role in many of the conceptualizations and applications of criterion-referenced measurement. They are fundamental to the frequently cited definition of a criterion-referenced measure provided by Glaser and Nitko (1971). They defined a criterion-referenced test as "...one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards" (p. 653). Classification of people as masters or non-masters with regard to a specific domain obviously implies a standard; so does the application of criterion-referenced measures in competency-based programs.

The prominence of standards in applications of criterion-referenced measurement has increased along with the emphasis on competency-based education, and with demands for accountability and for minimum standards for high school graduation. Glass (1976) has shown how the meaning of criterion-referenced measurement has shifted from the time Glaser (1963) originally introduced the term. As noted by Glass, the emphasis in Glaser's original discussion was on determining what an individual can do along a carefully defined and very specific performance continuum. Over the years, however, the emphasis has shifted so that "... the term is now taken to mean tests that relate performance to absolute standards rather than to the performance of others" (Shepard, 1977, p. 3).

Although of fundamental importance, the problem of setting standards has received considerably less systematic attention than other problems such as reliability estimation, determination of test length, or procedures for generating items. This relative lack of attention is unfortunate because a set standard is often the foundation on which an application is based.

Many of the statistical and psychometric formulations which are discussed in other sections of this paper, depend on the acceptance of a standard. If the standard lacks adequate justication the techniques will be of little, if any, value. Glass (1976) put it more strongly. He suggested that the statistical and psychmetric treatments of problems of criterion-referenced measurement represent "...misdirected precision and axiomatization" (p. 36). The prior issues involve the setting and justification of the standard.

The problem of standards is not unique to criterion-referenced measurement. There were standards for graduation or for passing from one grade to the next long before the notion of criterion-referenced measurement was introduced. A standard is implicit, albeit somewhat fuzzier, when children in given schools are reported to be performing a year below grade level. Thus, criterion-referenced measurement did not introduce performance standards to education but it did make them more explicit and has reinforced a tendency to form dichtomies such as master-non-master, competent-incompetent, or pass-fail.

The creation of sharp dichotomies is seen by some as a negative attribute of the criterion-referenced measurement movement. Certainly, if set capriciously, standards are apt to detract from the potential advantages associated with other aspects of criterion-referenced measurement. The National Education Association's Guidelines and Cautions for Considering Criterion-Referenced Testing (1975) caution that "'Minimal comptency' or 'mastery' cut-off points should be viewed with some suspicion" (p. 11). In the discussion of this caution it is argued that "...the setting of such standards is extremely arbitrary." (NEA, 1975, p. 11). The concern that standards are often arbitrary and apt to be a negative aspect of criterion-referenced measurement was well-articulated by Dyer (1977) who wrote:

> "Now there is a new term, criterion-referenced testing, which all lovers of new jargon proudly hail as the great new device for mitigating the ego threat commonly attributed to odious comparisons inherent in norm-referenced testing. ...It seems to me rather ironic that when the notion of norms first found its way into the vocabulary of testing, it represented an attempt to get away from the odium of the pass-fail criterion. Norms, properly considered were to be neutral indicators of relative position; they were to say nothing whatever about passing or failing, winning or losing. Yet the notion of passing or failing is precisely the one that, in a painfully refined and exagerated form, the idea of criterion-referenced testing has brought back to the language of measurement. Just last month, one of our local papers was reporting the percentage of "passes" that each of various school districts had achieved on the reading tests administered in New Jersey this year. And in this case the criterion is the familiar old fashioned one of getting right answers on 65 per cent of the items." (p. 18-19).

We desperately need to do better than arbitrarily setting a standard at 65 per cent or 80 percent for no better reason than habit. Investigations of various methods of setting standards and of their effects are needed. This is largely unexplored territory. Some recent work (e.g., Block, 1972; Huynh, 1976b; Meskaukas, 1976; Millman, 1973; Jaeger, 1976; Shepard, 1976) may provide some guidance on how to proceed. But, there are flaws in all of the approaches that have been suggested (see, for example, Burton, 1976; Glass, 1976).

Judgmental Nature of Standard Setting

The above introductory comments are not intended to imply that standards are always set without thought or effort. Nor, are they meant to imply an absence of suggested procedures for establishing standards. Procedures have been suggested and there are examples where great care and effort has gone into the establishment of standards. Unfortunately such cases are the exception rather than the rule and even the best examples have some arbitrariness.

Millman (1973) reviewed a variety of procedures for setting standards and classified them according to whether they were dependent on (1) a comparison to the "performance of others," (2) an analysis of "item content,"

(3) an evaluation of their "educational consequences," (4) an evaluation of their "psychological and educational costs," and/or (5) adjustments for "errors due to guessing and item sampling." More recently, Glass (1976), suggested a system of six categories, five of which roughly correspond to Millman's. The additional category involves a process of "bootstrapping on other criterion scores". These categories are not mutually exclusive and hybrid procedures cutting across categories can be imagined. Regardless of category or category system, all the procedures share one important feature. They all involve judgment. They differ in terms of the sources of information and the type and extent, if any, of empirical evidence that is needed in order to make the judgment. But, "None of the procedures eliminates the need for judgment." (Millman, 1973, p. 206).

The pre-eminence of human judgement in standard setting is widely acknowledged. For instance, Messick (1975) claimed that "...all procedures for establishing performance standards require judgment at some point" (p. 957) and according to Jaeger (1976):

> "All standard-setting is judgmental. No amount of data collection, data analysis, and model building can replace the ultimate judgmental act of deciding which performances or which levels of performance are meritorious or acceptable and which are unacceptable or inadequate." (p. 22).

This does not imply that empirical data are irrelevant to the process. On the contrary, they may be of great value in facilitating judgment making, "...but they cannot be used to ferret out standards as if they existed independently of human opinions and values" (Shepard, 1976, p. 29). Thus, the procedures that are briefly described below are best considered as aids to judgment and not substitutes for it. Methods for making the judgmental process more systematic and for evaluating the process are also needed.

Types of Models

Meskaukas (1976) distinguished between two broad classes of mastery models which he refers to as "continuum models" and "state models." In a continuum model, performance levels are conceived of as varying along a continuum, but the continuum is dichotomized into two regions Persons in the upper region are referred to as masters and those in the lower region as non-masters. The dichotomization is considered useful, or essential, for certain educational decisions. There is no "natural" or obviously preferred way of forming a dichotomy when it is recognized that the performance of interest varies along a continuum, however.

In a "state model," learning is conceptualized in an all-or-none fashion. Students are either masters and therefore can perform the task, or they are not masters and cannot perform it. They either know the answers to all the items in the content domain or they don't know any of them.

In state models, the setting of standards is natural. The measurement task is simply to provide the means of deciding which of two states a person is in at a particular point in time. In a continuum model, on the other hand, the task is two-fold. First, one must decide how to dichotomize the continuum, then the measure must be used to decide in which region of the continuum the person belongs.

Meskaukas (1976) reviews procedures for setting cutting scores on tests and obtaining mastery decision rules for the two types of models. Our main concern will be with procedures for continuum models. There are two reasons for this focus: (1) only continuum models require both the judgmental task of setting a standard as well as the procedural task of specifying the decision rules, and (2) it is our contention that relatively

few areas of achievement that are broad enough to correspond to important content domains are truely all-or-none tasks. Meskaukas argued that "... a great deal of what is learned, particularly in situations where errorless replication is required, follows...[the state] model" (1976, p. 155). But errorless replication is only one half of the coin. A state model would also require that the only other possibility would be consistently replicated errors. At least while learning a content domain, some intermediate outcomes are usually possible. In some situations, arguments for a state model may actually be arguments that the mastery region should contain only perfect performance.

If the domain consists of a single addition problem, say 2 + 2, then a state model is most appealing. If the domain is expanded to include the addition of all possible pairs of one-digit numbers, however, knowing all the answers or knowing none of them are no longer the only logical possibilities. We concur with Millman's (1974) conclusion that "Because these models assume that partial knowledge or skill does not exist, they are considered unrealistic...." (P. 355).

It has been suggested that domains should be subdivided to the point that the items they contain are so homogeneous that a state model is reasonable (Macready & Merwin, 1973; Macready & Dayton, 1977). If this is done, then setting the criterion level is no longer an issue, and the problem then is reduced to assessing the probability that a person is a master and models for doing this have been developed by Macready and Dayton. The extent to which a state model can provide a useful approximation for domains that are not too narrow to be of interest is somewhat problematic, however.

Procedures

Item Content

Domain specifications and/or item content may be the primary source of information used in judging the level at which to set standards. The formal procedures reviewed by Millman (1973) for using item content to set standards involve the analysis of the individual items on a specific test. These approaches ignore the issue of setting a standard for the domain and move directing to setting a passing score for the specific sample of items that comprise the test.

The distinction that is made in discussions of test length (see below) between "criterion level" and "passing score" are often obscured in the actual setting of standards. Criterion level refers to the minimum proportion, π_0, of all items in the domain that are required for a person to be a "master", "acceptable", or "competent". The passing score, n_0, on the other hand, is the score on a particular test that is required to be classified as a master. Since the items on the test do no exhaust the domain, classification errors will be made regardless of the type of correspondence between the passing score and the criterion level. Frequently the distinction between the criterion level and the passing score is ignored. Procedures are used to establish n_0 and the π_0 is implicitly assumed to be equal to n_0 divided by the total number of items on the test, n. The two are conceptually distinct, however, and it is not necessarily desirable to set n_0 equal to n times π_0.

Ideally, standard setting procedures would be used to establish π_0 and then n_0 would be determined in view, not only of π_0, but also of the disutility associated with false-positive and false-negative errors. None of the item content procedures reviewed by Millman do this. Instead, n_0 is set directly by reviewing the items on a test and in one form or another, deciding the number of items that a "minimally acceptable person" should get right. The approaches vary in detail (see Millman, 1973 for

more description) and may involve a direct estimate of n_o (e.g., Science Research Associates, 1966) or indirect estimates of n_o by summing estimates of "...the proportion of minimally acceptable persons who would answer each item correctly" (Angoff, 1971, p. 515). Some would combine judgments of item importance and expected proportion of minimally qualified persons passing each item (e.g., Ebel, 1974; Educational Testing Service, 1976) Yet another approach (Nedelsky, 1954) involves the elimination of distractors that the minimally qualified person should be able to eliminate.

The role of judgment is obvious in all these methods. More needs to be known about the sensitivity of the results of these methods to the nature and variety of audiences involved in the judgments, to the nature and training for the task, and to the availability and nature of empirical data on the performance of various groups. Evidence regarding the inter- and intrajudge consistency is also to be desired.

The different procedures can be expected to yield different standards. (Andrews & Hecht, 1976; Glass, 1976). Such differences reveal that there is an arbitrariness to the standards that may result from a given procedure. Furthermore, the very notion of "minimal competence, on which the procedures depend, is questionable. After a consideration of the logical and psychological bases for the concept of minimal competence, Glass concluded that "the idea of minimal competence is bad logic and even worse psychology" (1976, p.32). For any non-trivally low "minimum" required for some other activity exception can surely be found. Such exceptions violate the logical basis for the label "minimum".

Educational Consequences

The educational consequences of setting standards at various levels may be used in place of, or in addition to, item content to set standards (Millman, 1973). For example, in a mastery based instructional program, the effect on learning in subsequent units of the program could be used as the primary basis for determining the desired passing score. If

experience shows that students with scores on a unit test of less than 85 percent have undue difficulty while those with higher scores generally progress well in the next unit, then 85 percent would be selected as the passing score.

Empirically investigating the educational consequences of various possible passing scores is a non-trivial task; one which has rarely been attempted. Block's (1972) investigation is relatively unique in this regard. He compared the subsequent performance of randomly formed groups of students that were required to meet various performance levels on a test of matrix algebra before moving to the next segment of instruction. The "best" cutting score was found to vary depending on the emphasis that was placed on cognitive or affective criteria in the instructional segments following the matrix algebra test. Judgement is required to determine the way in which to weigh the criteria. Indeed, judgement is required to decide what criteria should be considered. There is bound to be some arbitrariness in these judgements.

Huynh (1976b) has developed a mathematical model that provides a way of setting the standard in terms of future performance on what he calls a referral task. The probability of success on the referral task, the distribution of true proportion correct, the loss associated with a false positive decision, and the loss associated with a false negative decision are all used in the determination of the optimal criterion level, if one exists. The optimal criterion level is in turn used to solve for the cutting score on the test.

Huynh's model provides a useful conceptualization. But, it is apt to prove very demanding in practice. Setting the loss associated with each type of error requires the completion of a difficult judgmental task. Obtaining the evidence needed to establish the probability of success on

the referal task for each unit of an individualized instructional program is formidable, if not an overwhelming task. It is still more difficult to imagine how one could obtain the empirical evidence that would support the claim that a given minimal standard for high school graduation was optimal in terms of future consequences of various possible standards.

Millman (1973) suggested that "In the absence of data about the educational consequences ..." that "... a logical analysis of the subject matter and the extent of the instructional system..." (p. 209) be used as the basis for setting the passing score. Relatively high passing scores would be required where future learning was seen as clearly dependent upon the skills and knowledge of a unit, while lower passing scores would be acceptable where there was not such a clear dependency. This approach obviously requires complex judgments the difficulty of which undoubtedly varies from area to area. Where more major decisions, such as those required to set standards for high school graduation, are to be made, the determination of necessary prerequisites becomes problematic since we lack clear answers to a prior question; i.e., prerequisites for what?

Comparative Data

The idea that comparative data on the performance of others might be used in the setting of standards would seem to be an anathema to a proponent of criterion-referenced measurement. Although possibly useful for selection (Millman, 1973), the setting of a passing score so that a fixed percent of the examinees pass would generally be eschewed by a criterion-referenced measurement devotee. Indeed, getting away from the comparison of performance of an individual to that of others is precisely what is often seen as one, if not the main, advantage of criterion-referenced measurement. But, it is apt to be impractical to set standards and apply them without regard for their normative implications.

The following example of an actual experience of a school district illustrates the type of problem that can result if normative results are completely ignored. The staff of a school district was given the mandate to construct a test that would be used to establish minimum standards for high school graduation. The teachers and other professionals within the school district devoted considerable time and energy to defining the domains to be covered, constructing the test items and making the judgments necessary to set the standards. The task was taken seriously and a conscientious effort was made to set a passing score in reading and one in math that represented minimal performance levels for high school graduates. A tryout of the tests revealed, however, that application of the passing scores would result in the failure of approximately 25 percent of the students on the reading test and 45 percent on the math test. Would failure rates such as these be educationally or socially desirable? Would they be politically feasible? We think not.

It might be argued that the teachers making the judgments must have used inappropriately stringent standards. But, why would others be expected to make more appropriate judgments? It might also be argued that, if the tests were given early enough, they could identify potential failures who could be given remedial instruction before they are given alternate forms of the tests. It seems unreasonable, however, to expect that a brief period of remediation is likely to teach students the minimal essentials when that couldn't be accomplished in the previous eleven years of schooling. Certainly, our experience with the results of compensatory education would hardly make us sanguine with regard to this possibility.

Another illustration of the importance of nomative feedback comes from the Michigan Assessment Program. The assessment provided information on the accomplishment of "minimal objectives" by school districts in the state. In their critique of Michigan Assessment

Program, House, Rivers, and Stufflebeam (1974) argued that there is "... considerable reason to believe that the objectives are not minimal" (p. 6) and cite in support of this the result that not a single district in the state was achieving "minimal objectives" as they were defined. Such an outcome defies the common sense meaning of the word "minimal", to say nothing of a strict interpretation of the word.

Knowing that 25 percent or 45 percent of the students would fail to meet a standard identifies a problem but doesn't tell us what to do about it either in terms of educational changes or in changes in the standard. It is information, however, that can hardly be ignored in the overall judgmental process of setting standards.

Threats to Validity of Inferences Based on Standards

Jaeger (1976) discussed several approaches to setting standards ranging from direct methods such as the item content procedure described above to "distal" methods involving dependence on normative data or the relationship of a test to an external measure. For each method he reviewed the ways in which the validity of inferences based on the standard might be threatened. A person who meets the standard is considered a master and should be able to do certain things. The inference that the person is competent to perform certain tasks may be erroneous for a variety of reasons.

A total of thirteen threats to validity were identified by Jaeger. Different methods of setting standards are subject to different threats. But, seven threats to validity are present regardless of the method used to set standards. Four of these threats are present when the test results are used to make inferences about domain performance. An inference about domain performance is made, for example, when the score on a test is used to infer that a person is above the criterion level for the domain from which the test items are sampled. The other three threats to validity,

that are present regardless of the method used to set standards, concern inferences about ultimate criteria. This latter type of inference will be considered more fully below.

According to Jaeger (1976), when inferences are limited to statements about the performance domain from which the test items are sampled, one is always faced with at least the following four threats to validity:

> "Bias in setting domain standard [i.e., criterion level] due to inadequate domain definition."
>
> "Random error among judges who set domain standard."
>
> "Inappropriateness of item sampling procedures: bias error in sample standard." (i.e., passing score).
>
> "Inadequate item sample size: random error in sample standard" (p. 26).

With rare exceptions, such as the domain of all addition problems with two addends of one digit each, the first and third threats to validity are apt to be particularly serious. As was discussed in previous sections of this paper, adequate domain definition is crucial. It is required not only for adequate test construction, i.e., construction such that the test is representative of the domain, but also to make appropriate judgments about the desired criterion level.

Although the threats to validity of inference about domain performance are serious, they pale by comparison to those for inferences about ultimate criteria and the latter inferences are usually the ones of primary interest. In isolation, inferences about domain performance "... are often uninteresting and insufficient" (Jaeger, 1976 p. 24). When someone meets the minimum standards on a test for graduation from high school, the types of inference might be limited to ones about performance level of the individual in the well-defined domains represented by the tests. The desire for this limited type of inference, however, was not the force that has led to the demands

for minimum standards in states throughout the country. Inferences about an individual's competency for various "life roles" are also desired. For example, Kohr (1977) described a bill before the Pennsylvania state legislature that deals with minimum competency requirements. These would include, among other things, evidence that a student had minimum competency to "....'understand and perform personal finance and consumer tasks including understanding consumer finance; computing interest rates, purchasing insurance, completing personal tax forms, knowing the basis of property and other taxes and comprehending residential leases and purchasing agreements'" (from Pennsylvania House Bill 770, as quoted by Kohr, 1977, p. 2). The inference that someone understands consumer finance or comprehends residential leases goes well beyond the inference that he or she can perform X percent of the items in a domain. Not only that, but the motivating force actually comes from a concern about future behavior; i.e., how the person will perform as an adult. Concern about future performance is not unusual. In fact, "Often our interest in current performance only substitutes for our true interest in later performance, perhaps years later" (Jaeger, 1976, p. 24). In short, inferences about "ultimate criteria" are often desired and these inferences go well beyond domain performance in terms of scope, time and the degree to which they can be translated into observable behavio .

For inferences about ultimate criteria that are based upon set standards, Jaeger identified the following three threats to validity:

> "Bias error due to invalidity of domain definition."
>
> "Error due to inconsistency of domain-criterion relationship."
>
> "Bias error due to invalidity of model for domain-criterion relationship." (1976, p. 26).

Combined, the threats to validity for the two types of inference produce a list that "... is discouraging, if not mind-boggling, and our knowledge of the magnitude of errors and the severity of such validity threats is extremely limited." (Jaeger, 1976, p. 25).

The process of setting standards and of making inferences based on those standards lacks a firm foundation. This creates problems that are apt to be of more than academic interest. The setting of standards and their use for certification has already been at issue in some recent court cases (e.g., U.S. v. State of South Carolina, 1977). The apparent trend toward increased use of tests to establish that standards of minimum competency have been met for various types of certification is apt to lead to more cases involving the use of standards. Typical current practice is not likely to provide a basis that most standard setters would be comfortable defending in court.

Empirical investigations involving real domains, inferences, and judges could provide answers to some important questions about the threats to validity (see, Jaeger, 1976). The results of these investigations coupled with considerable thought might lead to a better theoretical foundation as was suggested by Jaeger. But, a concerted effort would be required and there are few signs that such an effort is underway, though calls for the needed research are not new (e.g., Airasian & Madaus, 1972; Quirk, 1974).

Are Standards Necessary?

Glass (1976) has argued that all existing methods of setting standards are so seriously flawed that they are apt to do more harm than good.. As has been noted several times, there is an arbitrariness in all of the methods of setting standards. Glass would want to avoid this arbitrariness by not using standards. He suggested that rather than trying to set absolute standards the focus should be on change. Improved performance is good and a decline in performance is bad. Absolute standards are viewed as unnecessary and, because of their arbitrariness, best avoided.

We agree with Glass' conclusion that every method of setting standards involves arbitrariness at some stage. We also endorse his suggestion that

comparisons over time have advantages over absolute judgements. We doubt, however, that all demands for standards can be so easily sidestepped. Many types of educational actions require something besides evidence of improvement. Knowledge that the performance of a student is better now than at sometime in the past is useful but it doesn't answer questions about whether the student is prepared to move on or still should receive some form of remediation.

Comparing the performance of graduates to that of previous graduates is useful. Agreement that a decline is bad, could probably be obtained from most people.* Unless all forms of certification were eliminated, however, there is still a need for a standard to determine whether the performance is sufficient to receive the certification. The standard will admittedly be based on some arbitrariness. It will not be an absolute. This does not imply, however, that a standard is worthless.

People will be categorized and differential actions taken on the basis of those categories with or without standards that are systematically set. The categorization will involve human judgement and a degree of arbitrariness with or without systematic methods. But, the more systematic methods have certain advantages. They are more explicit and can therefore more readily be made public and be subjected to debate. They need not be viewed as fixed once they are established. Indeed, as will be suggested below, part of the systematic procedure should provide for frequent review and revision.

Standards, set by any method, may be arbitrary, but they need not be capricious. Systematic procedures can reduce the chances of capriousness. Glass (1976) concluded that "setting performance standards on tests and exercises by known methods is a waste of time" (p. 49). We disagree. The setting of standards is an important problem, one that deserves more

* The recent controversy over the meaning and importance of the decline in scores on the Scholastic Aptitude Test suggests that not everyone would even agree that a decline is bad.

time and effort that it normally receives. The methods all have defects, but are better than nothing because in this case, "nothing" really means hidden or unknown standards.

Interim Suggestions

In the absence of an adequate theory on which standard setting procedures can be developed, only tentative guidelines and opinions can be offered. What we judge to be a very reasonable set of recommendations has recently been offered by Shepard (1976). She suggested that standard setting should be considered an iterative process and one that involves various audiences. The involvement of various audiences in the task allows for differences in points of view of these audiences. It gives explicit recognition to the complexity of setting a single standard when different standards would be selected by different groups of judges. While this may complicate the process when the groups are found to differ substantially it is far preferable to the misleading simplicity of using only one relevant audience to set a standard.

Shepard's suggestion that the process be an iterative one permits adjustments in the judgments based on the accumulation of information over time. Thus, in the example used above where the application of graduation standards would have resulted in an unacceptably high failure rate, this information would be fed back to the judges who might want to modify their judgments in light of the additional information.

Another of Shepard's recommendations is that normative data ought to be provided to judges for use in their deliberations. The norms are not to be used to set the standards; merely to inform the judges.

Shepard's remaining recommendation is more relevant for assessment or accountability systems than for mastery based instruction or the certification of individuals. This recommendation is that improvement rather than current performance be the standard that is employed.

Shepard's suggestions are sensible. They do not avoid the dependence on human judgment, nor could they. Neither do they eliminate Jaeger's threats to validity of the inferences based on the use of standards. We think, however, that they provide the broad guidelines within which defensible methodologies for setting standards may be developed. We also concur with Conway's (1976) conclusion that "Unless standards are established by some defensible methodology which involves careful human judgment, they will not serve their intended purposes nor will they stand up against the careful scrutiny of those who doubt their validity"(p. 35).