DOCUMENT RESUME

ED 365 147 FL 021 760

AUTHOR Hamp-Lyons, Liz; Prochnow, Sheila

TITLE The Difficulties of Difficulty: Prompts in Writing Assessment.

PUB DATE 91

NOTE 20p.; In: Sarinee, Anivan, Ed. Current Developments in Language Testing. Anthology Series 25. Paper presented at the Regional Language Centre Seminar on Language Testing and Language Programme Evaluation (April 9-12, 1990); see FL 021 757.

PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Cues; Difficulty Level; English (Second Language); *Language Tests; *Second Languages; Testing; *Test Items; *Writing Exercises; *Writing Tests

IDENTIFIERS *Michigan English Language Assessment Battery

ABSTRACT

This study investigated the effect of writing task topic on learner performance in a second-language writing test, in this case the Michigan English Language Assessment Battery designed to test proficiency in English as a Second Language. The 64 topics or "prompts" used in the test (offered as pairs of options) were categorized according to writing task type (expository/private; expository/public; argumentative/private; argumentative/public; and a combination of two or more of the previous types) and the categories assigned a level of difficulty. Scores received on the test were then correlated with topic difficulty. Contrary to expectation, the mean writing score increased rather than decreased as topic difficulty increased. Implications for test construction and for rater judgment are examined. (MSE)

THE DIFFICULTIES OF DIFFICULTY: PROMPTS IN WRITING ASSESSMENT

Liz Hamp-Lyons an:' Sheila Prochnow

INTRODUCTION

In the field of writing assessment, a growing educational industry not only in the United States but also worldwide, it is often claimed that the "prompt", the question or stimulus to which the student must write a response, is a key variable. Maintaining consistent and accurate judgments of writing quality, it is argued, requires prompts which are of parallel difficulty. There are two problems with this. First, a survey of the writing assessment literature, in both L1 (Benton and Blohm, 1986; Brossell, 1983; Brossell and Ash, 1984; Crowhurst and Piche, 1979; Freedman, 1983; Hoetker and Brossell, 1986, 1989; Pollitt and Hutchinson, 1987; Quellmalz et al, 1982; Ruth and Murphy, 1988; Smith et al, 1985) and L2 (Carlson et al, 1985; Carlson and Bridgeman, 1986; Chiste and O'Shea, 1988; Cummings, 1989; Hirokawa and Swales, 1986; Park, 1988; Reid, 1989 (in press); Spaan, 1989; Tedick, 1989; Hamp-Lyons, 1990), reveals conflicting evidence and opinions on this. Second (and probably causally prior), we do not yet have tools which enable us to give good answers to the questions of how difficult tasks on writing tests are (Pollitt and Hutchinson, 1985). Classical statistical methods have typically been used, but are unable to provide sufficiently detailed information about the complex interactions and behaviors that underlie writing ability (Hamp-Lyons, 1987). Both g-theory (Bachman, 1990) and item response theory (Davidson, in press) offer more potential but require either or both costly software and statistical expertise typically not available even in moderate-sized testing agencies, and certainly not to most schools-based writing assessment programs.

An entirely different direction in education research at the moment, however, is toward the use of judgments, attitude surveys, experiential data such as verbal protocols, and a generally humanistic orientation. Looking in such a direction we see that language teachers and essay scorers often feel quite strongly that they can judge how difficult or easy a specific writing test prompt is, and are frequently heard to say that certain prompts are problematic because they are easier or harder than others. This study attempts to treat such observations and judgments as data, looking at the evidence for teachers' and raters' claims. If such claims are borne out, judgments could be of important help in establishing prompt difficulty prior to large-scale prompt piloting, and reducing the problematic need to discard many prompts because of failure at the pilot stage.

II. BACKGROUND

The MELAB, a test of English language proficiency similar to the TOEFL but containing a direct writing component, is developed by the Testing Division of the University of Michigan's English Language Institute and administered in the US and in 120 countries and over 400 cities around the world. In addition to the writing component, the test battery includes a listening component and a grammar/cloze/vocabulary/reading component (referred to as "Part 3"). There is also an optional speaking component, consisting of an oral interview. Scores on the 3 obligatory components are averaged to obtain a final MELAB score, and both component and final scores are reported. Scores are used by college or university admissions officers and potential employers in the United States in making decisions as to whether a candidate is proficient enough to carry out academic work or professional duties in English.

The writing component of the test is a 30-minute impromptu task, for which candidates are offered a choice of two topics. Topics are brief in length, usually no more than three or four lines, and intended to be generally accessible in content and prior

assumptions to all candidates. Topic development is an ongoing activity of the Testing Division, and prompts are regularly added to and dropped from the topic pool. In preparation of each test administration, topic sets are drawn from the topic pool on a rotating basis, so as to avoid repeated use of any particular topic set at any test administration site. Currently, 32 topic sets (i.e. 64 separate topics) are being used in MELAB administrations in the US and abroad and it is these topic sets, comprising 64 separate prompts, which examined in this study.

MELAB compositions are scored by trained raters using a modified holistic scoring system and a ten-point rating scale (see Appendix 1). Each composition is read independently by two readers, and by three when the first two disagree by more than one scale point. The two closest scores are averaged to obtain a final writing score. Thus, there are 19 possible MELAB composition scores (the 10 scale points and 9 averaged score points falling in between them). Compositions from all administration sites are sent to the Testing Division, where they are scored by trained MELAB raters. Inter-rater reliability for the MELAB composition is .90.

II. METHOD

Since research to date has not defined what makes writing test topics difficult or easy, our first step toward obtaining expert judgments had to be to design a scale for rating topic difficulty. Lacking prior models to build on, we chose a simple scale of 1 to 3, without descriptions for raters to use other than 1=easy, 2=average difficulty and 3=hard. Next the scale and rating procedures were introduced to 2 trained MELAB composition readers and 2 ESL writing experts, who each used the scale to assign difficulty ratings to 64 MELAB topics (32 topic sets). The four raters' difficulty ratings were then summed for each topic, resulting in one overall difficulty rating per topic, from 4 (complete agreement on a 1=easy rating) to 12 (complete agreement on a 3-hard rating). We then compared "topic difficulty" (the sum of judgments of the difficulty of each topic) to actual writing scores obtained on those topics, using 8,497 cases taken from MELAB tests administered in the period 1985-89.

Next, we categorized the 64 prompts according to the type of writing task each represents. We began with application of the topic type categories developed by Bridgeman and Carlson (1983) for their study of university faculty topic preferences. However, judges found that of Bridgeman and Carlson's nine categories, three were not usable because there were no instances of such topic types in the dataset; further, only about half of the dataset fit in the remaining six categories. The remaining half of the topics were generally found to call either for expository or for argumentative writing. The expository/argumentative distinction is of course one which has been made in many previous studies (Rubin and Piche, 1979; Crowhurst and Piche, 1979; Mohan and Lo, 1985; Quellmalz et al, 1982; etc). Another noticeable difference between topics is that some call for the writer to take a public orientation toward the subject matter to be discussed whereas others call for a more private orientation. Similar distinctions between prompts were noted by Bridgeman and Carlson (1983), who discuss differences in their various topic types in terms of what they call "degree of personal involvement", and by Hoetker and Brossell (1989) in their study of variations in degree of rhetorical specification and of "stance" required of the writer.

Based on these distinctions, we created a set of 5 task type categories: (1) expository/private; (2) expository/public; (3) argumentative/private; (4) argumentative/public, and (5) combination (a topic which calls for more than one mode of discourse and/or more than one orientation; an example of such a topic might be one which calls for both exposition and argumentation, or one which calls for both a personal and public stance, or even one which calls for both modes and both orientations). Examples of the five types are shown in Appendix 2. All 64 topics were independently assigned to the category, and then the few differences in categorization were resolved through discussion. Following a commonly held assumption often found in the literature (Bridgeman and Carlson, 1983; Hoetker and Brossell, 1989), we hypothesized that some topic type categories would be judged generally more difficult than others, and that expository/private topics would, on average, be judged least difficult, and argumentative/public topics most difficult. To test this prediction, we used a two-way analysis of variance, setting topic difficulty as the dependent variable and topic type as the independent va~iable.

III. RESULTS and INTERPRETATIONS

Topic Difficulty

When we displayed the summed topic difficulties based on four judges' scores for each of the 64 prompts, we obtained the result shown in Table 1:

Table 1
Topi: Difficulty for 64 MELAB Prompts

Topic Difficulty	Topic Set	No	Topic Difficulty	Topic Set	No.
4	11	A	8	42	A
4	27	A	8	43	B
4	31	B	8	44	B
4	33	B	8	45	A
4	34	B	8	49	B
5	46	B	9	12	A
5	49	A	9	18	B
6	30	B	9	21	B
6	35	B	9	22	B
6	41	A	9	23	A
6	47	A	9	24	B
7	12	B	9	31	A
7	22	A	9	33	A
7	29	B	9	35	A
7	34	A	9	46	A
7	37	B	9	50	B
7	38	A	10	11	B
7	40	A	10	13	A
7	40	B	10	24	A
7	43	A	10	29	A
8	10	A	10	30	A
8	21	A	10	39	B
8	23	B	10	42	B
8	26	A	10	45	B
8	28	A	10	47	B
8	28	B	11	10	B
8	33	A	11	13	B
8	32	B	11	18	A
8	37	A	11	26	B
8	38	B	11	27	B
8	39	A	11	44	A
8	41	B	12	50	A

Most prompts had a difficulty score around the middle of the overall difficulty scale (i.e. 8). This is either because most prompts are moderately difficult, or, and more likely, because of the low reliability of our judges' judgments. The reliability of the prompt difficulty judgments, using Cronbach's alpha, was .55.

And here was our first difficulty, and our first piece of interesting data: it seemed that claims that easy readers and language teachers can judge prompt difficulty, while not precisely untrue, are also not precisely true, and certainly not true enough for a well-grounded statistical study. When we looked at the data to discover whether the judgments of topic difficulty could predict writing score, using a two-way analysis of variance, in which writing score was the dependent variable and topic difficulty was the dependent variable, we found that our predictions were almost exactly the reverse of what actually happen (see Table 2).

Table 2: Difficulty Judgments and Writing Scores

```
ANALYSIS OF VARIANCE OF 8.CATSCOR  N= 8583 OUT OF 8583

SOURCE              DF  SUM OF SQRS  MEAN SQR  F-STATISTIC SIGNIF

BETWEEN              8    413.31     51.663      5.2529     .0000
WITHIN            8574    84327.     9.8352
TOTAL             8582    84740.     (RANDOM EFFECTS STATISTICS)


ETA= .0698  ETA-SQR= .0049  (VAR COMP= .46927 -1  %VAR AMONG= .47)

SUMDIFF       N    MEAN    VARIANCE   STD DEV

(4)         679  8.9455    8.4439     2.9058
(5)         113  8.9823    6.5533     2.5599
(6)         737  9.1045    9.3872     3.0638
(7)        1539  9.4048    10.579     3.2526
(8)        2325  9.4705    9.5634     3.0925
(9)        1501  9.5776    10.851     3.2941
(10)       1040  9.6519    9.1242     3.0206
(11)        577  9.7660    10.763     3.2807
(12)         72  9.4028    7.1453     2.6731

GRAND      8583  9.4394    9.8742     3.1423
```

Mean writing score increased, rather than decreased, as topic difficulty increased, except for topics in the group judged as most difficult (those whose summed rating was 12, meaning all four judges had rated them as 3=difficult). As shown in Figure 1, topic difficulty as measured by "expert" judgment is unable to explain any of the variance in MELAB writing score.

Figure 1: ANOVA

Topic Difficulty and Writing Score

```
ANALYSIS OF VARIANCE OF 8 CATSCOR  N= 8583 OUT OF 10447

      SOURCE               DF   SUM SQRS    MEAN SQR    F-STAT     SIGNIF

      REGRESSION            1   372.05      372.05      37 841       0000
      ERROR              8581   84368.      9.8320
      TOTAL              8582   84740.

      MULT R=  06626   R-SQR= .00439 SE= 3.1356

      VARIABLE        PARTIAL     COEFF       STD ERROR     T-STAT     SIGNIF

      CONSTANT                    8.5291       .15179       56.190     0.
   16.SUMDIFF         .06626      .11456       .18623 -1    6.1515      .0000
```

Further, while the effect of judged topic difficulty on writing score is significant (p=.0000), the magnitude of the effect is about 18 times smaller than would be expected, considering the relative lengths of the writing and topic difficulty scales. That is, since the writing scale is approximately twice as long as the topic difficulty scale (19 points vs. 11 points), we would expect, assuming "even" writing proficiency (i.e. that writing proficiency increases in steps that are all of equal width) that every 1-point increase in topic difficulty would be associated with a 2-point decrease in writing score; instead, the coefficient for topic difficulty effect (.11456) indicates that a 1-point increase in topic difficulty is actually, on average, associated with only about a 1/10-point increase in writing score. Also, it should be noted that such an increase is of little practical consequence, since a change of less than a point in MELAB writing score would have no effect either on reported level of writing performance or on final MELAB score.

Task Type Difficulty

We had hypothesized that when topics were categorized according to topic type, the topic type categories would vary in judged difficulty level, and that the overall difficulty level of categories would vary along two continua: "orientation" (a private/public continuum), and "response mode" (an expository/argumentative continuum) (see Figure 2).

Figure 2: Response Mode, Orientation and Topic Difficulty

Predictions

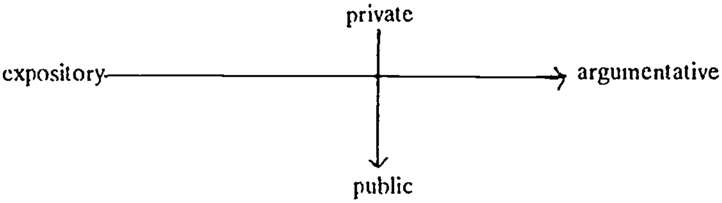

Table 3 shows the difficulty ratings for each category or "response mode":

Table 3: Response Modes and Difficulty Ratings

Topic Category Groupings

ExpPers		ExpPub		ArgPers		ArgPub		Comb.	
#	Diff	#	Diff	#	Diff	#	Diff	#	Diff
11A	4	40B	7	49A	5	37A	8	30B	6
27A	4	10A	8	12B	7	39A	8	34A	7
29B	4	32A	8	38A	7	43B	8	28A	8
31B	4	41B	8	38B	8	21B	9	45A	8
33B	4	49B	8	42A	8	22B	9	26B	11
34B	4	12A	9	35A	9	24B	9		
46B	5	18B	9	24A	10	31A	9		
35B	6	23A	9	29A	10	33A	9		
41A	6	10B	11	39B	10	46A	9		
47A	6	18A	11	45B	10	11B	10		
22A	7					13A	10		
37B	7					42B	10		
21A	8					13B	11		
23B	8					27B	11		
26A	8					44A	11		
28B	8					50A	12		
32B	8								
50B	9								
30A	10								
47B	10								
$\bar{x}$ diff=6.528 $\bar{x}$ wr =9.063		$\bar{x}$ diff=8.746 $\bar{x}$ wr =9.398		$\bar{x}$ diff=8.440 $\bar{x}$ wr.=9.359		$\bar{x}$ diff=8.932 $\bar{x}$ wr =9.904		$\bar{x}$ diff=7.713 $\bar{x}$ wr.=9 517	

overa.l $\bar{x}$ diff=7.9455 overall $\bar{x}$ wr=9.4709

We conducted an ANOVA, shown in Figure 3, which showed that our predictions were correct: prompts categorized as expository/private by judges are, on average, judged easiest and those categorized as argumentative/public are judged hardest.

Figure 3: ANOVA

Topic Difficulty Judgments and Response Mode Difficulty

Judgments

```
ANALYSIS OF VARIANCE OF 16.SUMDIFF  N= 8497 OUT OF 8497

SOURCE                DF  SUM OF SQRS  MEAN SQR  F-STATISTIC SIGNIF

BETWEEN                4     8635.0     2158.8      998.42     0.
WITHIN              8492     18361.     2.1622
TOTAL               8496     26996.     (RANDOM EFFECTS STATISTICS)

ETA=  5656  ETA-SQR= .3199  (VAR COMP= 1.3219  %VAR AMONG= 37 94)


CATEGORY      N     MEAN     VARIANCE    STD DEV

EXPPRI     2538   6.5284     2.8666      1.6931
EXPPUB     1210   8.7463     1.6618      1.2891
ARGPRI     1543   8.4407     1.7447      1.3209
ARGPUB     2417   8.9326     1.6482      1.2838
COMBIN      789   7.7136     3.0549      1.7478

GRAND      8497   7.9854     3.1775      1.7826


  CONTRAST
  OBSERVED     PREDICTED     F-STAT     SIGNIF

  -2.0986      -0.           892.49     0.
  -2 7098      -0.           1488.0     0.
  -1.7261      -0.           603.74     0.
```

Since the two sets of judgments were made by the same judges, albeit six months apart, such a finding is to be expected.

Judgments and Writing Scores

When we looked at the relationships between our "expert" judgments of topic difficulty and task type, and compared them with writing scores, our predictions were not upheld by the data. We had hypothesized that topics in the category judged most difficult (argumentative/public) would get the lowest scores, while topics in the category judged least difficult (expository/private) would get the highest scores, with topics in the other categories falling in between. To test this hypothesis, we conducted a two-way analysis of variance, in which writing score was the dependent variable and topic type the independent variable. The results of the ANOVA, shown in Figure 4, reveal that our predictions were exactly the reverse of what actually happened: on average, expository/private topics are associated with the lowest writing scores and argumentative/public the highest.

Figure 4: ANOVA

Writing Performance for Prompt Categories

```
ANALYSIS OF  VARIANCE OF  8.CATSCOR   N= 8497 OUT OF 8497

SOURCE                 DF  SUM OF  SQRS   MEAN SQR   F-STATISTIC SIGNIF

BETWEEN                 4      896.71       224.18       22.899      .0000
WITHIN               8492      83137.       9.7900
TOTAL                8496      84034.      (RANDOM EFFECTS STATISTICS)

ETA=   1033   ETA-SQR= .0107   (VAR COMP= .13141   %VAR AMONG= 1.32)


 CATEGORY        N     MEAN     VARIANCE    STD DEV

 EXPPRI       2538   9.0634      8.9849      2.9976
 EXPPUB       1210   9.3983      11.348      3.3887
 ARGPRI       1543   9.3597      9.9127      3.1484
 ARGPUB       2417   9.9040      9.8762      3.1107
 COMBIN        789   9.5171      10.100      3.1781

 GRAND        8497   9.4462      9.8910      3.1450

CONTRAST
OBSERVED     PREDICTED     F-STAT      SIGNIF

-.80192      -0.           28.781      .0000
-.87924      -0.           34.599      .0000
 .20941      -0.           1.9627      .1613
```

We then looked at the combined effects of topic difficulty and prompt categories, predicting that topics with the lowest difficulty ratings and of the easiest (expository/private) type would get the highest writing scores, and that topics with the highest difficulty ratings and of the hardest (argumentative/public) type would get the lowest writing scores. To test this, we again used a two-way analysis of variance, this time selecting writing score as the dependent variable and topic difficulty and topic type as the independent variables. It should be noted that in order to be able to use ANOVA for this analysis, we had to collapse the number of difficulty levels from 9 to 2, in order to eliminate a number of empty cells in the ANOVA table (i.e. some topic types had only been assigned a limited range of difficulty ratings). The results of this analysis are shown in Figure 5.

Figure 5: ANOVA

Topic Difficulty Judgments, Prompt Categories, and Writing

Performance

diffic	type	COUNT	CELL MEANS	ST DEV
1	expri	1647	8.99454	3.01525
1	expub	215	8.27442	3.26895
1	argpri	290	9.60690	3.11886
1	argpub	431	9.97680	3.08627
1	combin	399	9.62406	3.28068
2	expri	891	9.19080	2.96185
2	expub	995	9.64121	3.34214
2	argpri	1253	9.30247	3.15372
2	argpub	1986	9.88822	3.11648
2	combin	390	9.40769	3.06995

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F	TAIL PROB
MEAN	451627.86938	1	451627.86938	46319.54	0.0
diffic	46.57869	1	46.57869		0.0289
type	769.24715	4	192.31179		0.0
dt	357.94852	4	89.48713		0.0000
ERROR	82750.52196	8487	9.75027		

As the ANOVA suggests and Table 4 shows clearly, our predictions were again almost the reverse of what actually happened: expository/private topics judged easiest (expri 1), as a group had the second lowest mean writing score, while argumentative/public topics judged most difficult, as a group had the second highest mean writing score.

Table 4:

Combined Effects of Topic Difficulty and Topic Type

x writing score	topic type & difficulty	
8.27442	expository/public	1
8.99454	expository/private	1
9.19080	expository/private	2
9.30247	argumentative/private	2
9.40769	combination	2
9.60690	argumentative/private	1
9.62406	combination	1
9.64121	expository/public	2
9.88822	argumentative/public	2
9.97680	argumentative/public	1

IV. DISCUSSION

Thus, patterns of relationship between topic difficulty, type and writing performance which we predicted based on commonly held assumptions were not matched by our writing score data. What we did find were unexpected but interesting patterns which should serve both to inform the item writing stage of direct writing test development, and to define questions about the effects of topic type and difficulty on writing performance which can be explored in future studies.

Several intriguing questions for further study arise from possible explanations for the patterns we did discover in our data. One possible explanation is that our judges may have misperceived what is and is not difficulty for MELAB candidates to write about. A common perception about writing test topics is that certain types of topics are more cognitively demanding than others, and that writers sill have more difficulty writing on these. Yet, it may be that either what judges perceive a. cognitively demanding to ESL writers is in fact not, or alternately, that is not necessarily harder for ESL writers to write about the topics judged as more cognitively demanding while some L1 studies have concluded that personal or private topics are easier for L1 writers than impersonal or public ones, and that argumentative topics are more difficult to write on than topics calling for other discourse modes, these L1 findings do not necessarily generalize to ESL writers.

Another possible explanation for the patterns we discovered is that perhaps more competent writers choose hard topics and less competent writers choose easy topics. In fact, there is some indication in our data that this may be true. We conducted a preliminary investigation of this question, using information provided by Part 3 scores of candidates in our dataset. The Part 3 component is a 75-minute multiple choice grammar/cloze/vocabulary/reading test, for which reliability has been measured at .96(KR21). The Pearson correlation between Part 3 and writing component scores is .73, which is generally interpreted to mean that both component are measuring, to some extent, general language proficiency. We assumed, for our investigation of the above question, that students with a high general language proficiency (as measured by Part 3) will tend to have high writing proficiency. In our investigation we examined mean indeed been chosen by candidates with higher mean Part 3 scores. We found this to be true for 15 out of 32--nearly half--of the topic sets; thus, half of the time, general language proficiency and topic choice could account for the definite patterns of relationship we observed between judged topic difficulty, topic type and writing performance. One of these 15 sets, set 27, was used in a study by Spaan (1989), in which the same writers wrote on both topics in the set (A and B). While she found that, overall, there was not a significant difference between scores on the 2 topics, significant differences did occur for 7 subjects in her study. She attributed these differences mostly to some subjects apparently possessing a great deal more subject matter knowledge about one topic than the other.

A further possible explanation for the relationship we observed between difficulty judgments and writing scores could be that harder topics, while perhaps more difficult to write on, push students toward better, rather than worse writing performance. This question was also explored through an investigation of topic difficulty judgments, mean Part 3 scores and mean writing scores for single topics in out dataset. We found in our dataset 3 topics whose means Part 3 scores were below average, but whose mean writing scores were average, and which were judged as "hard"(11 or 12, argumentative/public). One of these topics asked writers to argue for or against US import restrictions on Japanese cars; another asked writers to argue for or against governments treating illegal aliens differently based on their different reasons for entering; the other asked writers to argue for or against socialized medicine. The disparity between Part 3 and writing performance on these topics, coupled with the fact that they were judged as difficult, suggests that perhaps topic difficulty was an intervening variable positively influencing the writing performance of candidates who wrote on these particular topics. To thoroughly test this possibility, future studies could be conducted in which all candidates write on both topics in these sets.

A related possibility is that perhaps topic difficulty has an influence, not necessarily on actual quality of writing performance, but on raters' evaluation of that performance. That is, perhaps MELAB composition raters, consciously or subconsciously, adjust their scores to compensate for, or even reward, choice of a difficult topic. In discussions between raters involved in direct writing assessment, it is not uncommon for raters to express concern that certain topics are harder to write on than others, and that writers should therefore be given "extra credit" for having attempted a difficult topic. Whether or not these concerns translate into actual scoring adjustments is an important issue for direct writing assessment research.

V. CONCLUSION

In sum, the findings of this study provide us with information about topic difficulty judgments and writing performance without which we could effectively proceed to design and carry out research aimed at answering the above questions. In other words, we must first test our assumptions about topic

difficulty, allowing us to form valid constructs about topic difficulty, allowing us to form valid constructs about topic difficulty effect; only then can we proceed to carry out meaningful investigation of the effect of topic type and difficulty on writing performance.

REFERENCES

Bachman, Lyle. 1990. Fundamental Considerations in Language Testing. London, England: Oxford University Press.

Benton, S.L. and P.J. Blohm. 1986. Effect of question type and position on measures of conceptual elaboration in writing. Research in the Teaching of English. 20: 98-108

Bridgeman, Brent and Sybil Carlson. 1983. A Survey of Academic Writing Tasks Required of Graduate and Undergraduate Foreign Students. TOEFL Research Report No. 15. Princeton, New Jersey: Educational Testing Service.

Brossell, Gordon. 1986. Current research and unanswered questions in writing assessment. In Greenberg, Karen S. Harvey S. Weiner and Richard S. Donovan (Eds). Writing Assessment: Issues and Strategies (168-182). New York: Longman.

Brossell, Gordon. 1983. Rhetorical specification in essay examination topics. College English,45: 165-173.

Brossell, Gordon and Barbara Hoetker Ash. 1984. An experiment with the wording of essay topics. College Composition and Communication, 35: 423-425.

Carlson, Sybiil and Brent Bridgeman. 1986. Testing ESL student writers. In Greenberg, Karen L., Harvey S Weiner and Richard A Donovan (Eds). Writing Assessment: Issues and Strategies(126-152). New York: Longman.

Carlson, Sybil.,Brent Bridgeman and Janet Waanders. 1985. The Relationship of Admission Test Scores to Writing Performance of Native and Nonnative Speakers of English. TOEFL Research Report 19. Princeton, New Jersey: Educational Testing Service.

Chiste, Katherine and Judith O'Shea. 1988. Patterns of Question Selection and Writing Performance of ESL Students. TESOL Quarterly, 22(4): 681-684.

Crowhurst, Marion and Gene Piche. 1979. Audience and mode of discourse effects on syntactic complexity of writing at two grade levels. Research in the Teaching of English, 13; 101-110.

Cummings,A. 1989. Writing expertise and second language proficiency. Language Learning,39(1): 81-141.

Davidson,Fred. Statistical support for reader training. In Hamp-Lyons, Liz (ed). Assessing Second Language Writing in Academic Contexts. Norwood, New Jersey: Ablex Publishing Company. In press.

Freedman, Sarah. 1983. Student characteristics and essay test writing performance. Research in the Teaching of English, '7: 313-325.

Greenberg, Karen L. 1986. The development and validation of the TOEFL writing test: a discourse of TOEFL research reports 15 and 19. TESOL Quarterly, 20(3): 531-544.

Hamp-Lyons, Liz. 1987. Testing Second Language Writing in Academic Settings. University of Edinburgh. Unpublished doctoral dissertion.

---------. 1988. The product before: task related influences on the writer. In Robinson, P (Ed). Academic Writing : Process and Product. London: Macmillan in Pauline association with the British Council.

--------. 1990. Second language writing: assessment issues. In Kroll Barbara, (Ed). Second Language Writing: Issues and Options. New York: Macmillan.

Hirokawa, Keiko and John Swales. 1986. The effects of modifying the formality level of ESL composition questions. TESOL Quarterly, 20(2): 343-345.

Hoetker, James. 1982. Essay exam topics and student writing. College Composition and Communication,33: 377-91

Hoetker, James and Gordon Brossell. 1989. The effects of systematic variations in essay topics on the writing performance of college freshman. College Composition and Communication,40(4): 414-421.

Hoetker, James and Gordon Brossell. 1986. A procedure for writing content-fair essay examination topics for large scale writing assignments. College Composition and Communication, 37(3): 328-335.

Johns, Ann. Faculty assessment of student literacy skills: implications for ESL/EFL writing assessment. In Hamp-Lyons, Liz(Ed). Assessing Second Language Writing in Academic Contexts. Norwood, New Jersey: Ablex Publishing Company. In press.

Lunsford, Andrea 1986. The past and future of writing assessment. In Greenberg, Karen L., Harvey S. Weiner and Richard a. Donovan (Eds). Writing Assessment: Issues and Strategies. New York: Longman.

Meredith Vana H. and Paul Williams. 1984. Issues in direct writing assessment: problem identification and control. Educational Measurement: Issues and Practice. Spring 1984, 11-15, 35.

Mohan, Bernard and Winnie An Yeung Lo. 1985. Academic writing and Chinese students: transfer and developmental factors. TESOL QUARTERLY, 19(3): 515-534.

Park, Young Mok. 1988. Academic and ethnic background as factors affecting writing performance. In Purves, Alan (Ed). Writing Across Languages and Cultures: Issues in Cross Cultural Rhetoric. Newbury Park, California: Sage Publications.

Pollitt, Alistair and Carolyn Hutchinson. 1987. Calibrating gradeu assessment: Rasch partial credit analysis of performance in writing. Language Testing, 4(1): 72-92.

Pollitt, Alistair, Carolyn Hutchinson, Noel Gutwhistle and 1985. What Makes Exam Questions Difficult: An Analysis of 'O' Grade Questions and Answers. Research Reports for Teachers, No. 2. Edinburgh: Scottish Academic Press.

Purves, Alan, Anna Soter, Sanli Takala and A Vahapassi. 1984. Toward a domain referenced system for classifying composition assignments. Research in the Teaching of English, 18: 385-409.

Quellmalz, Edys. 1984. Toward a successful large scale writing assessment: where are we now? where do we go from here? Educational Measurement: Issues and Practice, Spring 1984: 29-32, 35.

Quellmalz, Edys, Frank Capell and Chih-Ping Chou. 1982. Effects of discourse and response mode on measurement of writing competence. Journal of Educational Measurement, 19(4): 241-258.

Reid, Joy. 1990. Responding to difference topic types: a quantitative analysis In Kroll, Barbara (Ed). Second Language Writing Assessment: Issues and Options. New York: Macmillan.

Ruth, Leo and Sandra Murphy. 1988. Designing Writing Tasks for the Assessment of Writing. Norwood, New Jersey: Ablex Publishing Company.

_________________. 1984. Designing topics for writing assessment: problems of meaning. College Composition and Communication, 35: 410-422.

Smith, W. et al. 1985. Some effects of varying the structure of a topic on college students' writing. Written Communication, 2(1): 73-89.

Spaan, Mary. 1989. Essay tests: What's in a prompt? Paper presented at 1989 TESOL convention, San Antonio, Texas, March 1989.

Tedick, Diane. 1989. Second language writing assessment: bridging the gap between theory and practice. Paper presented at the 89th Annual Convention of the American Educational Research Association, San Francisco, California, March 1989.

APPENDIX 1

MICHIGAN ENGLISH LANGUAGE ASSESSMENT BATTERY

COMPOSITION GLOBAL PROFICIENCY DESCRIPTIONS

(See reverse for composition codes)

97

Topic is richly and fully developed. Flexible use of a wide range of syntactic (sentence level) structures, and accurate morphological (word forms) control. There is a wide range of appropriately used vocabulary. Organization is appropriate and effective, and there is excellent control of connection. Spelling and punctuation appear error free.

93

Topic is fully and complexly developed. Flexible use of a wide range of syntactic structures. Morphological control is nearly always accurate. Vocabulary is broad and appropriately used. Organization is well controlled and appropriate to the material, and the writing is well connected. Spelling and punctuation errors are not distracting.

87

Topic is well developed, with acknowledgement of its complexity. Varied syntactic structures are used with some flexibility, and there is good morphological control. Vocabulary is broad and usually used appropriately. Organization is controlled and generally appropriate to the material, and there are few problems with connection. Spelling and punctuation errors are not distracting.

83

Topic is generally clearly and completely developed, with at least some acknowledgement of its complexity. Both simple and complex syntactic structures are generally adequately used; there is adequate morphological control. Vocabulary use shows some flexibility, and is usually appropriate. Organization is controlled and shows some appropriacy to the material, and connection is usually adequate. Spelling and punctuation errors are sometimes distracting.

77

Topic is developed clearly but not completely and without acknowledging its complexity. Both simple and complex syntactic structures are present; in some "77" essays these are cautiously and accurately used while in others there is more fluency and less accuracy. Morphological control is inconsistent. Vocabulary is adequate, but may sometimes be inappropriately used. Organization is generally controlled, while connection is sometimes absent or unsuccessful. Spelling and punctuation errors are sometimes distracting.

73

Topic development is present, although limited by incompleteness, lack of clarity, or lack of focus. The topic may be treated as though it has only one dimension, or only one point of view is possible. In some "73" essays both simple and complex syntactic structures are present, but with many errors; others have accurate syntax but are very restricted in the range of language attempted. Morphological control is inconsistent. Vocabulary is sometimes inadequate, and sometimes inappropriately used. Organization is partially controlled, while connection is often absent or unsuccessful. Spelling and punctuation errors are sometimes distracting.

67

Topic development is present but restricted, and often incomplete or unclear. Simple syntactic structures dominate, with many errors; complex syntactic structures, if present, are not controlled. Lacks morphological control. Narrow and simple vocabulary usually approximates meaning but is often inappropriately used. Organization, when apparent, is poorly controlled, and little or no connection is apparent. Spelling and punctuation errors are often distracting.

63

Contains little sign of topic development. Simple syntactic structures are present, but with many errors; lacks morphological control. Narrow and simple vocabulary inhibits communication. There is little or no organization, and no connection apparent. Spelling and punctuation errors often cause serious interference.

57

Often extremely short; contains only fragmentary communication about the topic. There is little syntactic or morphological control. Vocabulary is highly restricted and inaccurately used. No organization or connection are apparent. Spelling is often indecipherable and punctuation is missing or appears random.

53

Extremely short, usually about 40 words or less. Communicates nothing, and is often copied directly from the prompt. There is little sign of syntactic or morphological control. Vocabulary is extremely restricted and repetitively used. There is no apparent organization or connection. Spelling is often indecipherable and punctuation is missing or appears random.

N.O.T.

N.O.T. (Not On Topic) indicates a composition written on a topic completely different from any of those assigned; it does not indicate that a writer has merely digressed from or misinterpreted a topic. N.O.T. compositions often appear prepared and memorized. They are not assigned scores or codes.

1/10/90

APPENDIX 1 (CONT'D)

MICHIGAN ENGLISH LANGUAGE ASSESSMENT BATTERY

COMPOSITION CODES

(See reverse for composition global proficiency descriptions)

NOTE: the codes are meant to indicate that a certain feature is ESPECIALLY GOOD OR BAD IN COMPARISON TO THE OVERALL LEVEL OF THE WRITING

CODE	INTERPRETATION
a	topic especially poorly or incompletely developed
b	topic especially well developed
c	organization especially inappropriate to material
d	organization especially uncontrolled
e	organization especially well controlled
f	connection especially poor
g	connection especially smooth
h	syntactic (sentence level) structures especially simple
i	syntactic structures especially complex
j	syntactic structures especially uncontrolled
k	syntactic structures especially controlled
l	especially poor morphological (word forms) control
m	especially good morphological control
n	vocabulary especially narrow
o	vocabulary especially broad
p	vocabulary use especially inappropriate
q	vocabulary use especially appropriate
r	spelling especially inaccurate
s	punctuation especially inaccurate
t	paragraph divisions missing or apparently random
u	handwriting illegible or nearly illegible
v	question misinterpreted or not addressed
w	reduced one score level for unusual shortness
x	other (write-in: see score report)

APPENDIX 2: Samples of Topic Categories

Type 1: EXPOSITORY/PRIVATE

When you go to a party, do you usually talk a lot, or prefer to listen? What does this show about your personality?

Type 2: EXPOSITORY/PUBLIC

Imagine that you are in charge of establishing the first colony on the moon. What kind of people would you choose to take with you? What qualities and skills would they have?

Type 3: ARGUMENTATIVE/PRIVATE

A good friend of yours asks for advice about whether to work and make money of whether to continue school. What advice would you give him/her?

Type 4: ARGUMENTATIVE/PUBLIC

What is you opinion of mercenary soldiers (those who are hired to fight for a country other than their own?)
Discuss.

Type 5: COMBINATION (ARGUMENTATIVE/EXPOSITORY/PUBLIC)

People who have been seriously injured can be kept alive by machines. Do you think they should be kept alive at great expense, or allowed to die? Explain your reasons.