ED 287 285 FL 016 941

AUTHOR Henning, Grant; Davidson, Fred
TITLE Scalar Analysis of Composition Ratings.
PUB DATE 87
NOTE 16p.; In: Language Testing Research; Selected Papers from the Colloquium (Monterey, California, February 27-28, 1986); see FL 016 938.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Cognitive Processes; *English (Second Language); Higher Education; *Language Tests; *Rating Scales; Second Language Instruction; *Test Interpretation; Test Reliability; *Writing (Composition); *Writing Evaluation
IDENTIFIERS Rasch Model; University of California Los Angeles

ABSTRACT

A study evaluated a composition grading scale used in the University of California at Los Angeles' program of English as a second language (ESL). It looked at (1) the comparative difficulty of the subscale categories; (2) the interrelationships of the subscale categories; (3) the incremental arrangement of the subscale categories and performance levels on a probabilistic writing performance continuum; (4) scalar reliability; (5) the fit of persons, categories, and performance levels with the predictions of the Rasch model; and (6) characteristics of the writing of misfitting persons. In general, results indicated that while high measurement accuracy was observed, misfit was present at the midpoints of the performance scales. Misfitting subjects were not disproportionately representative of any one sex or language group background, but did show writing characteristics involving low content ratings relative to other subscale categories and having abnormally low cohesion, which may represent disordered language and thought patterns. Further research with a larger subject sample is recommended. (MSE)

SCALAR ANALYSIS OF COMPOSITION RATINGS

Grant Henning
Fred Davidson
University of California, Los Angeles.

Introduction

Inasmuch as there is reliance on various kinds of rating scales in evaluating language learner performance in writing and speaking, it seems appropriate to investigate efficacy of such scales for use with particular examinees for particular decision-making purposes. Classical analyses of language performance scales have usually been limited to considerations of interrater reliability and of concurrent, predictive, or construct validity (Henning, 1983, 1984).

In general, additional information is needed regarding the accuracy and incremental nature of language performance scales at all points along the scoring continua which they define. Further information is required about the interrelatedness and appropriate combinatory weightings of oral performance categories such as fluency and pronunciation accuracy, and of written performance categories such as content and mechanics. Measures of response validity are needed to enable judgments regarding the particular persons, performance categories, and performance levels for which fair measurement decisions can be made. Performance scales should also be evaluated in terms of the consistent match of the difficulty of performance they are purported to gauge to the ability range of the examinees with whom the scales are used. Reliability of scales should be estimated--not merely as a global interrater correlation coefficient--but also as an index of the sensitivity of the scale to differentiate strength of person performance at all levels of the performance continuum. These needs in the evaluation of performance scales have suggested the areas of investigation for the present study of a composition rating scale currently in use to evaluate writing at all levels of an English-as-second-language (ESL) instructional program at the University of California,Los Angeles (see Appendix A). Specifically, the following questions are being investigated with regard to the performance scale considered:

1. What is the comparative difficulty of the subscale categories (Content, Organization, Expression, Structure, and Mechanics)?

2. How are subscale categories interrelated and, thus, what combinatory weightings can be suggested to produce a total score that is most predictive of performance variance?

3. How are subscale categories and performance levels incrementally arranged on a probabilistic writing performance continuum (with regard to equality or non-equality of intervals)?

4. What estimates of scalar reliability can be derived, both as a global measure of person separability, and with regard to the error associated with each estimate of difficulty and ability?

5. How do persons, categories, and performance levels fit the probabilistic predictions of the Rasch Model, thereby exhibiting a kind of response validity?

6. What are the characteristics of the writing of misfitting persons, i.e., persons for whom ratings should not be interpreted with the same confidence in decision making?

These questions along with several related questions are addressed in the present study of a given composition grading scale as it was employed in the rating of compositions produced by students in the advanced levels of the ESL Service Courses at UCLA.

Method

Subjects

143 Subjects applied to enter the UCLA ESL writing courses at the beginning of Fall Quarter, 1985. Any student seeking entry was required to take a composition placement test. The results of the test (in addition to other enrollment criteria) determined eligibility for one of three levels of ESL writing courses. Generally, these 143 students were at a high level of ESL proficiency (above 550 on the TOEFL). The three levels of writing courses span the non-native speaker (NNS)/native speaker (NS) transition at the freshman level of college writing. That is, the lowest course, English 35, was intended for NNSs who do not quite exhibit the command of written English necessary for coping with a first-year writing load. The next course, English 36, is equivalent to first year college writing for native speakers. The final course, English 106J, is intended for students who still require refinement of written skills beyond English 36 or who for various reasons wish to study more advanced ESL composition. Of the 143 students who took the test, and on whom this paper reports, 10 were placed in English 35, 87 in English 36 and 17 in English 106J. The remainder were placed on a waiting list for subsequent terms.

Instrumentation

All subjects wrote for 75 minutes on the first day of class. Administration, including instructions, completion of an information sheet, a brief error-detection task, and the 75-minute composition test itself were all designed to require exactly two hours--the normal writing class period. The students were advised to check with the ESL office prior to the second class meeting so that they could determine which course level they were assigned. The composition topic was selected and approved by instructors of the ESL writing courses prior to the exam.

Rating Procedure

Following the administration on the first class day, the ESL writing instructors met with several other experienced ESL writing teachers to score the compositions. In all, nine raters participated. Raters were paid for the scoring task, over and above their normal instructor salaries. Their first task was to participate in a norming session to ensure uniform interpretation and application of the composition rating scale. The actual scale employed along with a set of related performance descriptors is provided in the appendices. (See Appendix A.)

Five papers were chosen at random, xeroxed, and scored by all raters. Then the raters met, shared their grades, and agreed on uniform standards and interpretation of the scoring terminology. It was decided that for placement purposes the five norming papers would be assigned the average score of all the ratings derived during the norming session, but that they would not be included in the study because they represented "pre-norming" ratings.

Following the norming, the 138 remaining papers were scored using the composition rating scale. The rating of these compositions employed the following established procedure. First, each rater took approximately 10-15 papers and read them independently. On finishing that set, the rater returned the set to a "second-rater-needed" box. Each rater then read another set from the unrated pile or from the second rater box. Each rater marked each completed composition with a code number to guarantee that no rater read the same paper twice. Furthermore, the score sheets were kept folded to ensure that no rater saw the score assigned by any previous rater.

If the first and second raters were four or more total points apart on a given composition (on the combined scale of 25), then a supervisor placed that essay into a "third-rater-needed" box. Once that box contained compositions, any returning rater had the option to choose from the unscored, the once-scored, or the twice-scored composition boxes. The four-point third rater cutoff was determined by conference prior to the beginning of the term and was based on prior experience and results with this scale. Thus, by the end of the rating session all 143 papers had been read at least twice, 25 (about 17%) of the papers required a third reader, and the five norming papers had been read by all raters.

Input Dataset

For each subject the following data were keyed into the scores, and the subscores for the first, second, and--if necessary--third rater. The raw rater data included decimal values to one place because raters were permitted to assign " + .5" scores at any level. Next, the rater scores were averaged for each case (whether involving two raters or more). Later in the analysis the averaged data were rounded for the Rasch analysis portion of the study to facilitate use of the Microscale procedure.

Table 1 below presents descriptive statistics for the unrounded, averaged data less the five compositions used for norming purposes and less five subject records misplaced between the score-reporting and data entry stages. Thus the final data set analyzed and reported contained 133 cases.

Table 1

Descriptive Statistics: Unrounded Rater Averages
Fall 1985 ESL Composition Placement Test, UCLA.

Subskill	Mean	S.D.	N
Content	3.46	0.78	133
Organization	3.53	0.69	133
Expression	3.66	0.54	133
Structure	3.60	0.59	133
Mechanics	4.04	0.48	133

Table 2 below presents descriptive statistics for the rounded, averaged data.

Table 2

Descriptive Statistics: Rounded R ater Averages
Fall 1985 ESL Composition Placement Test, UCLA.

Subskill	Mean	S.D.	N
Content	3.57	0.83	133
Organization	3.70	0.74	133
Expression	3.79	0.62	133
Structure	3.75	0.63	133
Mechanics	4.10	0.55	133

As can be seen, rounding the data to achieve a five-point (non-decimal), averaged scale (suitable for the Rasch analysis reported below) had a minimal effect on the magnitudes of the means and standard deviations of composition ratings. Although the means and standard deviations increased slightly, the rank-ordering of the means remained unchanged.

Analysis Procedures

The analysis proceeded in two stages. First, correlation and multiple regression analyses were run on the unrounded data set to determine the contribution of each subscore to the total score for the purpose of suggesting appropriate subscale weights. Second, the rounded data were Rasch analyzed to address problems of person ability and category/continuum difficulty scaling and fit to the model.

The multiple regression analysis was done on SPSS Version H, using tne "New Regression" procedure (Hull and Nie, 1981: 94-121). The solution employed the stepwise procedure. This routine was run on an IBM 3090 computer under the MVS operating system. Rasch analysis was done using Microscale Version 1.0 (Wright and Linacre, 1984). The Microscale analysis was done on an IBM PC (640K, two drives). This analysis generated logit scale values, fit statistic and standard error values (as explained in Wright and Stone, 1979) for all 133 persons, the five subscale categories, and the upper three of the five scale values. Because this data set was taken from the upper end of the ESL ability continuum in the UCLA ESL program, there were no low scale values of 1 and very few scale values of 2. Since Microscale does not analyze the bottom-most value of a scale (a feature somewhat analogous to the loss of a degree of freedom in inferential statistics), only the Rasch results for scale values 3, 4, and 5 are reported in this paper for each subscale.

Category

Mention needs to be made of the Rasch modeling assumption of unidimensionality. Unless this assumption is met with regard to the subscale categories, it would not be possible to position these categories along a single scoring continuum. The assumption would require that the hypothetical construct of composition writing performance be describable in a one-dimensional latent space incorporating the constructs of Content, Organization, Expression, Structure and Mechanics as defined by the scale. The usual method of testing this assumption involves use of principal components analysis to decompose the matrix of subscale category correlations into principal components. The finding of a primary, dominant factor accounting for at least 20 per cent of the matrix variance is taken as support for the presence of unidimensionality for purposes of application of item response theory (Reckase, 1979; Hattie, 1985).

In the present case this method is not applicable, since it is not considered a productive procedure to attempt to decompose a matrix of as few as five variables into other than a unidimensional solution (Cattell and Vogelman, 1977). This leaves alternative possible methods such as eyeballing of the correlation matrix (armchair factoring), consideration of estimates of internal consistency reliability, and examination of Item Response Theory (IRT) fit statistics. These procedures were followed in the present study. Although some correlational clustering was present, a high estimate of internal consistency (0.92) was obtained, no noticeable violation of category fit constraints appeared, and, thus, there was no convincing evidence that the rating data violated unidimensionality assumptions. Similarly, assumptions of non-speededness and local independence were not violated for the present data. A more rigorous test of unidimensionality might have been applied had there been additional subscale rating categories, say, 10 to 15 categories. Since all of the categories are language-related, and since they all involve the skill of writing, and because tests involving other, more diverse language skills have produced unidimensional solutions for students from this same population (Henning, Hudson and Turner, 1985) it is reasonable to anticipate a unidimensional solution.

Results

Observed Correlations

Correlational results are reported in Table 3 below. Note that there is a slight clustering tendency for subscales of Content and Organization and also for Structure and Expression. Note also that Content, Organization and Expression clearly bear the highest relation to total score, and that relation is similar in magnitude for each of these subscale categories.

Table 3

Subscore Correlation Matrix.

	C	O	E	S	M	T
Content	1.00					
Organization	.71	1.00				
Expression	.51	.39	1.00			
Structure	.38	.32	.73	1.00		
Mechanics	.22	.38	.36	.39	1.00	
Total	.81	.79	.79	.73	.58	1.00

Regression Analysis

The regression equation resolved on the final step had the standardized beta coefficients reported in Table 4 below.

Table 4

Standardized Regression Beta Coefficients and Suggested Weighting Factors.

Subscale Category	Standardized Beta	Suggested Weighting
Content	.337	1.261
Organization	.296	1.111
Expression	.235	.882
Structure	256	.961
Mechanics	.208	.781
Total	1.332	4.996

The t-values associated with each of the standardized betas reported above was significant ($p < .05$). The suggested weights would be applied to observed scores in each subscale category to achieve maximal prediction of total score variance. Note that the rating total across all subscales was used as the best available dependent global estimate of writing ability. We are aware of possible deficiencies in this measure, and thus input of an overlap-corrected correlation matrix was also used in the regression procedure. Standardized betas are reported rather than raw score betas since the goal was not merely to predict or reproduce the total score, but to ascertain relative proportional contributions of subscales for derivation of weights.

Rasch Analysis

Microscale is unique in that, in addition to the typical person and item logit measures (cf. Rasch, 1960; Wright and Stone, 1979; Wright and Masters, 1982) it permits logit measurement for each of the five response scale values employed in the composition scale. Hence, the Wright and Masters (1982) refinement of the Rasch family of formulas can be applied to all three elements: persons, items, and response scale values. Potentially, analysis of response data can not only position these three elements along an ability/difficulty continuum, but can also identify persons, items and response categories for which the probabilistic predictions of the model are not met. When predictions of the model are not met, elements are said to misfit the model or exhibit a kind of response invalidity. Subsequent tables will report logit difficulty and fit statistics for subscores and rating scale values.

Table 5

Subscore Calibrations and Fit Statistics
UCLA ESL Composition Placement, Fall 1985.

Subscore	Logit Diff.	Error	Infit	Outfit
Expression	−.02	.14	−2.09	−1.19
Content	.57	.13	1.13	1.65
Organization	.23	.13	.22	.99
Structure	.08	.14	−.58	−.09
Mechanics	−.87	.14	−.74	.02

Table 6

Rating Scale Value Calibrations and Fit Statistics
UCLA ESL Composition Placement, Fall 1985.

Scale Value	Logit Diff.	Error	Infit	Outfit
3	−2.01	.17	15.00	8.39
4	−.85	.09	10.12	15.00
5	2.86	.11	1.74	−.62

Reliability Estimation

An added feature of the Microscale analysis is that reliability is calculated both as a global person separability index (0.92 for the overall composition grading scale) and as standard errors of measurement at each point along the scoring continuum. Note from Tables 5 and 6 that measurement error ranged from a low of .09 to a high of .17 logits. Both the global estimate of reliability and the individual standard error magnitudes were indicative of a hiᴐh degree of measurement accuracy for this rating scale.

Person Misfit

Wright and Linacre (1984) discuss the difference between infit and outfit in the Microscale manual:

> We try to correct mismatches between items and people because they degrade the functioning of our measuring system. When item difficulty matches person ability (e.g., both are at 1.0 logits) we obtain the maximum amount of information about the variable and the person. The greater the difference between item difficulty and person ability, the less information the item provides. In order to minimize the influence of contacts between persons and items that are remote from one another, Microscale calculates an information weighted mean square residual, infit. The deviations used in calculating the outfit statistic are here weighted by the amount of information provided by each response . . . [cf. Wright and Masters, 1982: 99-100].
>
> In contrast to outfit, the infit statistic is not sensitive to outliers but focuses instead on the fit situation in the region where responses are delivering the most information [p. 4-18].

Careful scrutiny of the person fit statistics revealed five out of 133 persons exhibiting positive misfit, whether infit, outfit, or both. It was originally planned that these individuals be interviewed to determine patterns of behavior predictive of misfit. On subsequent consideration it became evident that person misfit could also derive as much from rater irregularities as from examinee irregularities. Therefore, all misfitting compositions were rerated independently of previous ratings. Rerating supported the consistency of previous ratings, and, therefore, attention was directed to the nature of the writing in the composition samples. The following observations were made:

1. The misfitting subjects represented a variety of language backgrounds (Spanish, Chinese and Iranian), so that the scale could not be said to be invalid for persons from any particular language background.

2. Misfitting subjects represented both sexes, so that there is no evidence of gender bias.

3. Each of the misfitting compositions was rated low on content with respect to the other subscale categories. Extreme problems with cohesion were evident in most cases. The misfitting performance resembled disordered language patterns as reported by Andreasen (1979). One of the students had been recommended for clinical counseling.

For such students, who may exhibit disordered thinking patterns, it may be the case that use of the composition grading scale would lead to invalid conclusions about student writing ability. No further patterns have as yet emerged from the analysis of misfit data.

Discussion of Subscores and Rating Scale Values

It can be seen from Tables 4, 5 and 6 that the primary operationalization of the composition variable is in the rating scale values. In other words, the rating values and not the subscale categories serve to bracket student ability. Indeed, without the facility of Microscale to analyze scale values, this might be overlooked. Table 4 indicates that the subscale scores differentially contribute to the total score. Table 5 shows that the range of logits for the subscale categories is quite narrow. Those logit values do not "bracket" the subjects. Rather, as seen in Table 6, the subjects are more nearly bracketed by the rating scale logits. This is an intuitively predictable finding, in keeping with previous work (Davidson and Henning, 1985). Generally, when one derives a scale such as this, there is a definite operationalization load on the rating scale value increments beyond that of the subscale categories.

This finding has implications for the use ot a scale such as this particular composition rating instrument. It suggests that a great deal of training and norming effort should go into the clear understanding and interpretation of the meaning of each scale value increment; for example, what is the difference between a 4 and a 5 in Content?

This suggestion is upheld when one notes the much larger fit values in Table 6 than in Table 5. The rating scale increments seem to misfit at the mid-range values of the scale; i.e., 3 and 4. Wright and Linacre (1984) make the following observations about the fit statistics output by Microscale 1.0:

> No absolute rules are possible for the evaluation of fit. The judgments must be adjusted to the situation and to your aims. Generally we assume that items with outfits falling outside a band of 2.0 units on the outfit scale should be examined more closely [p. 4-13].
>
> As a general principle, whenever one of the fit statistics (outfit or infit) shows item fit values less than -2.0 we should look for factors that tie the items together more tightly than expected. Usually these factors will be influences on the measuring system which we want controlled. Sometimes the causes are a source of test bias, and should definitely be removed from the test [p. 4-17].
>
> Most psychometric procedures limit their attention to items which fit poorly because they contain too much noise (e.g., low item-by-test biserials would also identify [positive INFIT or OUTFIT items] as deviant). But items that fit too well because they contain too little noise, and hence signal an unexpected interdependence, also indicate unexpected irregularities in the measuring process. Overfit can often be explained by the presence of a second, unnoticed and usually unintended variable, positively correlated with the first, that, operating additively, creates a better fit than would occur if the items were more 'pure' in the measure of the intended main variable [pp. 4-15 to 4-17].

In the present analysis, the fit values for the subscale categories are all tolerable. The only marginally questionable fit statistic is the infit for the Expression subscore. Since it is negative, and bearing in mind the above comments, it may safely be assumed that the subscale categories fit model expectations well enough. However, the picture is somewhat more troubled with regard to the fit statistics for the scale values. They are intolerably high. This finding seems to suggest that the scale does not fit model predictions well along its scale value dimension, which further emphasizes the need for careful training and norming of raters concerning scale value differences.

One further comment can be made about the rating scale values. The sample under analysis was truncated in terms of the intended operation of this scale. The scale has been phased into the entire range of ESL courses at UCLA. At present, the only data that are available are from the upper courses, which are reported here. However, it is interesting to speculate on how the scale might operate at the lower end of the course spectrum, and indeed, such a study (a similar composition task across all levels) is in the planning stages.

Microscale 1.0 provides several interesting graphic displays through the use of its SuperCalc-3 spreadsheet capabilities (Sorcim, 1983). Figure 1 (see Appendix B) gives the relation of rating response probabilities to logit difficulty values. It is clear that the upper end of the scale, especially scale value 4, accommodates the majority of the response probability in this sample, as one might expect from a knowledge of the high general language proficiency represented.

What is more interesting is the comparatively low span of the latent difficulty continuum for which scale value 3 would be the most probable choice of the rater. This is shown in Figure 1 by the comparatively small unique, unshared probability area for scale value 3. This finding complements the finding of general misfit for scale value 3, and suggests that raters were not sufficiently uniform in their interpretation and application of the composition rating scale at that point. There may be a tendency for raters to identify extremes of performance with comparative ease and accuracy, but the midpoints of the scale encourage subjective judgments that are less reliable or valid.

Conclusion

In the introductory section of this study it was proposed that investigative results be reported for the UCLA ESL composition grading scale with regard to comparative difficulties of subscale categories, appropriate subscale combinatory weights, the nature of subscale-difficulty and performance-level intervals, global and local reliability estimates, person/category/level fit validity, and the nature of misfitting individual writing performance. This has been done in the present study. Tables 5 and 6 address the issue of comparative subscale difficulty (research question 1 above). Table 4 suggests subscale weightings (question 2). Figure 1 and its discussion address the probabilistic writing continuum in this dataset (question 3). Reported reliability values discussed above address scalar consistency (question 4). Tables 5 and 6, as well as the above discussions of fit address any misfit in subscale categories, scale values, or persons (question 5). Finally, the discussion of person misfit above identifies some interesting characteristics of student essays that should not be interpreted with the same decision confidence as the rest of the sample (question 6).

In general, results indicated that, while high measurement accuracy was observed, as reflected in global reliability estimation and local measurement error indices, misfit was present at the mid points of the performance scales. While it was readily possible for raters to identify highest performers, raters seemed to have difficulty distinguishing between performance ratings of three and four. This finding may suggest the need for a larger study to incorporate more learners at the lowest end of the ability continuum.

Misfitting subjects were not disproportionately representative of any one sex or language group background. Misfitting subjects did, however, exhibit interesting writing characteristics involving low content ratings relative to other subscale categories and demonstrating abnormally low cohesion, as would be representative of disordered language and thought patterns. This observation also merits follow-up and investigation with a larger sample of participants.

References

Andreasen, N. C. 1979: Thought, language, and communication disorders. Arch. gen. psychiatry, Vol. 36, Pp. 1315-1321.

Cattell, R. E. and Vogelman, S. 1977: A comprehensive trial of the scree and KG criteria for determining the number of factors. The Journal of Multivariate Behavioural Research. 12: 289-325.

Davidson, F. and Henning, G. 1985: A self-rating scale of English difficulty: Rasch scalar analysis of items and rating categories. Language Testing. Vol. 2, No. 2, Pp. 164-169.

Hattie, J. 1985: Methodology review: Assessing unidimensionality of tests and items. Applied Psychological Measurement, Vol. 9, No. 2, Pp. 139-164.

Henning, G. 1983: Oral proficiency testing: Comparative validities of interview, imitation, and completion methods. Language Learning. Vol. 33, No. 3, Pp. 315-332.

Henning, G. 1984: Advantages of latent trait measurement in language testing. Language Testing. Vol. 1, No. 2., Pp. 123-133.

Henning, G ., Hudson, T. and Turner, J. 1985: Item response theory and the assumption of unidimensionality for language tests. Language Testing. Vol. 2, No. 2, Pp. 141-154.

Hull, C. Hadlai and Norman H. Nie. 1981. SPSS update 7-9. New York: McGraw Hill.

Rasch, G. 1960: Probabilistic models for some intelligence and attainment tests. 1980 expanded edition. Chicago: The University of Chicago Press.

Reckase, M. D. 1979: Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics. Vol. 4, No. 3, Pp. 207-230.

Sorcim Corporation. 1983: SuperCalc-3 user's guide and reference manual: Documentation 1.0. Third edition. San Jose, Ca.: Sorcim Corporation.

Wright, B. and Linacre, J. M. 1984: Microscale manual for microscale version 1.2. Westport, Ct.: Mediax Interactive Technologies.

Wright, B. D. and Masters, G. N. 1982: Rating scale analysis. Chicago: MESA Press.

Wright, B. D. and Stone, M. H. 1979: Best test design. Chicago: MESA Press.

Appendix A

COMPOSITION GRADING SCALE

LANGUAGE

WRITING

FUNCTION

Expression

____

5 PRECISE AND VARIED VOCABULARY/IDIOM 'USAGE CONSISTENT AND APPROPRIATE VARIETY OF SENTENCE STRUCTURE CONSISTENT AND APPROPRIATE REGISTER CONCISE

4 GOOD AND VARIED VOCABULARY/IDIOM USAGE, A FEW MISUSED WORDS/EXPRESSIONS. APPROPRIATE VARIETY OF SENTENCE STRUCTURE FLUENT EXPRESSION BUT INCONSISTENT OR INAPPROPRIATE REGISTER. FAIRLY CONCISE

3 USES BASIC VOCABULARY WELL. IDEAS NOT ALWAYS PHRASED IDIOMATICALLY SOME VOCABULARY IDIOMS MISUSED, ATTEMPT AT SENTENCE STRUCTURE VARIETY LACKS AWARENESS OF REGISTER. MAY BE TOO WORDY

2 VOCABULARY TOO BASIC TO EXPRESS IDEAS EFFECTIVELY VERY LITTLE SENTENCE STRUCTURE VARIETY, LITTLE AWARENESS OF REGISTER

1 INAPPROPRIATE USE OF VOCABULARY. NO SENTENCE VARIETY

Organization

____

5 APPROPRIATE TITLE. EFFECTIVE INTRODUCTORY PARAGRAPH. THESIS MADE CLEAR WITH CLEAR CONTROLLING IDEA EACH PARAGRAPH WELL FOCUSED AND WELL DEVELOPED. NECESSARY TRANSITIONS BETWEEN AND WITHIN PARAGRAPHS CONCLUSION APPROPRIATELY TIES POINTS OF THE COMPOSITION TOGETHER

4 GOOD TITLE. INTRODUCTION AND CONCLUSION CEARLY ALL PARAGRAPHS HAVE A WELL-FOCUSED TOPIC SENTENCE FULL SUPPORT AND LOGICAL DEVELOPMENT. TRANSITIONS BETWEEN AND WITHIN PARAGRAPHS MAY BE MISSING OR INAPPROPRIATE

3 SOME PORTION OF THE COMPOSITION INCORRECTLY DEVELOPED. SOME ATTEMPT MADE TC STATE PURPOSE OF ESSAY BUT CONTROLLING IDEA(S) UNCLEAR PROBLEMS WITH ORDERING OF IDEAS IN BODY TRANSITIONS BETWEEN AND WITHIN PARAGRAPHS MISSING OR INAPPROPRIATE. SOME EVIDENCE OF PLAN

2 MINIMAL OR INAPPROPRIATE INTRODUCTION OR CONCLUSION. PURPOSE OF ESSAY/CONTROLLING IDEA MISSING POORLY STATED OR UNCLEAR. IDEAS IN BODY ORDFRED INAPPROPRIATELY LITTLE EVIDENCE OF PLAN

1 LACK O INTRODUCTION AND CONCLUSION. NO APPARENT PLAN

Content

____

5 GENERALIZATIONS ARE FULLY SUPPORTED AND MAKE SENSE EXCEPTIONAL ANALYSIS OR INSIGHT SHOWN

4 GENERALIZATIONS ARE WELL SUPPORTED THROUGHOUT

3 NOT ENOUGH EXAMPLES OR SUPPORTING EVIDENCF FOR MAIN POINTS 'N THE 3ODY NEEDLESS REPETITION MAJORITY OF MATERI/' ON TOPIC

2 LITTLE OR INAPPROPRIATE DEVELOPMENT OF 1 SIS ENOUGH MATERIAL OFF TOPIC TO DISTRACT READER PURPOSE OF ESSAY NOT CLLARLY IDENTIFIED

1 IDEAS NOT DEVELOPED INSUFFICIENT SUPPORTING DETAILS IDEAS TOO FEW

FORM

Structure

____

5 NATIVE LIKE USE OF VERB-MODAL FORMS TENSE SEQUENCING. SUBJECT VERB AGREEMENT ARTICLES PREPOSITIONS RELATIVE/ADVERBIAL CLAUSES, PHRASES. NO SENTENCE FRAGMENTS OR RUN ONS

4 SOME GRAMMAR PROBLEMS WHICH DO NOT ADVERSELY AFFECT COMMUNICATION e g VERB-MODAL FORMS ARTICLES PREPOSITIONS PRONOUNS RELATIVE ADVERBIAL CLAUSES PHRASES ONE OR TWO SENTENCE FRAGMENTS OR RUN ONS

3 IDEAS UNDERSTANDABLE DESPITE NUMEROUS VARIED GRAMMAR PROBLEMS SUCH AS INCORRECT VERB TENSE WORD FORMS, ARTICLE USAGE WHICH HAVE A NEGATIVE EFFECT ON THE READER. MANY SENTENCE FRAGMENTS OR RUN ONS

2 SOME EVIDENCE OF KNOWLEDGE OF GRAMMAR BUT SERIOUS GRAMMAR POROBLEMS INTERFERE WITH COMMUN CATION OF WRITER S IDEAS

1 GRAMMAR PROBL MS OBSCURE THE MESSAGE MINIMAL CONTROL OF SENTENCE STRUCTURE

Mechanics

____

5 NEAT AND LEGIBLE WITH CORRECT USE OF MARGINS INDENTATION SPELLING, PUNCTUATION, AND CAPITALIZATION

4 NEAT AND LEGIBLE WITH GENERALLY CORRECT USE OF MARGINS/INDENTATION SPELLING PUNCTUATION AND CAPITALIZATION

3 VARIOUS PROBLEMS W'TH LEGIBIL.TY MARGINS INDENTATION, SPELLING, PUNCTUATION AND CAPITALIZATION

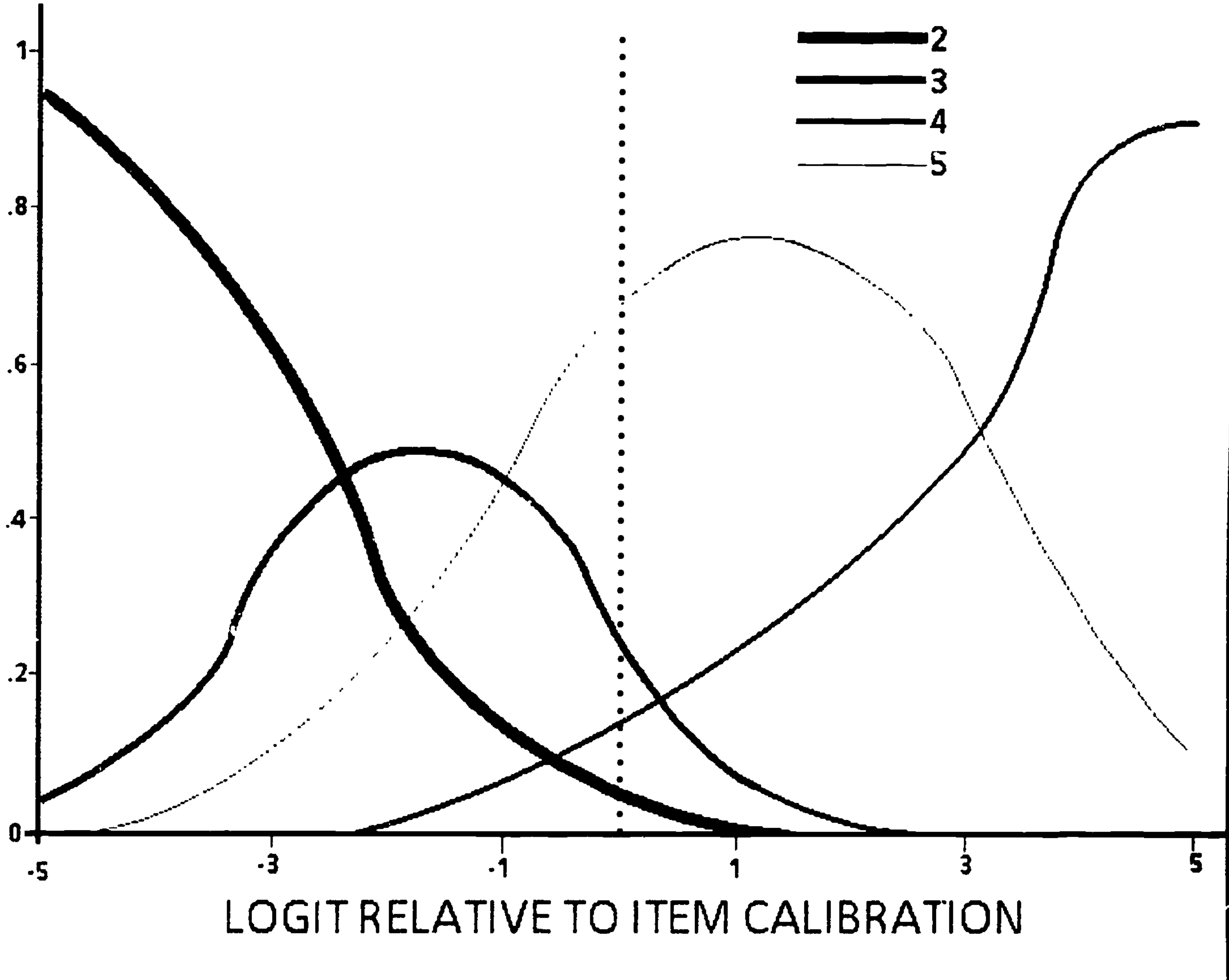
2 MAY BE DIFFICULT TO READ FREQUENT PROBLEMS WITH MARGINS INDENTATION SPELLING PUNCTUATION AND CAPITALIZATION

1 MAY BE ILLEGIBLE SERIOUS PROBLEMS W'TH MARGINS INDENTATION. SPELLING. PUNCTUATION AND CAPITALIZATON

FIGURE 1

COMPOSITION RATING SCALE - F85 PLACEMENT
PROBABILITY OF RESPONSE BY CATEGORY

Appendix B


LOGIT RELATIVE TO ITEM CALIBRATION