DOCUMENT RESUME

ED 388 717 TM 024 174

AUTHOR Henning, Grant
TITLE Scalar Analysis of the Test of Written English. TOEFL Research Reports. Report 38.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-92-30
PUB DATE Aug 92
NOTE 35p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *English (Second Language); Equated Scores; *Essays; *Interrater Reliability; Psychometrics; *Rating Scales; *Scaling; *Scoring
IDENTIFIERS Rasch Model; *Scale Analysis; *Test of Written English; Writing Prompts

ABSTRACT

The psychometric characteristics of the Test of Written English (TWE) rating scale were explored. Rasch model scalar analysis methodology was employed with more than 4,000 scored essays across 2 elicitation prompts to gather information about the rating scale and rating process. Results suggested that the intervals between TWE scale steps were surprisingly uniform and that the size of the intervals was appropriately larger than the error associated with assignment of individual ratings. The proportion of positively misfitting essays was small (approximately 1% of all essays analyzed) and was approximately equal to the proportion of essays that required adjudication by a third reader. This latter finding, along with the low proportion of misfitting readers detected, provided preliminary evidence of the feasibility of employing Rasch rating scale analysis methodology for the equating of TWE essays prepared across prompts. Some information on characteristics of misfitting readers was presented that could be useful in the reader training process. Appendixes present the TWE Scoring Guide and the mathematical specification of the rating model. (Contains 9 tables and 26 references.) (Author/SLD)

TOEFL

TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

REPORT 38
August 1992

Scalar Analysis of the Test of Written English

Grant Henning

ETS
EDUCATIONAL TESTING SERVICE

Scalar Analysis of the Test of Written English

Grant Henning

Educational Testing Service
Princeton, New Jersey

RR-92-30

Educational Testing Service is an Equal Opportunity/Affirmative Action Employer.

Abstract

The present research was conducted to explore the psychometric characteristics of the Test of Written English (TWE) rating scale. Rasch model scalar analysis methodology was employed with more than 4,000 scored essays across two elicitation prompts to gather the following information about the TWE rating scale and rating process:

1. the position and size of the interval on the overall latent trait that could be attributed to behavioral descriptors accompanying each possible integer scoring step on the TWE scale

2. the standard error of estimate associated with each possible transformed integer rating

3. the fit of rating scale steps and individual rated essays to a unidimensional model of writing ability and, concurrently, the adequacy of such a model, including the proportion of misfitting essays as a portion of all essays analyzed

4. the fit of individual readers to a unidimensional model of writing ability and to the expectations of a chi-square contingency test of independence of readers and ratings assigned, along with information on some characteristics of misfitting readers

5. comparative scalar information for two distinct TWE elicitation prompts, including nonparametric tests of the independence of readers and scale steps assigned and the feasibility of equating of scales.

Results suggested that the intervals between TWE scale steps were surprisingly uniform and that the size of the intervals was appropriately larger than the error associated with assignment of individual ratings. The proportion of positively misfitting essays was small (approximately 1% of all essays analyzed) and was approximately equal to the proportion of essays that required adjudication by a third reader. This latter finding, along with the low proportion of misfitting readers detected, provided preliminary evidence of the feasibility of employing Rasch rating scale analysis methodology for the equating of TWE essays prepared across prompts. Some information on characteristics of misfitting readers was presented that could be useful in the reader training process.

The Test of English as a Foreign Language (TOEFL®) was developed in 1963 by a National Council on the Testing of English as a Foreign Language, which was formed through the cooperative effort of more than thirty organizations, public and private, that were concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board assumed joint responsibility for the program, and in 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations (GRE) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a Policy Council that was established by, and is affiliated with, the sponsoring organizations. Members of the Policy Council represent the College Board and the GRE Board and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

A continuing program of research related to the TOEFL test is carried out under the direction of the TOEFL Research Committee. Its six members include representatives of the Policy Council, the TOEFL Committee of Examiners, and distinguished English as a second language specialists from the academic community. Currently the Committee meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Research Committee serve three-year terms at the invitation of the Policy Council; the chair of the committee serves on the Policy Council.

Because the studies are specific to the test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. However, many projects require the cooperation of other institutions, particularly those with programs in the teaching of English as a foreign or second language. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. All TOEFL research projects must undergo appropriate ETS review to ascertain that the confidentiality of data will be protected.

Current (1991-92) members of the TOEFL Research Committee are:

James Dean Brown	University of Hawaii
Patricia Dunkel (Chair)	Pennsylvania State University
William Grabe	Northern Arizona University
Kyle Perkins	Southern Illinois University at Carbondale
Elizabeth C. Traugott	Stanford University
John Upshur	Concordia University

Table of Contents

List of Tables

Background and Purpose of the Study

The current six-point rating scale in use for scoring the Test of Written English (TWE), reproduced in Appendix A, was chosen with great care on the basis of expert recommendation and common practice in the field (Educational Testing Service, 1989). Nevertheless, it was thought that more information would be useful about the operational properties of the TWE® test scale. For example, it had not been fully determined whether the steps on the scale defined equal intervals or not, or, if not, what the actual intervals might be. Also, it was not known whether the probability of assignment to each step on the rating scale corresponded uniformly and appropriately to the distribution of writing ability said to be measured at that level. It was considered useful to gather more information about how accurate or valid ratings are at the various points on the scoring continuum. It had not yet been fully determined whether the range of ability extending between any two adjacent points on the scale exceeded the standard error associated with the assigning of those points and thus whether or not a true scale was defined. There was no systematic test of reader fit to performance expectations at the various steps of the rating scale. Although it was known that the reporting of scores on a unitary scale or continuum promotes the desirability of psychometric unidimensionality in the response data matrix (Henning, 1988a, 1989, 1992), it was not known what the expected proportion of misfitting writing samples might be when unidimensional models of analysis were applied to the TWE rating scale. Nor was it known how readily an item response theory approach to the analysis of TWE essays might contribute to the equating of TWE topics and topic prompts.

It is fair to point out that these kinds of information have only infrequently been provided for other scales commonly used to rate language performance in the various skill areas (e.g., Hamp-Lyons & Henning, 1991; Henning & Davidson, 1987; Pollitt & Hutchinson, 1987). However, some of these research needs in the TWE context were foreseen by Stansfield and Ross (1988)--especially those related to hitherto seemingly intractable problems of essay topic and prompt equating. Careful research into these and other questions is possible by means of a family of rating scale analysis procedures commonly referred to in the literature as Poisson Counts models, Binomial Trials models, Rating Scale models, and Partial Credit models, all of which are extensions of Rasch Dichotomous models (Andrich, 1978a-d, 1979; Davidson & Henning, 1985; Engelhard, 1991; Henning & Davidson, 1987; Linacre, 1989; Muraki, 1991; Pollitt & Hutchinson, 1987; Rasch, 1960, 1980; Wright & Masters, 1982). The present study was intended to apply appropriate candidates from among these analysis procedures to TWE ratings to provide information related to the problems mentioned above.

Among the further purposes for this study was identification of points on the scale at which more thorough descriptors might be needed

to guide raters in making correct assessments. Typically in the rating of writing performance according to scales like the TWE scale, raters experience less difficulty in making judgments at the extremes of the scale (e.g., points 1, 2, 5, and 6) and greater difficulty in differentiating performance in the middle (e.g., points 3 and 4) (Henning & Davidson, 1987). If this pattern were found to persist in the case of the TWE scale, it was hoped that areas of scalar refinement could be suggested, including modification of weak descriptors so identified and special training of raters. Also, provided appropriate statistical requirements were satisfied and the application of scalar modeling procedure to the analysis of TWE scores was found to be feasible, it was recognized that use of Rasch model rating scale analysis might also provide an appropriate means of TWE topic equating similar to the item response theory equating already in use for the individual sections of the TOEFL test.

Application of Rasch model rating scale analysis requires statistical unidimensionality and local independence of ratings for score interpretation and equating (Henning, 1988a, 1989). It was hoped that, by testing fit to a unidimensional latent trait model, such a study would provide further insight into the psychometric dimensionality of ratings of ESL/EFL writing performance at various points along the scale and could possibly help in the identification of patterns of performance that contribute to unidimensional and multidimensional solutions. It is important to note here that there is evidence that "psychologically multidimensional" behavior such as writing behavior can often be found to exhibit "psychometrically unidimensional" statistical characteristics that are useful for the purposes of reporting construct-valid scores on a unitary scale (Henning, in press). Finally, it was hoped that such a study could provide a means for comparing the function of at least two different essay prompts considered simultaneously.

Method

Subjects and Instrumentation

Subjects included in the study were drawn from the May 1990 administration of the Test of Written English. In all, scores from 4,116 essays as rated by the 59 most frequently paired readers in that administration were analyzed. These essays were written on two separate essay elicitation prompts that will hereafter be identified as prompt B and prompt C. Accordingly, 2,572 essays were gathered and scores analyzed for prompt B, and 1,544 essays were gathered and scores analyzed for prompt C.

Essay sampling was done systematically so as to maximize the frequency of paired ratings. Thus, separately within each of the two prompt distributions, essays were selected that had been read most frequently by the same reader pairs. This was done purposely to permit certain statistical analyses that required frequent reader pairing. For

prompt B, the 29 most frequently paired readers and the essays they had read were selected and organized within reader pairs (see Table 3). Similarly, for prompt C, the 30 most frequently paired readers and the essays they had read were selected and organized within reader pairs (see Table 6). Further subsampling was done to permit several analyses based on optimal paired reader frequencies. Due to the paucity of disclosed writing prompts from the relatively young TWE testing program, the actual wording of the prompts analyzed is not reported here. Suffice to note that both prompts were of the compare/contrast discourse genre.

Data for the study consisted of actual TWE response data. Thus, no deviations from usual administrative procedures were observed. In no case was the name of any reader or essay writer revealed to the researcher prior to or throughout the conduct of the study, and no violation of privacy or confidentiality occurred.

In addition to the two TWE essay prompts mentioned, further instrumentation was provided in the form of the Rasch model software program MICROSCALE 2.0 (Wright & Linacre, 1985), which, although not the latest generation of such programs, was found suitable to perform the required analyses for the comparatively large samples considered in the study.

Procedure and Analyses

The two data sets to be used in the study were drawn systematically from existing TWE rating data to maximize frequency of rater pairs.

Descriptive statistics were derived using traditional statistical analyses available in the software program SYSTAT, and IRT rating scale modeling was conducted via the software program MICROSCALE 2.0 (Wright & Linacre, 1985). Both rating scale analysis and partial credit modeling procedures were initially employed in the analyses; but eventually, after several iteration outcomes and analysis results were compared, and after more thorough consideration of the philosophy underlying application of the TWE rating scale, preference was given to the Rasch model rating scale analysis procedure for the remainder of the study (Wright & Masters, 1982). Mathematical specification of this model is provided in Appendix B.

For most frequent reader sets, separate chi-square contingency analyses were conducted to test the independence of reader and rating scale categories across the two essay prompts. For each analysis, the six most frequent readers were compared with regard to the frequency of assignment of every possible rating for 1,919 essays prepared on prompt B, and for 967 essays prepared on prompt C. Thus it was possible not only to establish the degree of independence of readers and ratings assigned, but also to examine the comparative fit to frequency expectation on the part of those readers and ratings assigned.

Results

Descriptive Statistics

Table 1 presents descriptive statistics for the data sets corresponding to scores assigned to 2,572 essays prepared on prompt B and to 1,544 essays prepared on prompt C. Note that the mean rating assigned by both primary readers for both essay prompts was almost exactly 4. Note also that 28 of the 2,572 essays on prompt B and 12 of the 1,544 essays on prompt C, or approximately 1% of all essays, required adjudication by a third reader. Adjudication of TWE essays is required when the ratings of the first and second readers differ by more than one point. Note also that adjudication was always in the middle of the scoring range, so that no essay with a rating of 1 or 6 required adjudication, suggesting that, confirming the findings of Henning & Davidson (1987), disparity in score judgment predictably is more likely to occur in the middle of the scoring range.

Because of the infrequency of recourse to a third reader, subsequent analyses are based only on the initial two readers. This means that some of the estimates of score reliability are somewhat conservative since discrepant ratings have not been adjusted. Table 1 reports correlations between first and second raters for prompt B of .818 and for prompt C of .821. When these coefficients are adjusted by means of the Spearman-Brown prophecy formula to reflect the reliability of combined ratings, the improved results correspond exactly to the interrater reliability coefficients reported in Table 2.

Rating Scale Calibrations

Table 2 reports the results of Rasch scalar analyses by scale step for the two essay prompts. Note that following each of the six possible ratings assigned are the count of total first and second ratings assigned at that level, the mean logit difficulty calibration, the standard error in logits associated with the mean logit calibration, the interval between successive logit calibrations, the gap reported for logit calibrations estimated, the alpha reliability, and the interrater reliability for each essay prompt. It is necessary to offer some interpretation of these values.

The rating count signifies that 4 was by far the most frequent rating assigned. The rating of 1 was so infrequent that it was not possible to estimate several of the other associated statistics. For those steps reported, mean logit calibrations ranged broadly from approximately -7 at the easy or incompetent end of the continuum to 7 at the difficult or competent end of the continuum. (Logits are logarithmically transformed raw scores that have the important characteristics of comprising equal-interval, sample-free scalar units with step difficulty and writer ability positioned on the same unitary scale [Wright & Masters, 1982; Wright & Stone, 1979]).

TABLE 1

Classical Descriptive Statistics for Scores Assigned to TWE Essays Based on Two Elicitation Prompts (N = 4,116 Essays; 59 Most Frequent Readers)

Prompt B

	Reader 1	Reader 2	Reader 3
N	2,572	2,572	28
Mean	4.058	4.077	4.143
Sd	.991	.998	.970
Minimum	1	1	2
Maximum	6	6	5
$r_{1,2}$	.818		

Prompt C

	Reader 1	Reader 2	Reader 3
N	1,544	1,544	12
Mean	3.982	3.981	4.583
Sd	.982	.938	.515
Minimum	1	1	4
Maximum	6	6	5
$r_{1,2}$	.821		

TABLE 2

Reliabilities and Rasch Model Rating Scale Calibrations
for Two Elicitation Prompts with Six Score Levels
(N = 4,116 Essays; 59 Most Frequent Readers)

Prompt B (N = 2,572 Essays)

Score	Rating Count	Mean Logit	SE	Logit Interval	Gap	α	Inter-rater
1	40	--	--	--	-3.283	.814	.900
2	235	-6.573	.137	3.008	-6.746		
3	1,076	-3.565	.060	3.392	-4.704		
4	2,143	-0.173	.035	3.577	8.468		
5	1,288	3.404	.034	3.503	5.694		
6	362	6.906	.052	--	.572		
Total	5,144	1.606					

Prompt C (N = 1,544 Essays)

Score	Rating Count	Mean Logit	SE	Logit Interval	Gap	α	Inter-rater
1	8	--	--	--	-.850	.752	.902
2	120	-7.694	.285	3.821	-4.079		
3	788	-3.873	.094	3.961	-5.883		
4	1,345	.088	.044	3.795	6.213		
5	660	3.883	.046	3.714	4.254		
6	167	7.597	.090	--	.346		
Total	3,088	1.547					

Note that the logit interval is approximately the same between all steps estimated. This suggests that the rating categories 1 through 6 (or at least 2 through 6 for which sufficient data were available) do tend to represent equal steps on the ability and difficulty continuum. This is important as a reflection that no one step is too inclusive of behaviors that would necessarily require further subdivision into still smaller steps. Also, notice in Table 2 that the standard error associated with mean logits was very small with respect to the interval defined between logits. This is an indication that a true scale has been defined by the score steps. However, the fact that the first rating category on the scale is used so infrequently makes it difficult to generalize about the properties associated with that step. Presumably, larger analysis samples would contain sufficient numbers of ratings at that level to permit generalizations.

The gap value reported is the difference between observation and expectation for estimated score output of the Microscale program. This should be viewed comparatively, since the magnitude of these scores can be adjusted manually as a means of determining the number of iterations required for run convergence. The alpha reliability reported is the ratio of observed score variance minus error of estimation to the observed score variance. This kind of reliability often tends to be more conservative than the interrater reliability that is also reported. In this case reliability estimates are especially conservative because discrepant ratings used in the analysis were not altered to correspond to the recommendation of the adjudication process.

Score Equating

Because of the properties of Rasch model logit scores, when statistical requirements are met it is readily possible to link or equate logit scores from one set of ratings to another set on a different topic or prompt, given some information known to be constant across administrations. For example, reader calibrations, or logit scores of repeating writers, or mean logit scores for steps can be used as translation constants or anchors to equate score sets from future administrations. The difference between the total mean logit calibration for prompts B and C in Table 2 (i.e., between 1.606 logits and 1.547 logits) could serve as a translation constant to equate the scores assigned to prompt B and prompt C. In this case, the equating relies neither on common writer nor on common reader but, rather, on common behaviorally defined steps employed across prompts. This difference between mean logit step difficulty estimates for prompts B and C is small (i.e., 0.059 logits) and is only slightly larger than the estimated standard error of equating prompt C essays to prompt B essays (i.e., about 0.034 logits; Wright & Stone, 1979). In cases where estimated mean differences are less than the estimated standard error of equating, no adjustment would be considered necessary. In the present example, equating of prompt C essays to prompt B essays would be accomplished by augmenting prompt C logit scores by the translation constant of 0.034.

Prompt B Reader Tabulations and Calibrations

Table 3 reports the reader identification numbers and numbers of essays read for the 29 most frequently paired readers of this particular essay reading session. The six most productive readers from among this group are further identified by letters A through F for subsequent analyses to be reported later.

Because the earlier analyses conducted did not attempt to maintain the same person as reader 1 or 2 throughout the data set, Table 4 reports findings when readers 1 and 2 were held constant over paired rating subsets of essays. For these analyses, data from the six most frequent pairings of readers were analyzed separately. Use of only the six most frequent reader pairs was dictated by a recognition that use of more than six reader pairs would result in essay rating subsets with too few essays for meaningful analysis. Note that distributions of raw ratings assigned are reported in Table 4 for each data set constructed. Note also that the comparative leniency or strictness of readers in each pairing is reflected in the logit scores reported below.

Prompt B Reader Fit to the Model

The infit and outfit estimates reflect the extent to which readers were found to fit the expectations of the Rasch scalar analysis, given the patterns of scores assigned in each data set. Such an analysis could be used to identify misfitting readers who might be provided additional orientation to the reading process or be asked not to participate in subsequent reading sessions. A fit value of positive 2.0 is frequently and conventionally used as a criterion for establishing misfit for items and persons (Wright & Stone, 1979). High negative fit values are also a concern, as they tend to reflect overfit to the expectations of the model. Infit represents an attempt to examine fit in the narrower region where most information is being supplied by the assigned score, and for this reason and because infit tends to be more sensitive to violations of unidimensionality, it is often more useful than outfit as a fit statistic (Henning, 1988a).

In practical terms, infit and outfit estimates help us identify readers who are not using the rating scale in the manner in which it was intended to be used. The estimates are estimates of the consistency with which each judge uses the rating scale across essays. The higher the infit or outfit value, the more inconsistent the reader is with regard to expectations of the model. In the present example, none of the readers exceeded a positive 2.0 infit value, so this outcome, along with the small size of the reader pairing data sets, would suggest that there is not sufficient evidence in Table 4 that any of these readers was necessarily performing in an unacceptable manner. The mean interreader correlation across the six data sets was .857. This comparatively high correlation also suggests a degree of consistency in judgments across readers.

TABLE 3

Tabulations of Essays Read by Most Frequent Readers 1 and 2 for Elicitation Prompt B (N = 2,572 Essays; 29 Most Frequent Readers)

Reader 1	N	Reader 2	N
312	92	311	65
313	72	314	73
*314 (B)	419	315	53
316	98	316	65
317	72	321	98
318	69	322	73
321	71	324	74
324	74	*325 (D)	284
*327 (F)	173	*326 (A)	536
328	73	327	144
330	127	328	212
331	65	331	71
332	74	336	72
335	75	337	69
336	74	338	75
337	138	341	73
338	98	343	72
340	138	344	173
*341 (E)	217	*345 (C)	290
343	73		
345	63		
346	146		
348	71		
Total	2,572		2,572

*Indicates six most frequent readers to be employed in subsequent analyses.
( ) Indicates reader label assigned.

TABLE 4

Score Frequencies and Reader Calibrations
for Most Frequent Reader Pairings for Elicitation
Prompt B (N = 428 Essays)

	Set 1 (N = 65)			Set 2 (N = 71)		
Score	Reader B	Reader A	Total	Reader E	Reader A	Total
1	0	0	0	0	0	0
2	3	5	8	1	2	3
3	21	18	39	11	18	29
4	22	26	48	31	26	57
5	17	16	33	15	19	34
6	2	0	2	13	6	19
Logit	-.284	.284		-.458	.458	
SE	.217	.217		.156	.157	
Infit	-1.778	-1.673		-.064	-.008	
Outfit	-2.033	-1.886		-.267	-.184	
Gap	-.098	-.061		.151	.166	

	Set 3 (N = 74)			Set 4 (N = 71)		
Score	Reader B	Reader C	Total	Reader B	Reader D	Total
1	0	1	1	0	0	0
2	3	3	6	2	2	4
3	12	10	22	18	16	34
4	29	35	64	23	27	50
5	21	19	40	21	18	39
6	9	6	15	7	8	15
Logit	-.595	.595		-.084	.084	
SE	.175	.182		.287	.287	
Infit	-2.455	-.790		-.825	-.785	
Outfit	-1.826	-1.593		-.970	-.962	
Gap	.719	1.237		-.096	-.091	

Table 4 (cont.)

	Set 5 (N = 73)			Set 6 (N = 74)		
Score	Reader E	Reader D	Total	Reader F	Reader C	Total
1	0	0	0	0	1	1
2	1	2	3	3	1	4
3	14	13	27	14	16	30
4	26	21	47	31	31	62
5	25	23	48	24	21	45
6	7	14	21	2	4	6
Logit	.294	-.294		.000	.000	
SE	.087	.086		.153	.153	
Infit	-1.343	-4.954		-3.474	-3.295	
Outfit	3.806	2.910		-3.793	-3.684	
Gap	2.569	2.073		1.201	1.201	

Although the positive infit criterion 2.0 was not exceeded for these frequently paired readers of prompt B essays, it is evident from Table 4 that readers D and E exceeded the positive outfit criterion in data set 5. Also, reader D exceeded the negative infit criterion in data set 5 and readers C and F exceeded all negative fit criteria in data set 6. These findings suggest that, while the most critical positive infit criterion was satisfied, readers C, D, E, and F exhibited some borderline unexpected rating behavior that merited closer examination.

Another way to examine misfit to expectation for rating assignments made by readers is to establish a chi-square contingency table such as that presented for essays prepared on prompt B in Table 5, and to test the independence of readers and rating categories. Because frequencies of essays within cells occasionally dropped below 5, Yates' correction for continuity procedure was used to compensate for this. Even after correction for continuity, it was found that the chi-square value 40.64 exceeded the critical value (37.653, 25 d.f., $p < .05$), suggesting that readers and rating categories assigned were not independent for this essay prompt and these 1,919 essays. It is possible to understand the reason for this lack of independence by examining the sums of absolute standardized residuals in the margins of the tables. It was clear that there was a high deviation from expectation (17.15) in the frequency of assignment of a rating of 6. Apparently these raters tended to show unexpected disagreement in what constituted an essay at the highest rated level. Some readers (e.g., C and F) tended to underassign a 6. Other readers (e.g., D and E) assigned this rating more frequently than expected. Perhaps these readers would have benefited from additional training in the assignment of ratings at the highest step of the scale, or perhaps the definition of this step needs to be clarified so judges will share a common understanding of what this scale step means in terms of writing behavior. If this single problem could be alleviated, the independence of reader and rating would be re-established for this data set. It is noteworthy that this chi-square analysis identified the same misfitting readers, C, D, E, and F, as were identified as borderline misfitting readers in the Rasch model scalar analysis. However, the chi-square procedure facilitated identification of the cause of misfit as overassignment or underassignment of a 6 rating. For this particular study, the chi-square procedure also held the advantage of allowing consideration of the entire group of most frequently paired readers in one combined analysis rather than just one pair of readers at a time.

Prompt C Reader Tabulations and Calibrations

Table 6 represents a summary of reader identification numbers and numbers of essays read for the 30 most frequently paired readers of essays prepared according to prompt C. In all, 1,544 essays were tallied for prompt C. This table represents a tally for prompt C corresponding to the tally provided in Table 3 for prompt B. Note again that the six most frequent readers (i.e., A-F) are identified and labeled for subsequent analyses. Although three readers are shown to have identical tabulations of 139 essays, reader number 432 was chosen

TABLE 5

Reader x Score Chi-Square Contingencies for the Six Most Frequent Readers of Prompt B Essays

	SCORE							
Reader	1	2	3	4	5	6	Total	$\frac{(O-E)^2}{E}$
A	1	22	119	212	149	33	536	
	-.24	.33	.28	-.12	.10	-.96		2.03
B	3	17	93	159	117	30	419	
	.32	.18	.21	-.66	.10	-.01		1.48
C	3	11	74	124	62	16	290	
	1.38	.00	2.54	.28	-3.18	-1.15		8.53
D	1	8	53	108	80	34	284	
	-.40	-.24	-.66	-.40	.11	7.42		9.23
E	0	2	37	101	53	24	217	
	-.18	-3.50	-1.47	1.75	-.44	3.45		10.79
F	0	8	28	75	57	5	173	
	-.07	.31	-1.72	.26	2.06	-4.16		8.58
Total	8	68	404	779	518	142	1919	
$\frac{(O-E)^2}{E}$	2.59	4.56	6.88	3.47	5.99	17.15		*40.64

* $p < .05$ with Yates' correction for continuity standardized residuals under cell frequencies, with sign indicating direction of deviation from expectation.

TABLE 6

Tabulations of Essays Read by Most Frequent Readers 1 and 2 for Elicitation Prompt C (N = 1,544 Essays; 30 Most Frequent Readers)

Reader 1	N	Reader 2	N
424	73	*414 (B)	199
431	75	*432 (F)	139
435	72	434	75
444	72	436	99
450	74	438	89
451	74	*441 (C)	149
*452 (E)	140	442	74
453	99	444	139
456	139	445	75
457	89	446	75
460	74	447	65
*462 (A)	200	454	74
*475 (D)	149	462	75
480	64	468	72
483	75	478	72
484	75	482	73
Total	1,544		1,544

*Indicates six most frequent readers to be employed in subsequent analyses.
( ) Indicates reader label assigned.

for subsequent analysis because of a higher observed pairing of readings with the other five most frequent readers.

Prompt C Reader Fit to the Model

Table 7 corresponds to Table 4, but presents information derived from the most frequent reader pairings with prompt C rather than with prompt B. Note that, because prompt C essays with frequently paired readers were about half the number of comparable prompt B essays, the total number of qualifying data sets for prompt C analysis reported in Table 7 was half the number of data sets for prompt B analysis reported in Table 4. Again, there is no evidence of positive reader misfit by the same criteria applied in the interpretation of Table 4. The overall fit to model expectation was even higher for prompt C essays than for prompt B essays. The mean interreader correlation across the three data sets in Table 7 was .852. This high coefficient suggests a high degree of interreader agreement similar to that witnessed for readers of prompt B.

Despite the fact that reader fit to the expectations of the Rasch scalar analysis model was even better for prompt C than for prompt B, it is useful to consider the further comparative results of the same chi-square analytic procedure for prompt C as was reported for prompt B. Table 8 reports the reader x score chi-square contingency table for the six most frequent readers of prompt C. This table corresponds to Table 5 for prompt B. In the case of Table 8, unlike Table 5, the chi-square value did not exceed the critical value, so we cannot assert that rating assignment overall was dependent on the readers. It is interesting, nevertheless, that there was a nonsignificant tendency to overassign a rating of 4 to prompt C, and this overall tendency was due primarily to unexpected behavior on the part of reader A. Because reader A was the reader who managed to evaluate the most essays in the time permitted, this unexpected outcome suggests the hypothesis that reader A may have achieved reading fluency by overassigning ratings at the mid-point of the scoring range. It may be desirable on the basis of this outcome for scoring administrators to caution some fluent readers against working too quickly at the expense of scoring accuracy. In particular, reader A might be encouraged to slow down and become more reflective and less compulsive in the reading of essays. It is also possible that the overuse of midrange values by reader A was in reaction to feedback that errors were being made in the assignment of scores outside the middle range. However, because the overall tendency to overassign midrange values was not statistically significant, it is also a distinct possibility that reader A was by chance supplied a disproportionate number of 4-level essays to read.

It is likely that this kind of simple chi-square contingency analysis could be easily implemented by computer at regular scoring intervals during training sessions or operational readings. This could provide readers and session leaders with rapid, detailed feedback on the appropriateness of reading judgments of individual readers. Over or underuse of particular rating values could also be identified.

TABLE 7

Score Frequencies and Reader Calibrations
for Most Frequent Reader Pairings for Elicitation
Prompt C (N = 275 Essays)

	Set 1 (N = 125)			Set 2 (N = 75)		
Score	Reader A	Reader B	Total	Reader D	Reader C	Total
1	1	0	1	0	0	0
2	3	5	8	2	7	9
3	31	36	67	19	23	42
4	65	49	114	30	17	47
5	19	29	48	19	18	37
6	6	6	12	5	10	15
Logit	.089	-.089		-.121	.121	
SE	.146	.146		.155	.155	
Infit	-1.596	-1.278		-2.251	-1.622	
Outfit	-1.696	-1.554		-2.282	-2.048	
Gap	-.115	-.065		.070	.098	

	Set 3 (N = 75)		
Score	Reader E	Reader F	
1	0	0	0
2	2	2	4
3	26	23	49
4	19	21	40
5	22	21	43
6	6	8	14
Logit	-.180	.180	
SE	.174	.174	
Infit	-1.553	-1.463	
Outfit	-1.672	-1.437	
Gap	-.199	.016	

TABLE 8

Reader x Score Chi-Square Contingencies for the Six Most Frequent Readers of Prompt C Essays

	SCORE							
Reader	1	2	3	4	5	6	Total	$\frac{(O-E)^2}{E}$
A	2	6	43	103	33	13	200	
	2.90	-.21	-2.15	7.06	-2.98	.00		15.30
B	0	7	61	79	42	10	199	
	.00	.00	.77	.00	-.12	-.65		1.54
C	0	9	40	48	37	15	149	
	-.12	1.26	.00	-1.80	.25	1.94		5.37
D	0	6	33	60	43	7	149	
	-.12	.02	-1.20	.00	2.36	-.66		4.36
E	0	2	41	50	36	11	140	
	-.16	-1.60	.16	-.40	.49	.11		2.92
F	0	8	47	45	29	10	139	
	-.17	.81	2.03	-1.59	-.11	.00		4.71
Total	2	38	265	385	220	66	976	
$\frac{(O-E)^2}{E}$	3.47	3.90	6.31	10.85	6.31	3.36		*34.20

* N.S. df=25 with Yates' correction for continuity standardized residuals under cell frequencies, with sign indicating direction of deviation from expectation.

Overall Essay Fit to the Model

One of the purposes of this study was to determine the feasibility of applying Rasch model scalar analysis to the analysis of TWE essays. One indication of the suitability of applying this analysis procedure is reflected in the percentage of essays found to misfit the expectations of the model. Rentz and Rentz (1979) reported that rejection rates ranging between 5 and 10% are usual in application of Rasch model procedure to dichotomously scored items, are to be expected, and can be considered acceptable. As Table 9 indicates, essay rejection rates in the TWE analysis of essays from two separate prompts were about 1% for positive misfit, and 4% for less critical negative misfit. Thus, the positive misfit rate for applying Rasch model rating scale analysis prior to adjudication was about the same as the rate of requirement of a third reader in the adjudication process as indicated in Table 1. Although it was not determined whether the misfitting essays were necessarily the same essays as those requiring adjudication, the nature of the fit estimation procedure makes it possible that considerable overlap existed between statistical misfit and need for adjudication.

Because the fit statistics reflect the degree of fit to a unidimensional model of analysis, the observed low rate of misfit also provides evidence of the basic psychometric unidimensionality of the data set. This supports the appropriateness of applying IRT methodology that requires such psychometric unidimensionality, and it further implies feasibility of equating. It is important to note, however, that satisfying the psychometric unidimensionality requirements does not imply that writing as assessed is not a psychologically complex phenomenon involving numerous and diverse abilities of the writers (Henning, in press).

Discussion and Conclusions

In order to provide information concerning psychometric properties of the TWE scoring scale and to examine reader, essay, and scale-step fit to patterns of expectation established for that scale, Rasch model rating scale analyses were applied in the analysis of 2,572 essays prepared on one TWE prompt and in the analysis of 1,544 essays prepared on a different TWE prompt. Results provided the following summarized information items:

1. Application of IRT-based Rasch rating scale analysis appeared feasible and appropriate for TWE essay data, even before adjudication of discrepant essay scores. Rates of essay misfit were extremely low and corresponded, in the case of positive misfit, to the rate for which third readers were required to adjudicate discrepant essays (i.e., 1%). However, the actual rate of overlap between misfitting essays and essays requiring adjudication was not reported.

TABLE 9

Frequency of Essay Misfit to Rasch Model
Rating Scale Score Predictors
(N = 4,116 Essays)

	Infit		Outfit	
	N	%	N	%
Prompt B				
Essays	2,572		2,572	
Mean	.060		.060	
SD	.644		.644	
Positive Misfit	28	1.08	28	1.08
Negative Misfit	110	4.28	110	4.28
Prompt C				
Essays	1,544		1,544	
Mean	.258		.258	
SD	.273		.273	
Positive Misfit	12	.78	12	.78
Negative Misfit	59	3.82	59	3.82

2. The high rate of essay fit to the expectations of the rating scale analysis procedure suggested the basic psychometric unidimensionality of the score data as is required by the rating scale analysis procedure. Although this suggestion of "psychometric" unidimensionality has many profound advantages from the perspective of reporting, interpreting, and equating scores, it does not imply that the writing process does not exhibit "psychological" multidimensionality, which is a demonstrably distinct proposition (Henning, in press).

3. Procedures were identified for the simple equating of TWE essays across prompts, and the feasibility of this process for the present data was shown. In the present study, mean scale-step difficulty estimates were employed as the basis for equating rather than alternative possibilities such as using common readers or common writers. Discrepencies across the two similar prompts examined were found to be predictably small (i.e., 0.059 logits) and only slightly exceeding one estimate of the standard error of equating (i.e., 0.034 logits). A procedure was described for using this estimated mean logit difference across steps as a translation constant in the equating. However, before such equating methodology can be operationally implemented for TWE essays, further study is required with more diverse prompt types than were employed in the present study. Such further study is particularly important as evidence grows that judgments of writing quality appear to be influenced by such variables as mode of discourse, experiential demand, and writer gender that were not systematically considered here (Engelhard, Gordon, & Gabrielson, 1991). Also, it would be advisable to employ more recent FACET software that would permit judgments of reader fit even when less rapidly scoring and less frequently paired readers are included in the sample (Linacre, 1989). Further study of this equating methodology is particularly attractive given the problems encountered with implementation of more traditional equating methodology with the TWE test (DeMauro, 1992) and given the need to ensure variety of prompts across TWE administrations (Golub-Smith, Reese, and Steinhaus, 1992).

4. Misfit of a subsample of paired readers for both prompts was found to be so small that, by some established criteria of interpretation, no particular reader was rejected by the analysis. However, subsequent chi-square contingency tests of the independence of readers and ratings assigned did provide insights into ways in which individual readers might be helped to improve their reading behavior. In particular, one fluent reader was indicated as possibly overassigning the rating of 4. It was hypothesized that the fluency of that reader might be related to the tendency to assign a preponderance of scores at the midrange position. Thus, the inaccuracy could be motivated by the desire to complete more readings in the assigned time. Another possible but untested hypothesis for this aberrant reader behavior was that readers who are cautioned in training that their ratings are inaccurate may adopt a more conservative approach of assigning midrange values when they are uncertain of the appropriate values.

5. In the case of essays prepared on prompt B, there was a significant undesirable chi-square dependency between readers and their assigned ratings. This was due primarily to unexpected disagreements in the frequency of the assignment of a rating of 6, with some readers overassigning and others underassigning this rating. For some readers, it was clear that further training in the identification of essays at the 6 level would be beneficial.

6. The rating scale defined by the TWE steps 1-6 appeared to be a true equal-interval scale with little standard error at each scale step relative to the breadth of the scoring intervals defined by those steps. This was also consistent with the finding of high Spearman-Brown adjusted interrater reliabilities estimated for essays on each prompt (i.e., B = .900 and C = .902). There was, however, comparative underuse of the rating scale category 1. The observed underuse of this rating category may disappear when samples larger than those employed in the present study are investigated.

References

Andrich, D. (1978a). A binomial latent trait model for the study of Likert-style attitude questionnaires. British Journal of Mathematical and Statistical Psychology, 31, 84-98.

Andrich, D. (1978b). A rating formulation for ordered response categories. Psychometrika, 43, 561-573.

Andrich, D. (1978c). Scaling attitude items constructed and scored in the Likert tradition. Educational and Psychological Measurement, 38, 665-680.

Andrich, D. (1978d). Application of a psychometric rating model to ordered categories which are scored with successive integers. Applied Psychological Measurement, 2, 581-594.

Andrich, D. (1979). A model for contingency tables having an ordered response classification. Biometrics, 35, 403-415.

Davidson, F., & Henning, G. (1985). A self-rating scale of English difficulty: Rasch scalar analysis of items and rating categories. Language Testing, 2(2), 164-179.

DeMauro, G. E. (1992). Investigation of the appropriateness of the TOEFL test as a matching variable to equate TWE topics (TOEFL Research Report No. 37). Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1989). TOEFL Test of Written English guide. Princeton, NJ: Author.

Engelhard, G., Jr. (1991, April). The measurement of writing ability with a many-faceted Rasch model. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Engelhard, G., Jr., Gordon, B., & Gabrielson, S. (1991, April). Writing tasks and the quality of student writing: Evidence from a statewide assessment of writing. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Golub-Smith, M., Reese, C., & Steinhaus, K. (1992). Topic and topic type comparability on the Test of Written English. Manuscript submitted for publication.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. Language Learning, 41(3), 337-373.

Henning, G. (1988a). The influence of test and sample dimensionality on latent trait person ability and item difficulty calibrations. Language Testing, 5(1), 83-99.

Henning, G. (1988b). A long-range plan for TOEFL program research. Princeton, NJ: TOEFL Research Committee, Educational Testing Service.

Henning, G. (1989). Meanings and implications of the principle of local independence. Language Testing, 6(1), 95-108.

Henning, G. (in press). Dimensionality and construct validity of language tests. Language Testing.

Henning, G. & Davidson, F. (1987). Scalar analysis of composition ratings. In K. M. Bailey, T. L. Dale, & R. T. Clifford (Eds.), Language testing research: Selected papers from the 1986 colloquium. Monterey, CA: Defense Language Institute.

Linacre, J. M. (1989). Many-faceted Rasch measurement. Chicago: MESA Press.

Muraki, E. (1991). Developing the generalized partial credit model. Paper presented at Educational Testing Service, Princeton, NJ.

Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assignments: Rasch partial credit analysis of performance in writing. Language Testing, 4(1), 72-92.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press, 1980. (Original work published by the Danish Institute for Educational Research, 1960).

Rentz, R. R., & Rentz, C. C. (1979). Does the Rasch model really work? Measurement in Education, 10, 1-8. (ERIC Document Reproduction Service No. ED 169137).

Stansfield, C. W., & Ross, J. (1988). A long-term research agenda for the Test of Written English. Princeton, NJ: Educational Testing Service.

Wright, B. D., & Linacre, J. M. (1985). Microscale manual. Version 2.0. Black Rock, CN: Mediax Interactive Technologies, Inc.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis: Rasch measurement. Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). Best test design: Rasch measurement. Chicago: MESA Press.

Appendix A

Test of Written English Scoring Guide

(Revised 2/90)

Readers will assign scores based on the following scoring guide. Though examinees are asked to write on a specific topic, parts of the topic may be treated by implication. Readers should focus on what the examinee does well.

Scores

6 **Demonstrates clear competence in writing on both the rhetorical and syntactic levels, though it may have occasional errors.**
A paper in this category
- -effectively addresses the writing task
- -is well organized and well developed
- -uses clearly appropriate details to support a thesis or illustrate ideas
- -displays consistent facility in the use of language
- -demonstrates syntactic variety and appropriate word choice

5 **Demonstrates competence in writing on both the rhetorical and syntactic levels, though it will probably have occasional errors.**
A paper in this category
- -may address some parts of the task more effectively than others
- -is generally well organized and developed
- -uses details to support a thesis or illustrate an idea
- -displays facility in the use of language
- -demonstrates some syntactic variety and range of vocabulary

4 **Demonstrates minimal competence in writing on both the rhetorical and syntactic levels.**
A paper in this category
- -addresses the writing topic adequately but may slight parts of the task
- -is adequately organized and developed
- -uses some details to support a thesis or illustrate an idea
- -demonstrates adequate but possibly inconsistent facility with syntax and usage
- -may contain some errors that occasionally obscure meaning

3 **Demonstrates some developing competence in writing, but it remains flawed on either the rhetorical or syntactic level, or both.**
A paper in this category may reveal one or more of the following weaknesses:
- -inadequate organization or development
- -inappropriate or insufficient details to support or illustrate generalizations
- -a noticeably inappropriate choice of words or word forms
- -an accumulation of errors in sentence structure and/or usage

Test of Written English Scoring Guide (continued)

2 **Suggests incompetence in writing.**
A paper in this category is seriously flawed by one or more of the following weaknesses:
- -serious disorganization or underdevelopment
- -little or no detail, or irrelevant specifics
- -serious and frequent errors in sentence structure or usage
- -serious problems with focus

1 **Demonstrates incompetence in writing.**
A paper in this category
- -may be incoherent
- -may be underdeveloped
- -may contain severe and persistent writing errors

Papers that reject the assignment or fail to address the question must be given to the Table Leader. Papers that exhibit absolutely no response at all must also be given to the Table Leader.

Appendix B

Mathematical Specification of the Rating Scale Model

Assuming

$$\delta_{ik} = \delta_i + \tau_k$$

Where δ_i is the location or "scale value" of item i on the variable and τ_k is the location of the k'th step in each item relative to the scale value of that item, and the pattern of item steps is described by the "threshold" parameters $\tau_1, \tau_2, \ldots, \tau_m$, and is estimated once for the entire item set, then

$$\phi_{nik} = \frac{\pi_{nik}}{\pi_{nik-1} + \pi_{nik}} = \frac{\exp\left[\beta_n - (\delta + \tau_k)\right]}{1 + \exp\left[\beta_n - (\delta_i + \tau_k)\right]} \quad k = 1, 2, \ldots, m$$

Where ϕ_{nik} is person n's probability of scoring k on item i, β_n is the ability of person n, which can be written as the probability of person n responding in category x to item i.

$$\pi_{nix} = \frac{\exp \sum_{j=0}^{x} \left[\beta_n - (\delta_i + \tau_j)\right]}{\sum_{k=0}^{n} \exp \sum_{j=0}^{k} \left[\beta_n - (\delta_i + \tau_j)\right]} \quad x = 0, 1, \ldots, m$$

Where $T_0 = 0$ so that

$$\exp \sum_{j=0}^{0} \left[\beta_n - (\delta_i + \tau_j)\right] = 1$$

Printed on Recycled Paper

57906-01201 • Y82M.5 • 275584 • Printed in U.S.A.