DOCUMENT RESUME

ED 400 293 TM 025 566

AUTHOR	Du, Yi; And Others
TITLE	Differential Facet Functioning Detection in Direct Writing Assessment.
PUB DATE	Apr 96
NOTE	23p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Age Differences; Elementary School Students; Elementary Secondary Education; *Essay Tests; High School Students; Identification; Interrater Reliability; *Item Bias; Item Response Theory; *Performance Based Assessment; Sex Differences; *Test Construction; Writing (Composition); *Writing Tests
IDENTIFIERS	*Direct Writing Assessment; FACETS Computer Program; *FACETS Model; Illinois Goal Assessment Program; Rasch Model

ABSTRACT

In the framework of performance assessment, because of the involvement of many facets, the development of ways to detect differential item functioning or differential facet functioning (DFF) has lagged beyond the practical needs of test developers. To monitor the validity and fairness of an assessment, it is critical to discover a method that can detect multiple sources of potential DFF from raters, item, topics, and other facets. Many-faceted Rasch modeling with the FACETS software provides a powerful way to detect DFF in performance assessment. This study focuses on raters and topic types as two sources of DFF using the FACETS model. Data came from 1,734 essays written by 867 students in grades 6, 8, and 10 as part of the Illinois Goal Assessment Program. A measurement model of eight facets was used. With the FACETS model, DFF analysis of raters identified biased raters. Evidence was also found that bias on the part of these raters affected students' writing ability estimates. DFF statistics for topic types and student demography showed effects of performance of topic types on student subgroups and provided evidence of gender and age impacts on different topic types. (Contains 3 figures and 12 tables.) (SLD)

DIFFERENTIAL FACET FUNCTIONING DETECTION IN DIRECT WRITING ASSESSMENT

Yi Du
Minneapolis Public Schools

Benjamin D. Wright
University of Chicago

William L. Brown
Minneapolis Public Schools

Paper for the presentation at the Annual Conference of American Educational Research Association, New York, April 1996

The last 25 years has witnessed the emergence of an arsenal of methods for analyzing and identifying items with Differential Item Functioning (DIF) characteristics in the context of standardized multiple-choice testing (Angoff, 1993, p.21). In the framework of performance assessment, however, because of the involvement of many facets (rater, examinee, topic, etc.), the development of DIF or DFF (differential facet functioning) methodology has not been so successful, and far behind the practical needs.

DFF (rater facet, task facet or other facets) may affect unfairly the observed measure of a student in the performance assessment. According to Angoff's definition, DIF is "referring to the simple observation that an item displays different statistical properties in different group settings (after controlling for differences in the abilities of the groups)" (Angoff, 1993, p.4). A similar definition may, in principle, apply to DFF in performance assessment. DFF refers to the simple observation that an item, a topic, a rater, or other testing facet displays different statistical properties in different group settings (after controlling for differences in the abilities of the groups).

Although several methods for detecting bias in performance assessment have been proposed, most of them can merely detect one source of DFF, for example, either rater DFF or topic DFF. In order to monitor the validity and fairness of assessment, it is critical to discover a method that can detect multiple sources of potential DFF from raters, item, topics and other facets of performance assessment.

Many-faceted Rasch modeling via the FACETS software (Linacre, 1989) provides a powerful way to detect DFF in performance assessment. As an extension of the Rasch model, FACETS models the probabilities of ordered-category ratings in terms of parameters for students, raters, topics, or other facets. Student parameters capture students' tendencies to receive high or low ratings; rater parameters, their severity or leniency; topic parameters, their difficulty level; etc..

Parameter interactions or facet interactions (raters and students, topics and students, topics and raters, etc.) are allowed in FACETS, which has the flexibility to define a wide range of models. This function makes it possible to detect DFF (interaction) in either standardized multiple-choice testing or performance assessment.

For example, to generate topic bias estimates for student ethnic groups, a five-facet model with topic-student ethnic group interaction can be used:

$$\log(\frac{P_{nijmk}}{P_{nijmk-1}}) = B_n - D_i - C_j - A_{gm} - E_{mg} - F_k \qquad (1)$$

where, P_{nijmk} is the probability of examinee n being graded in category k by rater j on item i and topic m, $P_{nijmk-1}$ is the probability of examinee n being graded k-1 by rater j on item i and topic m, B_n is the writing ability measure of examinee n, D_i is the difficulty calibration of item i, C_j is the severity measure of rater j, F_k is the difficulty calibration of grading category k-1 relative to category K. The rating scale is k=0, K; A_{gm} is the difficulty of each topic m for each student ethnic group g; and E_{mg} is the ability of each student ethnic groups on each topic m.

This interaction allows every student group to be estimated based on every topic. After that estimation has been completed and with all measures and calibrations for the model anchored, estimates of DFF measures for the interactions between each topic (A_m) and each student group (E_g) across the whole data set are conducted by FACETS.

This study focuses on detecting two sources of DFF (raters and topic types) using the FACETS model. First, this study proposes the procedures for defining different interaction models and detecting DFF, and uses data from a large-scale performance assessment of writing to illustrate these procedures. Second, this study analyzes differential rater functioning and identifies potentially biased raters. Third, this study analyzes differential topic type functioning and provides evidence of gender and age impacts on different topic types. Finally, this study provides information about the writing assessment to help teachers, administrators and test developers to identify how student characteristics (such as maturity level, gender and ethnic background) influence examinee writing skills, in order to provide the best conditions for understanding student achievement.

Data

The study used 1,734 essays written by 867 students. These essays were randomly selected from 150,000 essays submitted as part of the 1993 writing assessment of the Illinois Goal Assessment Programs (IGAP). The sample included students at grades six, eight and ten. Fifty-one percent of the selected students were male and 49 percent were female. The racial/ethnic distribution was: 74 percent white, and 26 percent minority (including black and Hispanic students). The percent by grade level were 27 percent grade 6 students, 24 percent grade 8 students and 49 percent grade 10 students. Eighty nine raters were used. Gender composition of the raters was 43 males and 46 females. Seventy-four raters were white and 15 were black.

Instrument

The IGAP writing framework focuses on students' abilities to write effectively for three purposes--narrative, expository and persuasive. Narrative writing encourages students to incorporate their imagination and creativity into the production of stories or personal essays. Persuasive writing focuses on the reader with the primary aim of “influencing others to take some action or bring about change." Expository writing "focuses primarily on the subject matter element in communication" (Writing on, Illinois! 1992).

The whole writing assessment used five writing topics representing persuasive, expository, and narrative discourse modes across grades six, eight and ten. Some of the prompts were grade specific. For example, the expository prompt "*trading places*" was used for grade 6 students, and "*change*" for grades 8 and 10 students. Of the persuasive prompts, "*space*" was used for grades 6 and 8, and "*invention*s" for grade 10. The narrative prompt about "*forget*" was used for all three grades.

Each student responded to two prompts: one was assigned and the other one was his/her own selection. Each essay was scored by two raters. Raters judged these essays across topics, topic types, and grades. Raters who graded the expository essays in the first scoring process may have graded the persuasive, or narrative, or even expository essays in the second scoring process. Raters who graded grade 6 student essays in the first scoring may have graded grades 8 or 10 student essays, or even grade 6 in the second scoring process. All raters received extensive training.

Scoring Scale

The writing assessment uses an integrated analytic/holistic scale for five features: Focus, Support/Elaboration, Organization, Integration and Convention. Description of the five features are:

Focus - the clarity with which a composition presents a clear main idea, point of view, theme, or unifying event.

Support/Elaboration - the degree to which the main point or event is elaborated and explained by specific detail and reason.

Organization - the clarity and/or coherence of the logical flow of ideas and the explicitness of the text structure or plan.

Integration - evaluation of the essay based on a judgment of how effectively the composition as a whole uses the basic features to address the assignment.

Conventions - use of standard written English.

Each feature is rated on a 6-point scale (except convention which is rated on a 2-point scale). Scores are summed to yield a total score:

Focus	+	Support	+	Organization	+	Convention	+	Integrating	=	Score
(1-6)		(1-6)		(1-6)		(1-2)		(1-6)		(5-26)

The writing assessment applies a developmental scale, intended to be uniform across grades. As a result, students in upper grades are expected to receive higher ratings than those in lower grades.

Methods and Procedures

To address all the facets to be analyzed, a measurement model with eight facets was used: writing ability, rater severity, item (scoring component) difficulty, topic difficulty, topic type difficulty, grade level ability, gender, and ethnicity. Each facet was estimated separately.

First Step--Primary FACETS Analysis

Four primary facets--student ability, rater leniency, writing topic, and writing features (items)--are defined, using the FACETS computer program for the primary analysis. Because the IGAP writing assessment uses two different scales for the items, i.e., a 2-point scale for item 4 and a 6-point scale for items 1, 2, 3, and 5, two FACETS models for the two scales are required. The two models for the four primary facets were defined as:

$$\log\left(\frac{P_{nijmk}}{P_{nijmk-1}}\right) = B_n - D_i - C_j - A_m - F_k \qquad (2)$$

where, P_{nijmk} is the probability of examinee n being graded in category k by rater j on item i and topic m, $P_{nijmk-1}$ is the probability of examinee n being graded k-1 by rater j on item i and topic m, B_n is the writing ability measure of examinee n, D_i is the difficulty calibration of item i, C_j is the severity measure of rater j, A_m is the difficulty calibration of topic k, F_k is the difficulty calibration of grading category k-1 relative to category K. The rating scale is k=0, K. Items i=1, 2, 3, 5; and scale categories k=1, 2, 3, 4, 5, 6.

$$\log\left(\frac{P_{nijm2}}{P_{nijm1}}\right) = B_n - D_i - C_j - A_m \qquad (3)$$

where, item i=4, and scale categories k=1, 2.

The convergence criteria for the joint maximum likelihood iterations was set at “no marginal score point residual greater than 0.5 score points, and no logit estimate changing faster than .01 logits.” Thus, the satisfied estimations for the four parameters can be obtained.

The first run of FACETS determines the four primary facets--student, rater, item, and topic--on a common logit scale. These writing ability (proficiency) distributions for each student on the logit scale are based on the topic, rater and item parameter estimates.
The second run of FACETS is conducted to calibrate student gender and ethnicity, as well as topic type difficulty.

Second Step– DFF Analysis

The second step is to define DFF models and to allow interactions between rater facet and student facets including grade, gender and ethnicity facets, between rater facet and topic facet, as well as between topic type facet and student facets. This analysis provides information of potential sources of bias (interaction) in assessment. In particular, this study focuses on rater bias and topic type bias for different groups of students.

Results and Interpretation

Student Measures

Figure 1 maps the elements of the eight facets of this examination on their common log odds scale. The eight facets are: students, raters, items, topics, topic types, student grades, students' gender and ethnic subgroups.

Because all facets are on a common scale, it is easy to compare elements within and between facets. For the student facet, high ability students are on top, and low ability students are at the bottom. For the rater facet, severe raters are on top, lenient raters at the bottom. For the item facet, "Integration," "Support," and "Organization" are harder items (that are on top), while "Convention" is easiest (which is at the bottom). For the topics, "*trading places*" is the hardest, while the other four are easier. For the topic types, narrative writing is easiest, while expository and persuasive writing are harder.

In the comparison of grades, the grade 10 is at top, indicating that grade 10 students have the highest writing abilities, while the grade 6 is at the bottom, which means that grade 6 students have the lowest writing abilities. In the comparisons of gender and ethnic groups, females did better than males, white students did better than black and Hispanic students. The last column maps the distances between categories of the 6-category scale. The distances between categories was unequal. Therefore, the original scale is nonlinear. For example, categories 4 and 6 take the largest space, while categories 2, 3 and 5 take the smallest space.

Figures 2 and 3 magnify important parts of Figure 1 to clarify the differences within each facet. These figures show clearer differences of elements within each facet than Figure 1 does.

Measr	+students	-raters	-items	-topics	-types	+grades	+gender	+ethnicity	s.1
11	*More able*	*Severe*	*Hard*	*Hard*	*Hard*	*More able*	*More able*	*More able*	(6)
10	.								
	.								
9	.								
	.								
8	.								
	.								
7	*.								
	**								
6	**.								
	*****.								
5	*******.								
	*********								5
4	*********.					G10	Female	White	
	*********.					G8			
3	********.						Male		
	*******.					G6			
2	********.	.						Black Hispanic	
	*******.	.							4
1	*****.	**							
	****.	*******.	Int Sup Org	trade places					
0	****.	******.		chng invn spce frget					
	****.	*********	Foc		exp pur				
-1	**.	**.	Con		nar				
	**.	.							3
-2	*	.							
	.								
-3	.								
	.								
-4	.								2
	.								
-5									
-6									
	.								
-7	*Less able*	*Lenient*	*Easy*	*Easy*	*Easy*	*Less able*	*Less able*	*Less able*	
Measr	* = 8	* = 3	-items	-topics	-types	+ grades	+ gender	+ ethnicity	s.1.

Figure 1. Calibrations of All Facets

Figure 2 magnifies the item, topic, and topic type facets. This figure provides a better picture of differences of elements within the facets of items, topics and topic types than Figure 1.

Measr	-items	-topics	-topic types
1	*Hard*	*Hard*	*Hard*
	Integration		
	Support		
	Organization		
		trade places(G6)	
0			
		invention (G10)	expository
			persuasive
		change (G8, 10) space (G6, 8)	
			narrative
		forget (G6, 8 and 10)	
	Focus		
-1			
	Convention		
-2	*Easy*	*Easy*	*Easy*
Measr	-items	-topics	-topic types

Figure 2. FACETS Map for Items, Topics, and Topic Types at the Range Between -2 and 1 Logits

Figure 3 magnifies the grade, gender and ethnicity facets. This figure provides a better picture of differences of students in terms of their grades, gender and ethnicity than Figure 1.

Measure	+Grade	+Gender	+Race/Ethnicity
4	*More able*	*More able*	*More able*
	Grade 10		
			White
		Female	
3	Grade 8		
		Male	
2			
	Grade 6		
			Black
			Hispanic
1	*Less able*	*Less able*	*Less able*
Measr	+Grade	+Gender	+Race/Ethnicity

Figure 3. FACETS Map for Grade, Gender and Ethnicity at the Range Between 1 and 4 Logits

Table 1 reports some student writing ability estimates, their standard errors, infit and outfit statistics, and the summary statistics for the student facet. A grade 10 student at the top in order of measures (10.10), has the highest ability, and a grade 6 student at the bottom (-6.53), has the lowest. The reliability of this student separation is 0.96. The mean infit is 1.0 and outfit is 0.9. The chi-square statistic, $\chi^2 = 24749.6$ with df=858, $p < 0.001$, indicates that these students are significantly different. The other chi-square statistic, $\chi^2 = 854.1$ with df=857, $p < .52$, supports the hypothesis that the distribution of students is normal.

TABLE 1 STUDENT MEASUREMENT REPORT

Obsvd Score*	Obsvd Count	Obsvd Average	Fair Average	Logit Measure	Model S.E.	Infit MnSq	Outfit MnSq	Num	Students
103	20	5.2	5.2	8.86	1.02	1.0	0.7	113	101923
103	20	5.2	5.2	8.86	1.02	1.0	0.7	35	112307
103	20	5.2	5.2	8.75	1.02	1.0	0.8	116	124568
103	20	5.2	5.2	8.75	1.02	1.0	0.7	176	125467
102	20	5.1	5.1	7.97	0.74	1.1	1.8	203	457123
102	20	5.1	5.1	7.97	0.74	1.0	0.8	256	454589
101	20	5.1	5.1	7.76	0.68	1.7	1.4	412	121201
102	20	5.1	5.1	7.57	0.74	1.0	0.8	120	415678
102	20	5.1	5.1	7.56	0.74	1.1	1.8	122	124598
101	20	5.1	5.1	7.36	0.62	0.9	0.6	77	121205
100	20	5.0	5.1	7.28	0.55	1.3	1.8	123	235104
102	20	5.1	5.1	7.24	0.74	1.1	1.8	115	107895
100	20	5.0	5.1	7.07	0.55	0.8	0.6	117	104589
99	20	5.0	5.1	6.96	0.51	1.5	1.8	293	084569
100	20	5.0	5.1	6.88	0.55	1.3	1.8	348	087412
100	20	5.0	5.1	6.88	0.55	0.9	0.6	387	089874
98	20	4.9	5.1	6.85	0.47	1.1	2.9	765	064574
98	20	4.9	5.0	6.81	0.47	0.8	2.9	101	054567
96	20	4.8	5.0	6.81	0.46	1.5	1.1	363	126598
101	20	5.1	5.0	6.80	0.62	0.9	0.8	630	126366
101	20	5.1	5.0	6.79	0.62	0.9	0.7	113	145556
100	20	5.0	5.0	6.76	0.55	1.1	1.2	274	012301
					......				
43	20	2.2	2.3	-3.10	0.35	1.4	1.7	460	601003
30	20	2.0	2.3	-3.19	0.33	2.6	3.4	782	601203
36	20	1.8	2.2	-3.27	0.33	0.2	0.2	455	601245
36	20	1.8	2.2	-3.35	0.33	1.5	1.2	577	604545
37	20	1.9	2.1	-3.47	0.33	1.5	1.3	733	804587
34	20	1.7	22.1	-3.63	0.34	0.5	0.4	449	802456
32	20	1.6	1.7	-4.32	0.35	1.9	1.6	569	801489
22	20	1.1	1.1	-6.53	0.71	0.7	0.3	507	894512
Obsvd Score	Obsvd Count	Obsvd Average	Fair Average	Measure	Model S.E.	Infit MnSq	Outfit MnSq	Num	Students
77.3	20	3.9	3.9	2.76	0.44	1.0	0.9	Mean(Count: 867)	
12.7	0	0.6	0.6	2.26	0.07	0.8	0.9	S.D.	
RMSE	0.44	Adj S. D.	2.22	Separation	5.03	Reliability	0.96		
Fixed (all same)			Chi-square:24,749.6		d.f.:858	.	Significance: 0.00		
Random (normal)			Chi-square:854.1		d.f.:857		Significance: 0.52		

Note: Maximum score is 4 × 26 = 104, minimum score is 4 × 5 = 20.

Table 2 reports ability estimates for student gender groups, standard errors, infit and outfit statistics, and summary statistics. Female students have higher writing ability than males. The reliability of the gender separation is 1.00 with separation 20.73. The chi-square statistic, $\chi^2 = 861.0$ with df=1, $p < 0.001$, indicates that the difference between female and male students in writing ability is significant.

TABLE 2 GENDER MEASUREMENT REPORT

Obsvd Score	Obsvd Count	Obsvd Aveage	Fair Avrge	Measure	Model S.E.	Infit MnSq	Outfit MnSq	N	Gender
33500	8420	4.0	4.1	3.43	0.02	2.3	3.1	2	Female
32885	8760	3.8	3.8	2.55	0.02	2.3	3.1	1	Male
33192.5	8590	3.9	3.9	2.99	0.02	2.8	3.4	Mean (count:2)	
307.5	170	0.1	0.1	0.44	0.00	0.5	0.3	S.D.	
RMSE	0.02	Adj S.D.	0.4	Separation:	20.73	Reliability	1.00		
Fixed	(all same)	Chi-square:	861.1	d.f.: 1		Significance	0.00		

Table 3 reports ability estimates for student ethnic groups, standard errors, infit and outfit statistics, and summary statistics. White students have the highest writing abilities. Hispanic students have the lowest writing ability. Black students are in between, but closer to Hispanic students than white students. The reliability of the ethnicity separation is 1.00 with separation 25.05. The chi-square statistic, $\chi^2 = 4021.1$ with df=2, $p < .001$, indicates that the differences among ethnic groups are significant.

TABLE 3 ETHNICITY MEASUREMENT REPORT

Obsvd Score	Obsvd Count	Obsvd Aveage	Fair Avrge	Measure	Model S.E.	Infit MnSq	Outfit MnSq	N	Ethnicity
53332	13360	4.0	4.1	3.45	0.02	2.3	2.8	1	White
9943	2900	3.4	3.5	1.15	0.04	3.0	3.0	2	Black
3110	920	3.4	3.5	0.86	0.07	3.3	3.2	3	Hispanic
22128.3	5726.7	3.6	3.7	1.82	0.04	2.8	3.0	Mean	(Count:3)
22240.0	5457.8	0.3	0.3	1.16	0.02	0.5	0.2	S.D.	
RMSE	0.05	Adj S.D.	25.1	Separation:	25.05	Reliability	1.00		
Fixed	(all same)	Chi-square:	4021.1	d.f.: 1		Significance	0.00		
Random	(normal)	Chi-square:	2.0	d.f.: 1		Significance	0.16		

DFF Analysis for raters

Different rater functioning (DRF) refers to a situation where individual students with the same underlying ability level have an unequal probability of obtaining the same level of ratings by the raters because of their group membership. Thus, a rater who has bias will favor or disfavor one particular student group compared to another group when rating students' essays. When topic responses are scored by raters who know the identity of each respondent or who can guess the respondent's gender or ethnicity, rater bias may occur. If respondents tend to receive higher scores from raters of their own race, then respondents who are scored by same-race raters may have an unfair advantage.

Table 4 reports the results of facet interactions between individual raters and students' ethnicities. The three panels present the conventional bias analysis, the Rasch bias analysis and measures for raters and student ethnic groups, respectively. The first panel reports the conventional statistics, including raters' observed scores, expected scores and counts of ratings, as well as the difference between the observed scores and the expected scores, which is obtained by subtracting expected scores from the observed count and dividing by the observed count. The second panel reports the Rasch bias analysis, including the magnitude of bias estimates in log odds units, the standard errors of the bias estimates, and z-score which is a standardized bias, respectively. Two directions of bias are reported in this table: the negative values of bias estimates indicate bias against student groups, and the positive values indicate bias for student groups. A criterion, z-score = 2, is selected (about p<.05 at the significant level) for this study. The last panel indicates rater ID, raters' demographic information, severity levels, as well as measures of student ethnic groups.

This DRF detection analyzed a total of 267 possible interactions between 89 individual raters and 3 student ethnic groups for these data. We found 29 significant rater biases and 238 insignificant interactions. These significant rater biases account for 11 percent of the total interactions. This implies that most raters in this study do not show any bias to student ethnicity. Furthermore, most of these significant raters' biases, because they are small, do not affect individual students' measures.

In order to analyze all significant rater biases and see whether they have consistent patterns, raters were divided into six groups by gender and ethnicity, reported following raters' ID as MW, FW and FB. MW represents male white raters, FW represents female white raters, and FB represents female black raters. Because only five black male raters were involved in the study and they did not show any significant bias with respect to students or topics, these black male raters were not reported in this table. The first two blocks consist of white male raters. Among them, four raters disfavored white students, while two others disfavored black students. In the second blocks, six white male raters favored white students, while only one white male favored black students. In the third block, three white female raters disfavored white students, while one disfavored black students. The fourth block shows that two white female raters favored white students and one favored black students. The fifth and sixth blocks show black female raters' bias for and against student ethnic groups.

TABLE 4 RATER BIAS TO STUDENT ETHNIC GROUPS

Observed Score	Expected Score	Count	Diff. Obs-Exp Average	Rater Bias Measure	S.E.	Z-Score	Rater ID	Rater Measr	Student Eth	Measr
1032	1094.0	300	-0.21	-0.93	0.12	-7.7	52MW	0.58	W	3.45
466	494.5	120	-0.24	-0.82	0.18	-4.6	76MW	-0.38	W	3.45
2807	2934.9	815	-0.16	-0.72	0.07	-9.7	46MW	0.53	W	3.45
839	876.6	220	-0.17	-0.64	0.14	-4.7	64MW	0.42	W	3.45
327	349.2	110	-0.20	-0.72	0.18	-4.1	50MW	0.55	B	1.15
519	542.6	155	-0.15	-0.67	0.16	-4.1	60MW	0.45	B	1.15
614	600.5	150	0.09	0.30	0.15	2.0	67MW	0.14	W	3.45
6255	6110.4	1540	0.09	0.32	0.05	6.9	60MW	0.45	W	3.45
1076	1050.0	260	0.10	0.33	0.11	3.0	39MW	0.01	W	3.45
474	455.5	115	0.16	0.54	0.17	3.2	84MW	0.18	W	3.45
658	629.9	160	0.18	0.59	0.14	4.2	89MW	0.17	W	3.45
5152	4936.6	1255	0.17	0.60	0.05	11.6	33MW	0.03	W	3.45
370	331.7	100	0.38	1.77	0.22	8.2	49MW	0.30	B	1.15
634	662.4	155	-0.18	-0.57	0.14	-4.0	31FW	-0.53	W	3.45
913	949.0	225	-0.16	-0.51	0.12	-4.2	74FW	-0.45	W	3.45
3636	3774.2	900	-0.15	-0.49	0.06	-8.1	71FW	-0.42	W	3.45
429	448.1	120	-0.16	-0.72	0.20	-3.7	23FW	-0.44	B	1.15
556	532.9	130	0.18	0.60	0.16	3.7	41FW	-0.43	W	3.45
957	895.4	225	0.27	0.91	0.12	7.6	38FW	0.03	W	3.45
660	608.9	160	0.32	1.19	0.15	8.2	34FW	0.03	W	3.45
567	520.6	130	0.36	1.19	0.16	7.5	42FW	-0.03	W	3.45
939	833.1	205	0.52	1.57	0.12	12.7	70FW	0.31	W	3.45
793	759.3	205	0.16	0.69	0.14	5.0	71FW	-0.42	B	1.15
607	687.8	165	-0.49	-1.84	0.17	-11.0	77FB	-0.29	W	3.45
1455	1522.9	360	-0.19	-0.59	0.10	-6.2	54FB	-0.16	W	3.45
416	430.9	120	-0.12	-0.58	0.20	-3.0	82FB	0.27	B	1.15
2483	2432.0	615	0.08	0.29	0.07	3.9	82FB	0.27	W	3.45
637	581.0	150	0.37	1.25	0.14	8.8	69FB	0.48	W	3.45
392	379.8	110	0.11	0.51	0.21	2.5	82FB	0.27	H	0.86
Observed Score	Expected Score	Count	Diff. Obs-Exp Average	Rater Bias Measure	S.E.	Z-Score	Rater ID	Rater Measr	Student Eth	Measr
1264.2	1256.0	319.8	0.03	0.09	0.13	0.7	Total Interaction: 267			
979.5	977.6	245.9	0.21	0.76	0.03	6.7				

In order to test if there are any significant differences between the bias for and against student ethnic groups, chi-square statistics were conducted for each group of raters. Two 2x2 crosstables for observed counts of white male raters and expected counts of white male raters were constructed in Table 5.

TABLE 5
TESTS OF BIAS OF WHITE MALE RATERS TO STUDENT ETHNIC GROUPS

(A). OBSERVED COUNTS

	Favor	Disfavor	Total
White Students	6	4	10
Black Students	1	2	3
Total	7	6	13

(B). EXPECTED COUNTS

	Favor	Disfavor	Total
White Students	5.38	4.62	10
Black Students	1.62	1.38	3
Total	7	6	13

Chi-square statistics were conducted based on the following formula,

$$\chi^2 = \Sigma \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \quad (4)$$

For the data, the result is,

$$\chi^2 = \Sigma \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

$$= \frac{(6-5.38)^2}{5.38} + \frac{(4-4.62)^2}{4.62} + \frac{(1-1.62)^2}{1.62} + \frac{(2-1.38)^2}{1.38}$$

$$= 0.67$$

The resulting $\chi^2 = 0.67$ with df=1 and $p > .10$, implies that these potentially biased white male raters did not show significant bias against some student group and for another group. The same chi-square statistics were conducted to test white female and black female rater groups. No significant differences were found between raters' bias against and for groups of students.

Table 6 shows significant raters' biases to student gender groups. There were 26 significant biases among 178 possible interactions. Only 3 raters, whose bias sizes are greater than 1 logit, need to recheck their ratings or regrade essays. The same chi-square procedures were conducted to test if these raters significantly favor one gender group and disfavor another group. These results also did not reveal any significant differences.

TABLE 6 RATER BIAS TO STUDENT GENDER GROUPS

			Diff.	Rater						
Observed	Expected		Obs-Exp	Bias			Rater	Rater	Student	
Score	Score	Count	Average	Measure	S.E.	Z-Score	ID	Measr	Gen	Measr
334	385.1	110	-0.46	-1.87	0.18	-10.3	51MW	0.41	F	3.43
2224	2353.7	670	-0.19	-0.81	0.08	-10.4	46MW	0.53	F	3.43
813	855.6	235	-0.18	-0.80	0.14	- 5.8	52MW	0.58	F	3.43
427	451.0	110	-0.22	-0.75	0.18	- 4.1	64MW	0.42	F	3.43
624	656.4	175	-0.18	-0.72	0.15	- 4.7	39MW	0.01	M	2.55
398	412.8	105	-0.14	-0.53	0.19	- 2.7	76MW	-0.38	M	2.55
942	962.9	280	-0.07	-0.33	0.12	- 2.7	49MW	0.30	M	2.55
2775	2731.0	730	0.06	0.24	0.07	3.2	33MW	0.03	M	2.55
3842	3742.1	990	0.10	0.39	0.06	6.4	60MW	0.45	M	2.55
402	390.9	100	0.11	0.39	0.18	2.1	67MW	0.14	M	2.55
601	575.3	140	0.18	0.59	0.15	3.9	39MW	0.01	F	3.43
421	399.3	100	0.22	0.69	0.17	3.9	89MW	0.17	F	3.43
988	930.2	260	0.22	0.93	0.12	7.6	49MW	0.30	F	3.43
1539	1606.9	400	-0.17	-0.60	0.10	-6.2	23FW	-0.44	M	2.55
2198	2226.8	560	-0.05	-0.18	0.08	-2.3	71FW	-0.42	M	2.55
708	737.3	175	-0.17	-0.54	0.14	-3.9	74FW	-0.45	F	3.43
1639	1697.6	405	-0.14	-0.48	0.09	-5.3	23FW	-0.44	F	3.43
451	438.7	105	0.12	0.35	0.17	2.1	70FW	0.31	F	3.43
469	447.3	110	0.20	0.68	0.18	3.8	41FW	-0.43	F	3.43
486	447.4	110	0.35	1.14	0.17	6.7	38FW	0.03	F	3.43
388	376.6	100	0.11	0.45	0.19	2.3	40FW	0.10	M	2.55
450	433.4	110	0.15	0.51	0.17	3.0	74FW	-0.45	M	2.55
525	501.4	135	0.17	0.70	0.17	4.2	37FW	-0.07	M	2.55
390	435.4	105	-0.43	-1.56	0.20	-7.9	77FB	-0.29	F	3.43
770	816.9	205	-0.23	-0.86	0.14	-6.0	54FB	-0.16	M	2.55
1788	1754.3	445	0.08	0.26	0.09	3.0	82FB	0.27	F	3.43
			Diff.	Rater						
Observed	Expected		Obs-Exp	Bias			Rater	Rater	Student	
Score	Score		Average	Measure	S.E.	Z-Score	ID	Measr	Gen	Measr
1022.8	1029.5	268.1	0.00	-0.10	0.10	-0.77	Total Interaction: 178			
680.4	685.1	179.6	0.20	0.70	0.00	4.78				

Table 7 shows 10 raters with 15 biases against and/or for topics. The first two blocks show that two white male raters have bias against the topic "*space*," but other two raters have bias for the topic "*change*." The third block shows two white female raters have bias against the topic "*invention*." The fifth block shows that two black female raters have bias against the topic "*forget*." This table shows that raters in different ethnic groups have differential topic bias. In addition, this table shows 50% of these biased raters (5 out of 10) have bias to more than one topic. White male rater 49 has bias both against the topic "*trade places*" and for the topic "*forget*." White male rater 33 has bias for both topics "*change*" and "*trade places*." Black female rater 54 has bias against both topics "*change*" and "*forget*." These raters above-mentioned may need more training.

TABLE 7 BIAS OF INDIVIDUAL RATERS TO TOPICS

```
-----------------------------------------------------------------------------------
|                           Diff.  |Rater                   |                      |
|Observed Expected          Obs-Exp|Bias                    |Rater  Rater Topic    |
|Score    Score     Count  Average|Measure  S.E.  Z-Score|ID     Measr Name    Measr|
-----------------------------------------------------------------------------------
|  585     637.6    185    -0.28 |-1.17   0.14    -8.1 |52MW   0.58  space  -0.07|
| 1545    1640.5    475    -0.20 |-0.84   0.09    -9.2 |46MW   0.53  space  -0.07|
|  495     518.1    150    -0.15 |-0.66   0.17    -4.0 |49MW   0.30  trade   0.27|
-----------------------------------------------------------------------------------
| 2221    2172.6    550     0.09 | 0.30   0.08     3.9 |60MW   0.45  change -0.05|
| 1817    1752.8    445     0.14 | 0.49   0.09     5.7 |60MW   0.45  invent  0.04|
|  929     896.0    250     0.13 | 0.58   0.13     4.5 |33MW   0.03  trade   0.27|
| 1220    1145.2    280     0.27 | 0.84   0.11     8.0 |33MW   0.03  change -0.05|
|  832     781.2    225     0.23 | 0.98   0.14     7.2 |49MW   0.30  forget -0.19|
-----------------------------------------------------------------------------------
|  713     757.1    180    -0.25 |-0.80   0.14    -5.7 |71FW  -0.42  invent  0.04|
|  775     819.3    195    -0.23 |-0.74   0.13    -5.6 |23FW  -0.44  invent  0.04|
-----------------------------------------------------------------------------------
|  392     381.7    100     0.10 | 0.39   0.19     2.0 |37FW  -0.07  forget -0.19|
-----------------------------------------------------------------------------------
|  578     622.6    150    -0.30 |-1.02   0.16    -6.5 |54FB  -0.16  change -0.05|
|  970     993.1    260    -0.09 |-0.35   0.13    -2.8 |82FB   0.27  forget -0.19|
|  525     537.9    130    -0.10 |-0.32   0.16    -2.0 |54FB  -0.16  forget -0.19|
-----------------------------------------------------------------------------------
|  869     829.0    215     0.19 |-0.66   0.12    -5.3 |82FB   0.27  space  -0.07|
-----------------------------------------------------------------------------------
|                           Diff.  |Rater                   |                      |
|Observed Expected          Obs-Exp|Bias                    |Rater  Rater Topic    |
|Score    Score     Count  Average|Measure  S.E.  Z-Score|ID     Measr Name    Measr|
-----------------------------------------------------------------------------------
| 940.4   965.7   252.7    -0.03 |-0.20    0.13    -1.2|Mean                      |
| 393.5   383.5    99.6     0.18 | 0.64    0.02     5.2|S.D.                      |
-----------------------------------------------------------------------------------
```

Table 8 summarizes results of rater bias detection based on the previous three tables. In this table, -B represents bias against some groups of students or topics, +B represents bias for some groups of students or topics. Because previous chi-square statistics do not reveal significant differences between raters' bias against and for student groups, bias analysis here does not differentiate the direction of bias. All rater groups reveal consistent patterns of bias: more raters have bias for or against white students than black students, more raters have bias for or against female students than male students. Although raters of different ethnic groups have bias to different topics, the topic "*forget*" received more bias than other topics.

TABLE 8 SUMMARY OF RATER BIAS DETECTION

Rater	Bias	Student Subgroups				Topic				
Background	Type	Male	Female	White	Black	T	S	C	I	F
	-B	3	4	4	2	1	2	0	0	0
White Male	A	35	34	31	38	37	39	39	40	40
	+B	3	3	6	1	1	0	2	1	1
% Biased Raters		15%	17%	24%	7%	5%	5%	5%	2%	2%
	-B	2	2	3	1	0	0	0	2	0
White Female	A	28	28	25	31	33	33	33	31	32
	+B	3	3	5	1	0	0	0	0	1
% Biased Raters		15%	15%	24%	6%	0%	0%	0%	6%	3%
	-B	1	1	2	1	0	0	1	0	2
Black Female	A	9	8	6	9	10	9	9	10	8
	+B	0	1	2	0	0	1	0	0	0
% Biased Raters		10%	20%	40%	10%	0%	10%	10%	0%	20%
	-B	0	0	0	0	0	0	0	0	0
Black Male	A	5	5	5	5	5	5	5	5	5
	+B	0	0	0	0	0	0	0	0	0
% Biased Raters		0%	0%	0%	0%	0%	0%	0%	0%	0%
Total		13%	16%	25%	7%	2%	3%	3%	3%	4%

N.B. -B indicates bias against some groups of students or topics;
+B indicates bias favor of some groups of students or topics;
A indicates no bias.

Generally, DRF detection identifies individual rater bias for some student groups and topics. The DRF functioning from the FACET analysis conceptualizes raters' behavior into the framework of the many-faceted Rasch response model. This approach provides both a sound theoretical basis and a practical way to detect rater bias.

DFF Analysis for Topic Types

Analysis for topic type DFF examines whether topic types are functioning differentially for different student groups. Table 9 reports the analysis between topic types and student ethnic groups. Three panels report different information. The first panel reports the observed scores, the expected scores, and the difference between observed and expected scores. The second panel reports the magnitude of bias measures, their standard errors, and their z-scores in a normal distribution to examine the significance level of the bias respectively. The third panel reports the topic types and calibrations as well as ethnic groups and group measures. Three blocks report performances of student ethnic groups in the three topic types: persuasive, expository and narrative.

Among the total of 9 interactions between topic types and student ethnicities, only 2 interactions are significant. In persuasive writing, a significant bias is shown against black students, and in narrative writing a significant bias is shown for black students. The magnitude of persuasive bias for black students is 0.31 with significance level at p < 0.001, z = 4.2. The magnitude of narrative bias for black students is 0.18 with significance level at p < 0.01, z = 3.0. Because these magnitudes are very small, these biases do not affect individual student ability estimates. No other bias is found against or for other student ethnic groups.

TABLE 9 TOPIC TYPE BIAS TO STUDENT ETHNIC GROUPS

		Diff.				Topic	Topic		
Observed	Expected	Obs-Exp	Bias+	Model		Type	Type	Student	
Score	Score	Average	Measure	S.E.	Z-Score	Name	Measure	Ethnicity	Measure
2425	2481.0	-0.08	-0.31	0.07	-4.2	PER	-0.39	Black	1.15
19267	19223.0	0.01	0.03	0.03	1.2	PER	-0.39	White	3.45
1074	1062.3	0.04	0.15	0.11	1.3	PER	-0.39	Hispanic	0.86
18020	18.3	0.00	-0.01	0.03	-0.3	EXP	-0.61	White	3.45
3586	3578.3	0.01	0.03	0.06	0.5	EXP	-0.61	Black	1.15
1095	1092.6	0.01	0.03	0.11	0.3	EXP	-0.61	Hispanic	0.86
941	955.1	-0.05	-0.21	0.12	-1.7	NAR	-0.91	Hispanic	0.86
16045	16078.9	-0.05	-0.03	0.03	-1.0	NAR	-0.91	White	3.45
3932	3883.7	0.04	0.18	0.06	3.0	NAR	-0.91	Black	1.15
7376.1	7376.1	0.00	0.01	0.07	0.1	Mean	(Count:9)		
7462.2	7459.2	0.04	0.15	0.04	1.9	S.D.			

Table 10 reports the analysis between topic types and student gender groups. The structure of this table is the same as Table 9. Among the total of 6 interactions, 4 significant biases are identified. In persuasive writing, biases are shown for males and against females at significance level p < 0.05. In narrative writing, biases are shown for females and against males at significance level p < 0.05. This means that female students did worse in persuasive writing and better in narrative than expected, while male students did better in persuasive and worse in narrative. Because the magnitudes of the biases are below 0.10 logit, these biases are too small to affect individual student ability estimates. The results may be explained better as gender impact than gender bias. Further study is needed to distinguish between gender difference (or gender impacts) and gender bias in direct writing assessment.

TABLE 10 TOPIC TYPE BIAS TO STUDENT GENDER GROUPS

		Diff.				Topic	Topic		
Observed	Expected	Obs-Exp	Bias+	Model		Type	Type	Student	
Score	Score	Average	Measure	S.E.	Z-Score	Name	Measure	Gender	Measure
9902	9966.9	-0.03	-0.09	0.04	-2.4	PER	-0.39	Female	3.43
12864	12799.5	0.02	0.07	0.03	2.2	PER	-0.39	Male	2.55
13150	13143.7	0.00	0.01	0.03	0.2	EXP	-0.61	Female	3.43
9551	9557.3	0.00	-0.01	0.04	-0.2	EXP	-0.61	Male	2.55
10448	10389.4	0.02	0.08	0.04	2.1	NAR	-0.91	Female	3.43
10470	10528.3	-0.02	-0.08	0.04	-2.1	NAR	-0.91	Male	2.55
11064.2	11064.2	0.00	0.00	0.04	0.0	Mean	(count:9)		
1412.0	1387.6	0.02	0.07	0.00	1.8	S.D.			

Table 11 reports the analysis between topic types and student grade groups. In persuasive writing, bias is found for the grade 8 students at significance level p < .05. In expository writing, biases are shown for the grade 8 students and against the grade 6 students. In narrative writing, biases are shown for the grade 6 students and against grade 8 students. Although 5 interactions are significant among the total of 6 interactions, all the magnitudes of biases are less than 0.5 logit. Therefore, these biases do not affect individual student ability estimates.

TABLE 11 TOPIC TYPE BIAS TO STUDENT GRADE GROUPS

		Diff.				Topic	Topic		
Observed	Expected	Obs-Exp	Bias+	Model		Type	Type	Student	
Score	Score	Average	Measure	S.E.	Z-Score	Name	Measure	Grade	Measure
5315	5351.6	-0.02	-0.10	0.05	-1.9	PER	-0.39	G6	1.55
11622	11645.9	-0.01	-0.03	0.03	-0.8	PER	-0.39	G10	3.56
5829	5768.9	0.04	0.15	0.05	3.0	PER	-0.39	G8	2.68
5507	5624.0	-0.07	0.05	0.05	-6.0	EXP	-0.61	G6	1.55
12280	12237.9	0.01	0.22	0.03	1.4	EXP	-0.61	G10	3.56
4914	4839.0	0.06	0.03	0.05	4.1	EXP	-0.61	G8	2.68
5172	5018.4	0.11	0.45	0.05	8.4	NAR	-0.91	G6	2.68
10361	10379.2	-0.01	-0.02	0.04	-0.7	NAR	-0.91	G10	3.56
5385	5520.1	-0.09	-0.37	0.05	-7.0	NAR	-0.91	G8	1.55
7376.1	7376.1	0.00	-0.01	0.05	-0.1	Mean	(Count:9)		
2906	2907.4	0.06	0.24	0.01	4.6	S.D.			

Table 12 summarizes bias statistics of topic types for students' ethnic, gender and grade backgrounds. In this table, +B indicates bias for student subgroups and -B indicates bias against. Regarding student subgroups, no bias is shown to white, Hispanic and the grade 10 students. Regarding topic types, no bias is shown in expository writing with respect to student gender and ethnic groups. Regarding gender, male students did better in persuasive writing and worse in narrative writing than expected, while female students did better in narrative writing and worse in persuasive writing than expected. Regarding student ethnic backgrounds, black students did better in narrative writing but worse in persuasive writing. Regarding different graders, the grade 6 students did better in narrative but worse in expository than expected, while the grade 8 students did better in persuasive and expository, but worse in narrative than expected. Because all these biases from topic types are less than 0.5 logit, none of them affect students' writing ability measures.

TABLE 12
SUMMARY OF BIAS DETECTION FOR TOPIC TYPE

Topic	Student Subgroups					Grade		
Type	Male	Female	White	Black	Hispanic	6	8	10
Persuasive	+B	-B		-B			+B	
Expository						-B	+B	
Narrative	-B	+B		+B		+B	-B	

N.B. -B indicates bias against a student subgroup;
+B indicates bias for a student subgroup.

DFF statistics for topic types and student demographic backgrounds provide a convenient means to examine effects of performance of topic types on students subgroups. This information from DFF detection is helpful for understanding different characteristics of subgroups of students and for constructing a fair and valid direct writing assessment.

Conclusion and Discussion

As direct writing assessment or other performance assessment grows in popularity, it will be increasingly important to monitor the validity and fairness of topic, item and raters' behavior (Zwick, Donoglue, & Grima, 1993). DFF detection procedures, as one component of this evaluation, can be helpful in investigating the effect on student groups of the introduction of topic, item and raters.

This study proposes procedures for defining different interaction models and detecting biases with a many-faceted Rasch (FACETS) model and illustrates these procedures using data from a large-scale performance assessment of writing. With the FACETS model, DFF analysis for rater identifies biased raters. Evidence is also found that these raters' bias effects students' writing ability estimates. Also, DFF statistics for topic types and student demography show effects of performance of topic types on student subgroups and provide evidence of gender and age impacts on different topic types.

Any kind of performance assessment must have DFF examination and identification. The FACETS model, because of its advantages in defining interaction models and flexible use in many situations, offers a potent approach to DFF identification in a wide variety of performance assessment.

REFERENCES

W. H. Angoff (1993). Perspectives on Differential Item Functioning methodology. In P. W. Holland & H. Wainer (Eds.). Differential Item Functioning (pp. 1-24). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Breland, H., & Jones, R. J. (1984). Perceptions of writing skills. Written Communication, 1, 101-109.

Brennan, R. L. (1992). Elements of Generalizability Theory. (Rev. ed.). American College Testing Program, Iowa City, Iowa.

Coffman, W. E. (1977b). On the reliability of ratings of essay examinations in English. Research in the Testing of English, 5.

Coffman, W. E. (1966). On the validity of essay tests of achievement. Journal of Educational Measurement, 3, 151-156.

Cooper, P. L. (1984). The assessment of writing ability: A review of research. (GRE Board Research Report No. 82-15R). Princeton, NJ: Educational Testing Service.

Du, Y. (1995). DIF adjustment. Rasch Measurement SIG Newsletter, 9, 414.

Engelhard, J. G. (1992a). The measurement of writing ability with a many-faceted Rasch model. Applied Measurement in Education, 5, 171-191.

Engelhard, J. G., Gordon, B., & Gabrielson, S. (1992). The influences of model of discourse, experiential demand and gender on quality of student writing. Research in the Teaching of English, 26, 315-336.

Guilford, J. P. (1954). Psychometric Methods. New York: McGraw-Hill Book Company.

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.). Test Validity (pp. 129-125). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Holland, P. W. & Wainer, H. (Eds.). (Eds.). (1993). Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Holzbach, R. L. (1978). Rater bias in performance rating: superior, self-, and peer ratings. Journal of Applied Psychology. 63, 579-588.

Houston, W. M., Raymond, M.R. & Svec, J. C. (1991). Adjustments for rater effects in performance assessment. Applied Psychological Measurement. 15, 409-421.

Illinois State Board of Education, (1992). Write On, Illinois!

Linacre, J. M. (1988). FACETS, Chicago: MESA Press.

Linacre, J. M. (1989). Many-facet Rasch Measurement. Chicago, IL: MESA Press.

Linacre, J. M. (1994a). Constructing measurement with a many-facet Rasch model. In Wilson M. (Ed.), Objective Measurement, Theory into Practice, 2, (pp. 129-144). Norwood, NJ : Ablex Publishing Co.

Linacre, J. M. (1994b). Measurement of judgment. International Encyclopedia of Education (2nd. ed.) Oxford: Pergamon Press.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley Publishing Company.

Rasch, G. (1960/1990). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 4, 321-333.

Wilson, M. (1992). (Ed.), Objective Measurement, Theory into Practice. 1. Norwood, New Jersey: Ablex Publishing Corporation.

Wilson, M. (1994). (Ed.), Objective Measurement, Theory into Practice. 2. Norwood, New Jersey: Ablex Publishing Corporation.

Wright, B. D. (1967). Sample-free test calibration and person measurement. Invitational conference on Testing Problems, ETS, Princeton, New Jersey.

Wright, B. D. (1968). Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service.

Wright, B. D., & Panchapakesan, N. A. (1969). A procedure for sample-free item analysis. Educational and Psychological Measurement, 29, 23-48.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.

Wright, B. D., & Douglas, G. A. (1977). Best procedures for sample-free item analysis. Applied Psychological Measurement, 1, 281-295.

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

Wright, B. D. (1980). Afterword. In G. Rasch (Ed.), Probabilistic models for some intelligence and attainment tests (pp. 186-195). Chicago: University of Chicago Press.

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

Wright, B. D. (1985). Additivity in psychological measurement. In E. Roskam (Ed.), Measurement and Personality assessment (pp. 101-112). Amsterdam: North-Holland.

Wright, B. D. & Douglas, G. A. (1986). The rating scale model for objective measurement. (Memorandum No. 35). MESA Psychometric Lab. Chicago:MESA Press.

Wright, B. D. (1988). The efficacy of unconditional maximum likelihood bias correction. Applied Psychological Measurement, 12, 315-318.

Wright, B. D. (1989). Deducing the Rasch model from the traditional requirement that counting right answers be sufficient. Rasch Measurement SIG Newsletter, 3.

Zwick, R., Donoghue, J. R., & Grima A. (1993). Assessment of differential item functioning for performance tasks. Journal of Educational Measurement, 30, 233-252.

TM025566

AERA April 8-12, 1996

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: DIFFERENTIAL FACET FUNCTIONING DETECTION IN DIRECT WRITING ASSESSMENT

Author(s): Yi Du, Benjamin D. Wright, William L. Brown

Corporate Source: MINNEAPOLIS PUBLIC SCHOOLS

Publication Date: N/A

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☑ ← **Sample sticker to be affixed to document**

Check here
Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document → ☑

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

or here
Permitting reproduction in other than paper copy.

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: Yi Du

Position: Testing & Evaluation Specialist

Printed Name: Yi DU

Organization: Minneapolis Public Schools

Address:
807 NE Broadway
Minneapolis Public Schools
Minneapolis, MN 55413

Telephone Number: (612) 627-2195

Date: 4-16-96