Running head: MULTI-FACET RASCH ANALYSIS OF ESSAY SCORING

A Multi-Facet Rasch Analysis of a Career Commitment Essay Scoring Process

Susan Gracia

Rhode Island College

Paper presented at the at the American Educational Research Association 2008 Annual Meeting, New York, NY.

Abstract

NCATE requires teacher education programs to implement assessment systems in collaboration with the professional community, conduct thorough studies to establish the reliability of its performance assessment procedures, and make changes in practice consistent with the results of these studies. In an effort to evaluate the reliability of collaborative performance assessment procedures used in the scoring of Career Commitment Essays used for teacher education program admissions decisions, essay data from faculty and practitioner evaluations of almost 500 teacher education program applicants were analyzed in this study. A Many-Facet Rasch Measurement (MFRM) approach was used to investigate the consistency of college faculty and PK-12 practitioner ratings, differences in levels of severity of ratings, the functioning of the essay rubric, and other key issues.

A Multi-Facet Rasch Analysis of a Career Commitment Essay Scoring Process

In 2002, the National Council for Accreditation in Teacher Education (NCATE) required that programs seeking NCATE accreditation develop and implement assessment systems that evaluate teacher candidates' knowledge, skills, and dispositions. Among other things, NCATE standards require that:

- Unit faculty collaborate with members of the professional community to design and implement the assessment system.
- Decisions about candidate performance are based on multiple assessments
- The unit conducts thorough studies to establish validity and reliability of its performance assessment procedures.
- The unit makes changes in its practices consistent with the results of these studies.

In response to these requirements, an educational institution in the Northeast U.S. designed a teacher education program Admissions Portfolio consisting of multiple assessments and work samples to be evaluated for admission into a teacher education program. Required components of the Admissions Portfolio include, among others: passing scores on the Praxis, a minimum grade point average, completion of certain courses, reference forms, a passing score on a Career Commitment Essay.

The Career Commitment Essay is the subject of this study. Applicants are instructed to prepare a well-organized, focused, two to three page essay describing why they want to become a teacher and the personal characteristics and skills they would bring to a career in teaching. Specifically, they are required to:

- Discuss their reasons for applying to a specific teacher preparation program and their commitment to teaching as a career.
- Use examples of their specific experiences which are related to a potential career in teaching to discuss their beliefs about the following:
  - Individual and cultural diversity
  - The potential for all children to learn
  - Professional collaboration
  - Teacher as lifelong learner
- Reflect on one of the above areas in which they need to alter or improve to become an effective teacher. Using specific example, candidates must explain how their attitudes or behaviors need to change and why. They are also to discuss how they might begin to work toward this change.

The purpose of the present study is to evaluate the reliability of the performance assessment procedures used in the essay scoring process and the functioning of the scoring rubric. Essay data from faculty and practitioner evaluations of almost 500 teacher education program applicants over four scoring sessions were analyzed in this study. A Many-Facet Rasch Measurement approach was utilized to address the following research questions:

- Do the raters (both faculty and practitioner) differ in the levels of severity they exercise, or do both groups of raters function interchangeably?
- Do faculty and practitioners rate essays in the same manner?

- Are there any inconsistent raters whose patterns of ratings show little systematic relationship to the ratings that other raters give?
- Do some teacher education program applicants exhibit unusual profiles of ratings across rubric dimensions, receiving unexpectedly high (or low) ratings on certain dimensions, given the ratings the candidate received on other dimensions?
- Are there any raters who cannot effectively differentiate between rubric dimensions, giving each teacher education program applicant very similar ratings across a number of conceptually distinct dimensions?
- Are the categories on the rating scale for each rubric dimension appropriately ordered? Are the rating scales functioning properly? Are all the scale categories clearly distinguishable?

Methods

Sample

Scorers. The sample of essay scorers was comprised of 40 individuals. Seventeen were School of Education faculty members, and 23 were PK-12 practitioners (classroom teachers and administrators) from cities and towns close to the college. These scorers participated in one or more Career Commitment Essay scoring sessions in October 2006, December 2006, February 2007, and/or March 2007.

Table 1 below displays each scorer by identification code, the session(s) in which each scorer participated, the total number of times each person scored Career Commitment Essays, and the total number of scorers per scoring session.

Table 1: Scorers by Scoring Session

Scorer ID	Oct	Dec	Feb	March	Total
1002	X				1
1019				X	1
1024			X		1
1025			X		1
1061	X				1
1062	X				1
1064	X				1
1066	X				1
1067	X				1
1068		X			1
1069		X			1
1070		X			1
1072				X	1
1073				X	1
1003		X	X		2
1007			X	X	2
1010			X	X	2
1014			X	X	2
1015	X		X		2
1016			X	X	2
1017			X	X	2
1020			X	X	2
1021	X		X		2
1023			X	X	2
1063	X	X			2
1071		X		X	2
1001		X	X	X	3
1004	X		X	X	3
1006	X		X	X	3
1009	X		X	X	3
1011	X	X	X		3
1012	X		X	X	3
1013		X	X	X	3
1018		X	X	X	3
1022	X	X	X		3
1050		X	X	X	3
1065	X	X		X	3
1005	X	X	X	X	4
1008	X	X	X	X	4
1060	X	X	X	X	4
Totals	19	16	25	23	83

As displayed in

Table 1, the number of scorers per session ranged from 16 to 23, with an average of 20.75 scorers per scoring session. Thirty-five percent of scorers (n=14) participated just once. Slightly fewer (12 scorers, or 30%) participated in two scoring sessions. Eleven scorers (27.5%) were scorers in three sessions, and just three individuals served as scorers in all four sessions. The mean number of times any particular individual participated in essay scoring was 2.07 times. Higher education faculty and practitioners participated at almost the same rate, with an average of 2.00 sessions for practitioners versus 2.17 sessions for faculty members.

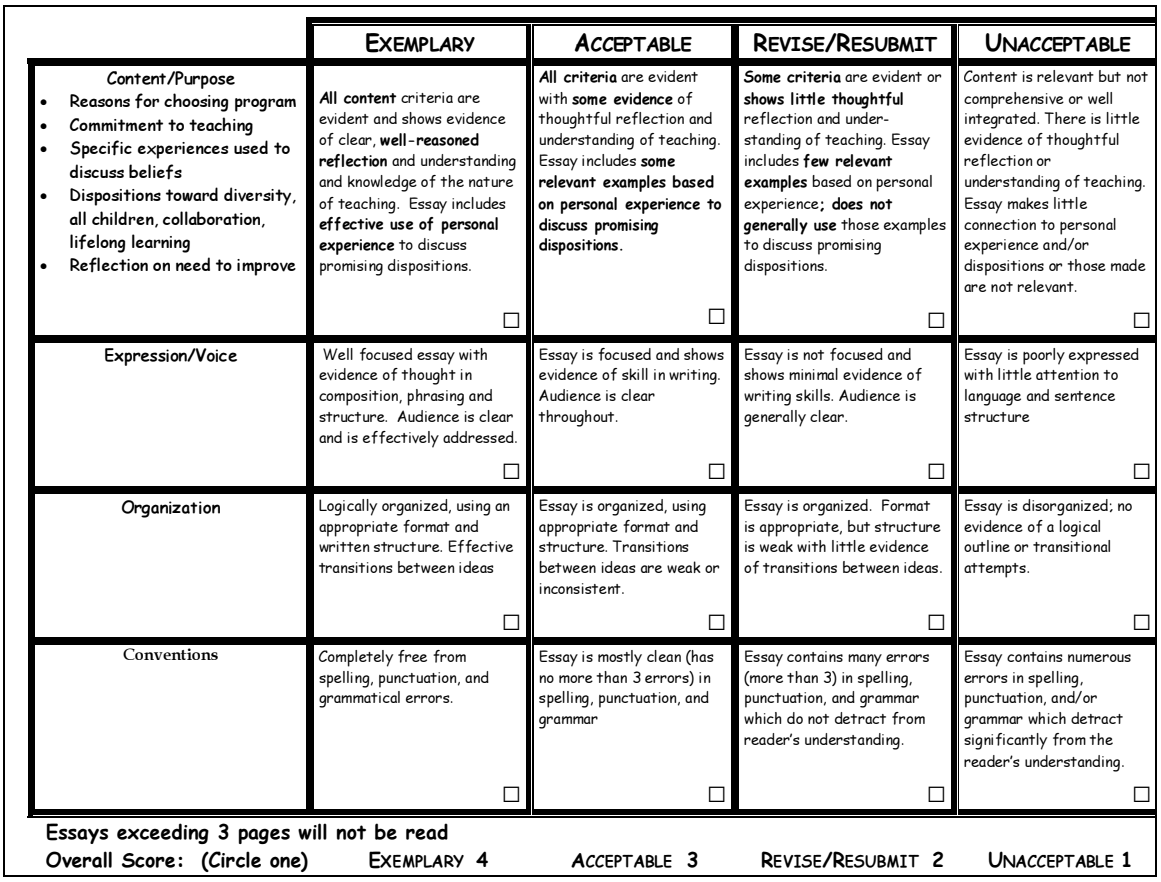
Students. Essays from 476 teacher candidates were analyzed for this study. All teacher candidates had completed at least 24 credits of college work, including one education course. Sixty-one students submitted a Career Commitment Essay more than one time over the four scoring sessions. Fifty-one students submitted essays in two different sessions; ten students submitted essays three times.

Over time, the Career Commitment Essay has evolved into a relatively "high stakes" assessment for teacher candidates, often holding many of them back from admission to the School of Education even though they meet other requirements for admission. In fact, 20% to 40% of teacher education applicants typically fail the Career Commitment Essay at any time point.

Procedures

Each scoring session took place on a Saturday morning from 8:30 am until 1 pm. The four sessions were conducted in October 2006, December 2006, February 2007, and March 2007. Each session began with an orientation to admissions to the School of Education, the Career Commitment Essay requirement, the task, and the rubric. This was followed by an examination and discussion of anchor papers. Then, scorers scored practice essays. Discussion of the correct scores for each essay, why the essay merited a particular score, and specific qualities of the essay ensued. After approximately two hours of orientation and calibration, participants began the actual scoring of real essays.

Essays are scored the four-point rubric displayed in Figure 1.



	EXEMPLARY	ACCEPTABLE	REVISE/RESUBMIT	UNACCEPTABLE
Content/Purpose • Reasons for choosing program • Commitment to teaching • Specific experiences used to discuss beliefs • Dispositions toward diversity, all children, collaboration, lifelong learning • Reflection on need to improve	**All content** criteria are evident and shows evidence of clear, **well-reasoned reflection** and understanding and knowledge of the nature of teaching. Essay includes **effective use of personal experience** to discuss promising dispositions. ☐	**All criteria** are evident with **some evidence** of thoughtful reflection and understanding of teaching. Essay includes **some relevant examples based on personal experience to discuss promising dispositions.** ☐	**Some criteria** are evident or **shows little thoughtful** reflection and under-standing of teaching. Essay includes **few relevant examples** based on personal experience; **does not generally use** those examples to discuss promising dispositions. ☐	Content is relevant but not comprehensive or well integrated. There is little evidence of thoughtful reflection or understanding of teaching. Essay makes little connection to personal experience and/or dispositions or those made are not relevant. ☐
Expression/Voice	Well focused essay with evidence of thought in composition, phrasing and structure. Audience is clear and is effectively addressed. ☐	Essay is focused and shows evidence of skill in writing. Audience is clear throughout. ☐	Essay is not focused and shows minimal evidence of writing skills. Audience is generally clear. ☐	Essay is poorly expressed with little attention to language and sentence structure ☐
Organization	Logically organized, using an appropriate format and written structure. Effective transitions between ideas ☐	Essay is organized, using appropriate format and structure. Transitions between ideas are weak or inconsistent. ☐	Essay is organized. Format is appropriate, but structure is weak with little evidence of transitions between ideas. ☐	Essay is disorganized; no evidence of a logical outline or transitional attempts. ☐
Conventions	Completely free from spelling, punctuation, and grammatical errors. ☐	Essay is mostly clean (has no more than 3 errors) in spelling, punctuation, and grammar ☐	Essay contains many errors (more than 3) in spelling, punctuation, and grammar which do not detract from reader's understanding. ☐	Essay contains numerous errors in spelling, punctuation, and/or grammar which detract significantly from the reader's understanding. ☐
Essays exceeding 3 pages will not be read				
Overall Score: (Circle one)	**EXEMPLARY 4**	**ACCEPTABLE 3**	**REVISE/RESUBMIT 2**	**UNACCEPTABLE 1**

Figure 1: Career Commitment Essay Scoring Rubric

Each essay was scored on content, expression/voice, organization, and conventions. Raters indicated their ratings on these four dimensions in the appropriate cell of the rubric. Additionally, raters assigned each essay an overall, holistic score. In scorer training, raters were informed that the Overall Score is not expected to be a simple average of the four trait scores. Rather, the Overall Score is meant to an independent, global, summary judgment of student performance in the essay.

Each essay was scored by at least two raters. Essays were randomly assigned by the Director of Assessment to faculty and practitioner scorers. Additionally, scorers did not know the identity of the authors of the essays that they score. Essays with Overall Scores that deviated by more than one point were sent to a third scorer for an additional review.

In practice, passing or failure of the Career Commitment Essay is based on an average of the two Overall Scores that are assigned to an essay. Any average score that is not a whole number is "bumped up" to the next highest score. For example, if an essay received an Overall Score of 2 from one rater and an Overall Score of 3 from another rater, the average score of 2.5 is converted to a score of 3. Only essays with average scores of 3 or 4 are considered passing. When the Overall Scores of two raters deviate by more than one point, the essay is sent to a third scorer. The candidate's final score is the average of the two highest scores s/he receives.

Analytic trait scores on the content, expression/voice, organization, and conventions dimensions are used to offer feedback to students, particularly those who fail the Career Commitment Essay

requirement. The intent of sharing these trait scores with candidates is to offer them an idea of the relative strengths and weaknesses of their essay so that they have a starting point for revision of the essay for a subsequent scoring session.

Data Analysis

Despite the fact that current judgments about students at the college involves averaged and "bumped" up scores, only the original scores assigned by raters were used in this study. The object was to explore properties of the scoring process at their most basic level.

The computer program, FACETS (Linacre, 1988), was used to analyze essay scoring data and furnish answers to the research questions identified earlier. In the first set of analyses, a three-facet model was used; the three facets were candidates, scorers (coded by unique scorer ID), and rubric dimensions. In the second set of analyses, a 3-facet model was employed once again; this time, the three facets were candidates, scorers (coded by scorer role: faculty or practitioner), and rubric dimensions.

Results

Raters

Do raters differ in the severity with which they rate examinees? How interchangeable are the raters? In this sample, rater severity measures ranged from -4.80 logits for the most lenient rater to 1.48 logits for the most severe rater (see Table 2). The spread of the severity measures is more than 6 logits. Using the Rater Separation Ratio, the Rater Separation Index was computed, revealing that 10.4 statistically distinct levels of rater severity could be discerned among the raters in this analysis. Considering that a Rater Separation Index of zero indicates interchangeability among raters, it was evident that all raters were not scoring with similar levels of severity. The Reliability of Rater Separation Index of .98 provided further evidence of this, suggesting that there was unwanted variation in rater severity that can affect examinee scores. Furthermore, the Fixed (All Same) Chi-Square statistic was significant, rejecting the null hypothesis that all raters are equally lenient (or severe).

It was also useful to examine the degree of exact inter-rater agreement among raters in the sample. Table 2 shows that Career Commitment Essay scorers were in exact agreement when presented with the same student just 34.4% of the time. This falls far short of the 90% inter-rater reliability that is generally accepted as required for high stakes decisions (Salvia and Ysseldyke, 1998).

Do raters use the rating scale(s) consistently? Among the 40 faculty and practitioner raters, analyses revealed that this mixed group of raters did not use the rating scales consistently. Mean square infit statistics ranged from .38 to 1.98. Mean outfit statistics ranged from .36 to 1.99. (see Table 2).

Table 2: Judges Measurement Report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Exact Agree. Obs %	Exact Agree. Exp %	Num	Judges
224	77	2.9	2.75	-.58	.20	1.98	4.6	1.99	4.2	.08	30.2	45.1	1002	1002
598	220	2.7	2.80	-.70	.12	1.72	6.5	1.69	6.3	.23	23.7	42.6	1012	1012
256	70	3.7	3.88	-4.80	.27	1.57	2.7	1.71	2.3	.46	13.5	17.9	1073	1073
149	69	2.2	1.98	1.48	.20	1.65	3.4	1.62	3.3	.24	25.5	35.1	1067	1067
463	180	2.6	2.57	-.06	.13	1.52	4.4	1.51	4.1	.41	37.4	43.9	1001	1001
174	61	2.9	2.81	-.73	.23	1.38	1.9	1.31	1.6	.64	25.9	45.6	1072	1072
544	200	2.7	2.90	-1.00	.12	1.33	3.1	1.28	2.3	.69	32.0	43.7	1013	1013
433	185	2.3	2.08	1.22	.12	1.33	3.0	1.31	2.9	.65	27.2	36.1	1022	1022
486	190	2.6	2.51	.10	.12	1.30	2.8	1.28	2.6	.69	29.6	42.0	1009	1009
443	150	3.0	3.26	-2.14	.14	1.23	1.9	1.25	2.1	.69	33.7	37.9	1071	1071
451	190	2.4	2.53	.04	.12	1.20	1.9	1.19	1.8	.76	38.0	42.3	1018	1018
272	100	2.7	2.68	-.35	.17	1.13	.9	1.08	.6	.90	38.1	43.8	1061	1061
251	116	2.2	2.35	.51	.16	1.07	.6	1.09	.7	.86	38.7	40.6	1010	1010
326	120	2.7	2.97	-1.25	.16	1.06	.5	1.08	.6	.90	36.4	40.4	1017	1017
882	305	2.9	2.92	-1.07	.10	.99	-.1	.99	-.1	1.03	26.5	41.1	1060	1060
161	63	2.6	2.76	-.61	.21	.98	.0	.96	-.1	1.06	43.0	43.5	1021	1021
174	65	2.7	2.32	.59	.22	.90	-.5	.95	-.1	1.10	48.2	43.7	1069	1069
242	110	2.2	2.57	-.06	.17	.94	-.4	.90	-.7	1.11	27.9	43.0	1023	1023
196	60	3.3	3.18	-1.89	.24	.98	.0	.89	-.5	1.09	30.2	35.0	1070	1070
285	137	2.1	2.04	1.31	.14	.88	-1.0	.88	-1.1	1.17	31.1	33.2	1015	1015
630	230	2.7	2.70	-.43	.11	.88	-1.3	.87	-1.4	1.14	42.5	43.0	1065	1065
353	120	2.9	3.18	-1.90	.16	.85	-1.1	.87	-1.0	1.15	36.8	40.1	1007	1007
216	80	2.7	2.80	-.72	.19	.85	-1.0	.87	-.8	1.15	41.0	45.3	1019	1019
508	179	2.8	2.58	-.09	.13	.87	-1.2	.85	-1.3	1.14	34.5	45.6	1063	1063
536	190	2.8	2.95	-1.18	.13	.82	-1.8	.85	-1.4	1.17	36.5	41.6	1050	1050
519	222	2.3	2.45	.26	.11	.83	-2.0	.82	-2.0	1.21	35.4	40.8	1005	1005
244	94	2.6	2.60	-.15	.17	.82	-1.2	.81	-1.3	1.21	34.9	43.6	1062	1062
328	125	2.6	2.97	-1.22	.15	.81	-1.6	.81	-1.6	1.20	41.1	40.3	1016	1016
668	238	2.8	2.69	-.41	.11	.78	-2.6	.77	-2.7	1.24	37.3	44.3	1011	1011
342	115	3.0	3.20	-1.96	.16	.77	-2.0	.78	-1.9	1.26	45.4	38.7	1020	1020
165	55	3.0	3.08	-1.59	.24	.76	-1.2	.92	-.2	1.19	46.3	40.2	1025	1025
828	305	2.7	2.79	-.68	.10	.75	-3.3	.75	-3.4	1.27	36.5	42.7	1008	1008
453	135	3.4	3.60	-3.30	.17	.77	-2.0	.73	-1.6	1.28	28.3	28.8	1014	1014
463	214	2.2	2.20	.91	.12	.71	-3.5	.72	-3.3	1.31	36.2	36.4	1006	1006
247	114	2.2	2.15	1.02	.15	.67	-2.9	.67	-2.9	1.37	42.3	37.6	1003	1003
507	184	2.8	2.90	-1.02	.13	.65	-3.9	.64	-3.9	1.37	40.1	42.1	1004	1004
214	64	3.3	3.45	-2.78	.24	.66	-2.3	.62	-1.7	1.45	28.5	33.7	1068	1068
189	79	2.4	2.20	.91	.19	.62	-2.7	.61	-2.8	1.42	41.4	39.9	1064	1064
109	50	2.2	2.48	.17	.23	.43	-3.8	.44	-3.7	1.63	36.8	30.8	1024	1024
254	90	2.8	2.55	-.01	.18	.38	-5.7	.36	-5.9	1.70	51.6	46.2	1066	1066
Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Exact Agree. Obs %	Exact Agree. Exp %	Num	Judges
369.6	138.8	2.7	2.73	-.60	.16	1.00	-.3	.99	-.3				Mean (Count: 40)	
185.3	68.1	.4	.42	1.27	.05	.35	2.7	.35	2.6				S.D. (Populn)	
187.6	69.0	.4	.42	1.29	.05	.35	2.7	.35	2.6				S.D. (Sample)	

Model, Populn: RMSE .17 Adj (True) S.D. 1.26 Separation 7.45 Reliability (not inter-rater) .98
Model, Sample: RMSE .17 Adj (True) S.D. 1.28 Separation 7.55 Reliability (not inter-rater) .98
Model, Fixed (all same) chi-square: 2065.1 d.f.: 39 significance (probability): .00
Model, Random (normal) chi-square: 38.1 d.f.: 38 significance (probability): .47
Inter-Rater agreement opportunities: 4802 Exact agreements: 1654 = 34.4% Expected: 1942.1 = 40.4%

Given the high stakes nature of the Career Commitment Essay, stringent upper- and lower-control limits for mean square fit statistics were utilized. The most stringent upper- and lower-control limits for mean square fit statistics generally used are an upper-control limit of 1.2 and a lower-control limit of .8

(Myford and Dobria, 2006). From such a point of view, slightly more than half of all raters (n=21 or 53%) were not internally consistent with regard to their application of the rating scales across all candidates and rubric dimensions. This is evidenced in mean square infit ratings higher than 1.2 and lower than .8. Similar findings occurred with respect to mean square outfit statistics, with the same proportion of raters (21%) displaying mean square outfit statistics higher than 1.2 or lower than .8.

More raters had mean square fit statistics under the lower limits (n=12) than over the upper limits (n=9). This shows that among raters not displaying adequate fit, rating tended to be overly consistent. Frequent reasons for this may include:

1. Some raters may only use certain categories of the rubric (e.g., ratings of 2 and 3) and avoid other categories
2. Some raters may give candidates similar ratings on all rubric dimensions even though the rubric dimensions are meant to assess quite different traits
3. As raters begin to tire, they may assign similar ratings across rubric dimensions or candidates (Myford and Dobria, 2006)

For raters who had mean-square infit statistics larger than the upper-control limit, the table of unexpected responses (

Table 3) was useful for revealing and analyzing misfitting responses.

Table 3 displays 20 misfitting responses. With the exception of just three responses, all unexpected responses (deemed so by residuals greater than or equal to three standard deviations) were more severe than expected (taking into account the rater's overall severity and the other ratings the candidate received). The rubric dimension with the most misfitting ratings was Conventions, which comprised eight of the twenty unexpected responses. This points to less consistency among Conventions ratings than ratings of the other dimensions. The Content dimension was rated in an unexpected manner second most frequently, accounting for six of the twenty unexpected responses.

The raters with the greatest numbers of unexpected responses were raters 1073 and 1060, with four unexpected responses each. These findings suggest a need to consult with these two scorers (one faculty member and one practitioner) regarding the nature of the scale categories, the conceptually distinct aspects of each dimension, and the context of their scoring (e.g., Are these raters sensitive to fatigue effects or inattention?). Rater 1073 was inconsistent with Students 325 and 416, as all four of his unexpected ratings were with these two students. Similarly, the four unexpected ratings for Rater 1060 occurred for with Students 42 and 915. In all situations, these two raters gave unexpectedly low ratings to the students in question. It would be interesting to explore why this was the case.

Table 3: Unexpected Responses

Cat	Step	Exp.	Resd	StRes	Num Stu	Num Judg	N	Item
3	3	3.9	-.9	-4.1	325 325	1073 1073	2	Expression
3	3	3.9	-.9	-3.7	416 416	1073 1073	1	Content
4	4	1.8	2.2	3.7	381 381	1012 1012	4	Conventions
1	1	3.1	-2.1	-3.7	914 914	1002 1002	4	Conventions
3	3	3.9	-.9	-3.7	280 280	1025 1025	4	Conventions
3	3	1.3	1.7	3.7	328 328	1001 1001	4	Conventions
1	1	3.0	-2.0	-3.6	81 81	1001 1001	1	Content
3	3	3.9	-.9	-3.6	186 186	1002 1002	4	Conventions
3	3	3.9	-.9	-3.5	325 325	1073 1073	4	Conventions
3	3	3.9	-.9	-3.5	416 416	1073 1073	5	Overall
1	1	3.0	-2.0	-3.5	927 927	1065 1065	4	Conventions
1	1	2.9	-1.9	-3.4	42 42	1060 1060	1	Content
1	1	2.9	-1.9	-3.4	915 915	1060 1060	1	Content
1	1	2.9	-1.9	-3.4	387 387	1012 1012	1	Content
1	1	2.9	-1.9	-3.3	42 42	1060 1060	5	Overall
1	1	2.9	-1.9	-3.3	915 915	1060 1060	5	Overall
4	4	2.0	2.0	3.3	987 987	1022 1022	3	Organization
4	4	2.1	1.9	3.1	991 991	1012 1012	4	Conventions
1	1	2.8	-1.8	-3.0	924 924	1006 1006	1	Content
1	1	2.8	-1.8	-3.0	906 906	1065 1065	2	Expression

Do faculty and practitioners rate essays in the same manner? A separate analysis was conducted in order to compare the rating patterns of college faculty rater versus PK-12 practitioner raters. Table 4 below reveals that rater severity measures were -.26 for faculty raters and -.58 for practitioner raters. Practitioner raters were clearly more lenient. This is supported by the significant Fixed (All Same) Chi-Square statistic, rejecting the null hypothesis that both groups of raters are equally lenient. Additionally, exact inter-rater agreement was just 34.3%.

Table 4: Judges Measurement Report for Faculty Vs. Practitioners

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Exact Agree. Obs %	Exact Agree. Exp %	N Judges
6663	2559	2.6	2.66	-.26	.03	.94	-2.3	.93	-2.6	1.08	34.3	43.1	1 Fac
8120	2992	2.7	2.79	-.58	.03	1.05	2.1	1.06	2.1	.93	34.3	43.1	2 Prac
7391.5	2775.5	2.7	2.72	-.42	.03	1.00	-.1	.99	-.2				Mean (Count: 2)
728.5	216.5	.1	.07	.16	.00	.06	2.3	.06	2.4				S.D. (Populn)
1030.3	306.2	.1	.10	.23	.00	.08	3.2	.09	3.4				S.D. (Sample)

Model, Populn: RMSE .03 Adj (True) S.D. .16 Separation 5.32 Reliability (not inter-rater) .97
Model, Sample: RMSE .03 Adj (True) S.D. .23 Separation 7.59 Reliability (not inter-rater) .98
Model, Fixed (all same) chi-square: 58.6 d.f.: 1 significance (probability): .00
Inter-Rater agreement opportunities: 2525 Exact agreements: 865 = 34.3% Expected: 1088.8 = 43.1%

It was also revealed that six of the nine scorers with unexpected responses were practitioner raters. Three scorers with unexpected responses were faculty members. Correspondingly, 15 of the 21 unexpected responses were made by practitioners. This points to a possible need to work more intensely with practitioners on issues related to essay scoring.

Rubric Dimensions

Do rubric dimensions differ in difficulty? Is it harder to get high ratings on some dimensions than on others? Table 5 displays Career Commitment Essay rubric scores in order of difficulty, as rated

by a combined population of higher education faculty and PK-12 practitioners. The difficulty measures of rubric dimensions in the Career Commitment Essay rubric range from .33 logits for the most difficult dimension (the Overall Score) to -.38 logits for the easiest dimension (Expression). This represents a spread of less than .71 logits.

Table 5: Item Measurement Report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	N	Item
2922	1147	2.5	2.65	.33	.05	.83	-4.3	.83	-4.2	1.18	5	Overall
2879	1108	2.6	2.70	.19	.05	1.05	1.1	1.04	.8	.96	1	Content
2945	1096	2.7	2.78	-.06	.05	1.21	4.6	1.24	5.0	.76	4	Conventions
2959	1098	2.7	2.79	-.08	.05	1.05	1.2	1.02	.5	.96	3	Organization
3078	1102	2.8	2.89	-.38	.05	.85	-3.8	.84	-3.7	1.17	2	Expression
2956.6	1110.2	2.7	2.76	.00	.05	1.00	-.2	.99	-.3			Mean (Count: 5)
66.5	18.9	.1	.08	.24	.00	.14	3.4	.15	3.4			S.D. (Populn)
74.3	21.1	.1	.09	.27	.00	.16	3.8	.17	3.8			S.D. (Sample)

Model, Populn: RMSE .05 Adj (True) S.D. .24 Separation 4.62 Reliability .96
Model, Sample: RMSE .05 Adj (True) S.D. .27 Separation 5.19 Reliability .96
Model, Fixed (all same) chi-square: 112.3 d.f.: 4 significance (probability): .00
Model, Random (normal) chi-square: 3.9 d.f.: 3 significance (probability): .28

Using the Item Separation Ratio provided in the Facets output, the Item Separation Index was calculated to be 7.25, suggesting the presence of seven statistically distinct levels of difficulty among the various rubric dimensions. The Reliability of Item Separation of .96 was very high, indicating that the dimensions in the rubric are well differentiated in terms of difficulty. Furthermore, the Fixed (All Same) Chi-Square statistic is significant, rejecting the null hypothesis that all rubric dimensions are equally difficult. In particular, it appears that it is much easier to score highly on the Expressions dimension than for any other rubric category. However, the difficulty values of the Conventions and Organization dimensions are nearly identical, raising the question of whether these two dimensions might be somewhat redundant—or indistinguishable among raters.

Do teacher candidates differ in proficiency? Do differences in rater severity affect candidates' scores? The Students Measurement Report, an abbreviated form of which is displayed in Table 6

Table 6, revealed that applicants to the teacher education program differed significantly in terms of proficiency as measured by Career Commitment Essay scores. The Reliability of Person Separation of .88 was fairly high, indicating that the candidates were well differentiated in terms of essay writing skills. The Examinee Separation Index (derived from the Examinee Separation Ratio) was 4.1, indicating that there were four statistically distinct levels of candidate proficiency among the teacher education candidates in the sample. Furthermore, the Fixed (All Same) Chi-Square statistic was significant, rejecting the null hypothesis that all persons were equally able.

Table 6: Students Measurement Report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M AvrageMeasure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Num Students
29	11	2.6	2.82	.16	.51	4.69	4.8	4.68	4.8	-2.60	141 141
21	7	3.0	3.33	1.76	.66	3.41	3.0	3.25	2.9	-1.30	982 982
43	15	2.9	2.86	.28	.45	3.29	3.8	3.29	3.8	-1.00	42 42
44	15	2.9	2.95	.57	.45	3.23	3.8	3.16	3.7	-1.07	387 387
60	21	2.9	3.05	.87	.38	3.18	4.4	3.20	4.4	-1.03	914 914
41	15	2.7	2.65	-.31	.44	3.18	3.9	3.13	3.8	-1.10	956 956
35	15	2.3	2.75	-.04	.42	3.15	4.2	3.15	4.2	-1.36	177 177
37	10	3.7	3.04	.84	.76	1.92	1.6	3.13	1.8	-.33	416 416
33	10	3.3	2.92	.47	.68	1.08	.3	3.08	1.9	.47	325 325
41	15	2.7	2.56	-.57	.44	3.07	3.7	3.04	3.7	-.90	403 403
21	10	2.1	2.23	-1.42	.51	2.90	3.1	2.90	3.1	-1.40	984 984
32	15	2.1	2.63	-.37	.42	2.84	3.7	2.86	3.7	-1.36	985 985
87	30	2.9	2.99	.70	.32	2.84	4.7	2.77	4.6	-.77	927 927
25	10	2.5	2.63	-.37	.54	2.78	2.8	2.66	2.7	-.85	987 987
22	10	2.2	2.43	-.92	.51	2.69	2.9	2.68	2.9	-1.00	418 418
54	20	2.7	2.71	-.16	.38	2.57	3.5	2.62	3.6	-.63	931 931
55	20	2.8	2.85	.27	.38	2.50	3.4	2.58	3.5	-.47	915 915
|											|
|											|
|											|
30	10	3.0	2.97	.63	.56	.02	-4.2	.02	-4.3	1.86	121 121
30	10	3.0	3.47	2.23	.56	.02	-4.2	.02	-4.3	1.86	199 199
15	5	3.0	2.96	.60	.79	.02	-2.9	.02	-2.9	1.86	303 303
15	5	3.0	2.85	.26	.79	.02	-2.9	.02	-2.9	1.86	158 158
18	6	3.0	3.15	1.20	.72	.02	-3.2	.02	-3.2	1.86	175 175
40	10	4.0	3.99(	6.52	1.85)	Maximum					48 48
40	10	4.0	3.93(	4.85	1.85)	Maximum					96 96
40	10	4.0	3.99(	7.06	1.86)	Maximum					119 119
24	6	4.0	3.94(	4.89	1.87)	Maximum					197 197
24	6	4.0	3.96(	5.40	1.87)	Maximum					200 200
40	10	4.0	3.99(	6.65	1.85)	Maximum					391 391
40	10	4.0	3.97(	5.85	1.86)	Maximum					407 407

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M AvrageMeasure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Num Students
31.6	11.8	2.7	2.74	.09	.55	.91	-.4	.91	-.4	Mean (Count: 476)
12.4	4.7	.5	.54	1.71	.19	.66	1.7	.67	1.7	S.D. (Populn)
12.4	4.7	.5	.54	1.72	.19	.66	1.7	.68	1.7	S.D. (Sample)

With extremes, Model, Populn: RMSE .58 Adj (True) S.D. 1.61 Separation 2.76 Reliability .88
With extremes, Model, Sample: RMSE .58 Adj (True) S.D. 1.61 Separation 2.76 Reliability .88
Without extremes, Model, Populn: RMSE .54 Adj (True) S.D. 1.47 Separation 2.71 Reliability .88
Without extremes, Model, Sample: RMSE .54 Adj (True) S.D. 1.47 Separation 2.71 Reliability .88
With extremes, Model, Fixed (all same) chi-square: 3750.4 d.f.: 475 significance (probability): .00
With extremes, Model, Random (normal) chi-square: 397.4 d.f.: 474 significance (probability): 1.00

Data in the Observed Average of Table 6 contains the average of raters' ratings across all of the rubric dimensions for each candidate. The number in the Fair Average column, in contrast, adjusts the candidates' observed average based on differences in the severity or leniency of raters who evaluated their essay. In fact, the Fair Average eliminates the effects of differential rater severity and shows the score the candidate would have received from raters of "average" severity. For example Student 141's observed average score across the Content, Expression/Voice, Organization, Conventions, and Overall Score ratings was 2.6. However, s/he must have been rated by some relatively severe raters because his/her Fair Average score was higher: 2.82. This suggests that the scores that students receive on the Career Commitment Essay might not be fair, given the differing levels of severity of raters. In fact, dependent sample t-tests comparing the means of the Observed Average (mean=2.69) and Fair Average scores (mean=2.73) revealed the two sets of scores to differ significantly (p=.006).

Finally, an examination of the information in Table 6 also reveals a large number of candidates with mean square fit statistics that are outside the lower and upper control limits. One hundred twenty-two students (26%) of candidates had mean-square infit statistics higher than the upper limit of 1.8. Variation in ratings for these candidates was greater than expected. It is possible that these students exhibited unusual performance on the Career Commitment Essay. On other hand, it is also conceivable that the raters of these students' essays were unusually severe or lenient.

In contrast, two hundred forty-five candidates (51%) had mean-square infit statistics lower than the lower-control limit of .8. These findings suggest little variation in candidates' ratings across the various rubric dimensions. This could be because raters evaluated students on all rubric dimensions based on an overall impression of the essay, rather than analyzing performance in the different categories. Alternatively, raters could have avoided the extreme performance levels in the scoring rubric (e.g., scores of 1 and 4), resulting in overly consistent scores. Finally, it is possible that candidates really did perform the same on all rubric dimensions. However, this unlikely given the large proportion of essay writers for with unusually low mean-square infit statistics.

Rating Scales

Are the rating scale functioning properly? The FACETS' 'yardstick' or variable map (Figure 2) displays the logit scores for raters, students, and rubric dimensions, all of which are on a common scale and directly comparable. The logit metric is shown in the far left column, with the average item difficulty set at zero. Raters are listed in order of severity in the second column, with severe, or strict, raters at the top and easier, more lenient raters at the bottom. Students are listed in the third column, with more able students at the top of the column and less able students at the bottom. Finally, the measures on the six rubric dimensions are shown in the fourth column. Harder dimensions on which to receive high ratings are shown at the top of the column, while easier dimensions are shown at the bottom. The last column shows the step structure for the four point rating scale for each rubric dimension.

In a well-developed assessment, items (dimensions in this example) will be spread out along the full length of the variable. Students will be appropriately targeted by the items; in other words, their ability levels should span the full length of the variable (as defined by the rubric dimensions). Along the continuum of the variable, there should be no gaps without items. The presence of such gaps suggests the level of difficulty of the items is not entirely appropriate, given the ability of the students being rated. Additionally, the scale structures for rubric dimensions with the same number of rating points should be comparable.

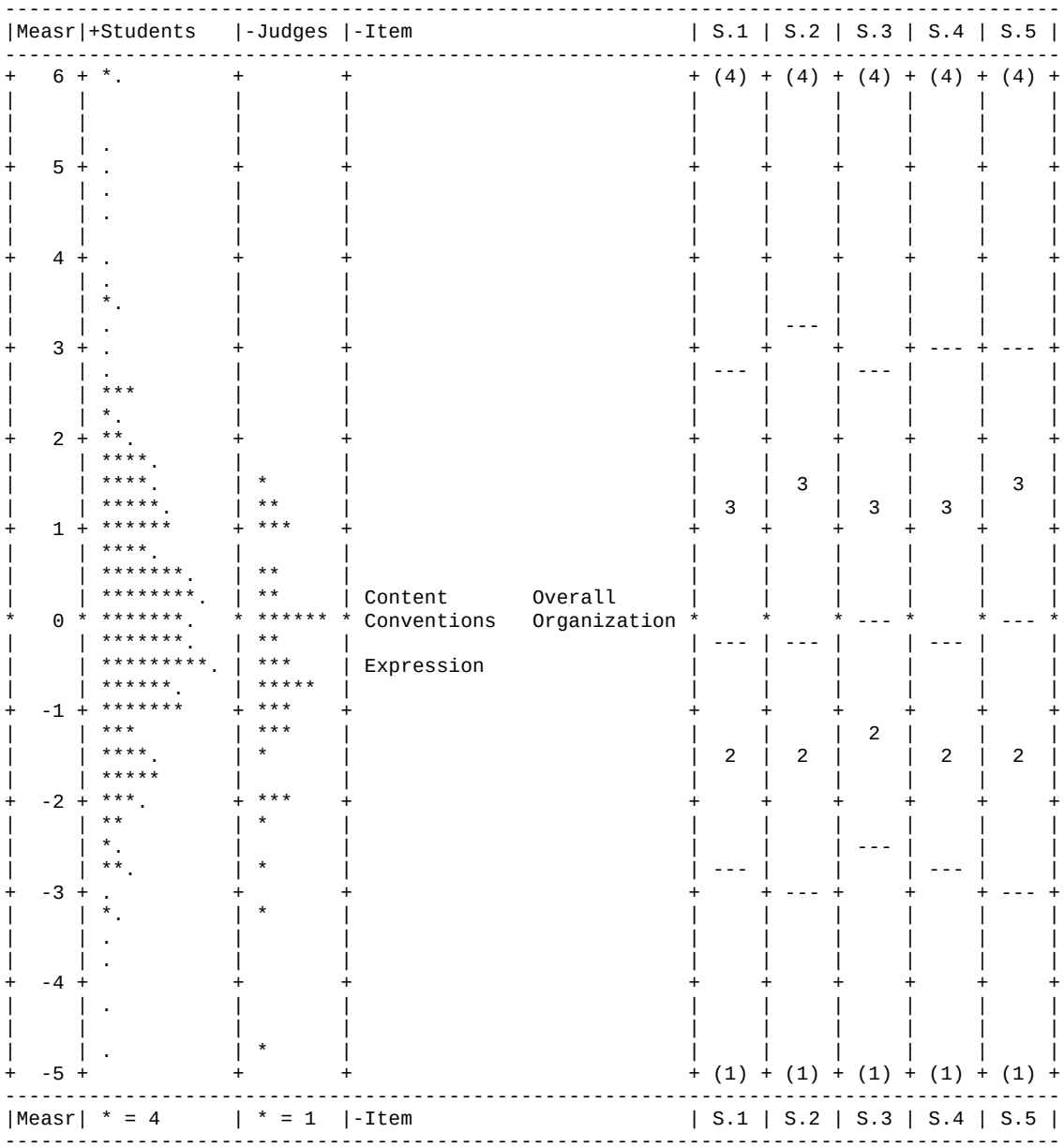


Figure 2: Career Commitment Essay Variable Map

The variable map above reveals that the Career Commitment Essay writing skills of teacher education candidates do, indeed, span the full length of the variable being measured. The rubric dimensions do not spread out along the full length of the variable, suggesting that they do not target all of the ability levels of students who write essays. There definitely are gaps along the continuum where no items are found. While this is a necessary prerequisite for a good test or assessment, the question remains as to whether different rubric dimensions should span a range of difficulty levels, particularly with a criterion-referenced task. Furthermore, the points where the horizontal lines for the five scales (one per rubric dimension) cross are similar, suggesting the existence of a comparable scale structure across items.

The rating categories for each rubric dimension were subsequently examined to determine whether each of them was functioning properly as a four-point scale. An examination of the data in Tables 7 through

11 reveals the following about the Content, Expression/Voice, Organization, Conventions, and Overall Score rating scales:

- For all dimensions of the Career Commitment Essay rubric, Average Measures increase with each Category Score (one through four). This indicates that candidates with higher ratings on any particular dimension are exhibiting "more" of the variable being measured. Hence, the scale categories are appropriately ordered and functioning properly.
- The "Most Probable from" thresholds are appropriately ordered as well (from low to high), as rating categories increase. This signifies every category is most probable to be observed at some point on the rating scale, the rating categories are appropriately ordered, and all categories are utilized.
- Mean-square outfit statistics for each rating category on every dimension are greater than .5 and less than 1.5, revealing that observed and expected examinee proficiency measures are very similar. Furthermore, mean-square outfit statistics in this range indicate that each rating category is contributing to meaningful measure of the particular rubric dimension.

Table 7: Category Statistics for CONTENT

DATA				QUALITY CONTROL			STEP		EXPECTATION		MOST	.5 Cumul.	Cat
Category	Counts		Cum.	Avge	Exp.	OUTFIT	CALIBRATIONS		Measure at		PROBABLE	Probabil.	PEAK
Score	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	at	Prob
1	110	10%	10%	-1.58	-1.92	1.4			(-3.64)		low	low	100%
2	371	33%	43%	-.76	-.64	.9	-2.51	.12	-1.37	-2.71	-2.51	-2.59	61%
3	481	43%	87%	.68	.71	.9	-.24	.08	1.27	-.13	-.24	-.19	69%
4	146	13%	100%	2.67	2.51	.9	2.74	.11	(3.86)	2.86	2.74	2.78	100%
									(Mean)		(Modal)	(Median)	

Table 8: Category Statistics for EXPRESSION/VOICE

DATA				QUALITY CONTROL			STEP		EXPECTATION		MOST	.5 Cumul.	Cat
Category	Counts		Cum.	Avge	Exp.	OUTFIT	CALIBRATIONS		Measure at		PROBABLE	Probabil.	PEAK
Score	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	at	Prob
1	46	4%	4%	-1.58	-1.61	1.0			(-4.03)		low	low	100%
2	315	29%	33%	-.42	-.33	.9	-2.92	.17	-1.55	-3.05	-2.92	-2.97	66%
3	562	51%	84%	1.12	1.11	.9	-.20	.08	1.46	-.12	-.20	-.17	72%
4	179	16%	100%	3.07	2.95	.9	3.11	.10	(4.21)	3.20	3.11	3.13	100%
									(Mean)		(Modal)	(Median)	

Table 9: Category Statistics for ORGANIZATION

DATA				QUALITY CONTROL			STEP		EXPECTATION		MOST	.5 Cumul.	Cat
Category	Counts		Cum.	Avge	Exp.	OUTFIT	CALIBRATIONS		Measure at		PROBABLE	Probabil.	PEAK
Score	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	at	Prob
1	94	9%	9%	-1.54	-1.72	1.1			(-3.53)		low	low	100%
2	340	31%	40%	-.59	-.47	.9	-2.39	.13	-1.28	-2.60	-2.39	-2.48	60%
3	471	43%	82%	.88	.85	.8	-.14	.08	1.22	-.07	-.14	-.11	65%
4	193	18%	100%	2.61	2.56	1.0	2.53	.10	(3.66)	2.69	2.53	2.59	100%
									(Mean)		(Modal)	(Median)	

Table 10: Category Statistics for CONVENTIONS

DATA				QUALITY CONTROL			STEP		EXPECTATION		MOST	.5 Cumul.	Cat
Category	Counts		Cum.	Avge	Exp.	OUTFIT	CALIBRATIONS		Measure at		PROBABLE	Probabil.	PEAK
Score	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	at	Prob
1	93	8%	8%	-1.59	-1.82	1.3			(-3.59)		low	low	100%
2	316	29%	37%	-.56	-.57	1.0	-2.43	.13	-1.42	-2.68	-2.43	-2.54	58%
3	528	48%	85%	.87	.80	1.4	-.41	.08	1.23	-.23	-.41	-.33	71%
4	159	15%	100%	2.21	2.61	1.3	2.84	.11	(3.94)	2.93	2.84	2.86	100%
									(Mean)		(Modal)	(Median)	

Table 11: Category Statistics for OVERALL SCORE

DATA				QUALITY CONTROL		STEP		EXPECTATION		MOST	.5 Cumul.	Cat
Category	Counts		Cum.	Avge	Exp.	OUTFIT	CALIBRATIONS	Measure at		PROBABLE	Probabil.	PEAK
Score	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	at	Prob
1	96	8%	8%	-1.87	-2.00	1.1			(-4.00)		low	low	100%
2	454	40%	48%	-.81	-.67	.8	-2.90	.12	-1.45	-3.02	-2.90	-2.95	68%
3	470	41%	89%	.80	.75	.8	-.01	.07	1.45	.00	-.01	.00	68%
4	127	11%	100%	2.80	2.56	.8	2.90	.12	(4.02)	3.03	2.90	2.94	100%
									(Mean)		(Modal)	(Median)	

Discussion

This study revealed good news and bad news about the Career Commitment Essay process that is taking place as part of candidate admission into the School of Education. The good news concerns the rubric used to evaluate Career Commitment Essays. For example, this study found that the four-point rating scales in all of the rubric dimensions (Content, Expression/Voice, Organization, Conventions, Overall Score) are functioning as intended. Furthermore, the dimensions in the rubric are well differentiated in terms of difficulty.

The "bad" news has to do with people: the people who score the essays and the people whose essays are being scored. Essay scorers vary significantly in terms of the degree of leniency and severity they exhibit toward essay writers. Exact inter-rater reliability hovers at approximately 34%, despite the fact that a high stakes assessment such as the Career Commitment Essay merits a much higher degree of inter-rater reliability. Differences in severity also affect the fairness of scores yielded from such a process. Mean differences between observed scores and "fair" scores adjusted to take into account differing levels of rater severity are statistically significant. It might appear that the actual practice at the School of Education of making decisions solely on the Overall Score, averaging the Overall Scores of the two highest raters of an essay, and bumping up average Overall Scores of 1.5, 2.5, and 3.5 to the next highest whole number is in fact a reaction to an intuitive understanding that scorers are not consistent. The practices just described in fact favor the teacher education candidate, making it more difficult for him/her to fail. In a way, this is an attempt to compensate for unreliable scoring.

While differences in severity and leniency among raters were associated with greater than expected variation in ratings for over one quarter of essay writers, the ratings of 51% of essay writers showed little variation, suggesting that raters did not analyzing candidate performance along the different rubric dimensions and/or raters avoided the extreme performance levels in the scoring rubric and confined themselves the middle of the rating scale for each rubric dimension.

What are the implications of such findings? Clearly, additional rater training is called for. Many raters need to develop a better understanding of what each rubric dimension purports to measure and how these dimensions differ from each other. Some raters also need to feel comfortable using the entire range of the four-point rating scales to so that ratings are not overly consistent. Furthermore, more scorers may need to be recruited to reduce the number of essays that scorers must read in a single session. This will reduce the probability of scorers falling prey to fatigue that may prevent them from observing differences among candidates and among dimensions of the Career Commitment Essay. The challenges in implementing these recommendations, however, concern time and money. Scorers are currently paid $150 per half day session. Can the School of Education afford to increase the stipend for

a longer training/scoring session? Will scorers be willing to work for a longer day for the same pay? These are questions that remain to be answered. At another level, there is ongoing discussion at the School of Education about the predictive validity of Career Commitment Essays. Does performance on the essay predict success as a teacher candidate? Does the essay merit such an important place in the unit's assessment system?

Nevertheless, studies such as this are important for establishing the validity and reliability of a school's performance assessment procedures. Without the knowledge that comes from such a study, it is impossible to make judgments or implement practices that are based on data, as opposed to intuition. The Multi-Facet Rasch Model is an excellent technique for accomplishing this, as it enables the researcher or assessment specialist to investigate scoring patterns, student performance, and scale functioning in ways that are directly relevant to the demands set forth by NCATE. The information yielded through analyses of this type also align with the goals of any responsible user of assessment.

References

Linacre, J.M. (1988). *FACETS*. Chicago: Mesa.

Myford, C.M. & Dobria, L. (2006). *Facets Workshop*. University of Illinois at Chicago.

National Council for Accreditation of Teacher Education. (2002). *Professional standards for accreditation of schools, colleges, and departments of education*. Washington, DC: Author.

Salvia, J., & Ysseldyke, J. E. (1998). *Assessment* (7th ed.). Boston: Houghton Mifflin.