Tabaran Institute of Higher Education
Iranian Journal of Language Testing
Vol. 3, No. 1, March 2013
ISSN 2251-7324

Received: Feb14, 2013 *Accepted: March 7, 2013*

Validation of an Analytic Rating Scale for Writing: A Rasch Modeling Approach

Susan Tan [1]

Abstract

Writing assessments often make use of analytic rating scales to describe the criteria for different performance levels. However, the use of such rating scales requires a level of interpretation by raters and if several raters are involved, the reliability of examinee scores can be significantly affected (Engelhard, 1992; McNamara 1996). Variability between raters is partly managed by rater training in the use of the rating scale and this necessarily means that the rating scale itself should be well constructed and can be accurately applied to discriminate examinee performance consistently. This paper reports on the use of the Many-facets Rasch model (MFRM, Linacre, 1989) to assess the validity of a proposed analytic rating scale. The MFRM is widely used to study examinee performance and rater behavior and is useful in rating scale validation to analyze sources of variation in tests (Schaeffer, 2008). Bias analysis allows systematic subpatterns of interactions between raters and the rating scale to be examined. In this paper, scores from a set of essays rated by a team using a revised analytic descriptor were analyzed and the indices for rater severity, rater consistency, rater bias, criteria difficulty and scale functionality were studied. The findings indicate that raters were able to use the revised rating scale to discriminate performances in a consistent manner. The MFRM can contribute to improvements in rater training and rating scale development.

Keywords*: Performance Assessments, Rating Scales, Many-Facets Rasch Model, Reliability, Validity*

1. Introduction

Measuring or rating performance assessments like essay writing is a complex cognitive process. This is because the use of such rating scales requires a level of interpretation and judgment by raters and if several raters are involved, the reliability of examinee scores can be significantly affected (Engelhard, 1992; McNamara 1996). To obtain reliable measurements of such performance, ratings must be independent of the particular raters that are used for the measuring, a concept Engelhard calls rater-invariant measurement of persons (2002). To help

1 National University of Singapore, Centre for English Language Communication, 10 Architecture Drive, Singapore 117511, elctans@nus.edu.sg.

achieve rater invariance, institutions often make use of marking schemes to describe different performance criteria and levels so that raters are more able to assign reliable scores to the performance that they are evaluating. Marking schemes, also commonly known as rubrics, descriptors or rating scales are particularly useful when there is a team of raters involved in rating many student performances. Marking schemes serve to align many raters to a common understanding of institutional or course expectations in the performance descriptions. Much time and effort are committed to developing clear marking schemes.
In writing assessments, either holistic or analytic rating scales (descriptors) to describe the criteria for different performance levels are used. While holistic descriptors describe overall proficiency on a single rating scale, analytic descriptors differentiate and make discrete different components (for e.g. ideas, mechanics, organization) of a text for rating on multiple scales. This paper will focus on the use of analytic descriptors for assessing essay writing.

The use of analytic descriptors allows for performance on specific criteria to be scrutinized separately yielding rich data about examinees' language abilities (Brown and Bailey, 1984; Kondo-Brown, 2002). Such information is useful to students as the breakdown of information and the different scores help students to understand their performance better. For teachers, such data can inform teaching as they are able to focus on areas which students are weak in. The use of analytic descriptors may also be helpful to focus the rater's attention on particular components to help improve inter-rater reliability.

Studying the way raters evaluate writing performance using a descriptor gives a means to assess if the descriptors are well constructed and can be used consistently. An analytic descriptor that does not describe or discriminate different performance categories well enough for raters or one that may cause raters to score in an erratic manner or to have different interpretations of performance levels would affect the reliability of examinee scores. Studies on how raters use descriptors are often examined through instruments like questionnaire surveys, interviews or through think-aloud protocols. In the latter approach, raters verbalize their thought processes while rating a set of essays. This verbalization process is studied to elucidate, among other things, how raters are using and interpreting the descriptors to arrive at a given score. In comparison to other approaches like questionnaires and interviews the think-aloud protocol has the advantage of being immediate and untainted by problems with recall and selection. However, think-aloud protocols are time-consuming and hard to administer. Raters may not find it natural to verbalize their thoughts while marking (Smagorinsky, 1994). Also, there is concern that verbalizing the thinking process may affect the rating itself (Stratman & Hamp-Lyons, 1994). An economical and analytic means to validate performance descriptors to ensure the validity and reliability of assessments would be beneficial. This study uses the Many-facets Rasch model to assess raters' use of a set of descriptors for scoring a placement test.

2. Test Instrument

The National University of Singapore requires students who do not meet entry requirements levels for English to take a placement test to assess if the students have the required language proficiency. If students do not, they will be required to take proficiency modules to help them improve. The placement test, administered by the Centre for English Language Communication, is a source-based essay task. Students respond to a prompt and write an essay of about 500 words after reading a few short passages. Scoring of the essay is based on an analytic descriptor. The descriptor describes performance at three levels and on three criteria, namely content, organization and language (items). Each item has a detailed

description of that performance at different levels – Low (1), Mid (2) and High (3). In this study, ten essays from a recent placement test were used to assess raters' use of a set of descriptors. The number of essays used is small but since this exercise was to test a revised descriptor, we felt that the number was adequate to provide useful indices to help us determine the functionality of these descriptors.

3. Participants

A team of four lecturers was tasked with refining and revising the analytic descriptors. The process of revision is not the focus of this paper and will not be described. A revised descriptor with four categories, namely content, organization, grammar accuracy and grammar fluency was developed. This revised descriptor had to be tested to see if raters are able to use it well to discriminate students' performance. It is important that raters with varying lengths of service and experience in rating placement tests at the Centre be able to use the revised descriptor. Thirteen (13) raters (operational raters) with varying lengths of service at the Centre were invited to test the revised descriptor by using it to score a set of ten essays from a recent placement test. Eleven (11) of the operational raters had used the old descriptor before and only two had not. The range of experience of raters in marking placement tests at the centre was taken as a facet to be examined in the analysis. Apart from the operational raters, the team of lecturers who had revised the descriptors also rated the ten scripts using the revised descriptors. This team will henceforth be known as the validation team.

4. Method

This study uses the Many-facets Rasch model (MFRM, Linacre, 1989), an extension of the Rasch model, to assess the validity of the revised analytic descriptor. The Rasch model (Rasch, 1960) and its extensions is a probabilistic model that meets the requirements for invariant measurement. Invariance is supported when acceptable model-data fit is observed. In this model, model-data fit indicates that the construct is being measured without interference by construct-irrelevant factors such as rater characteristics (for example, race, gender, and severity) or specific assessment characteristics like difficulty of items or time of tests.

The MFRM may be represented as:

$$\log (Pnijk/ Pnij(k-1)) = Bn- Cj - Di- Fk$$

where P_{nijk} is the probability of essay n being rated *k* by Rater *j*; and $P_{nij(k-1)}$ is the probability of essay *n* being rated *k-1* by Rater *j*. B_n is the ability of the student as reflected in the quality of the essay while C_j is the severity of the rater, D_i is the difficulty of the item and F_k is the difficulty of the rating scale step relative to the previous step.

The MFRM is widely used to study examinee performance and rater behavior (such as severity and consistency) and is useful in rating scale validation to analyze sources of variation in tests (Schaeffer, 2008). Also, bias analysis allows systematic sub-patterns of interactions between raters and the rating scale to be examined.

5. Research questions

This study is motivated by these research questions:

Q1. Do the revised descriptors function well to discriminate performance categories and performance levels?
Q2. Can the operational raters use the revised descriptors in a reliable and consistent manner?
Q3. Did construct irrelevant factors like the rating experience affect the manner in which the revised descriptor was used?

6. Data Analysis

Ten scripts from a recent placement test were marked and the scores assigned by the operational raters were analyzed using the computer program Facets (Linacre, 2010). Seven indices reported by Facets were used to examine the results: student ability, rater severity, rater fit, rater experience, item difficulty, scale functionality and bias interaction. The section that follows describes the various findings.

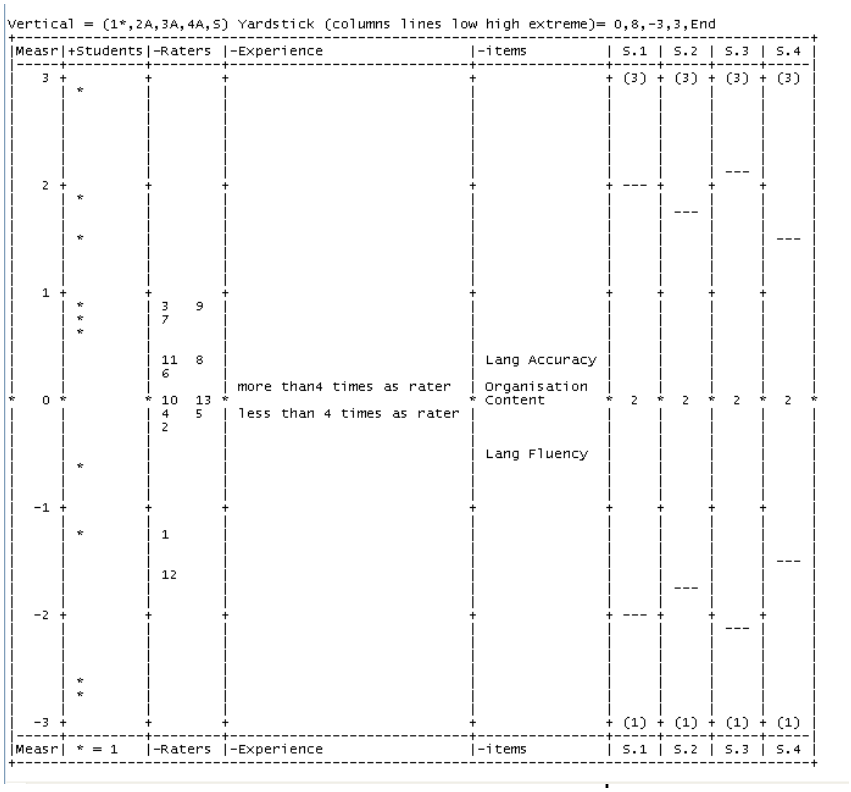
7. Results

7.1. Reliability and fit to model

The data summary shows that the data obtained is a good fit to the many-facets Rasch model with the mean of standardized residuals at 0.1 and the standard deviation (*SD*) at 0.99. The mean should be 0 and the SD should be 1.

Figure 1 shows the variable map which provides a simple yet clear view of the placement of the students, raters, experience, items (criteria in the descriptors) on a "ruler-like scale". The first column is the measure expressed in logits (i.e., the units of the "ruler"). Unlike raw scores, logits represent true interval scale where the distance between intervals is equal. A common ruler for measurement is therefore available for measuring relevant probabilities for all four facets – student ability, rater severity, experience and the rating scale. Larger logit number (items on the top half) will have more of the trait being measured – ability, severity, difficulty.

Figure 1. Variable map



The second column shows the measures of the 10 students each represented by a '*'. Students are ordered with the most able at the top, and the least able at the bottom. The students' ability measures range from +2.8 to -2.8 logits indicating a good spread of students at different ability levels. A detailed analysis of student ability, infit and outfit statistics are presented in Table 1. The purpose of fit statistics in the Rasch model is to ascertain the suitability of the data to construct variables and to identify intra- and inter-rater judge consistency in order to make measures and the recommended range is 0.5 to 1.5 logits (Lunz & Stahl, 1990). Put simply, fit statistics are quality control indices.

The third column shows the severity levels of the thirteen raters. The most severe rater is at the top and the least severe at the bottom. The severity measures for raters are spread over a range, the most severe being +0.87 logits and the least severe at -1.59 logits. Most raters are clustered at or around the 0 logit measure. A detailed analysis of rater severity infit, standard error, infit and outfit statistics are presented in Table 3.

The fourth column shows the experience of the raters and how it affects the way they rate. It can be seen that raters with more experience are more severe in their rating than raters with less experience but the difference does not seem to be large.

The fifth column shows the difficulty level for items. We note that language accuracy is highest on the logit scale and this means that it is harder to score high on this item. Conversely, language fluency is lowest on the logit scale with content and organization close together in the middle of the scale. The difference between language accuracy and language fluency is about 1 logit in measure. A detailed analysis of item difficulty with standard error measurement, infit statistics are presented in Table 5.

The last four columns on the right show the three band rating scale used to score student responses. Each item (Content = S1, Organization = S2, Language Accuracy = S3, Language Fluency = S4) has its own rating. The horizontal dash lines or the threshold levels show the point at which the likelihood of scoring the next higher rating starts to exceed the likelihood of scoring a lower rating for that item. For example, for Item 4 (Language Fluency), students with ability measures from above -1.6 logits to slightly below +1.5 logits are more likely to receive a rating of 2 than any other rating on the scale. A detailed analysis of rating scale functionality measurement is presented in Table 6. A more detailed discussion of the results for the four facets is presented in the next section.

7.2. Students

Table 1 shows the students ability measures expressed in logits in the second column. Positive logits show higher ability and negative indices show lower ability. Students' ability levels range from -2.76 to 2.87. Reliability of separation is 0.97. Separation reliability (similar to KR-20 internal consistency statistic) demonstrates the proportion of observed variance not due to measurement error (Wright & Masters, 1982). It is important that the ratings should achieve good examinee separation in order to distinguish students of different ability levels. A higher examinee separation ratio indicates a more discriminating rating scale (Knoch, 2009). The separation index is 8.53 indicating that the raters are separating the students into at least 8 levels. The separation of students across different ability levels indicates a well-constructed rating scale.

Table 1 also shows that students' fit statistics are all within the acceptable range of 0.5 to 1.5. As explained, fit indices are quality control indices and acceptable indices suggest that

the students' performance is within model expectations. Student 5 has a rather small infit and a check of all the raters' scores show that the ratings across the different criteria are similar, which explains the small infit. Interestingly, the validation team had also scored the script as 2,2,2,2 across all the items. This confirms the quantitative analysis from the Facets output. The acceptable fit statistics of these students indicate that the individual rater's scores are not deviating much from one another, i.e., that raters' scores for the essays are quite close to each other.

Table 1. Students' measures, infit, and outfit indices

Student No.	Measure	InfitMS	InfitZ	OutfitMS	OutfitZ
3	-2.76	1.16	0.84	1.23	1.03
5	0.84	0.57	-2.54	0.55	-2.57
7	-2.57	0.98	-0.08	1.11	0.62
11	2.87	1.14	0.84	1.02	0.16
17	0.7	0.97	-0.1	0.93	-0.25
18	-1.26	0.74	-1.81	0.74	-1.75
19	1.51	1.1	0.63	1.1	0.56
20	1.82	1.02	0.17	1	0.06
22	-0.57	0.88	-0.59	0.87	-0.69
28	0.62	1.28	1.32	1.33	1.47

Table 2 compares students' Fair Measures with the raw score assigned by the validation team. This comparison is done to ascertain if the Facets Fair Measures are close to those assigned by the validation team who revised the descriptors. The Fair Measure is the Facets calculated score which has been "adjusted" to take into account construct irrelevant facets that could possibly be intervening in the "true measure" or the score that the student should get if raters were all equally lenient. In other words, the Facets program has partialed out the effect of raters' severity on the students' measures to obtain the students' Fair Measure.

Table 2. Comparison of validation team score and fair measure

Student No.	Validation Team	Facets Fair Measure
3	1	1.28
5	2	2.22
7	1.5	1.32
11	2.75	2.75
17	2.25	2.18
18	1.75	1.66
19	2.25	2.41
20	2	2.5

22	1.75	1.83
28	2.5	2.16

The validation team's score is a weighted raw score of the four items. We can see that most of the scores of the validation team are close to those produced by Facets from the operational raters' scores. This indicates that the raters are likely interpreting the descriptors in a manner consistent with that by the validation team. It also demonstrates that Facets could be effective software to handle the issues related to inter-rater differences as the scores produced by Facets is close to those by the validation team who are the benchmark raters.

7.3. Raters

The Facets analysis shows that the 13 raters have varying levels of severity but all raters are performing consistently (intra-rater reliability) which suggests that they are interpreting the revised descriptors in a consistent manner.

Table 3 shows the raters' measures of severity range from 0.87 (most severe) to -1.59 logits (least severe). Reliability of separation is 0.80. Rater separation is "a measure of the spread of the rater severity measures relative to the precision of those measures" (Myford & Wolfe, 2004, p.195). This means that the differences between rater severities are 0.8 times greater than the error with which these severities are measured. The rater separation index is 1.98. This index presents the number of statistically distinct levels of rater severity among the raters. The Mean Standard error of measurement is 0.32. The rater separation index is not large (possibly two distinct levels of rater severity) and this indicates that raters are quite alike in their rating. Standard deviation is also rather small at 0.64.

Similar to Table 1 for students' measures, the fit statistics (infit and outfit) for raters are listed in Table 3. The fit statistics indicate the degree to which each rater's ordering of candidates is consistent with the estimated candidate ability measures (Lunz, Stahl, Wright & Linacre, 1989). We can see that there is no misfitting rater. Infit mean squares were all between 0.5 to 1.5, i.e., they are within the region of acceptable fit. An infit mean square value that is close to 1 means that the rater is closer to model expected ratings. Fit statistics of 1.5 or greater indicate too much unpredictability in raters' scores, while fit statistics of 0.5 or less indicate overfit, or not enough variation in scores.

Taken together, these indices in Table 3 point to a well-functioning rating scale and indicate that raters are able to use the revised descriptors in a consistent manner, despite their differences in severity levels.

Table 3. Raters' measures and infit and outfit indices

Rater No.	Measure	S.E.	InfitMS	OutfitMS
1	-1.26	0.33	0.77	0.77
2	-0.23	0.32	1.14	1.18
3	0.85	0.33	0.84	0.9
4	-0.13	0.32	0.75	0.73
5	-0.16	0.32	0.73	0.68
6	0.24	0.32	1.23	1.34

7	0.74	0.32	1.01	1.2
8	0.36	0.32	1.26	1.22
9	0.87	0.32	0.89	0.97
10	-0.06	0.32	1.43	1.44
11	0.44	0.32	0.89	0.85
12	-1.59	0.34	0.9	0.8
13	-0.06	0.32	0.76	0.75

7.4. Bias Interactions

Facets provides bias interaction analysis to show if raters are exhibiting bias that cannot be accounted for under the model. Bias analysis in facets measures rater variability in relation to other facets in the model (Schaeffer, 2008) and refers to rater severity or leniency in scoring. Engelhard defines it as "the tendency on the part of raters to consistently provide ratings that are lower or higher than is warranted by students' performances" (1994, p.84) Non-significant bias may point to, among other things, a well-constructed rating scale. In this study, bias measures are all less than ±2 and the interactions are non-significant (all p values are > 0.05) for all operational raters.

7.5. Raters' Experience

The facet of raters' experience shows that raters with more experience are more severe than raters' with less experience but the difference is non-significant (p = 0.21). This indicates that rater's experience with marking the test does not have a significant contributory effect on the students' placement. Insignificant bias and raters' fit statistics (within acceptable range) suggest that the operational raters are able to use the descriptors well in scoring, regardless of the level of experience (see Table 4). It also shows that experienced and inexperienced raters alike are able to use the revised descriptors consistently.

Table 4. Raters' experience and infit and outfit indices

Experience	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ
> 4 times	0.12	0.11	0.96	-0.53	0.99	-0.07
< 4 times	-0.12	0.14	0.99	-0.08	0.98	-0.16

7.6. Items

In Table 5 we see the items (criteria for marking) had varying levels of difficulty and the reliability of separation is 0.64 and p = 0.01 (significant). Whereas positive measures in students' data indicate higher ability and for raters they indicate greater severity, positive measures here means that the item was harder to score well on. The item Language Accuracy is the hardest to score well on, while Language Fluency is easiest to score well on. The items, Content and Organization lie between them. The fact that these items have different difficulty measures, especially for the Language Accuracy and Language Fluency which categories underwent substantial revision, indicate that the raters are not treating them in the same way when they are rating. In other words, they are able to discriminate among these items. The fit statistics are also within the acceptable range and there is no overfitting item. Overfitting items indicate that raters may have difficulty separating the different items and suggests that

items may be merged instead. In sum, the difference in difficulty levels of the items and fit statistics indicates that raters are able to attend to the separation of the different criteria in the revised descriptor.

Table 5. Item difficulty and infit and outfit indices

Items	Measure	S.E.	InfitMS	InfitZ	OutfitMS	OutfitZ
Content	-0.03	0.18	0.96	-0.35	0.98	-0.13
Organization	0.14	0.18	1.07	0.63	1.14	1.13
Lang Accuracy	0.35	0.19	0.94	-0.5	0.92	-0.6
Lang Fluency	-0.46	0.17	0.92	-0.68	0.91	-0.62

7.7. Rating scale functionality

To examine if the rating scale is functioning well, we studied the average student ability measure and outfit mean-square index of the rating scale. To obtain the average student ability measure, the student ability measure for all students receiving a rating in that category on that item is averaged. We expect to see that the average student ability measures will increase as the rating scale categories increase. An example of category statistics for the language accuracy criteria is shown in Table 6. In our study, we note that average student ability measures increase for all items as the rating categories increase. The average student ability measure values or the observed measure and expected ability measure values are close and the outfit mean square is near the expected value of 1.0. For all three items, the outfit mean square is close to 1, and this suggests that the rating scale is functioning well.

Table 6: Category statistics for language accuracy

Scale	% used	Average Measure	Expected Measure	Outfit MnSq
1	31%	-2.54	-2.33	0.8
2	76%	0.11	-0.01	0.9
3	23%	1.58	1.70	1.0

8. Conclusion

This study has shown that the MFRM is a useful tool to validate the functionality of descriptors. The Facets output shows that the operational raters were able to use the descriptors to discriminate students' performance categories and performance levels. Fit indices also show that raters were able to use the descriptors to score the essays in an internally consistent manner despite differences in severity levels. It also shows that raters' experience in scoring did not have a significant difference in raters' performance. The analysis indicates that the revised descriptor is functioning well and suitable for operational use.

This small-scale study has demonstrated that the MFRM can be used to examine raters' scoring behavior in order to assess the validity of an analytic rating scale. The data generated allows us to easily determine if students are reliably separated and most

importantly if raters are rating consistently and analytically. The data also allows us to assess the items to determine if categories of a rating scale are well separated and discriminated by raters. In other words, the MFRM can be used to help test developers see if marking descriptors are being used as intended. Where resources are limited and other more labor intensive methods of validation might be difficult to employ, the MFRM offers researchers and teachers an efficient method to generate robust quantitative data to analyze the use of rating scales.

References

Brown, J.D. & Bailey, K.M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning*, *34*, 21-24.

Engelhard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*(3), 171-191.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93-112.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal, *Large-Scale Assessment programs for All Students: Validity, Technical Adequacy, and Implementation* (pp. 261-187). Mahwah: Lawrence Erlbaum Associates.

Knoch, U. (2009). *Diagnostic assessment of writing: The development and validation of a rating scale*. Frankfurt: Peter Lang

McNamara**,** T. (1996) *Measuring Second Language Performance*. London and New York: Addison Wesley Longman

Myford, C.M. & Wolfe, E.W. (2004). Detecting and measuring rater effects using many-facets Rasch measurement. *Journal of Applied Measurement 5*(2), 189-227.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*, 3-31.

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

Linacre, J. M. (2010). Facets Rasch measurement computer program, version 3.67.1. Chicago: Winsteps.com.

Lunz, M.E. & Stahl, J.A. (1990). Judge consistency and severity across grading periods. *Evaluation & The Health Professional*. *13*(4), 425-444.

Lunz, M.E., Wright, B.D. & Linacre, J.M. (1990) Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*. 3(4), 331-345.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. (Expanded edition, Chicago: University of Chicago Press, 1980).

Schaeffer, E. (2008). Rater bias patterns in an EFL writing assessment.*Language Testing 25*(4), 465-493.

Smagorinsky, P. (1994). Think-aloud protocol analysis: Beyond th eblack box. In P. Smagorinsky (ed.), *Speaking about writing: Reflections on research methodology*, (pp. 3-19). Thousand Oaks, CA: Sage.

Stratman, J.F. & Hamp-Lyons, L. (1994). Reactivity in concurrent think-aloud protocols:Issues for research. In P. Smagorinsky (ed.), *Speaking about writing: Reflections on research methodology*, (pp. 89-111). Thousand Oaks, CA: Sage.

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