

# STUDY OF SOURCES OF SCORE VARIABILITY IN PERFORMANCE ASSESSMENT USING MFRM : A CASE OF SPEAKING TEST IN PETS BAND3

Zhang Jie

Guangdong University of Foreign Studies / Zhejiang University

He Lianzhen

Zhejiang University

## Abstract

As direct measure of learners communicative language ability , performance assessment ( typically writing and speaking assessment ) claims construct validity and a strong power for predictive utility of test scores . However , it is also of common concern that the subjectivity of rating process and the potential unfairness for test takers who encounter different writing prompts and speaking tasks would constitute threats to reliability and validity of test scores , especially in those large scale and high stakes tests . Therefore , appropriate means for quality control of subjective scoring should be held essential in test administration and validation . Based upon raw scores from one administration of speaking test in PETS Band3 held in Hangzhou , the present study investigates and models possible sources of score variability within the framework of Many Facet Rasch Model ( MFRM ) . MFRM conceptualizes the possibility of a examinee being awarded a certain score as a function of several facets —examinee ability , rater severity , domain difficulty and step difficulty between the adjacent score categories and provides estimates of the extent to which the examinee's test score is influenced by those facets . Model construction and data analysis was carried out in FACETS Version 3 .58 , computer program for conducting MFRM analysis . The results demonstrate statistically significant differences within each facet . Despite the generally acceptable rater consistency across examinees and rating domains , fit statistics indicate some unexpected rating patterns in certain raters such as inconsistency and central tendency , to be avoided through future rater training . Fair scores for each examinee are also provided , minimizing the variability due to facets other than examinees ability . MFRM manifests itself as effective in detecting whether each test method facet functions as intended in performance assessment and providing useful feedback for quality control of subjective scoring .

## Key words

PETS ; speaking test ; quality control ; many facet Rasch model ( MFRM )

## 1 . Introduction

### 1 .1 New challenges to reliability and validity in performance assessment

With increasing emphasis on teaching and assessing learners communicative language ability , as well

---

The present study is part of Educational Measurement Research Project sponsored by 2006 National Education Science Research Plan of National Education Examinations Authority . ( 教育部考试中心全国十一五教育科学规划“教育考试科学专设课题”课题编号 2006JKS3052 )

as criticism against the traditional paper and pencil test in multiple choice format, performance assessment (also called alternative assessment) is gaining momentum in the recent decades (Shohamy 1995; McNamara 1996, 1997; Norris et al. 1998). Large scale performance tests, typically on writing or speaking proficiency, has been widely incorporated into many important and well-recognized test batteries (TOEFL, IELTS, PETS, CET, NMET etc.). While performance testing is often automatically regarded as authentic, direct and valid indicators of examinees communicative language competence in real-life language use, it is still indispensable for a performance test to go through the process of test validation to justify the interpretation and use based upon test scores (McNamara 1996; Bachman 2002a; Fulcher 2003).

As a type of performance testing, speaking tests intend to measure examinees proficiency in oral communication by eliciting oral discourse (the actual performance) through a language speaking task or a series of tasks on certain topics. Like the majority of performance tests, speaking tests require subjective judgment by the raters against an established rating scale on test performance. The introduction of subjective judgment in scoring and other task facets, therefore, brings sources of score variability other than that from examinees underlying ability of interest, which constitute new challenges for ensuring reliability and validity of the test.

Figure 1. Adapted from Bachman's (2002) model of oral test performance

Figure 1 is a graph adapted from Bachman's model of oral test performance to illustrate the facets involved in a performance test, particularly in a speaking test. We might see that facets like the examiners and task characteristics would influence examinees test performance, but raters following rating rubrics to give the final score would act as medium intervening between test performance and the final score. Therefore, how accurately and consistently raters carry out their job determine to a large degree the accuracy of the scores as indicator of examinees underlying speaking proficiency.

## 1.2 Necessity of quality control for subjective scoring

As is shown above, in speaking tests or other performance tests, raters play the decisive role of translating the test performance into test score. When engaged in the act of rating, raters do not simply record what they see. Rather, their ratings are rooted in observation, interpretation, and perhaps most importantly, the exercise of personal and professional judgment (Myford & Wolfe 2003). Historically, most of the research is concerned with rater consistency, including inter-rater consistency and intra-rater consistency (Fulcher 2003) and other relatively systematic and stable differences among raters. Among those systematic effects, difference in rater severity, central tendency, halo effect and drift over time are the most discussed and researched in the field of performance testing (Lumley & McNamara 1995; Weigle 1998; Linn & Gronlund 2000; Fulcher 2003; Myford & Wolfe 2003, 2004; Hoskens & Wilson 2001). Individual differences in rater performance may result from the raters unfamiliarity with or inadequate training in the rating scale, their different interpretations of the rating criteria, tendency to

play safe (central tendency), halo effect (the tendency to allow impression on one aspect of examinees performance to influence ratings on other distinct aspects) or their personal beliefs that conflict with the criteria espoused by the rating rubric (Wolfe et al., 1997; Myford & Wolfe 2003). That is why rater training and the effect of it are drawing more and more attention (Weigle 1994, 1998; Lumley and McNamara 1995; Fulcher 2003).

However, there is also substantial evidence from previous studies that raters do differ significantly in severity even if they rate self consistently after rater training (Weigle 1998; Lumley, McNamara 1995; Wolfe et al., 1997; Lunz, Wright & Linacre 1990). As Weigle (1998) put it, idiosyncracies are inevitable and hard to remove in operational rating sessions. Therefore, appropriate and effective post hoc quality control of the accuracy of ratings should be conducted to guarantee scoring reliability and test fairness.

The objectives of quality control are as follows:

- 1) To detect raters who might be rating in a deviant way;
- 2) To provide specific feedback for raters in question and thus advice for them to improve scoring the next time;
- 3) To pinpoint examinees whose ratings might be inaccurate so that experts can conduct efficient check on the ratings and make necessary adjustment.

## 2. Using Many-Facet Rasch Model as a Means of Quality Control

Many facet Rasch Model (MFRM) is an extension of the one parameter Rasch model in Item Response Theory (IRT) which conceptualizes response on a specific test item as a function of the examinees ability and the item difficulty. MFRM (Linacre 1989, 1994) inherits the essence of this logistic model but further develops this prototype to include investigation into more complicated situations where test performance is not dichotomously scored but judged against a defined rating scale and is thus influenced by more factors (facets) than the examinee's ability and test item difficulty. This advance in IRT greatly promotes and facilitates analysis of more facets in the process of test administration, especially in the context of performance assessment with more test method facets than just examinee and item, as I have shown previously in the graph.

FACETS (Linacre 2005), a Rasch based computer program for MFRM analysis, could handle a model with maximum facets of nine, which introduces flexibility into the experimental design. Different analysis models can be constructed according to the specific situation and research questions. Researchers could define different facets they would like to investigate. Technically, MFRM does not require fully crossed data for analysis but it is still necessary to guarantee adequate connectivity in data so as to eliminate ambiguity in estimation and interpretation of results (Myford & Wolfe 2000).

In data analysis, MFRM provides rich statistical output including both facet level and individual level statistics for each facet and each element in that facet. For each facet, we could get overall estimates concerning the significance of separation among element measures within the facet relative to measurement error, including separation ratio, reliability, and chi square statistics. These estimates indicate how reliably and consistently elements in one facet separate with each other. For each element (e.g. individual examinee, rater, etc.), MFRM provides a measure, a standard error, and fit statistics, among which fit statistics, including Outfit and Infit indices, are very informative and important for both diagnostic and quality control purposes. Both Outfit and Infit statistics have an expected mean of 1.0 and those values which are beyond certain acceptable range would be labeled as exhibiting unexpected inconsistency with or over dependency of other elements. The misfitting/overfitting data are then further examined to provide useful feedback for expert re-rating, rater training, test construction and administration. Apart from these, MFRM can also conduct bias analysis to reveal some sub patterns of how data might deviate from what the model predicts.

Given the effectiveness of MFRM in handling problems in the context of performance assessment, there are quite a number of previous researches and studies dealing with issues in both writing and speaking assessment. (Lynch & McNamara 1998; Park 2004; Bachman et al., 1995; Bonk & Ockey 2003; Engelhard 1994)

### 3 . Speaking Test of Public English Testing System ( PETS ) Band 3

The speaking test is a compulsory component of PETS ( Public English Testing System ) Band 3 in China ,held twice every year . Examinees are assessed in a group of two or three . There are altogether three parts in the whole test procedure . Part 1 requires each candidate to make a brief self introduction followed by answering a question related to certain topic area . In Part 2 ,each candidate is first given a visual prompt on a card and is asked to have a 3 -minute group discussion on the given topic related to the visual prompt .In Part 3 ,the interlocutor gives each examinee a different picture and asks them to talk on their own something about what they see for 1 .5 -minute . After that ,the other examinee in the same group may have some comments on what their partner has said within half minute .

All the examinees are assessed according to their overall performance in the whole task . Each test room has two raters ,one is responsible for delivering test requirements and interacting with examinees while at the same time assesses examinees speaking ability on a 5 point holistic rating scale ;the other rater focuses exclusively on scoring but is required to assess the examinees analytically in four domains ( Grammar and vocabulary ,Discourse management ,Pronunciation and Interactional Competence ) . The four sub criteria are also defined on a 5 point rating scale with specific descriptors at each level .

### 4 . The Present Study

#### 4 .1 Aims

The present study ,based on the raw scores from one administration of speaking test in PETS Band 3 held in Hangzhou ,aims to achieve the following objectives :

- 1 ) To investigate whether raters could follow the rating rubrics to discriminate different levels of speaking proficiency of examinees and conduct rating self consistently ;
- 2 ) To detect raters who might exhibit deviant manner of rating ( too lenient /severe or with inadequate consistency ) ;
- 3 ) To detect examinees who might have been awarded inaccurate ratings and thus need further check and re rating ;
- 4 ) To illustrate how MFRM could be utilized for quality control purpose in operational rating session in a speaking test .

#### 4 .2 Data collection

Data collected for analysis are one holistic score and four analytic scores for each of the total 322 examinees who took part in the speaking test of PETS Band 3 held in Hangzhou in March 2007 ,together with their scores of the paper based part of the test ( involving sections of listening ,use of English , reading comprehension and writing ) . In test administration ,examinees were randomly assigned to 17 test rooms each of which had two raters ,one responsible for delivering the test and scoring holistically and the other exclusively for assigning scores analytically . All ratings from the 34 raters were then collected for data analysis .

#### 4 .3 Methods

##### 1 ) Constructing models for analysis in MFRM

Model construction and data analysis was carried out in FACETS Version 3 .58 ,computer program for conducting MFRM analysis . Since two scoring modes were adopted in the test ,different mathematical models for MFRM analysis are constructed .

For holistic scoring mode ,the model takes the following mathematical form ,

$$\text{Log} ( P_{nik} / P_{ni(k-1)} ) = B_n - D_i - F_k$$

The left side of the equation stands for the possibility the examinee  $n$  being awarded the score  $k$  whereas the right side are linear addition of factors which might contribute to this possibility . Specifically , $B_n$  stands for the ability measure of examinee  $n$  , $D_i$  stands for the severity measure of rater  $i$  who awards the score to examinee  $n$  ,and  $F_k$  indicates the step difficulty from score category  $(k-1)$  to  $k$  .

It should be noted that the minus signs before  $D_i$  and  $F_k$  indicate that those factors are defined as challenges for certain examinee to be awarded higher scores while  $B_n$  is positive in that it will add to the chances of the examinee to obtain better scores.

The equation for analytic scoring mode is different from that for holistic mode only in that there is one more facet involved in the right side of the equation.

$$\text{Log} \left( \frac{P_{nik}}{P_{ni(k-1)}} \right) = B_n - D_i - E_j - F_{jk}$$

Here  $E_j$  stands for the difficulty level of  $j^{\text{th}}$  rating domain in the analytic rating scale. Another difference is that partial credit model is adopted because we assume the rating scale of each rating domain is different from each other so as to investigate whether raters would exhibit different level of severity or consistency in rating different domains of speaking proficiency.

## 2) Establishing external criterion for comparison

As is mentioned above, to estimate rating accuracy there should be an external criterion whether it being other ratings from a second or third rater or other measures of examinees' relevant ability. In this case, although the scoring procedure involves two raters for each examinee, they are following totally different rubrics to award the scores. Therefore, the scores awarded by these two raters could not be readily and directly compared.

With this in mind, we tentatively propose a solution, that is, to conceptualize scores from the examinees' paper-based part of the same test as ratings awarded by a common "monitoring rater". This monitoring rater is relatively objective in assessing the examinees' overall English proficiency, which should generally agree with the ratings obtained from the speaking test. The introduction of examinees' paper-based scores as scores awarded by a third rater could on the one hand compensate for the lack of connectivity in data and on the other hand establish an external criterion for the estimation of rating accuracy and consistency of raters.

For data analysis, we assume that if raters conduct rating in an accurate and consistent way their pattern of ratings should be generally in accordance with the monitoring rater. Under this assumption, if the severity measure and infit statistics (consistency) of certain raters are in large discrepancy with the monitoring rater or beyond some pre-defined acceptable limits, we may well conclude that this rater is probably inadequately carrying out his/her job. In addition, examinees who have unacceptable fit statistics or have large discrepancy in their observed ratings and fair scores provided by MFRM analysis would be flagged for further investigation and expert re-rating.

## 5. Results and Discussion

### 5.1 Facet map

Facet map is a graphical summary for the statistical output from MFRM analysis. Figure 2 is facet map for holistic scoring mode and Figure 3 for analytic scoring mode.

For each map, the leftmost column is the common scale in the unit of logit against which all the measures are calibrated. For each following column, the elements in each facet are located according to their own measures on the scale. Examinees with higher ability level, raters more severe and items more difficult are positioned at the top of the map. It is worth noting that the facet map for analytical scoring mode has four separate columns of scales (from  $S_1$  to  $S_4$ ), which represent the four subscales for Grammar and vocabulary, Discourse management, Pronunciation and Interactional Competence, respectively. The horizontal lines across each column indicate the point at which the likelihood of getting the next higher rating begins to exceed the likelihood of getting the next lower rating for a given item.

The facet maps could serve as a visual aid to the interpretation of the statistical results, from which we can get a quick overview of the distribution of those variables on the common ground and obtain a rough idea about how the elements in each facet relate to one another.

Figure 2 . Facet map for holistic scoring

Figure 3 . Facet map for analytical scoring

5 .2 Rater report

Table 1 . Rate report for holistic scoring mode

Rater	Measure	S .E .	Infit Mnsq	Zstd
9	3 .57	0 .41	1 .02	0 .1
1	1 .48	0 .41	0 .93	0
12	1 .43	0 .38	1 .32	1
14	1 .14	0 .34	1 .2	0 .7
13	1 .07	0 .4	0 .82	-0 .4
3	.92	0 .4	0 .79	-0 .5
6	.68	0 .37	0 .78	-0 .6
2	.39	0 .43	0 .9	-0 .1
17	-.03	0 .33	0 .83	-0 .5
18	-.15	0 .09	0 .94	-0 .7
7	-.27	0 .38	0 .53	-1 .7
8	-.32	0 .44	0 .80	-0 .4
4	-.53	0 .45	0 .69	-0 .7
16	-.63	0 .36	0 .94	-0 .2
15	-1 .39	0 .43	1 .36	1 .1
10	-1 .45	0 .39	1 .23	0 .7
5	-2 .16	0 .41	0 .99	.0
11	-3 .74	0 .41	0 .73	-0 .4

Separation 4 .08 Reliability ( not inter-rater ) .94 chi square :275 .1 d .f . :17 significance (probability) :.00

Table 1 summarizes statistics for rater facet in holistic scoring mode. At the bottom of the table are some overall facet level statistics. The separation index and reliability are 4.08 and 0.94, respectively. Generally speaking, a separation ratio of more than 2 and reliability index larger than 0.9 could well indicate statistically significant difference among all the elements, which in the rater facet means that there is significant and consistent difference in rater severity among the raters.

For each rater, the measure column refers to their level of severity across all the examinee they rate. Note that the 18th rater represents the monitoring rater as mentioned above. It is anchored in the estimation so that its measure would be around the middle of the logit scale to provide a common ground on which the severity of other raters could be compared. From this column, it is found that the raters from 9th and 11th test rooms are extremely severe or lenient toward examinees compared with others. The column of Infit Mnsq indicates to what extent the raters could rate self-consistently, as the model expects. There is no hard and fast rule for interpreting fit statistics. Considering the small number of ratings each rater awarded, a lenient criteria (0.6-1.4) is adopted in the present study for detecting raters with rating problems.

As Table 1 indicates, the fit statistics of the holistic raters show that the majority could rate rather consistently and independently. Only the rater from the 7th test room has the infit index less than 0.6, suggesting that the ratings awarded by him exhibit less variability compared to model expectation which might be an indicator of central tendency or restriction of range during scoring.

Table 2. Rate report for analytic scoring mode

Rater	Measure	S.E.	Infit Mnsq	Zstd
9	2.27	0.17	1.45	2.4
12	1.23	0.16	1.52	3
1	0.79	0.18	1.02	0.1
14	0.79	0.15	1.32	1.9
13	0.54	0.17	1.13	0.8
6	0.44	0.16	0.89	0.7
3	0.37	0.17	0.86	0.8
18	0.02	0.04	0.78	6.2
2	-0.01	0.17	1.42	2.5
4	-0.01	0.18	1.3	1.6
17	-0.20	0.14	0.81	1.4
15	-0.34	0.19	1.48	2.4
16	-0.38	0.16	1.15	1
8	-0.41	0.19	0.86	0.7
10	-0.69	0.17	1.49	2.6
7	-0.71	0.16	0.71	2
5	-1.40	0.18	0.92	0.4
11	-2.30	0.18	1.09	0.5

Separation 6.02 Reliability (not inter rater) .97 Fixed (all same) chi square :588.1 d.f. :17  
significance (probability) :.00

From the statistics for rater facet in analytic scoring mode shown in Table 2, we may find that the pattern of variation in rater severity in analytic mode is similar with that in holistic mode, both for the overall facet level statistics and the severity measure. However, the infit indices show that there are 5 raters with their infit value larger than 1.4, the upper limit for quality control, indicating that the ratings

awarded by these five analytic raters have larger variance compared to the model expectation .It might be inferred that these raters may rate inconsistently across the four rating domains and examinee group . Therefore ,it is worth further investigation that whether more raters exhibit inconsistent rating manner in analytic rating mode is due to the incompetence of raters themselves or the inadequacies in analytic rating rubrics and rater training .

5 .3 Item report

Table 3 . Item statistics in analytic scoring mode

Item	Measure	S .E .	Infit Mnsq	Zstd
2	.15	.06	.91	1 .7
1	.09	.06	.92	1 .5
4	-.03	.06	.99	-.1
3	-.21	.06	1 .01	.2

Separation 2 .17 Reliability .83 chi square :22 .7 d f . :3 significance (probability) : .00

The four items in Table 3 correspond to the four distinct rating domains as above mentioned . The facet level statistics show that the four rating domains are statistically different in their difficulty measure . However ,the absolute difference in values is quite small ,indicating that raters are similarly severe /lenient when rating the four domains . The infit statistics all fall into the quality control limit , manifesting considerable overall consistency of raters when they assessed the four domains .

5 .4 Examinee report

Table 4 . Examinee report (abridged ) for analytical scoring mode

Examinee	Obsvd Average	Fair Average	Measure	Model S .E .	Infit MnSq	Zstd
175	4 .9	4 .83	4 .31	1 .04	1 .02	0 .3
85	4 .8	4 .79	4 .07	0 .76	0 .79	0 .1
185	4 .9	4 .77	3 .97	1 .06	1 .30	0 .6
...	...	...	...	...	...	...
17	2 .1	2 .12	1 .07	0 .47	2 .36	2 .5
118	1 .3	1 .17	4 .21	0 .75	0 .68	0 .3
266	1 .1	1 .10	4 .75	1 .03	0 .81	0 .1

Separation Ratio =3 .49 reliability =0 .92 chi square = 3820 .1 d f . :321 ,p = .00

Given the large number of examinees ,we just list some examinee statistics for illustration ,as shown in Table 4 . For each examinee ,MFRM provides estimates of fair average score and infit mean square index . The former refers to scores expected by the model after making some adjustment due to influences from facets other than examinees ability . For instance ,if the examinee encounters an especially severe rater ,his fair average would be correspondingly higher than his observed average score ,after compensation for the influence from the severe rater . Infit Mnsq is a statistical index describing the extent to which an examinee s ratings are in agreement with what the model predicts . Similar with rater facet ,Infit Mnsq larger or smaller than certain limits would indicate that problems or inaccuracies exist in the ratings awarded .

For quality control purpose ,these two output statistics could be used as criteria for picking out examinees whose ratings might be inaccurate and need further check or expert re rating . That is ,examinees who have large difference between their observed average and fair average score and those whose Infit values are beyond some pre defined limits could be flagged as targets for review and further investigation .

In the present study ,we have flagged 10 examinees (322 in total ) whose Infit value are the largest /

smallest with  $|Z_{std}| > 2$  in holistic scoring mode and 15 in analytic scoring mode. We also picked out 15 examinees for each mode whose observed score and fair score has the largest difference. Compared with others, those examinees are more likely to have been awarded higher or lower scores they deserve. The specific number to be flagged, of course, depends much upon the total number of examinees and the maximum examinees quality control team can afford to re-examine.

Table 5. Facet statistics for holistic and analytic scoring

	Separation Ratio	Reliability	Chi square
Holistic Scoring	1.97	0.8	chi square = 1365.9 d.f. : 321, p = .00
Analytic Scoring	3.49	0.92	chi square = 3820.1 d.f. : 321, p = .00

For examinee facet, there is also some difference detected in the two scoring modes as shown in Table 5. In the holistic scoring mode, although the chi square test shows that examinees are significantly different in their proficiency, the separation ratio is only 1.97 and reliability index 0.8, indicating that holistic ratings could not very well discriminate among examinees, whereas the analytic scoring mode, as indicated by its separation ratio and reliability index, is comparatively better in discriminating candidates' ability.

## 6. Conclusion

Based on the above statistical output from FACETS, conclusions can be drawn as follows:

- 1) Generally speaking, the speaking test could effectively discriminate examinees according to their underlying speaking proficiency. However, the two scoring modes seem to have different power in doing so. As indicated by examinee facet level statistics, analytic scoring is comparatively better in discriminating candidates' ability than holistic scoring.
- 2) Raters differ significantly in their severity in both scoring modes. Two holistic raters were flagged as being too severe or lenient toward examinees. The majority of holistic raters could rate rather consistently and independently. Only one rater has Infit index under the quality control limit indicating that the ratings awarded by him exhibit less variability compared to model expectation which might be an indicator of central tendency or restriction of range during scoring. In analytical scoring mode, there are 5 raters with their Infit values larger than 1.4, the upper limit for quality control, indicating that the ratings awarded by these five analytic raters have larger variance compared to the model expectation. It is therefore worth further investigation that whether more raters exhibit inconsistent rating manner in analytic rating mode is due to the incompetence of raters themselves or the inadequacies in analytic rating rubrics and rater training.
- 3) Infit indices and fair average scores provided for each examinee can be utilized as criteria for picking out examinees whose ratings might be inaccurate and need further check or expert re-rating.
- 4) MFRM manifests itself as effective in detecting whether each test method facet functions as intended in performance assessment and providing useful feedback for further rater training and quality control of subjective scoring.

## 7. Implications and Further Study

By using MFRM, the present study examines and estimates main sources of score variability to investigate the extent to which variance in test scores are due to different facets involved in the context of performance assessment. By providing detailed statistics for individual raters and examinees, MFRM helps to pinpoint where the potential problems may lie and provide information for quality control of subjective scoring in the speaking test of PETS Band 3.

In the next stage of the present study, we would base on these statistical findings and use tape records from the operational test as well as post hoc expert re-rating to further check:

- 1) whether the raters labeled as too severe / lenient or lacking adequate consistency by MFRM analysis really have the detected problems or not ;
- 2) whether the examinees picked out by the two criteria mentioned above really need re rating or not .

By doing so , we could find whether statistics from MFRM analysis are accurate and sensitive indicators of rater behavior and whether it is feasible to use such statistical indices as criteria to pick out raters who need further training as well examinees whose scores require further check and re rating .

## References

- Bachman ,L .F . 2002 . Some reflections on task based language performance assessment . *Language Testing* 19 / 4 :453 -476 .
- Bachman ,L .F . ,B .K .Lynch &M .Mason .1995 . Investigating variability in tasks and rater judgments in a performance test of foreign language speaking . *Language Testing* 12 :238 -257 .
- Bonk ,W .J . &G .J .Ockey .2003 . A many facet rasch analysis of the second language group oral discussion task . *Language Testing* 20 :89 -110 .
- Engelhard ,G .1994 . Examining rater errors in the assessment of written composition with a many faceted rasch model . *Journal of Educational Measurement* 31 :93 -112 .
- Fulcher ,G .2003 *Testing Second Language Speaking* . Longman /Pearson Education :London .
- Hoskens ,M . &M .Wilson .2001 . Realtime feedback on rater drift in constructed response items :An example from the Golden State Examination . *Journal of Educational Measurement* 38 /2 :121 -145 .
- Lumley ,T . &T .F .McNamara .1995 . Rater characteristics and rater bias :Implications for training . *Language Testing* 12 :54 -71 .
- Lunz ,M .E . ,J .A .Stahl &B .D .Wright .1996 . The invariance of judge severity calibrations . In M .R .Wilson &G .Engelhard Jr .(eds .) . *Objective Measurement Theory into Practice* ,3 ,pp .99 -112 . Norwood , NJ :Ablex .
- Linacre ,J .M .1989 ,1994 . *Many facet Rasch Measurement* . MESA Press :Chicago .
- Linacre ,J .M .2005 . *A User s Guide to FACETS :Rasch -Model Computer Program* . Chicago :MESA Press .
- Linn ,R .L . &N .E .Gronlund .2000 *Measurement and Assessment in Teaching* (8th ed .) . Columbus ,OH :Merrill .
- Lynch ,B .K . &T .M .McNamara .1998 . Using G theory and many face rasch measurement in the development of performance assessments of the ESLspeaking skill of im migrants . *Language Testing* 15 :158 -180 .
- McNamara ,T .F .1996 . *Measuring Second Language Performance* . London ;New York :Long man .
- McNamara ,T .F .1997 . Performance testing . In C .Clapham (ed .) . *Language Testing and Assessment* . Vol .7 of the *Encyclopedia of Language and Education* ,Kluwer Academic Publishers .
- Myford ,C .M . &E .W .Wolfe .2000 . Strengthening the ties that bind :improving the linking network in sparsely connected rating designs . TOEFL Technical Report TR 15 . Princeton ,NJ :Educational Testing Service .
- Myford ,C .M . &E .W .Wolfe .2003 . Detecting and measuring rater effects using many facet Rasch measurement :Part I . *Journal of Applied Measurement* 4 /4 :386 -422 .
- Myford ,C .M . &E .W .Wolfe .2004 . Detecting and measuring rater effects using many facet Rasch measurement :Part II . *Journal of Applied Measurement* 5 /2 :189 -227 .
- Norris ,J .M . ,J .D .Brown ,T .Hudson &J .Yoshioka .1998 . Designing second language performance assessment (vol . SLTCC Technical Report #18 ) . Honolulu :Second Language Teaching &Curriculum Center ,University of Hawaii at Manoa .
- Park ,T .2004 . An investigation of an ESL placement test of writing using many facet rasch measurement . Teachers College ,Columbia University , Working Paper in TESOL & Applied Linguistics 4 /1 .
- Shohamy ,E .1995 . Performance assessment in language testing . *Annual Review of Applied Linguistics* 15 :188 -211 .
- Weigle ,S .C .1994 . Effects of training on raters of ESLcompositions . *Language Testing* 11 :97 -223 .
- Weigle ,S .C .1998 . Using FACETS to model rater training effects . *Language Testing* 15 /2 :263 -287 .
- Wolfe ,E .W . ,Chiu &W .T .Chris .1997 . Detecting rater effects with a Multi faceted rating scale model . Paper presented at the annual meeting of the National Council on Measurement in Education . Chicago , IL ,March 25 -27 ,1997 .