DOCUMENT RESUME

ED 365 091 FL 021 398

AUTHOR	Lumley, Tom; McNamara, T. F.
TITLE	Rater Characteristics and Rater Bias: Implications for Training.
PUB DATE	93
NOTE	17p.; Paper presented at the Language Testing Research Colloquium (15th, Cambridge, England, United Kingdom, August 1993).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	English (Second Language); *Experimenter Characteristics; Foreign Countries; *Interrater Reliability; *Performance Tests; Role Playing; Second Language Learning; *Testing
IDENTIFIERS	Rasch Model; Rater Effects

ABSTRACT

Recent developments in multi-faceted Rasch measurement (Linacre, 1989) have made possible new kinds of investigations of aspects of performance assessments. Bias analysis, interactions between elements of any facet, can also be analyzed, which permits investigation of the way a particular aspect of the test situation may elicit a consistently biased pattern of responses from a rater. This research investigates the use of these analytical techniques in rater training for the speaking sub-test of the Occupational English Test (OET) administered in Australia, a specific purpose English-as-Second-Language performance test for health professionals. Data are presented from two rater training sessions (30 raters, 10 candidates) separated by a 6-month interval and an intervening operational test administration session (12 of the above raters, 100 candidates). The analysis is used to establish: (1) consistency of rater characteristics over the two or three occasions; and (2) rater bias in relation to role play materials and/or candidate type. The use of feedback to raters of the results of this analysis is also reported. (Contains 23 references.) (Author)

RATER CHARACTERISTICS AND RATER BIAS: IMPLICATIONS FOR TRAINING[1]

Tom Lumley and T.F. McNamara
NLLIA Language Testing Centre, University of Melbourne

Address for correspondence: NLLIA Language Testing Centre, Department of Applied Linguistics and Language Studies, The University of Melbourne, Parkville, Victoria 3052, AUSTRALIA
Phone: #61 3 344 4207 *Fax*: #61 3 344 5163
e-mail: tom_lumley@muwayf.unimelb.edu.au; t.mac@unimelb.edu.au

Abstract

Recent developments in multi-faceted Rasch measurement (Linacre, 1989) have made possible new kinds of investigation of aspects (or 'facets') of performance assessments. Relevant characteristics of such facets (for example, the relative harshness of individual raters, the relative difficulty of test tasks) are modelled and reflected in the resulting person ability measures.

In addition, bias analyses, that is, interactions between elements of any facet can also be analyzed. (For the facet 'person', an element is an individual candidate; for the facet 'rater', an element is an individual judge, and so on.) This permits investigation of the way a particular aspect of the test situation (type of candidate, choice of prompt, etc.) may elicit a consistently biased pattern of responses from a rater. Lunz and Stahl (1992) used these techniques to produce judge performance reports, which provide individual raters with information on their relative characteristics as raters, their consistency and any individual biased ratings, in a judge-mediated examination of histotechnology.

The purpose of the research is to investigate the use of these analytical techniques in rater training for the speaking sub-test of the Occupational English Test (OET), a specific purpose ESL performance tests for health professionals. The test involves a role-play based, profession-specific interaction, involving some degree of choice of role-play material. Data are presented from two rater training sessions (30 raters, 10 candidates) separated by a 6 month interval and an intervening operational test administration session (12 of the above raters, 100 candidates). The analysis is used to establish (1) consistency of rater characteristics over the two or three occasions and (2) rater bias in relation to role play materials and/or candidate type. The paper reports on the use of feedback to raters of the results of this analysis as part of the rater training process. It also addresses the question of the stability of rater characteristics over rating occasions, which has practical implications in terms of the accreditation of raters and the requirements of data analysis following test administration sessions. The paper also has research implications concerning the role of multi-faceted Rasch measurement in understanding rater behaviour in performance assessment contexts.

---

[1] This paper was originally presented at the Language Testing Research Colloquium, University of Cambridge, August. 1993. The research was made possible by a grant from the National Languages and Literacy Institute of Australia.

Introduction

The aim of this paper is two-fold: to address the question of the consistency of rater characteristics over time; and to investigate the potential of multi-faceted Rasch measurement in exploring this question. The paper thus has a substantive and a methodological focus. The substantive focus is on rater characteristics and their stability over time. The methodological focus is on the comparative advantages of routine analysis with anchoring versus bias analysis as analytical tools in addressing this question.

Rater variability

It has long been recognized (for at least a century, in fact) that variability in test scores associated with rater factors is extensive. For example, Edgeworth (1890: 653, quoted in Linacre, 1989: 10) states:

> *I find the element of chance in these public examinations to be such that only a fraction - from a third to two-thirds - of the successful candidates can be regarded as safe, above the danger of coming out unsuccessfully if a different set of equally competent judges had happened to be appointed.*

Huot (1990: 250-251) summarizes a study by Diederich *et al.* (1961), who

> *...analyzed 300 papers on two topics written by freshman at four Northeastern colleges. These 300 papers were scored on a 9-point scale by 53 readers... Ninety-four percent of the papers received at least 7 grades, and no paper received less than five separate scores.*

Cason and Cason (1984, quoted in Linacre, 1989: 28) found that differences in judge severity can account for as much variance in ratings as differences in examinee ability.

Nature of variability among judges

Traditional theory conceptualized rater characteristics in terms of the difference between an idealized judge (the 'perfect' examiner) and actual judges ('ordinary' examiners); these differences were essentially seen as something to be regretted; the 'shortcomings' of ordinary examiners were the problem. Differences between judges could be understood in terms of overall severity (or leniency) on the one hand, and randomness (error) on the other. Of these two elements, Harper and Misra (1976: 15) (quoted in Linacre, 1989: 15) found that the extent of error was as great as the extent of the differences between the mean scores allocated by judges (an indication of their overall severity), and more problematic (because it was harder to eliminate or compensate for; equating the mean scores given by judges is a fairly simple matter). Linacre (1989: 48-49; 51) uses the term *severity* to refer both to the overall severity of the rater and to differences between raters in the way they interpret rating scale thresholds for particular items; McNamara and Adams (1991: 3) suggest use of the term *rater characteristics* to cover both overall severity and more specific effects.

One way of dealing with error is to see if it can be further broken down to reveal sub-patterns in the behaviour of raters which may be systematic in some way, that is, predictable, and thus able to be compensated for. Raters may display particular patterns of harshness or leniency in relation to only one group of candidates, not others, or in relation to particular tasks, not others, or on one rating occasion, not the next. That is, there may be an *interaction* involving a rater and some other aspect of the assessment setting. Such an interaction, between rater and occasion of rating will be examined in this paper. In multi-

faceted measurement, such an interaction is termed *bias* (note that this is a particular use of this term in this context, and is not test bias in the more familiar sense).

Usefulness and limitations of rater training

Typically, in performance assessments, attempts are usually made to reduce the variability of judges' behaviour. For example, Ruch as early as 1929 stated that[2]:

> *Subjectivity of marking may be reduced about one-half by the adoption of and adherence to a set of scoring rules when essay examinations are to be graded.*

Traditionally, rater training attempted to reduce both variability associated with differences in overall severity, and randomness. The usual form of these attempts is rater training sessions, in which raters are introduced to the assessment criteria and asked to rate a series of carefully selected performances, usually illustrating a range of abilities and characteristic issues arising in the assessment. Ratings are carried out independently, and raters shown the extent to which they are in line with other raters and thus achieving a common interpretation of the rating criteria. The rating session is usually followed by additional follow-up ratings, and on the basis of these an estimate is made of the reliability and consistency of the rater's judgements, in order to determine whether the rater can participate satisfactorily in the rating process.

Surprisingly, the effectiveness of rater training has been little studied. Recently, however, our assumptions about the nature and effect of the rater training process have had to be reconsidered. For example, the elimination of differences between raters has itself been questioned as a desirable goal. Constable and Andrich (1984) raise this as an issue:

> *It is usually required to have two or more raters who are trained to agree on independent ratings of the same performance. It is suggested that such a requirement may produce a paradox of attenuation associated with item analysis, in which too high a correlation between items, while enhancing reliability, decreases validity.*

In the Japanese director Kurosawa's classic film *Rashomon*, the accounts of four witnesses to a dramatic incident are presented; they are profoundly different. Where does the truth lie? Each of the accounts is plausible, each deceptive, all frustratingly at odds with each other, but also, paradoxically, mutually illuminating. The same may be said (more trivially!) of assessments of human performance: in a matter of some complexity, no one judgement may be said to be definitive, although there is likely to be a considerable area of overlap between judgements. These differences and similarities need to be taken into account in determining the best estimate of a candidate's ability.

From the point of view of practicality, recent research has demonstrated the following:

1) In terms of *overall severity*, rater training can reduce but by no means eliminate the extent of rater variability. Rater training has the effect of reducing extreme differences - outliers in terms of harshness or leniency are brought into line (McIntyre, 1993). But significant and substantial differences between raters persist (see for example Tables 3, 4 and 6 below, where the reliability of the differences between raters ranges between .87 and .94; these figures are typical of raters in a range of performance assessment contexts). Lunz and Stahl (1990) argue that

---

[2] Quoted in Linacre, 1991: 7.

Judges often sense that they have unique standards, and it is hard for them to alter their standards.

This being the case, then attempts to deal adequately with differences in rater severity through rater training are bound to be only partially successful, in which case compensation for rater characteristics needs to be built into the rating process.

2) The main contribution of rater training is to reduce the random error in rater judgements. Rater training is successful in making raters more *self-consistent*, the most crucial quality in a rater, according to Wiseman (1949)[3]. Without this self-consistency, no orderly process of measurement can be conducted. Cushing (1993) shows that it is difficult to derive usable measures of the ability of candidates from untrained raters, even when attempts are made to adjust for rater characteristics using multi-faceted measurement techniques, because of the large randomness associated with the ratings of such raters.

Multi-faceted Rasch measurement

One of the most promising recent developments in understanding and controlling rater variability is multi-faceted Rasch measurement (Linacre, 1989), implemented through the computer program FACETS (Linacre and Wright, 1992). In this approach, the chances of success on a performance task are related to a number of aspects of the performance setting. These aspects, or facets, will include the ability of the candidate and the difficulty of the task, but also the characteristics of the rater and other characteristics of the context in which the performance is elicited and rated. These facets are related to each other as increasing or reducing the likelihood of a candidate of given ability getting a given score on a particular task. This is expressed in the following way (Figure 2):

Figure 2: Multi-faceted Rasch Measurement

Probability of a given score on a rating scale = B - D - J - K - O (etc)

where B = ability of candidate
D = difficulty of task
J = severity of judge
K = 'step' difficulty for the particular score point on the rating scale
O = other aspect (facet) of the assessment situation.

All of the terms in the equation are estimated as probabilities, expressed mathematically in units called logits.

The number of facets of potential interest is large, and research in the field at the moment is marked by a phase of exploration, in which various aspects of the assessment setting are being conceptualized and modelled using multi-faceted measurement. This research is motivated by two factors: a research motivation, to try to identify aspects of the assessment context which can be shown to significantly affect scores; and a practical motivation, to build in a compensation for those facets which can be shown to exert a significant influence on the chances of success in an examination. This paper, which examines the facet of occasion of rating, is a contribution to this ongoing task.

[3]'The efficiency of raters should be judged primarily by their self-consistency' (Wiseman, 1949: 208).

The stability of rater characteristics over time

The question of the stability of rater characteristics even over relatively short periods has been little considered in published research. Using traditional methods, Coffman and Kurfman (1968) and Wood and Wilson (1974) produced evidence of instability in marking behaviour in the course of an extended marking period when a large number of scripts are involved. Using multi-faceted measurement, Lunz and Stahl (1990) showed, in the context of an essay examination, a clinical examination and an oral examination, inconsistencies in judges' level of severity across half-day grading periods, within grading sessions of between one and a half and four days.

The question of the stability of rater characteristics over time is in fact made more pressing by the existence of the new technical possibilities of multi-faceted measurement (McNamara and Adams, 1991), the question arises as to whether it is reasonable to build what is known of a rater's characteristics at the time of rater training into the estimation of candidates' abilities at the time of the analysis of data from actual test administrations. Or do such characteristics need to be recalibrated in relation to the new data set, a procedure which will involve relatively complex design of the analysis? An additional feature of multi-faceted measurement which has the potential to be of use in the investigation of such issues is its capacity to investigate interactions between elements of facets, that is, interactions between particular raters and particular conditions of each facet of interest. It is possible, for example, that only certain raters may vary their characteristics across occasion, but not others, and that no overall or general pattern emerges across raters. In this case, instead of an across the board compensation, an appropriate strategy may be to give feedback to individual raters on these interactions, in the hope that this feedback will remove the unwanted interaction effect. But even here, further questions arise. If a rater's characteristics are successfully modified by training, are these changes stable over time, or does the rater revert to old habits? How often do raters need to be re-trained?

Aims of the research

This paper investigates the potential of the new analytical techniques offered by the program FACETS, in the context of rater training for the speaking sub-test of the Occupational English Test, a specific purpose ESL performance test for health professionals administered on behalf of the Australian Government. In this paper we consider the stability of ratings by a group of assessors on three occasions over a period of 20 months.

The specific substantive issues considered are:

1) whether or not trained raters of spoken performance demonstrate consistency in the level of severity of their assessments over time, and

2) what implications the findings might have for rater training.

The methodological issue of interest are:

1) to what extent multi-faceted Rasch measurment assists in investigating the issue of the stability of rater characteristics over time, and

2) specifically, the relative usefulness of the techniques of *anchoring* and of *bias analysis* in the investigation of this question.

Table 1: Items - OET speaking sub-test

OVERALL COMMUNICATIVE EFFECTIVENESS		
Near-native flexibility and range	____ : ____ : ____ : ____ : ____ : ____	Limited
INTELLIGIBILITY		
Intelligible	____ : ____ : ____ : ____ : ____ : ____	Unintelligible
FLUENCY		
Even	____ : ____ : ____ : ____ : ____ : ____	Uneven
COMPREHENSION		
Complete	____ : ____ : ____ : ____ : ____ : ____	Incomplete
APPROPRIATENESS OF LANGUAGE		
Appropriate	____ : ____ : ____ : ____ : ____ : ____	Inappropriate
RESOURCES OF GRAMMAR AND EXPRESSION		
Rich, flexible	____ : ____ : ____ : ____ : ____ : ____	Limited

The test

The Occupational English Test (McNamara 1990) is administered in Australia and overseas to members of 11 different health professions (doctors, nurses, dentists, vets, dietitians and physiotherapists, among others) who have obtained their professional qualifications overseas and who wish (after being accepted as immigrants or refugees) to practise in Australia. This study concerns itself with the speaking component of this test, which uses materials specific to the profession of each candidate. These materials take the form of simulated conversations, two per candidate, between the interlocutor, who adopts the role of patient/client, or the relative of a patient/ client, and the candidate, who assumes his or her professional role.

Candidates are rated on a six-point rating scale for each of six linguistic categories (Table 1). In this paper, these assessment categories will be referred to as items.

Assessment is carried out by raters who have participated in a training session followed by rating of a series of audio-taped recordings of speaking test interactions to establish their reliability. The assessment is either carried out live, during the test, by a trained rater acting as interlocutor, or later, by a trained rater using an audio-tape of the interaction. In either case the interaction is recorded.

Method

Data from two rater training sessions, eighteen months apart, and a subsequent operational test administration (approximately 2 months after the second training session), were used. Thus, three occasions are represented in the data, as shown in Table 2a.

Table 2a

Time	*Date*	*Material*
1	Sept 1991	Rater training session tapes
2	March 1993	Rater training session tapes (as at Time 1)
3	April/May 1993	Test administration tapes

Thirteen raters (Group A) gave ratings at Times 1 and 2 only; of these, six raters (Group B) gave ratings on all three occasions; The tapes rated were identical at Times 1 and 2, and different at Time 3. The data for these two groups of raters formed part of a much larger (and therefore richer) data matrix which was used in the calibration of the characteristics of these particular groups of raters (cf Table 2b).

Table 2b

Times	*No of raters studied*	*No of tapes rated*	*Total number of raters in matrix*
1&2	13 (Group A)	10	Time 1= 55
			Time 2 = 32
3	6 (Group B)	73	6

Two different kinds of analysis were attempted. In Approach A, rater harshness was calibrated on each of the three occasions, and compared. In Approach B, occasion was treated as a facet in the analysis, and rater by occasion interactions were examined using the bias analysis facility in FACETS.

Results

Approach A: comparison of measures of rater harshness from rating times 1, 2 & 3

FACETS produces three statistics for each element of each facet analysed. For the facet 'rater', this takes the form of 1) an estimate of rater harshness, 2) a standard error associated with this estimate, and 3) a model fit assessment. In addition, a number of indicators of the degree to which different levels of the facet are defined is given, the most straightforward being *separation reliability* , expressed as a reliability coefficient.

The first approach used was to compare the measures of harshness for individual raters over the 20-month period of the three rating sessions.

Analysis 1 - Group A, Time 1

Results from September 1991 ratings, made during the week following the rater training sessions (55 raters, 10 candidates) were analysed[4], and measures of rater harshness, candidate ability and item difficulty were derived from this analysis. **Table 3** shows individual rater measurement for the 13 raters in Group A The mean logit value for rater harshness (of all 55 raters involved) for this occasion was -0.27. The level of error was small and varied little amongst raters. No raters were identified as misfitting. Significant variations in harshness were shown to exist between the 55 raters (Reliability of rater separation = 0.89).

Table 3: Rater measurement report, Group A, Time 1

Measure Logit	Model Error	Infit MnSq	Infit Std	Outfit MnSq	Outfit Std	Rater ID
-0.80	0.19	0.7	-1	0.7	-1	1
-0.41	0.19	1.2	1	1.3	1	2
-1.14	0.19	0.8	0	0.7	-1	3
-1.78	0.19	0.6	-2	0.5	-2	6
0.65	0.19	0.9	0	0.9	0	7
-0.80	0.19	1.0	0	1.0	0	8
-0.64	0.19	0.6	-2	0.7	-2	9
-0.94	0.19	1.0	0	1.0	0	10
-0.96	0.18	0.9	0	0.9	0	14
0.24	0.18	1.1	0	1.3	1	20
-0.29	0.21	1.0	0	1.0	0	24
0.04	0.20	1.1	0	1.1	0	30
0.12	0.21	0.8	0	0.8	0	32

Measure Logit	Model Error	Infit MnSq	Infit Std	Outfit MnSq	Outfit Std	Rater ID
-0.27	0.20	1.0	-0.2	1.0	-0.1	**Mean** (Count: 55)
0.60	0.02	0.4	1.8	0.4	1.8	S.D.

Separation 2.85 **Reliability of separation** 0.89
Fixed (all same) chi-square: 492.21 d.f.: 54 significance: .00

[4]In FACETS analyses, the facet of candidate ability is by convention made the non-centred facet, with the mean logit measure of candidate abilities varying according to the sample tested, while the other facets (rater harshness, item difficulty, etc.), will have a mean set arbitrarily at 0. However, the facet of interest in this analysis was rater harshness. It is therefore not possible simply to compare the logit values provided by the FACETS analysis for the two occasions, even though the candidates and items were common. In the analyses using this approach, therefore, mean rater harshness was made to vary from the mean (it was the non-centred facet), with the other facets (candidate ability and item difficulty) being centred on 0. Values of individual rater harshness will vary according to the composition of the group; establishment of a common mean was therefore necessary.

Analysis 2 - Group A, Time 2

Results from March 1993 ratings (32 raters, including 13 from the previous occasion, using tapes of the same 10 candidates) were then analysed. In this analysis, the estimates of candidate ability and item difficulty were anchored to the measures derived from the previous analysis. This was so as to provide a common frame of reference for the data, and hence allow a comparison of rater harshness, as represented in logit values. The mean logit value for rater harshness for this occasion was -0.17 (**Table 4**). Error values were comparable to the previous occasion. Of the 13 raters in Group A, one (R 7) was identified on this occasion as misfitting. Again, significant differences in rater harshness were found (Reliability of rater separation = 0.87)

Table 4: Rater measurement report, Group A, Time 2

Measure Logit	Model Error	Infit MnSq	Infit Std	Outfit MnSq	Outfit Std	Rater ID
-1.15	0.20	0.8	-1	0.7	-1	1
-0.01	0.19	0.9	0	0.9	0	2
-0.88	0.18	1.3	1	1.4	1	3
0.08	0.18	1.2	0	1.2	0	6
0.63	0.20	1.8	3	2.1	4	7
-0.30	0.21	0.8	0	0.8	0	8
-0.79	0.20	0.5	-2	0.5	-2	9
-0.92	0.21	1.1	0	1.0	0	10
-0.66	0.16	1.2	0	1.1	0	14
-0.70	0.19	1.3	1	1.3	1	20
-0.52	0.18	1.3	1	1.6	2	24
-0.32	0.18	1.0	0	1.0	0	30
-0.12	0.20	0.9	0	0.9	0	32

Measure Logit	Model Error	Infit MnSq	Infit Std	Outfit MnSq	Outfit Std	Rater ID
-0.17	0.20	1 0	-0.2	1.0	-0.1	**Mean** (Count: 32)
0.54	0.02	0.3	1.5	0.4	1.8	S.D.

Separation 2.55 **Reliability of separation** 0.87
Fixed (all same) chi-square: 247.99 d.f.: 31 significance: .00

Table 5 shows a comparison between harshness of these 13 raters in terms of logits, on the two occasions. It will be seen from Tables 3 and 4 that the mean rater harshness for the groups of raters contained in the calibration differed on each occasion (Time 1: -0.27 logits; Time 2: -0.17 logits). The logit values for Time 2 therefore need to be adjusted to make them comparable with Time 1 values. By adding -0.10 to individual rater harshness values from this second analysis the two sets of data were then made comparable, and logit values of the 13 individual raters common to both rating periods could be compared (Table 5).

Table 5 Stability of rater characteristics, Times 1 & 2 (Group A)

Rater ID	Time 1	Error	Time 2*	Error	Change
6	-1.78	.19	-0.02	.18	1.76
20	0.24	.18	-0.80	.19	-1.04
30	0.04	.20	-0.42	.18	-0.46
1	-0.80	.19	-1.25	.20	-0.45
8	-0.80	.19	-0.40	.21	0.40
32	0.12	.21	-0.22	.20	-0.34
24	-0.29	.21	-0.62	.18	-0.33
2	-0.41	.19	-0.11	.19	0.30
9	-0.64	.19	-0.89	.20	-0.25
14	-0.96	.18	-0.76	.16	0.20
3	-1.14	.19	-0.98	.18	0.16
7	0.65	.19	0.53	.20	-0.12
10	-0.94	.19	-1.02	.21	-0.08

* Values have been adjusted to make them comparable with those of Time 1.

It will be seen that the error associated with each logit measure of rater harshness accounts for all or almost all of the change in harshness for all of the raters, with the exception of the first two raters shown here, rater 6, who has become harsher by 1.76 logits, and rater 20, who has become more lenient by 1.04 logits. There has otherwise been relatively little shift in rater harshness. It appears, then, that there is non-uniform variation in severity amongst the 13 raters over the two occasions and 18 months for which we have data so far. However, the significant variations appear to be restricted to a relatively small number of raters (2 out of 13).

Analysis 3 - Group B, Time 3

A further analysis, following the same procedure, was carried out with an additional data set, with ratings produced in April and May 1993, that is, *following* the second rating occasion described above.

This set (Time 3) comprised ratings from 6 of the raters considered so far, with a new cohort of 73 candidates' tapes, taken from 1992 test administrations, each rated twice. There were no common candidates from earlier sessions, but in order to create a common frame of reference the item difficulties (the mean difficulty values obtained for each of the six categories of language assessed - **Table 1**) were anchored to the values obtained in the first (1991) analysis (and the raters were again the non-centred facet). The results of the analysis are presented in Table 6.

Table 6: Rater measurement report, Group B, Time 3

Measure Logit	Model Error	Infit MnSq	Infit Std	Outfit MnSq	Outfit Std	Rater
0.15	0.42	0.7	-2	0.5	-1	**1**
-0.40	0.18	0.8	-1	0.7	-1	**2**
-1.45	0.15	0.7	-1	0.7	-2	**6**
1.19	0.33	0.6	-1	0.6	-1	**7**
0.71	0.13	0.9	-1	0.9	-1	**8**
-1.49	0.15	0.9	0	0.9	0	**10**
-0.22	0.23	0.8	-1.6	0.7	-1.6	**Mean** (Count: 6)
1.01	0.11	0.1	0.5	0.1	0.4	S.D.

Separation 3.87 **Reliability of separation** 0.94
Fixed (all same) chi-square: 201.53 d.f.: 5 significance: .00

The logit values for rater harshness for this occasion were derived from the analysis, and adjusted as before, to make them comparable with the previous analyses. On this occasion the mean rater harshness was -0.22, and -0.05 was added to individual values, so that this set could be compared with the previous times. The increased error on this occasion for raters 1 and 7 is due to the fact that the data matrix this time is much sparser (only two ratings per tape), and these two raters rated fewer tapes than the other raters did.

The stability of rater characteristics over the three time periods are presented in Tables 7.1, 7.2 and 7.3.

Table 7.1: Stability of rater characteristics, Times 1 & 3 (Group B)

Rater ID	Time 1	Error	Time 3*	Error	Change
8	-0.70	.19	0.66	.13	1.36
1	-0.70	.19	0.10	.42	0.80
10	-0.84	.19	-1.54	.15	-0.70
7	0.75	.19	1.14	.33	0.39
6	-1.68	.19	-1.50	.15	0.18
2	-0.31	.19	-0.45	.18	-0.14

Table 7.2: Stability of rater characteristics, Times 2 & 3 (Group B)

Rater ID	Time 2	Error	Time 3*	Error	Change
6	0.08	.18	-1.50	.15	-1.58
1	-1.15	.20	0.10	.42	1.25
8	-0.30	.21	0.66	.13	0.96
10	-0.92	.21	-1.54	.15	0.62
7	0.63	.20	1.14	.33	0.51
2	-0.01	.19	-0.45	.18	-0.44

Table 7.3: Changes in rater severity over the 3 occasions

Rater ID	Change Times 1-2	Change Times 1-3	Change Times 2-3
1	-0.45	0.80	1.25
2	0.30	-0.14	-0.44
6	1.76	0.18	-1.58
7	-0.12	0.39	0.51
8	0.40	1.36	0.96
10	-0.08	-0.70	-0.62

These tables partially confirm the picture obtained from the earlier analyses, suggesting non-uniform change in levels of rater severity over the three occasions. However, the degree of change appears larger between the two rating periods in 1993, Times 2 and 3, than between Times 1 and 3, and it also appears that there may be wider variation in raters' consistency across rating times than suggested by the earlier analysis. Nevertheless, it has been shown to be possible to identify the raters who show significant variation, provided an analysis of rater severity is carried out for each rating occasion, with appropriate anchoring.

Turning to individual raters, it appears that:
- rater 6 has reversed the change shown between times 1 and 2;
- rater 10 is noticeably more lenient at time 3 than on either of the two previous occasions;
- raters 8 is noticeably harsher at time 3 than on either of the two previous occasions;
- rater 1 is noticeably harsher at time 3 than at time 2.

Approach B: bias analysis

A second approach to the question was therefore employed, using the bias analyses that FACETS offers.

These model interactions between elements of any facet. For the facet 'person', an element is an individual candidate; for the facet 'rater', an element is an individual judge, etc. A bias analysis permits investigation of the way a particular aspect of the test situation (rating occasion, type of candidate, etc.) may elicit a consistently biased pattern of responses from a rater. Lunz and Stahl (1992) used these techniques to produce judge performance reports, which provided individual raters with information about their relative characteristics as raters, their consistency, and any individual biased ratings, in a judge-mediated examination of histotechnology.

FACETS has the advantage that it can model time as a facet, and hence provide an estimate of the difficulty associated with a particular time of assessment. A second advantage is that a bias analysis of the interaction between the facets of rater and rating time can give information about whether individual raters are rating consistently harshly or leniently on any particular occasion.

The bias analysis seemed likely also to yield more accurate information about any changes in raters' on different occasions, as the program would model probabilistically all the information available to it, rather than relying on the somewhat crude averaging employed in the first approach.

Table 8: Rater/Time bias, Times 1 and 2 (significant bias only)

Obsvd Score	Exp. Score	Obs-Exp Average	Bias+ Logit	Model Error	Z-Score	Infit MnSq	Rater	Time
223	210.8	0.20	-0.43	0.19	-2.3	2.2	7	2
258	245.3	0.21	-0.43	0.18	-2.3	1.1	20	2
293	261.7	0.52	-0.93	0.18	-5.2	0.5	6	1
199	211.2	-0.20	0.40	0.18	2.3	0.8	7	1
233	245.7	-0.21	0.42	0.18	2.3	1.2	20	1
230	261.3	-0.52	0.92	0.17	5.3	1.0	6	2
Obsvd Score	Exp. Score	Obs-Exp Average	Bias+ Logit	Model Error	Z-Score	Infit MnSq	Rater	Time
237.6	237.6	0.00	-0.00	0.20	0.0	1.0	Mean (Count: 87)	
17.4	16.4	0.10	0.19	0.02	1.0	1.0	S.D.	

Fixed (all = 0) chi-square: 92.97 d.f.: 87 significance: .31

A bias analysis of the interaction between all raters and rating Times 1 and 2 was performed. (It was not possible to produce bias analyses involving Time 3, because of the lack of common candidates across the three occasions). The output (**Table 8**) provides an estimate of the extent (measured in logits) to which an individual rater was biased on a particular occasion; this logit value is then standardized to a *z*-score, and any *z*-scores exceeding 2.0 indicate significant bias.

This analysis identifies raters 6, 7 and 20 as biased. The output shows both the extent and direction of the bias. A negative bias logit / *z*-score indicates the rater was more lenient than the model predicted, given all the information provided about this rater, and a positive measure indicates the rater was harsher than expected. Unexpectedly harsh ratings at one rating time are matched by unexpectedly lenient ratings at the other time.

The conclusion reached earlier, using the overall estimates of rater harshness, is thus confirmed, that raters 6 and 20 showed significant changes in their severity between times 1 and 2 (cf Table 5); in addition, the analysis identifies rater 7 as changing in severity.

The fact that an additional rater is identified as being biased in this analysis requires discussion. First, the extent of bias is not large ($z = 2.3$). Second, the rater's behaviour is inconsistent. Table 4 revealed that Rater 7 was misfitting at Time 2 (Infit MnSq = 1.8, Infit $t = 3$). Furthermore, the bias for this rater at Time 2 is inconsistent (Bias Infit MnSq = 2.2). Examination of individual misfitting ratings involving Rater 7 (Table 9) reveal that the bias report at Time 2 is likely to have been strongly influenced by one or two rather unexpected ratings, and is thus less certainly a general pattern.

Table 9 Individual misfitting ratings, Times 1 and 2

Cat	Step	Exp.	Resd	StRes	Ca	Rtr	Tm	Item
6	6	4.0	2.0	3	9	3	2	2
4	4	5.7	-1.7	-3	13	3	2	5
5	5	2.5	2.5	3	12	10	2	2
3	3	4.9	-1.9	-3	17	30	2	4
5	5	5.9	-0.9	-3	13	24	2	4
5	5	1.6	3.4	4	12	7	2	5
3	3	5.0	-2.0	-3	13	7	2	2
5	5	2.8	2.2	3	12	6	2	4
3	3	1.2	1.8	3	12	7	1	5
2	2	4.9	-2.9	-4	17	30	1	4
4	4	2.3	1.7	3	12	20	1	5
6	6	3.6	2.4	3	17	20	1	2

Cat	Step	Exp.	Resd	StRes				
4.0	4.0	4.0	-0.0	0.0	Mean (Count: 5214)			
1.1	1.1	0.9	0.7	1.0	S.D.			

On the whole, it seems that bias analysis is a more sensitive measure of alterations in rater characteristics over time, but its data requirements are restrictive.

Finally, it is worth noting that each of the rating times analysed here covered a period of about a week (Times 1 and 2) or more (Time 3), much longer than the periods considered by Lunz and Stahl (1990); there appears to be reasonable consistency for all raters (except rater 7 for time 2) within these rating periods, with the significant variation coming over much longer periods.

Implications of the study

The study further reveals the potential of the new technology of multi-faceted measurement for research on performance tests (cf earlier studies by McNamara and Adams, 1991, and recent papers by Elder, 1993, Wigglesworth and O'Loughlin, 1993, Brown, 1993 and McNamara and Lumley, 1993). By producing rater calibrations that are independent of the data used to derive them, comparison across different rating occasions becomes possible. Multi-faceted measurement has made possible the close examination of an issue that has long been recognized.

One point that emerges consistently and very strongly from all of these analyses is the substantial variation in rater harshness, which training has by no means eliminated, nor even reduced to a level which should permit reporting of raw scores for candidate performance. This appears clear enough justification for using FACETS analysis of performance test data where no more than 2 raters are involved in assessing each candidate, since it is able to take relative severity of judges into account and make adjustments to estimates of candidate ability.

With regard to rater training, since the rating occasion has been shown to influence different raters in different ways, it is proposed that Lunz and Stahl's (1992) suggestion for

rater performance reports be taken up. They recommended the use of performance reports as feedback to judges focusing on judge by item interactions. In the context of the OET, performance reports could be produced for each rating time and given to individual raters identified as rating harshly or leniently on particular occasions. A follow-up study could then be carried out to determine whether rater characteristics stabilize as a result of such feedback. A study investigating this issue has recently appeared (Wigglesworth, 1993), suggesting that such reports may indeed have the desired effect.

Such reports could be complemented by protocol or interview analysis, prior to the FACETS analysis, to see if it is possible to identify beforehand the likelihood of personal circumstances influencing a rater's severity or leniency on a particular occasion.

The variability that has been discovered in the study, particularly between the rater training session and the actual test administration, means that we should call into question the practice sometimes adopted, e.g. by IELTS and the ASLPR, of certifying raters and then basing judgements of candidates on single ratings by such certified raters. Just as the analyses confirm yet again that judge differences survive training, so intra-rater differences are likely to be an issue for at least some raters over different rating occasions. It seems that at every administration, new calibrations of rater characteristics are required; failing that, the traditional technique of double and if necessary multiple ratings seems amply justified.

References

Brown, A. (1993) *The effect of rater variables in the development of an occupation-specific language performance test.* Paper presented at the 15th Language Testing Research Colloquium, Cambridge, August.

Cason, G.J. and C.L. Cason (1984) A deterministic theory of clinical performance rating. *Evaluation and the health professions* 7,2: 221-247.

Coffman, W.E. and D. Kurfman (1968) A comparison of two methods of reading essay examinations. *American Educational Research Journal* 5,1: 101-120.

Constable, E. and A. Andrich (1984) Inter-judge reliability: Is complete agreement among judges the ideal? Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans.

Cushing, S. (1993) Paper presented at AAAL Conference, Atlanta, GA, April.

Diederich, P.B., J.W. French and S.T. Carlton (1961) *Factors in judgments of writing ability* (Research Bulletin 61-15). Princeton, NJ: Educational Testing Service. [ERIC Document Reproduction Service # ED 002 172]

Edgeworth, F.Y. (1890) The element of chance in competitive examinations.*Journal of the Royal Statistical Society* 53: 460-475 and 644-663.

Elder, C. (1993) *Are raters' judgements of language teacher effectiveness wholly language based?* Paper presented at the 15th Language Testing Research Colloquium, Cambridge, August.

Harper, A.E. Jr and V.S. Misra (1976) *Research on examinations in India.* New Delhi: National Council of Educational Research and Training.

Huot, B. (1990) The literature of direct writing assessment: major concerns and prevailing trends. *Review of Educational Research* 60,2: 237-263.

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

Linacre, J.M. (1991) *Constructing measurement with a many-facet Rasch model.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois, April. [ERIC Document Reproduction Service # ED 333 047].

Linacre, J.M. and B. Wright (1992) *Facets: Rasch Measurement Computer Program, version 2.6.*

Lunz, M.E. and J. Stahl (1990) Judge consistency and severity across grading periods. *Evaluation and the health professions* 13,4: 425-444.

Lunz, M.E. and J. A. Stahl (1992) *Judge Performance Reports: Media and Message.* Paper presented at American Educational Research Association, San Francisco, 1992.

McIntyre, P.N. (1993) *The importance and effectiveness of moderation training on the reliability of teacher assessments of ESL writing samples.* Unpublished M.A. thesis, University of Melbourne.

McNamara, T.F. (1990) *Assessing the second language proficiency of health professionals.* Unpublished Ph.D. thesis, University of Melbourne.

McNamara, T.F. and R.J. Adams (1991) *Exploring rater characteristics with Rasch techniques.* Paper presented at the Language Testing Research Colloquium, Princeton, NJ, March [ERIC Document Reproduction Service # ED 345 498].

McNamara, T.F. and T. Lumley (1993) *The effects of interlocutor and assessment mode variables in offshore assessment of speaking skills in occupational settings.* Paper presented at the 15th Language Testing Research Colloquium, Cambridge, August.

Wigglesworth, G. (1993) Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing* 10,3.

Wigglesworth, G. and K. O'Loughlin (1993) *An investigation into the comparability of direct and semi-direct versions of an oral interaction test.* Paper presented at the 15th Language Testing Research Colloquium, Cambridge, August.

Wiseman, S. (1949) The marking of English composition in English grammar school selection. *British Journal of Education Psychology* 19,3: 200-209.

Wood, R. and D. Wilson (1974) Evidence for differential marking discrimination among examiners of English. *The Irish Journal of Education* 8,1: 37 *et seq.*