Systematic effects in the rating of second-language speaking ability: test method and learner discourse

John A. Upshur and Carolyn E. Turner

Language Testing 1999; 16; 82
DOI: 10.1177/026553229901600105

Systematic effects in the rating of second-language speaking ability: test method and learner discourse

John A. Upshur *Concordia University, Montreal*
Carolyn E. Turner *McGill University, Montreal*

Major differences exist in two approaches to the study of second-language performance. Second-language-acquisition (SLA) research examines effects upon discourse, and is typically unconcerned with scores. Language-testing (LT) research investigates effects upon scores, generally without reference to discourse. Within a general framework of test taking and scoring, we report research from these two fields as it relates to questions of systematic effects on second-language tests. We then examine findings incidental to a test-development project.

The findings were consistent with LT research into systematic effects of task and rater on ratings, and with SLA research into systematic effects of task on discourse. Using empirically derived scales as indicators of salient features of discourse, we infer that task type influences strategies for assessing language performance. Explanations for these joint findings are not afforded by either standard LT or SLA perspectives. There is no theory of method to explain how particular aspects of method affect discourse, how those discourse differences are then reflected in ratings and how task features influence the basis for judgement. We conclude that a full account of performance testing requires a paradigm that incorporates relationships that are not specified in either the major language-testing research tradition or the tradition of second-language-acquisition research.

I Introduction

Concern with systematic error effects has been reflected in language-testing (LT) research at least since early applications of Campbell and Fiske's (1959) multitrait-multimethod approach to construct validation (see, e.g., Brütsch, 1979; Clifford, 1981; Corrigan and Upshur, 1982). The early studies examined tests of locally independent items scored dichotomously. These differ greatly from tests consisting of a very few performances rated on scales with a number of levels. More recently, studies focusing on the systematic effects upon ratings can be found in the research domain of language testing, but these studies

Address for correspondence: Carolyn E. Turner, McGill Univesity, Department of Second Language Education, Faculty of Education, 3700 McTavish Street, Montreal, QC, Canada H3A 1Y2; e-mail: cx9x@musica.mcgill.ca

Language Testing 1999 16 (1) 82–111 0265-5322(99)LT162OA

do not consider the effects upon the discourse that influence scores. This consideration appears in studies in the domain of SLA. In general, however, SLA research on task-related variation in discourse is not concerned with scores.

In this paper we consider research from the fields of language testing and second-language acquisition as it relates to questions of systematic effects on second-language tests. In order to illustrate this research, we first develop a framework of test taking and test scoring, that is general enough to accommodate a variety of traits and methods: from discrete-point, multiple-choice tests to rated samples of communicative performance. We then review the locations and relations in the frameworks that have been addressed by research in language testing and in SLA. Next we report some findings incidental to a test-development project. These confirm some findings on method effects reported in the LT research literature and show some discourse features that are differentially important for assessing performance on different performance test tasks.

II Test taking and scoring

The minimum requirements for a test are an examinee and a task. The attempt by the examinee to solve the task results in a score. That is, the interaction between the test taker and the test task leads to a score on that task. This is illustrated in Figure 1. This interaction can be expressed in terms of a psychometric model: the likelihood of getting a score of 1 rather than 0 on a dichotomously scored item is a function of the difference between the ability of the person and the difficulty of the item.[1]

The process of test taking and scoring differs from the psychometric model. This process, assuming the absence of method effects, is illustrated in Figure 2. For a multiple-choice item, the interaction of an examinee with a given ability and a test task with a given ability requirement (i.e., difficulty) lead the examinee to make a mark on an answer sheet. This is the test taking phase. The marked answer sheet is then compared by a human or machine scorer to an answer key;

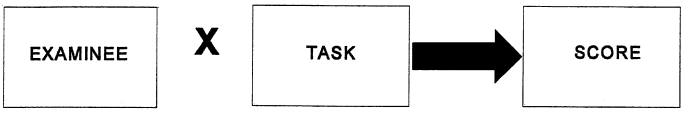

Figure 1 Generalized model of a test

[1]The Rasch one-parameter model is one illustration of this. Other more complex models exist.

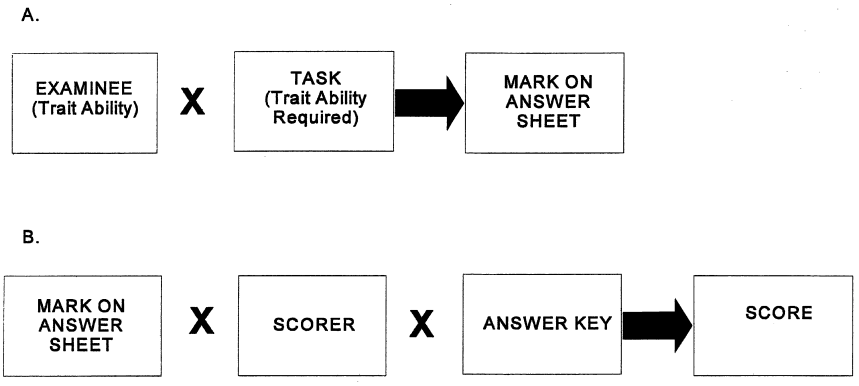

Figure 2 Objective test taking and scoring, no method effects

if the two match, a score of 1 is given. Otherwise, the score is 0. This is the test-scoring phase.

Figure 2 describes the process applicable to standardized tests made up of discrete-point tasks (i.e., locally independent, objective items). Under the assumptions of no method effect and perfectly reliable answer marking and scoring, an examinee's score on a task is a function of (1) how much of the trait the examinee possesses and (2) how much of the trait the task requires for correct performance.

When one assumes the presence of method effects, however, a score is a function not only of trait ability and trait requirements, but also of other non-trait abilities and requirements. This is illustrated in Figure 3. Tests are never perfectly valid. Every test method requires a variety of abilities for satisfactory performance. Test performance depends, therefore, upon a number of non-trait abilities of the test taker. These constitute systematic errors of measurement.

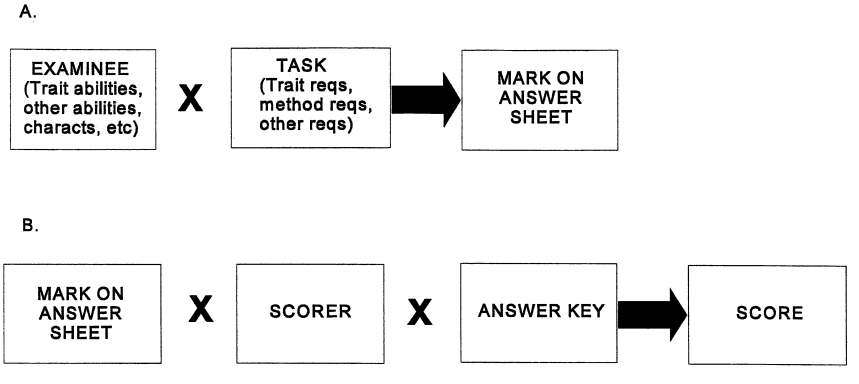

Figure 3 Objective test taking and scoring, with method effects

Figure 3 illustrates the process of objective testing when method effects are included. In addition to ability in the trait of interest, an examinee's performance is affected by other abilities and characteristics that are related to non-trait requirements of the task. For example, items in a test of grammatical knowledge might be influenced systematically by an examinee's reading ability, visual acuity, general intelligence, gender, cultural affiliation, risk tolerance, vocabulary knowledge, etc.

For much of modern language testing this model is too limited. The assumptions of unambiguous answer marking and error-free dichotomous scoring no longer hold, even approximately. Instead of filling in squares on an answer sheet, the examinee produces, perhaps together with an interlocutor, an oral or written discourse. This discourse is heard or read by a human judge who refers to a scoring guide or rating scale in order to select a score to represent the examinee's ability in the trait of interest. Figure 4 illustrates this extended model of systematic effects on performance test scores.[2] Examinee and task interaction produces a discourse, not a score as in the simpler model. Then the three-way interaction of discourse, rater (or, judge, as often referred to) and scale yields a score that is interpreted as a measure of the examinee's trait ability. This elaborated model presents a graphic indication of systematic effects to be studied in order to better understand performance tests.

Bachman *et al.* (1995: 239) note that performance testing brings with it 'potential variability in tasks and rater judgements, as sources of measurement error.' This has been recognized, and studies of some of these additional effects have been reported (e.g., Lumley and

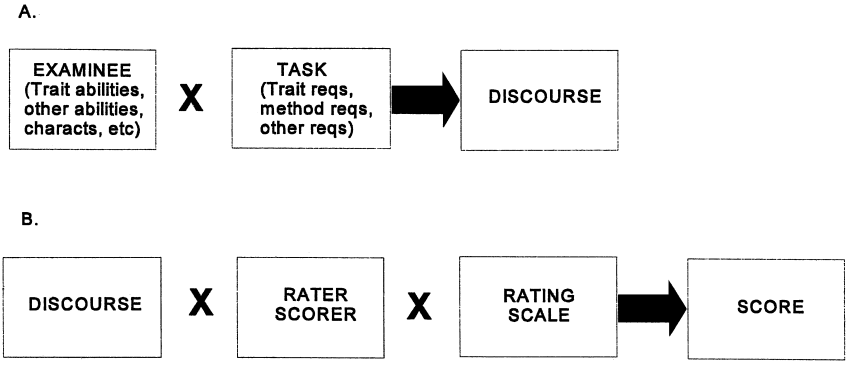

Figure 4 Performance test taking and scoring (with method effects)

[2]This is similar to the representation independently developed by McNamara (1996: 86).

McNamara, 1995; Tyndall and Kenyon, 1996). Most of the LT studies focus upon systematic effects upon scores. They do not generally take account of the discourse that is a determiner of scores.

There is another research domain, SLA, that examines test factors that affect discourse. That literature tends not to be concerned with raters and scores, however. There is some cross-referencing of works in LT and SLA (e.g., Bachman and Cohen, in press; Chalhoub-Deville, 1995), but we have been unable to locate any comprehensive synthesis of findings from the two domains that addresses the question of systematic effects on performance tests.

III Systematic error effects: the research literature

Systematic effects on the rating of writing and speaking ability have been investigated. The typical approach in the field of LT is to examine effects on scores. On the other hand, the typical approach in the field of SLA is to examine effects on the discourse generated in task performance. Studies exemplifying both approaches are cited below.

1 Systematic effects on scores

a Examinee characteristics Studies of examinee effects are relatively infrequent. Simon (1994) provides a discussion on differential item functioning studies within a bilingual context (viz, Canada); Elder (1995) has preliminary results concerning the use of common assessment instruments and scales to assess the language skills of learners from different language backgrounds; Sunderland (1995) has demonstrated ways in which gender bias may manifest itself. Test bias has been studied in educational contexts, focusing on minority groups taking tests where the test language is not their first language (cf. Green, 1994; Samuda *et al.*, 1989). Norming studies document effects of examinee group membership (e.g., geographical region, first language), but effects of these demographic variables cannot be disentangled from the many other variables with which they are confounded. Kunnan (1995) has reported a rigorous investigation of examinee effects. Using structural equation modelling with two culturally different groups, he evaluated systematic effects of informal exposure and instruction, location of exposure, and monitoring on proficiency test performance. Speaking ability was not well accounted for by these factors.

b Task attributes Research documenting the task effect on test scores is more abundant. In oral proficiency testing, this effect is most often examined in terms of task difficulty. Kenyon (1995), for

example, in a validation study of performance-based tests of oral proficiency explores the ability demanded by different tasks. Empirical ordering of the tasks by difficulty support an *a priori* ordering. Nonetheless, Fulcher (1994) includes the effect of task difficulty on scores as a priority area for research on oral language testing. Chalhoub-Deville (1995) points out that a construct can be represented by several tasks, but that construct dimensions are manifested differently depending on how the construct is operationalized. Therefore, ratings are context-specific with regard to tasks. She calls for empirically derived dimensions according to the specific task and audience.

A further direction of study concerning the method effect is the investigation into the effects of direct versus semi-direct versions of oral proficiency tests. Stansfield and Kenyon (1992) compared direct and semi-direct interviews. O'Loughlin (1995) found a high correlation between test scores ($r = .92$), but observed that the semi-direct version produced a more literate type of language than the direct or live version. It was the degree of interactiveness in the direct version more than the difference in test format with the semi-direct version that had salient effects on the discourse produced. The more interaction, the lower the degree of lexical density in the discourse.

c Discourse qualities Most of the work relating discourse characteristics to scoring has been done in the field of writing assessment, and indirectly in the process of rating scale development. Vaughan (1991), for example, found that holistic raters of ESL compositions frequently reported six different text features that influenced their judgements. Weigle (1994a) documented the criteria applied by raters in assessing writing ability before and after training. A number of writers have described empirical methods for constructing rating scales (Pollitt and Murray, 1993; Stansfield, 1986; Upshur and Turner, 1995). Examination of samples of learner performance reveal qualities of the discourse that affect judgements and, ultimately, the scores assigned. Another approach to relating discourse qualities to scores has been to investigate the relation of subtraits to holistic scores (see, e.g., Henning and Davidson, 1986).

d Rater characteristics A major concern with performance testing is that the tasks require subjective assessments by raters. The rater is not only an additional source of measurement error but, as a method facet, may also exert systematic – although unwanted – effects upon scores. Investigating effects and ways to control them have been the focus of several studies.

The developments in many-facet Rasch measurement (Linacre, 1989–1993a) have provided improved techniques for investigating

rater effects. It is now well established that raters differ in the severity of their judgements of ability (Lumley and McNamara, 1995; Wigglesworth, 1993). The seriousness of this effect has been questioned, however. Bachman *et al.* (1995: 253) using both many-facet Rasch measurement and generalizability theory found that, although there was a wide range of judge severity, this was unlikely to produce an overall effect on test scores that are based upon double ratings of multiple tasks. Going beyond questions of severity, McNamara and Adams (1991/1994) note further that raters can demonstrate a variation in behaviour depending on the particular group of examinees, the particular task, and the particular occasion. Linacre (1989–1993a) uses the term *bias* to describe this interaction.

Rater training, as an attempt to control rater variability, has also been investigated. Lunz *et al.* (1990) suggest that training cannot make judges equally severe, but it can increase the consistency with which individual judges rate all subjects (i.e., intrarater reliability). Weigle (1994b) found changes in rater behaviour after training. Even though differences in severity across raters still remained, the training appeared to bring extreme raters within a range of tolerable severity. Individual rater consistency improved across the inconsistent raters. Lumley and McNamara (1995) suggest that results of training may not last for long after a training session, thus demonstrating a need for renewed training before each test administration. It has been suggested that information obtained from many-facet Rasch analyses could be used effectively for feedback in rater training (Tyndall & Kenyon, 1996). Using information from a bias analysis, Wigglesworth (1993) found that raters were responsive to feedback and were able to incorporate it into subsequent ratings so that bias was reduced.

In spite of rater training, however, each rater has a unique background that may effect judgements. Elder (1993) studied rater behaviour in the assessment of English proficiency of non-native-speaker graduate students training as secondary maths and science teachers. Raters were of two groups: ESL teachers and maths/science subject specialists. They gave similar ratings for overall communicative effectiveness, but differences were found in their ratings of particular dimensions of language use. Brown (1995) explored effects of rater occupation and language in an occupation-specific oral-language test. Results were similar to Elder's (1993). She found no significant differences in overall grades awarded by different rater groups. Differences were found, however, on ratings for individual criteria. This is interpreted as a demonstration of different perceptions of what constitutes good performance.

e Scale types There are a number of ways to classify rating scales.[3] These may refer to the physical form or layout of the printed scale, to the latent variables assessed or to the content of descriptors. One way of classifying second-language rating scales is of especial relevance to a consideration of systematic method effects: scale types represent the underlying assumptions about possible systematic effects upon ratings. Three different assumptions about scaling language performance are represented in existing scale types. The first notion is that of 'absolute proficiency rating'. It assumes, implicitly at least, that there is no task effect, or that the rating scale can itself compensate for task effects.

The second notion is that of 'task proficiency rating'. It views language performance tasks as more difficult or less difficult, but holds that a single rating scale will apply to all tasks of a kind (e.g., speaking tasks). Within this view, examinee abilities are estimated from ratings of their performances on a task, *adjusted by* the calibrated difficulty of that task.

The third notion is that of rating according to a task/scale unit. Task effects are acknowledged; it is further assumed that the qualities of discourse that reflect progressive-ability levels will differ across tasks and populations. For example, the qualities of discourse that mark competence in telemarketing will differ from those that mark competence in psychotherapy. Rating scales developed according to this notion are, therefore, specific to a population and task, and are often generated from empirical data.

We have found no studies that examine effects upon test performance by scales reflecting these three sets of assumptions about systematic effects. Most studies utilize one of the first two types of rating scales. They do not, however, analyse effects.

2 Systematic effects on discourse

Two kinds of effect upon discourse have been examined, those involving learner abilities and those involving such method effects as elicitation procedure, interlocutor characteristics, etc. In this paper we are not concerned with associations between proficiency (the trait of speaking ability) and performance, but rather with other systematic effects that may obscure our estimates of learner ability. A few studies of effects upon discourse in language testing have been conducted (e.g., O'Loughlin, 1995; Lazaraton, 1996). These are rare, however,

[3]In this paper we consider a scale as an ideally unidimensional measure of a single (albeit complex) construct or ability. We do not consider multi-scale scoring schemes such as profiles, their considerable diagnostic and prescriptive educational values notwithstanding.

and do not as yet provide substantial links between method, discourse and scores.

Examination of systematic effects upon discourse quality has been studied most frequently as research in systematic variation in the field of second-language acquisition. Ellis (1994: 138 and 142) divides this research into two general classes: (1) investigations of task-induced variation that documents the effects that context has upon variation but does not identify the effective contextual factors; and (2) investigations of context-induced variation that have 'examined the effects of different aspects of context: linguistic, sociolinguistic, and psycholinguistic'. Studies of this first sort are exemplified by Rose and Ono (1995) who found that discourse completion tasks and multiple-choice tasks elicit different forms of interrogation by Japanese learners of English.

a Examinee characteristics A large number of studies have documented effects of examinee characteristics upon the discourse produced by second-language learners. Among these characteristics are: ethnicity and culture (Beebe and Zuengler, 1983; Takahashi, 1989); first language (Reid, 1988); gender and status (Porter, 1991).

b Task attributes Most of the task effect studies have examined relations between common elicitation procedures and the incidence of specific forms in elicited speech.

Phonology has been a common focus of study. Dickerson (1975), for example, found relation between formality and pronunciation of /z/ by Japanese learners of English. Beebe (1980) found formality effects for pronunciation of Thai and English variants of /r/ in a study with Thai speakers. Sato (1985) found a clear relation between task and pronunciation in a ten-month study of a Vietnamese learner, but could not explain it in terms of task formality.

In studies of morphology and syntax, Tarone (1985) found relation between task (grammatically judgement, oral narration, interview) and accuracy of three morphemes. Like Sato she could not explain the effects according to formality of the task. Ellis (1987) studied accuracy of past-tense forms produced in three tasks. He found a relation that was explainable in terms of planning time.

Another set of investigations of task effects focuses upon interlocutor attributes and behaviour. Studies of interlocutor effects have often looked not only at specific linguistic forms, but also at broader discourse characteristics. Beebe and Zuengler (1983), for example, found that the amount of speech produced by children during interviews was positively related to the amount used by their interviewers. In addition to the effect of amount of interlocutor speech, documented

interlocutor effects include the support provided by the interlocutor to the examinee (Lazaraton, 1996).

Although there is a considerable literature in LT relating method and scores and in SLA relating method and discourse, there is yet no theory of method to explain how particular aspects of method have systematic effects upon discourse that are reflected in test scores. Neither is there a developed explanation of how rater and examinee characteristics interact with one another and with discourse characteristics to yield ratings. Some tentative hypotheses have been offered to explain the effects of method facets upon produced discourse: task formality, planning time and social accommodation, for example (Ellis, 1994). These hypotheses are far from comprehensive, however. Descriptions of effects of discourse features and examinee characteristics upon rater judgements are virtually non-existent and, so, consequently, are any attempts to explain those effects.

IV Incidental findings

In this section we report some findings incidental to a test-development project. This is not a report on the validity of the developed tests for their intended purpose. The report is intended, rather, to illustrate the framework introduced above, to report findings about method effects that confirm previous LT research results, to extend somewhat the method effect findings, and finally to present new information about the salience for raters of task-related features of discourse. This information arises from a type of analysis different from the more usual discourse analysis favored in SLA research.

1 The test-development project

The purpose for the project was to provide a school board with an efficient, standard measure of speaking ability. It was assumed by the board that individual teachers were providing reliable assessments of student ability within their own classrooms, but that a supplementary measure, common for all classes, was needed. A detailed description of the project is given in Turner and Upshur (1996). An abbreviated account of the project is given below.

a Setting and participants The study took place in a large urban school board near Montreal, Quebec. The language of instruction is French; English as a second language (ESL) is a required subject for 120 minutes per week starting in Grade 4. At the end of Grade 6, students are graduated to secondary schools and are streamed into ESL classes as determined by their Grade 6 ESL marks. According

to the curriculum adviser of the school board, the Grade 4–6 ESL program is based upon principles of communicative language teaching.

There were two groups of participants: (1) 12 elementary school teachers who served both as test developers and raters, and (2) 255 Grade 6 ESL learners. Teacher participants were volunteers. Student participants were enrolled in the classes of the teacher participants. Standard practices for use of human subjects were followed.

With the collaboration of the school board's ESL curriculum advisor, we undertook a project to develop a standardized instrument of speaking ability which could serve as one indicator for secondary-school ESL class placement.

b Instruments This study employed two measures of speaking ability developed specifically from the population. A measure is defined here as a speech elicitation task and a rating scale. The development process involved the 12 teachers and two researchers. The EBB scaling procedure was followed (i.e., the scale is *empirically* derived, requires *binary* choices by raters, and defines the *boundaries* between score levels; see Turner and Upshur, 1996).

Elicitation tasks Following a meeting with the 12 teachers, the curriculum advisor and the two researchers, two tasks were agreed upon and developed for use. This included piloting and revision. The two tasks were: Story Retell (SR) and Audio-Pal (AP).

- Story Retell (SR) – The students watch a two-and-a-half minute video. They are instructed to draw a picture after viewing the video to help them remember the story. They are asked to save their pictures. Individually they go to a quiet place where they find a tape recorder and an instruction sheet. They retell the story in their own words, using their pictures.
- Audio-Pal (AP) – Students go individually to a recording site. They are informed that English-speaking exchange students of the same grade from other provinces are coming to live and attend school with them for a month. Each student is then instructed to compose a letter in the form of a tape recording to an exchange student. They are guided to talk about topics like: 'yourself and family, your hobbies and interests, things you might do together, what school is like'.

Rating scale Rating-scale development took place after the teachers had administered the elicitation tasks to a sample of 36 Grade 6 learners from across the school board and recorded their performances.

A sample set of 12 performances was identified for each task (i.e., one sample set from SR responses and one sample set from AP responses).

Scales were constructed in accordance with the third notion of rating scales as outlined above, namely that tasks affect both score and type of discourse. Two rating scales were empirically developed from the sample sets: one scale specific to the Story Retell task, and the other specific to the Audio-Pal task. Six teachers and one of the researchers developed one scale, while the other six teachers and second researcher developed the other scale. For both tasks the teachers found that they were able to distinguish six ability levels. They then devised rating procedures for two different six-level scales. The scale development procedure is described briefly below in the section on task effects on discourse.

c Testing procedure Near the end of the school year, teachers used the two tasks to test their students. Students were individually called out from their regular classrooms activities for testing. All responses were taperecorded for subsequent scoring.

d Scoring procedure Soon after testing, a scoring day session was set up. The teachers brought in the tape recordings of their students. They rated the students from their own classes and the students from two other classes taught by different teachers. In this way, each student recording was rated independently by three of the participating teachers.

e Analysis The data set was analysed using many-facet Rasch measurement with the program FACETS, Version 2.75 (Linacre, 1994). To examine the measurement characteristics of the tests, three facets were specified: subject, rater and task. The partial-credit model was chosen instead of the rating scale model because the scoring criteria for the two scales were qualitatively different (see also Pollitt and Hutchinson, 1987). In addition, a bias analysis was performed to examine the interactions between rater and task.

The model for this analysis was:

$$\log\left(\frac{P_{nijk}}{P_{nijk-1}}\right) = B_n - D_i - C_j - F_k$$

Where:

P_{nijk} is the probability of examinee *n* being awarded on task *i* by judge *j* a rating of *k*;
P_{nijk-1} is the probability of examinee *n* being awarded on task *i* by judge *j* a rating of *k–1*;

B_n is the ability of examinee *n*;
D_i is the difficulty of task *i*;
C_j is the severity of rater *j*;
F_k is the difficulty of the step up from category *k–1* to category *k*.

The use of FACETS permitted a second analysis in which students abilities for the two genders and across the different schools in the district were compared. These were matters of concern to the local school commission, but will not be reported here.[4]

A final, two-facet analysis (subjects and raters) was performed with the AP data set in order to compare the harshness of teachers who had developed the scale with the harshness of those who had developed the other rating scale. There was insufficient data to perform a comparable analysis with the SR data. Two hundred and fifty-five students performed either SR or AP. Forty-two of those students did both tasks. For reasons discovered later, the majority of teachers administered AP, thus resulting in an imbalance in data collection across the two tasks.

The FACETS analysis was performed on a total of 805 ratings given by 12 raters to 297 speech performances produced by 255 children. Seventy-two of the responses were rated in the lowest (1) or highest (6) score level and could not be used for estimating the parameters of the measurement model. An initial concern with a FACETS analysis is whether the mathematical measurement model fits the data one is working with. Convergence was achieved in 63 iterations using the program default values. Three of the 733 measurable responses were unexpectedly high or low. These were ratings given by three different raters to speech samples produced by three different students; two ratings were unexpectedly high and one low. There was, therefore no apparent pattern to the few unexpected responses. Abilities of 250 of the 255 students were adequately estimated.[5] In brief, the data fitted the model quite well, and most of the children's abilities were well measured.

FACETS provides a graphical summary of all facets and their elements (Figure 5). They are positioned on a common logit scale which facilitates comparisons across and within facets. This common scale appears as the first column in Figure 5. The second column shows the subjects by open and filled circles, the latter representing three subjects and the open circle representing one or two. Subjects

[4]No substantive or statistical differences were found between genders. The difference in mean logits was .02; the probability for a fixed (both the same) chi-square was .87. Mean logits for schools had a standard deviation of .24. Fixed chi-square = 17.9; $df = 13$; $p = .16$.

[5]Of the 255 students, five had high ($> |2.0|$) standardized infit indices. For four of these five children standardized outfit indices were also extreme.

Measure	Subject	Task	Rater	SR Scale	AP Scale
6 –	–●●●●●			– (6)	– (6)
	○				
5 –	–●○			–	– -------
	●				
	●				
4 –	–○			– -------	–
	●●				5
	●○				
3 –	–●●●●○			–	–
	●●				
	●●●●			5	-------
2 –	–●●			–	–
	○				
	●●●●●			-------	
1 –	–●●●●●	–	– 4	–	– 4
	●●●●●	SR	9		
	●●●●○		11 2 5	4	
0 –	–●●●○		– 1 6 7 8	–	– -------
	●●●●●○		12	-------	
	●●●○	AP	–		
–1 –	–●●●○	–	10 3	– 3	–
	●●●●				3
	●			-------	
–2 –	–●○			–	–
	●●			2	-------
	●●				
–3 –	–●●○			–	–
	○			-------	2
	●				
–4 –	–●●○			–	–
	●○				
	○				
–5 –	–○			–	– -------
	○				
–6 –	–			–	–
	○				
–7 –	–●●●			– (1)	– (1)
Measure	**Subject**	**Task**	**Rater**	**SR Scale**	**AP Scale**

● = 3 subjects; ○ = 1 or 2 subjects

Figure 5 All facets summary

are ranked by ability, with high ability at the top portion of the column and low ability at the bottom. The third column shows task difficulty with more difficult tasks at the top of the column. Rater severity is indicated in the fourth column, with the more severe raters at the top of the column and the more lenient at the bottom. The last two columns graphically describe the two rating scales. Each task has its own scale. Abilities are represented by the different scale levels across each task. On the scales, numerical values are positioned at integer-expected scores and the horizontal lines are positioned at half-score points (i.e., the point at which the likelihood of getting the next higher rating begins to exceed the likelihood of getting the next lower rating; Myford *et al.*, 1996: 21). To summarize, the most likely scale score for each ability level is shown.

FACETS provides several indications of the reliability of differences among the elements of each facet. Helpful ones to examine are: Separation, Reliability and Fixed (all same) chi-square. The separation index is a measure of the spread of the estimates relative to their precision. It is the ratio of the adjusted standard deviation of element measures to the root mean-square standard error. One may think of this as follows. If two subjects differ in estimated ability by one standard deviation and the standard error is one-third as large (separation index = 3.0), one would be quite confident that there was a true difference in real subject ability. The reliability coefficient indicates how well the analysis distinguishes among the elements. It is the Rasch equivalent to the KR20 or Cronbach's α statistic, that is, the ratio of true variance to observed variance (Linacre, 1989–1993b: 65). Wright and Masters (1982: 105–6) characterize it as the test reliability of (element) separation, that is, the proportion of the observed variance in measurements of ability (severity, etc.) which is not due to measurement error. The fixed (all same) chi-square tests the null hypothesis that all elements of the facet are equal. These three statistics are reported below as appropriate.

2 Measurement characteristics

The measurement characteristics of SR and AP are reported here to provide evidence that the data from the project are reliable, that findings based upon them are credible. Two specific questions were addressed in the FACETS analysis in order to evaluate the metric quality of the tests: (1) Is student ability effectively measured? and (2) Are scales efficient and consistent with assumptions about distributions of student ability?

1) *Is student ability effectively measured?* As shown in Figure 5, subject ability estimates range from a high of approximately 6 logits

Table 1 Summary of FACETS analysis

FACET	Number	Reliability[a]	Variance (in logits)
Subject[b]	255	0.85	5.02
Task[b]	2	0.98	0.53
Rater[b]	12	0.87	0.27

Notes:
[a]Variance ratio, the FACETS equivalent of Chronbach's α.
[b]Significant at $p < .01$.

to a low of –7 logits, a spread of 13 logits in terms of student ability. The reliability index of these estimates was .85 (see Table 1), which demonstrates it is possible to achieve reliable ability scores using these tasks and raters. The sum of variances for all facets (including subjects) was 5.82 logits. Variance in subject ability was 5.02 logits. That is, ability differences were much greater than differences in task difficulty or in rater severity.

2) *Are scales efficient and consistent with assumptions about distributions of student ability?* Figure 5 shows that the two rating scales are functioning differently. Examination of the scale category statistics, however, reveals that both scales are satisfactory. Each score level was most probable for a range of ability; all score levels were utilized although the lowest level for AP was infrequent (see Table 2, the percentage of ratings at each scale level). The scale for AP yielded a near normal distribution of scores, the scale for SR produced a flatter distribution. The ability ranges for the non-extreme score levels of the AP scale are remarkably similar. The ranges for the SR scale are less similar and are smaller, with the exception of level 5.

The two tests reported in this study, Audio-Pal and Story Retell, were developed to provide efficient estimates of speaking ability in a school district where student abilities differed from school to school. Audio

Table 2 Percentage of ratings at each score level

Scale level	Story retell (%)	Audio-Pal (%)
6	8	10
5	27	24
4	25	34
3	17	19
2	13	10
1	11	3

Pal could be recommended for that purpose. Reliability was satisfactory; the rating scale functioned well; teachers rated well with little training; it was an easy test to administer and was well accepted by all of the teachers who participated in the study. We suspect, moreover, that differences in rater performance could be reduced with some group scoring experience (cf. Tyndall and Kenyon, 1996; Weigle, 1994a; Wigglesworth, 1993). More important to this paper, we are justified in accepting the results obtained incidentally to development of the tests.

3 Systematic effects on scores

Three questions, answerable through the FACETS analysis, relate to systematic method effects upon test scores. A further question investigates a possible source for severity differences among raters.

1) How much do tasks (i.e., tests) that are designed to be equivalent actually differ in difficulty?
2) Are teacher-raters equally severe? Are they individually consistent?
3) Are teacher-raters biased with respect to test tasks?
4) Do teachers who have been scale constructors rate differently from other raters?

1) *How much do tasks (i.e., tests) that are designed to be equivalent actually differ in difficulty?* The analysis reports a 1.48 logit difference in the difficulty level of the two tasks. SR is .74 logits and AP is –.74 logits. This difference is shown in Figure 5. The separation index is 7.76 and the reliability coefficient is .98. The analysis reliably distinguishes between different levels of task difficulty. The fixed (all same) chi-square is 122.3 with $df = 1$ and $p = .00$, therefore the null hypothesis of no difference must be rejected. Furthermore, not only do the tasks present different levels of difficulty, but differences in the respective scale steps lead to different scoring on the two tasks. Although most students, regardless of ability level, would be rated at the same or the adjacent scale level on the two tasks, the two tests examined here cannot be considered equivalent.

2) *Are teacher-raters equally severe? Are they individually consistent?* Rater behaviour can be analysed in terms of relative severity, and also in terms of consistency within individual raters.

Figure 5 reveals that, although eight raters appear to cluster around the center of 0, three raters are near the extremes of 1 and –1 logits, approximately 2 logits difference in severity. The fixed chi-square for rater severity is 93.2 with $df = 11$ and $p = .00$. In other words, the raters are not equally severe. Separation of raters is 2.54, with a

reliability of .87. This reliability coefficient indicates that the analysis is fairly reliably separating raters into different levels of severity.

Within-rater consistency measures are represented in FACETS by two measures of fit: the infit and the outfit. The infit is the weighted mean-squared statistic which is sensitive to unexpected responses near the point where decisions are made. The outfit has the same form but is the unweighted mean-squared statistic and is more sensitive to outliers (extreme scores). For the purposes of this study the infit statistic will be analysed. The literature provides no hard and fast rules for setting upper and lower limits. These would depend on the nature of the study. As guidelines, however, Lunz and Stahl (1990) have suggested lower and upper limits of .5 and 1.5 respectively. Lower than .5 indicates too little variation, lack of independence, or overfit. Greater than 1.5 indicates too much unpredictability in rater scores. Linacre (1989–1993b) suggests .7 or .8 and 1.3 or 1.2. The rater measurement report of infit statistics shows four who are extreme by Linacre's standards: raters 6 and 12 highly conforming at .6; rater 1 at 1.4 and rater 7 at 1.5 tend towards unpredictability. None of these raters exceed the limits proposed by Lunz and Stahl, however. FACETS also provides standardized infit statistics with an expected mean of 0 and standard deviation of 1. This shows the degree of variability in individual rater scores relative to the amount of variability in the entire data set and proves helpful in comparing elements of a facet. Greater than 2 or less than –2 are considered indications of misfit. None of the raters have standardized infit statistics outside of this range.

In summary, raters are not equally harsh, but they rank students in much the same way and they are reasonably consistent in their own rating behaviour.

3) *Are teacher-raters biased with respect to test tasks?* To further explore rater behaviour, a bias analysis was conducted between rater and task. After correcting for bias, there were 3 unexpected responses that remained. This is an acceptably low number.

The Bias/Interaction Calibration Report lists the extent of bias for each rater. Rater 8 had significant bias, being unexpectedly severe on SR (Z-score of 3.3, with $Z > 2.0$ indicating significant bias). Raters 2 and 10 showed no bias on either SR or AP. All other raters demonstrated bias, but the direction of bias differed. Severity bias against SR (i.e., judging SR more harshly than AP) was shown by Raters 3, 6, 8 and 12. There was no severity bias for AP. Leniency bias was shown for SR by Raters 1, 4, 5 and 7. Leniency for AP was shown by Raters 9 and 11. In summary, 10 of the raters showed bias, with only Rater 8 showing significant bias. Eight of the 10 raters showed

a greater degree of bias towards SR than AP. To summarize, one of the raters showed significant bias, giving unexpectedly severe ratings for SR. Although there were no other significant rater-by-task biases, there was a clear tendency for raters to agree on AP scores, but to give unexpectedly high or low ratings on SR.

4) *Do teachers who have been scale constructors rate differently from other raters?* A two-facet analysis, subjects and raters, was run on the AP data set. Only the AP results provided a sufficient amount of data for meaningful analysis, so a comparable analysis on the SR data set was not undertaken. The range in harshness of the 12 raters is 3.44 logits, with the most severe at 1.74 and the most lenient at –1.70. The six most severe raters were the developers of the scale. They ranged from 1.74 to .35 logits. The remaining raters, who did not construct the scale, were more lenient with a range from .00 to –1.70 logits. See Table 3. The probability for such a split is less than .005. We have no evidence pointing towards a particular explanation for this finding. It seems, however, that raters tend to be lenient when rating instructions or scale descriptors are less well understood or internalized.

Summary of findings on method effects The Facets analysis of the test development data modelled the trait of speaking ability and three different method effects: task difficulty, rater severity and scale step. The results confirm the findings of others that task differences affect scores (cf. Brütsch, 1979; Chalhoub-Deville, 1995; Clifford, 1981; Corrigan and Upshur, 1982; Fulcher, 1994; Kenyon, 1995; O'Loughlin, 1995; Stansfield and Kenyon, 1992). It also confirms general findings that judges differ in severity (cf. Bachman *et al.*, 1995; Elder, 1993; Lumley and McNamara, 1995; Myford *et al.*, 1996; Weigle,

Table 3 Teacher severity on Audio-Pal in relation to the scale constructed

Teacher	Severity (in logits)	Scale construction team
4	1.74	AP
5	0.87	AP
1	0.49	AP
2	0.44	AP
7	0.38	AP
9	0.35	AP
6	0.00	SR
8	–0.37	SR
11	–0.38	SR
12	–0.69	SR
10	–1.13	SR
3	–1.70	SR

1994b; Wigglesworth, 1993). Although severity may be related to permanent characteristics of raters, we found evidence that it might be also related to a rater's involvement in construction of the scale used for ratings. We also found rater by task bias (cf. Brown, 1995; Elder, 1993; Wigglesworth, 1993 for other findings of rater bias).

4 Task effects on discourse

At issue in our examination of the project is: What features of discourse are salient for the assessment of speech generated by two different tasks?

The procedure for developing the rating scales for SR and AP yielded an analysis of those features. This 'analysis' differs considerably from the types of analysis generally employed in SLA research in which an investigator either looks for the existence of preselected features or attempts to provide a comprehensive account of a text.

Before creating a rating scale, the construction team agree in general terms upon the ability, construct or attribute that they wish to measure. The essentials of the scale-making procedure itself can be described as a repeated sequence of three operations: first, a group of scale constructors individually divide a sample of performances (e.g., voice recordings or compositions, etc.) into a better half and a poorer half; then, as a group, they discuss their divisions of the sample and reconcile any differences; finally, they find some characteristic of the samples that distinguishes the two halves, and they state that characteristic in the form of a binary question that can be used in sorting other samples. The procedure is applied to successive subsamples of the original sample. The outcome is a set of questions reflecting characteristics of the performances that are salient for distinguishing among the different ability levels of the group. Figure 6 illustrates formally a set of five binary questions devised by this procedure to sort a set of performances into six levels.

Rating scales developed for different tasks may reveal different features that are salient for demonstrating ability to perform those tasks. For example, the feature that distinguished upper half performance on SR was 'Did the learner produce a coherent story with all three story elements without long pauses?' Better performance on AP was distinguished by 'Did the learner use a variety of structure (at least 2 sentence patterns, with expansions)?' The complete set of salient features for the two tests appear as Figure 7.[6]

---

[6]The rating procedure involves a hierarchical comparison of scale questions with student performance. In this way a performance is rated by considering only two or three features. It is possible, therefore, for a higher rated performance to lack some quality that is present in a lower rated performance. The full scales are presented in Turner and Upshur (1996).

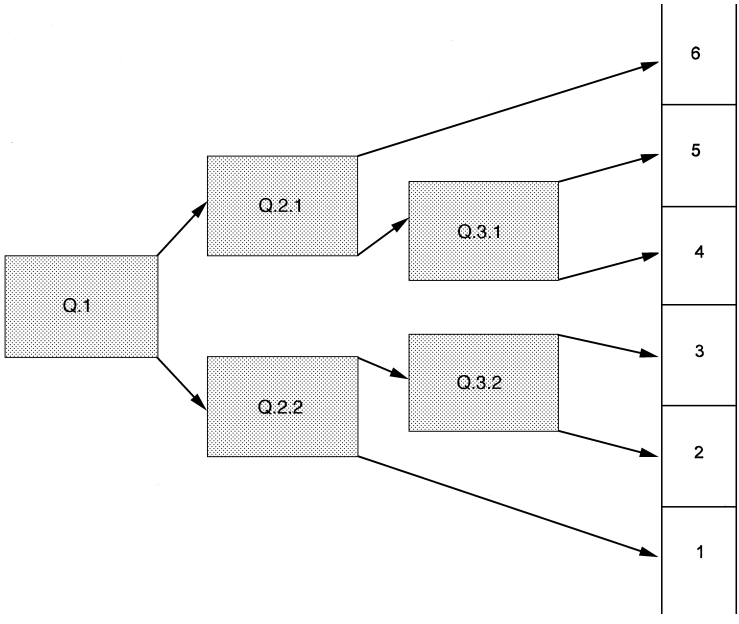

Figure 6 Sample and subsamples for developing scale questions

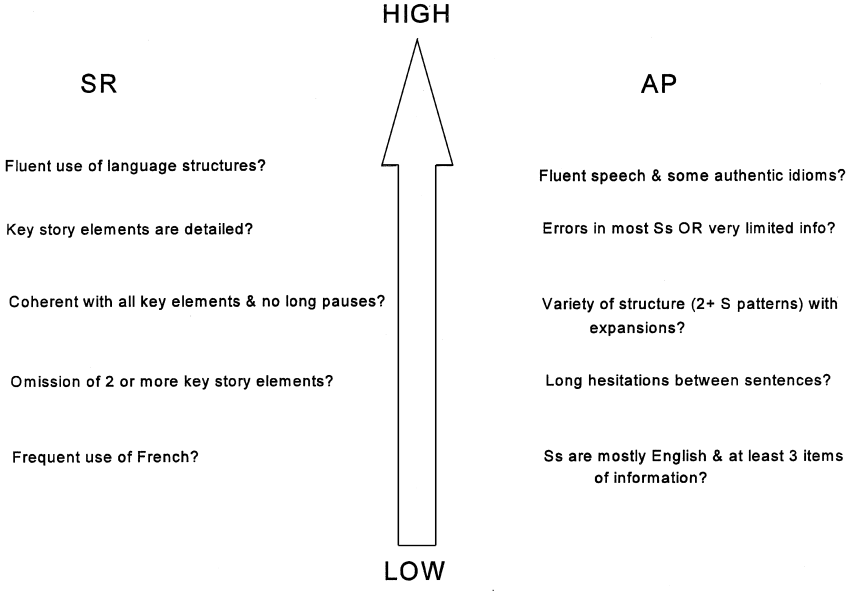

Figure 7 Salient discourse characteristics for two tests at different ability levels

The highest level students are distinguished from the next highest level by fluency regardless of task or scale. Lowest students are typically marked by reliance upon their L1. Otherwise, SR students at intermediate levels on the scale are differentiated by considerations of content in their responses. AP students differ according to formal featues of phonology and grammar.

V Conclusions and implications

1 Scale making as analysis of discourse

In addition to confirming findings by other investigators, the study shows how the tasks stimulate discourse that is judged according to quite different criteria. This difference appeared during scale development. The development teams had to identify salient characteristics of the two sets of discourses that were significant indicators of speaking ability. These different characteristics are incorporated in the two task-specific rating scales. It has not been determined, however, whether the different sets of criteria are themselves ordered in terms of how difficult they are to satisfy.

Although this type of analysis of salient features is useful for scale construction and may be useful for some other purposes (e.g., identifying developmental sequences), it should be recognized that it is by no means comprehensive. Features that do not distinguish learners at different ability levels do not emerge in the scale construction procedure. There may also be other features that distinguish among learners at different ability thresholds, but which were not recognized because of the thresholds selected for scale-making. Also, there may be alternatives to the identified features at the selected level boundaries, alternatives that just happened to be less salient to the scale construction team.

2 Strategies for judgement: bias and rating of pseudo-communication

The bias analysis of our test development data showed that raters were unexpectedly severe or lenient in scoring Story Retell. There are two possible explanations for this finding that have occurred to us. One is related to the overall severity of scale developers in rating Audio-Pal. The second assumes that raters employ different assessment strategies when judging communicative success when they know what a student is trying to convey.

The overall severity explanation is that most of the data are generated by performance on the Audio-Pal task. Therefore, rater severity is determined largely by ratings of Audio-Pal. Any bias is demonstrated

in FACETS for or against Story Retell, the test with the smaller data subset. In fact, however, it seems that the experience of constructing the Audio-Pal scale may influence the scale makers to be more severe. Because of data limitations, it is not possible to know whether the developers of the scale for Story Retell are more severe in their use of that scale. Without this information we cannot generalize about a relation between scale construction and rating severity.

The assessment strategy explanation is based upon consideration of the observed bias and also upon qualitative differences in the rating scales for the two tests. In Story Retell a student's intentions are fully predictable by the rater; the rater knows the story. In Audio-Pal, in contrast, the rater cannot predict precisely what a student might choose to say. For this reason it is impossible for a judge to be certain if a given interpretation corresponds to speaker intentions. Furthermore, it is often impossible to verify student assertions. These differences in predictability and possibilities for judgement seem to be reflected in the two sets of scale descriptors. The most important quality in rating Story Retell was the inclusion of elements of the story. This may be because it is possible to determine whether they are included. In rating Audio-Pal it was variety of sentence structure. This seems to be a fall-back procedure for inferring communicative language ability. In summary, it seems that if raters know what the child wants to say (possible with Story Retell), they can check to see if it is said. However, if they do not know what is intended (as in Audio-Pal), they tend to estimate the formal resources at the speaker's command, and then from that estimate they infer success in communicating.

Different rating strategies can account for some raters being unexpectedly lenient in judging Story Retell and others being unexpectedly severe. Their own knowledge of the story allowed some raters to better understand what a student was trying to say and to rate the performance according to this enhanced understanding. For other raters, their story knowledge seemed to provide a standard by which they could identify a student's communicative infidelities. Thus, there may be two sorts of strategies that define judges of pseudo-communicative task performance. Knowledge of communicative intent allows, probably unconsciously, some to overrate for success and others to underrate for failures.

In this way we are able to infer a plausible difference in strategy for rating pseudo-communicative discourse that could account for the bias noted in this study.

3 Nature of performance-test scales

None of the literature cites direct comparisons among the three scale types described above: absolute proficiency ratings, task proficiency

ratings and task-specific rating. The test development project also employed only one type of scale. Tentative conclusions may be drawn, however.

First, we noted important qualitative differences in the salient characteristics of discourse produced in the two tasks and also found reliable differences in task difficulty. These findings agree with SLA studies on linguistic variation and with the LT studies that have examined task effects: one cannot ignore systematic task effects upon performance test scores. Confidence in absolute proficiency ratings does not appear warranted.

Second, the findings also cast doubt upon the validity of task proficiency ratings. One reason for doubt is the differences between the salient qualities of discourse that emerged in the construction of the scales. As noted above, these differences appear to be related to the predictability of speaker intentions in the tasks, and are therefore not likely to be artifacts of having two-scale development teams. A second reason is the functional difference between the scales as shown in Figure 5. This reason is less persuasive as it might possibly be caused by differences in student ability levels in the scale-development samples.

The weight of evidence suggests, therefore, that rating scales should be task-specific, not just population-specific. If this is true for brief performance tasks designed to measure a single, higher-order construct, it may be true as well when one is designing lengthier tasks to test for multiple constructs. On the basis of our evidence we do not believe that a more general scale-type should be assumed. A further implication of our findings is that effective rating scales may reflect task demands as well as discourse types. This relation between task and rating scale has not been studied in either the LT or the SLA traditions. Reliance upon task-specific rating scales does nothing to solve the problem of generalization. One would clearly like to have general rating scales that could be applied to a wide range of tasks in a manner such that a given rating would always have the same interpretation. Claims that some scales do actually have such a general nature should be received with skepticism. One can speculate that raters who employ a single, standard scale to rate performance on a variety of tasks may in fact reinterpret that scale for each different task. In this way the actual rating scales may reflect task influences even though the ostensible rating scale is invariant.

4 Test taking and rating

In this article we have reported findings consistent with (1) LT research on systematic effects on ratings and (2) SLA research on

systematic effects of task on discourse. These consistencies give confidence that our findings are valid. An adequate explanation for them is lacking, however, because there is no theory of method to explain how particular aspects of method affect discourse and how those discourse differences are then reflected in test scores. We proposed the presence of different rating strategies as a result of the way task type determines the importance of various discourse characteristics. We find no theory to explain this relationship. Nor is there a developed explanation of how rater and examinee characteristics interact with one another and with discourse characteristics to yield ratings, or how tasks relate to well-functioning rating scales. We find that simply concatenating the traditional research approaches of LT and SLA, as schematized in Figure 4, fails to consider a number of relationships that must be addressed in a full account of performance testing. A somewhat more complex model is sketched in Figure 8. We have included a relation between task and rater to indicate our finding that judges seem to adapt their strategies to task demands. We have not included other relations that we have speculated upon in this paper. However, we think that they should be investigated. Developing and evaluating a more sophisticated model of performance testing appears an important challenge for LT and for SLA.

We have demonstrated that the use of task-specific rating scales is integral to a more comprehensive view of the process of performance

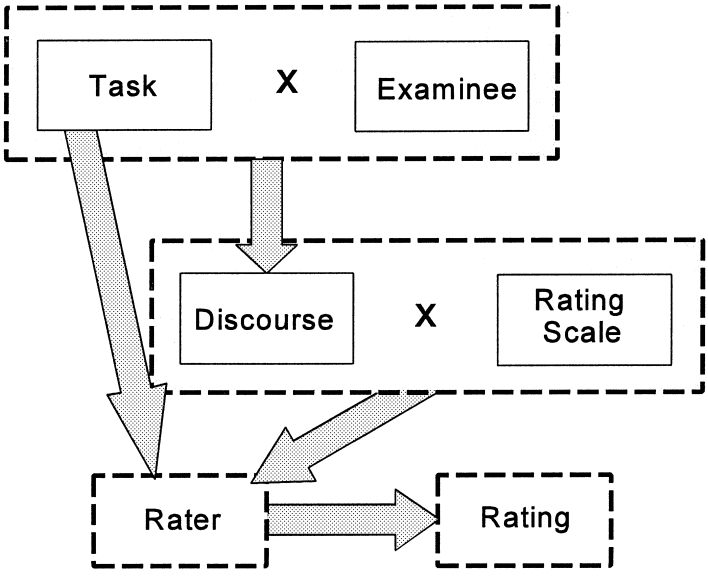

Figure 8 Augmented model: Performance test taking and scoring

test taking and scoring. We are encouraged that others are suggesting that empirical procedures be used in the development of rating scales and that rating scales be task-specific (e.g., Chalhoub-Deville, 1995; Fulcher, 1987; 1988; Shohamy, 1990). Moreover, some are even putting this into practice (Fulcher, 1996). Based on his own research, Fulcher (1996: 228) suggests '. . . that a data-based approach to rating scale development appears to be promising, and that further research should be carried out into the description and operationalization of constructs for language testing, reinforcing the necessary link between applied linguistics, second language acquisition research and language testing theory and practice'. Within this view there are still many unresolved issues. When using empirical methods of scale construction, the composition of construction teams and the make-up of samples of performances may have effects that deserve study. A more pressing issue relates to the tension between the needs for accuracy in assessing a particular performance and generalization to broader domains of language use.

Acknowledgements

We should like to acknowledge gratefully the assistance of Dorry Kenyon, Michael Laurier, Patsy Lightbown and Nina Spada who read earlier versions of this manuscript and provided useful advice and encouragement. They are absolved from responsibility for any shortcomings that remain. Special thanks are due to Dorry for critical advice on the best way to employ FACETS in the analysis of our data. We thank also the anonymous reviewers for *Language Testing* whose insightful comments have guided us in making needed revisions.

VI References

Bachman, L. 1990: *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L.F. and **Cohen, A.D.,** editors, in press: *Interfaces between second language acquisition and language testing research*. New York: Cambridge University Press.

Bachman, L.F., Lynch, B.K. and **Mason, M.** 1995: Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing* 12, 238–52.

Beebe, L. 1980: Sociolinguistic variation and style-shifting in second language acquisition. *Language Learning* 30: 433–47.

Beebe, L. and **Zuengler, J.** 1983: Accommodation theory: an explanation for style-shifting in second language dialects. In Wolfson, N. and Judd,

E., editors, *Sociolinguistics and second language acquisition*. Rowley, MA: Newbury House, 195–213.

Brown, A. 1993: The role of test-taker feedback in the test development process: test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing* 10, 277–303.

Brown, A. 1995: The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing* 12, 1–15.

Brütsch, S. 1979: Convergent-discriminant validation of prospective teacher proficiency in oral and written French by means of the MLA Cooperative Language Proficiency Tests for Teachers (TOP and TWP), and self-ratings. Unpublished PhD dissertation, University of Minnesota, Minneapolis, MN.

Campbell, D. and **Fiske, D.** 1959: Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin* 56, 81–105.

Chalhoub-Deville, M. 1995: Deriving oral assessment scales across different tests and rater groups. *Language Testing* 12, 16–33.

Clifford, R. 1981: Convergent and discriminant validation of integrated and unitary language skills: the need for a research model. In Palmer, A., Groot, P. and Trosper, G., editors. *The construct validation of tests of communicative competence*. Washington DC: TESOL, 62–70.

Corrigan, A. and **Upshur, J.** 1982: Test method and linguistic factors in foreign language tests. *IRAL XX*, 313–21.

Dickerson, L. 1975: The learner's interlanguage as a system of variable rules. *TESOL Quarterly* 9, 401–7.

Douglas, D. and **Selinker, L.** 1985: Principles for language tests within the discourse domains: theory of interlanguage. *Language Testing* 2, 203–26.

Elder, C. 1993: How do subject specialists construe language proficiency? *Language Testing* 10, 235–54.

Elder, C. 1995: The effect of language background on 'foreign' language test performance: Problems of classification and measurement. *Language Testing Update* 17, 36–38.

Ellis, R. 1987: Interlanguage variability in narrative discourse: style-shifting in the use of the past tense. *SSLA* 9, 1–20.

Ellis, R. 1994: *The study of second language acquistion*. Oxford: Oxford University Press.

Fulcher, G. 1987: Test of oral performance: The need for data-based criteria. *English Language Teaching Journal* 41, 287–91.

Fulcher, G. 1988: *Lexis and reality in oral testing*. Washington DC: ERIC Clearing House on Languages and Linguistics (ED 298 759).

Fulcher, G. 1994: Some priority areas for research in oral language testing. *Language Testing Update* 15, 39–47.

Fulcher, G. 1996: Does thick description lead to smart tests? A data-based approach to rating sclae construction. *Language Testing* 13, 208–38.

Green, B.F. 1994: Differential item functioning: techniques, findings and prospects. In Laveault, D., Zumbo, B.D., Gessaroli, M.E. and Boss,

M.W., editors, *Modern theories of measurement: problems and issues*. Ottawa, ON: University of Ottawa, 141–62.

Henning, G. and **Davidson, F.** 1986: Scalar analysis of composition ratings. In Baily, K., Dale, T. and Clifford, R., editors, *Language testing research, selected papers from the 1986 colloquium*. Monterey, CA: Defense Language Institute, 24–38.

Kenyon, D.M. 1995: An investigation of the validity of the demands of tasks on performance-based tests of oral proficiency. Paper presented at the 17th annual Language Testing Research Colloquium, Long Beach, CA.

Kunnan, A.J. 1995: Test taker characteristics and test performance: a structural modelling approach. In Milanovic, M., series editor, *Studies in language testing, Number 2*. Cambridge: University of Cambridge Local Examinations Syndicate.

Lazaraton, A. 1996: Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing* 13, 151–72.

Linacre, J.M. 1989–1993a: *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J.M. 1989–1993b: *Users guide to Facets: Rasch measurement computer program*. Chicago, IL: MESA Press.

Linacre, J.M. 1994: *FACETS (Version 2.75)* [Computer software]. Chicago, IL: MESA Press.

Lumley, T. and **McNamara, T.F.** 1995: Rater characteristics and rater bias: implications for training. *Language Testing* 12, 54–71.

Lunz, M.E. and **Stahl, J.A.** 1990: Judge severity and consistency across grading periods. *Evaluation and the health professions* 13, 425–44.

Lunz, M.E., Wright, B.D. and **Linacre, J.M.** 1990: Measuring the impact of judge severity on examination scores. *Applied Measurement in Education* 3, 331–45.

McNamara, T.F. 1990: Item Response Theory and the validation of an ESP test for health professionals. *Language Testing* 7, 52–75.

McNamara, T.F. 1996: *Measuring second language performance*. London: Longman.

McNamara, T.F. and **Adams, R.J.** 1991/1994: Exploring rater characteristics with Rasch techniques. In *Selected papers of the 13th Language Testing Research Colloquium (LTRC)*. Princeton, NJ: ETS (ERIC Document Reproduction Service ED 345 498).

Myford, C.M., Marr, D.B. and **Linacre, J.M.** 1996: *Reader calibration and its potential role in equating for the TWE* (TOEFL) Research Report No. 52). Princeton, NJ: Educational Testing Service.

O'Loughlin, K. 1995: Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing* 12, 217–37.

Pavesi, M. 1986: Markedness, discoursal modes and relative clause formation in a formal and an informal context. *SSLA* 8, 38–55.

Pica, T., Holliday, L., Lewis, N., Berducci, D. and **Newman, J.** 1991: Language learning through interaction: what role does gender play *SSLA* 13, 343–76.

Plough, I. and **Gass, S.** 1993: Interlocutor and task familiarity: effects on interactional structure. In Crookes, G. and Gass, S., editors, *Tasks and language learning: integrating theory and practice* (35–56). Cleveden, UK: Multilingual Matters.

Pollitt, A. and **Hutchinson, C.** 1987: Calibrated graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing* 4, 72–82.

Pollitt, A. and **Murray, N.** 1993: What raters really pay attention to. Paper presented at the Language Testing Research Colloquium, Arnhem.

Porter, D. 1991: Affective factors in the assessment of oral interaction: gender and status. In S. Anivan, editor, *Current developments in language testing*. Singapore: SEAMEO RELC, 99–102.

Reid, J. 1988: Quantitative differences in English prose written by Arabic, Chinese, Spanish and English writers. Unpublished doctoral dissertation, Colorado State University, Fort Collins, CO.

Rose, K.R. and **Ono, R.** 1995: Eliciting speech act data in Japanese: the effect of questionnaire type. *Language Learning* 45, 191–223.

Samuda, R.S., Kong, S.L., Cummin, J., Lewis, J. and **Pascual-Leone, J.,** editors, 1989: *Assessment and placement of minority students*. Kingston, ON: C.J. Hogrefe.

Sato, C. 1985: Task variation in interlanguage phonology. In Gass, S. and Madden, C., editors, *Input in second language acquisition*. Rowley, MA: Newbury House, 181–96.

Shohamy, E. 1990: Discourse analysis in language testing. *Annual Review of Applied Linguistics* 11, 115–28.

Simon, M.G. 1994: Differential item functioning: applicability in a bilingual context. In Laveault, D., Zumbo, B.D., Gessaroli, M.E. and Boss, M.W., editors, *Modern theories of measurement: problems and issues.* Ottawa, ON: University of Ottawa, 163–69.

Stansfield, C. 1986: A history of the Test of Written English: the development year. *Language Testing* 3, 224–34.

Stansfield, C. and **Kenyon, D.** 1992: Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *SYSTEM* 20, 347–64.

Sunderland, J. 1995: Gender and language testing. *Language Testing Update* 17, 24–35.

Takahashi, T. 1989: The influence of the listener on L2 speech. In Gass, S., Madden, C., Preston, D. and Selinker, L., editors, *Variation in second language acquisition, volume I: discourse and pragmatics*. Clevedon, UK: Multilingual Matters, 245–79.

Tarone, E. 1985: Variability in interlanguage use: a study of style-shfiting in morphology and syntax. *Language Learning* 35, 373–403.

Turner, C. and **Upshur, J.** 1996: Developing rating scales for the assessment of second language performance. In Wigglesworth, G. and Edler, C., editors, *Australian review of Applied Lingustics, series S, number 13: the language testing cycle: from interception to washback.* Melbourne: ARAL, 55–79.

Tyndall, B. and **Kenyon, D.M.** 1996: Validation of a new holistic rating

scale using Rasch multi-facted analysis. In Cumming, A. and Berwick, R., editors, *Validation in language testing*. Clevedon, UK: Multilingual Matters, 39–57.

Upshur, J. and **Turner, C.** 1995: Constructing rating scales for second language tests. *ELT Journal* 49, 3–12.

Varonis, E. and **Gass, S.** 1985: Non-native/non-native conversations: a model for negotiation of meaning. *Applied Linguistics* 6, 71–90.

Vaughan, C. 1991: Holistic assessment: what goes on in the reader's mind? In Hamp-Lyons, editor, *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex Publishing Corporation, 111–25.

Weigle, S.C. 1994a: Effects of training on raters of ESL compositions. *Language Testing* 11, 197–223.

Weigle, S.C. 1994b: *Using FACETS to model rater training effects*, Paper presented at the 16th annual Language Testing Research Colloquium, Washington DC.

Wigglesworth, G. 1993: Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing* 10, 305–35.

Wright, B.D. and **Masters, G.N.** 1982: *Rating scale analysis*. Chicago, IL: MESA Press.