Title
Weighting Patterns and Rater Variability in an English as a Foreign Language Speaking Test

Author
Cai, Hongwen

Publication Date
2012

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Weighting Patterns and Rater Variability in an English as a Foreign Language Speaking Test

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Applied Linguistics

by

Hongwen Cai

2012

ABSTRACT OF THE DISSERTATION

Weighting Patterns and Rater Variability

in an English as a Foreign Language Speaking Test

by

Hongwen Cai

Doctor of Philosophy in Applied Linguistics

University of California, Los Angeles, 2012

Professor Lyle F. Bachman, Chair

This study is an attempt to measure the weighting patterns of the raters in a large-scale English as a Foreign Language (EFL) speaking test, classify these raters according to their weighting patterns, characterize the different types of raters in the rating process, and associate the rater types with different patterns of rater variability. The context was the Test for English Majors - Band 4, Oral Test (TEM4-Oral), a high-stakes certification test administered to college EFL majors in China toward the end of their sophomore year. To quantify the weighting patterns, 126 nonnative-speaking college teachers of English who served as TEM4-Oral raters in 2010 were requested to judge the EFL oral proficiency of 120 hypothetical test-takers with computer-generated score profiles featuring strengths and weaknesses in various criteria. Their relative weights on the criteria were derived from regression analyses, and then fed into cluster analyses to classify the raters into three types. To characterize different types of raters, a sample

of 21 raters were involved in verbal protocols and requested to rate the performance of five real test-takers and justify their ratings. To associate the rater types with the different patterns of rater variability, the real ratings of 33 raters including all three types were analyzed through Many-Facet Rasch Measurement, Hierarchical Linear Modeling, Generalizability Theory, and Confirmatory Factor Analysis. The cluster analyses classified the raters into three types according to whether they gave the largest weights to form-related criteria or content-related criteria, or were balanced in the weighting patterns, and the three types were named form-oriented, content-oriented, and balanced respectively. In the verbal protocols, the form-oriented raters were found to be most severely subject to the anchoring and masking effects of pronunciation and intonation whereas the content-oriented raters displayed the strongest mitigation of such effects. The balanced raters came in between, but shared more similarity with the content-oriented raters. In association with rater variability, the form-oriented raters were found to be most severe among the three types and the content-oriented raters most lenient. On specific TEM4-Oral subscales, the form-oriented raters were unexpectedly severe in pronunciation and intonation, but unexpectedly lenient in grammar and vocabulary, whereas the content-oriented raters were unexpectedly lenient on the content-related subscale of discussion but unexpectedly severe on the subscale of grammar and vocabulary. However, no clear-cut relationship was found in reliability and restriction of range, and mixed results were reported in terms of halo effect.

The dissertation of Hongwen Cai is approved.

Peter M. Bentler

Frederick D. Erickson

Noreen M. Webb

Lyle F. Bachman, Committee Chair

University of California, Los Angeles

2012

To Weihua, my doctoress.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

There are many people to whom I wish to express my gratitude, for making my doctoral study at UCLA possible, for motivating and inspiring me during my doctoral years, and for supporting me in this study. The one individual who has played all these roles is Prof. Lyle F. Bachman, my advisor. To avoid making this a biography, let me limit my description of "the professor" (as Dr. Anthony Kunnan refers to him in front of us) to three words: strict, nice, and wise. I am sure that all Lyle's students will agree to the first two adjectives, so I will only quote two of his sayings as evidence of his wisdom. "The most important thing is to get things done." "Write something down each day". Simple as they may sound, these principles have been my most effective weapon against procrastination in the completion of my doctoral study and this dissertation in particular.

Let me also give a sketch of the other professors in my committee, in the sequence of my acquaintance with them. These are all celebrities in their fields, but here I will only describe how they have impressed and inspired me. Prof. Frederick D. Erickson successfully washed my brain and convinced me of the value of qualitative studies, especially through the term paper requiring us to compare qualitative to quantitative studies. The verbal protocol analysis in this dissertation resulted directly from the brain washing. Prof. Peter M. Bentler surprised me in his feedback to my second qualifying paper with his comments on specific grammatical errors and the format of my references. He also referred me to Dr. Albert Satorra for a new statistical procedure concerning the Satorra-Bentler scaled chi-square difference test. I am glad to add here that my manuscript based on this paper has now been accepted by *Language Testing*. Peter's comment on my dissertation proposal resulted in Figure 5.1 of this dissertation. Similarly, Prof. Noreen M. Webb's comment on my proposal has directly motivated my discussion of the interaction between rater weighting patterns and test-taker profiles in Section 6.5. I have only audited Noreen's G-Theory course, but her lucid explanation of complex statistical concepts has inspired me in the drafting of this dissertation. During my preliminary oral exam, Peter emphasized the importance of a clear account of the data structure, while Noreen and Fred doubted that my research plan was "too ambitious". I hope that I have lived up to their expectations in general.

Beyond this dissertation, the professors of UCLA and my supervisors in CRESST have supported, inspired, and impressed me in various ways. I would like to quote Prof. Olga T. Yokoyama and Dr. John L. Lee here to show how willing to help these nice people have been. "You can always count on my support, ever since your paper on language universals" (Olga). "We will support you until you graduate" (John). Let me also give two quotations to show how these people have inspired and impressed me. "You don't take a fish out of water and test how well it swims" (Prof. Charles Goodwin, who inspired my thinking inadvertently). "Let's do the right thing" (Dr. William L. Bewley, who decided that we should report the poor fit of a CFA model as it was). Beside wisdom and integrity, these people have a kind heart. I still remember Fred's tears during class for our presumptive misunderstanding of a primary school teacher featured in a video, who was regarded as impatient due to her repeated sighs in addressing her students, which turned out to be the symptom of a heart disease. I also remember Olga's words after she made sure that the medical personnel took good care of her student who was struck by a sudden illness at the end of a class. "This is more important than linguistics," she said.

My inspiration has also come from my fellow students in UCLA. These intelligent and diligent people are the best companions I could expect during these years. Witnessing their

graduation, advancement to candidacy, or presentation on various conferences has spurred me on in my own efforts. Thank you, Hoky Min, Yasuhiro Imao, Huan Wang, Youngsoon So, Hsin-min Liu, Yujie Jia, Ikkyu Choi, and Jonathan Schmidgall.

My family has been most supportive during these years. My parents and parents-in-law have helped with all the housework, my wife Weihua has shouldered more than twice the burdens of a mother, plus loneliness, and my sons have grown with much less care from me. I owe them more than I can imagine.

This is an expensive study, which wouldn't have been possible but for the financial aid and assistance from the following benefactors: the late Mr. Harry Lenart and Mrs. Yvonne Lenart (Harry and Yvonne Lenart Graduate Travel Fellowship), Educational Testing Service (TOEFL Small Grant for Doctoral Research in Second or Foreign Language Assessment), and the Institute of Oral English Studies of Nanjing University. I am deeply indebted to these persons and organizations for their kind support. My UCLA years have been supported financially in the forms of an aid package and a GSR position in CRESST, and I was supported by the Dissertation Year Fellowship from UCLA and financial aid from the Department of Applied Linguistics during the study, especially during the drafting of this dissertation, for which I am not only grateful, but have a further reason for my pride of being a Bruin.

I am grateful for the TEM4-Oral raters who participated in this study, especially those people who were involved in the verbal protocols. They have remained anonymous, and have been described only statistically in this dissertation, but for me they are real people worthy of respect. At least, however, I can mention the names of the following people, who have helped me otherwise in this study: Prof. Qiufang Wen, Prof. Xuexi Zhao, Dr. Wenyu Wang, Dr. Ling Wang, Ms. Yan Tang, Ms. Qiyun Li, and Ms. Lan Guo.

I am most grateful for the kind permission to reuse the following published materials: Figure 2.4 and the relevant excerpt on p. 28 from Wolfe (1997) (© Copyright by Elsevier, License Numbers: 2957560698107 and 2957561255330), Table 2.3 and the relevant excerpt on pp. 29-30 from Cumming, Kantor, & Powers (2002) (© Copyright by John Wiley and Sons, License Number: 2957570364506), and Table 2.1 from Saal, Downey, & Lahey (1980) (© Copyright by American Psychological Association, who requires no formal requests for a maximum of three figures or tables from a journal article or book chapter).

VITA

Degrees Earned

1993, BA in English Language and Literature, Guangzhou Institute of Foreign Languages

1998, MA in Linguistics and Applied Linguistics, Guangdong University of Foreign Studies

Professional Employment

1993-1995, translator, Guangzhou Institute of Foreign Languages

1995-1998, assistant lecturer, Guangdong University of Foreign Studies

1999-2008, lecturer, Guangdong University of Foreign Studies

2009-2011, graduate student researcher, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles

Publications

Cai, H. (in press). Partial dictation as a measure of EFL listening proficiency: Evidence from confirmatory factor analysis. *Language Testing.*

Bewley, W. L., Lee, J. L., Jones, B., & Cai, H. (in press). Assessing cognitive readiness in a simulation-based training environment. In H. F. O'Neil, R. S. Perez, & E. L. Baker (Eds.), *Teaching and measuring cognitive readiness*. New York: Springer.

Cai, H., Guo, L., & Pan, C. (2009). *Breakthrough in listening (I).* Chongqing, China: Chongqing University Press.

Cai, H., Guo, L., Jiang, L., & Pan, C. (2009). *Breakthrough in listening (II).* Chongqing, China: Chongqing University Press.

Presentations

Cai, H., Min, S., & Yang, H. (2011). Validity issues in linking language assessments to reference levels. Paper presented at the 14th Annual Conference of Southern California Association for Language Assessment Researchers, California State University, Los Angeles, May 7.

Cai, H. (2010). An AUA view on diagnostic language assessment. Paper presented at the 13th Annual Conference of Southern California Association for Language Assessment Researchers, University of California, Los Angeles, May 1.

Cai, H. (2009). Clustering to inform standard setting in an oral test for EFL learners. Paper presented at the 31st Language Testing Research Colloquium, Denver, Colorado, March 17-20.

Cai, H. (2009). The effect of rater background on inter-rater reliability in language test validation: A generalizability study. Paper presented at the 12th Annual Conference of Southern California Association for Language Assessment Researchers, California State University, Fullerton, May 2.

Chapter 1 Introduction

1.1 Background

The use of performance assessment in language assessment has enjoyed increasing popularity during the past decades (McNamara, 1996; Brown, 2004). Performance assessment has been used in large-scale language assessment programs that are of a high-stakes nature, such as tests for student selection, for professional certification, and for graduation. Justifying the use of performance assessment in these contexts is the fundamental responsibility of the developers and users of the assessment.

From the perspective of the Assessment Use Argument (AUA) framework (Bachman, 2005; Bachman & Palmer, 2010), the use of a performance language assessment, like any other language assessments, is only justified if equitable and values-sensitive decisions are made to bring about beneficial consequences. This, in turn, depends on consistent reports/scores of test-taker performance and meaningful and impartial interpretations of these reports/scores that are generalizable to the Target Language Use (TLU) domain. For a researcher in language assessment, consistent scores and meaningful interpretations of scores are of particular interest.

One of the issues involving both the consistency and meaningfulness problems is the issue of rater variability, also referred to as rater effects, rater bias, or rater errors in the literature of psychological and educational measurement (Engelhard, 1994; Guilford, 1954; McNamara, 1996; Myford & Wolfe, 2004; Saal, Downey, & Lahey, 1980). The term *rater variability* is an umbrella term that covers a various set of phenomena, typically the following: A) *rater leniency* or *severity*, the general disposition of a rater to give high or low scores; B) *restriction of range*,

the tendency of a rater to limit his ratings within a small range of possible scores; C) *rater reliability* or *agreement*; the degree of agreement of a rater with himself or other raters; and D) *halo effect*, which is evident when a rater makes little distinction between the various criteria in a scoring rubric.

In psychological and educational measurement, various methods for detecting rater variability have been developed, on the basis of the Classical Test Theory (CTT, Gulliksen, 1950; Saal et al., 1980), the Multitrait-Multimethod (MTMM) paradigm (Campbell & Fiske, 1959; Jöreskog, 1974), Generalizability Theory (G-Theory, Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991), and the Rasch Model and its extensions (Andrich, 1978a, 1978b; Rasch, 1960), especially the Many-Facet Rasch Measurement (MFRM) approach (Linacre, 1989; Myford & Wolfe, 2004).

The conception of rater variability and the methods for detecting this in psychological and educational measurement have been extended to the field of language assessment. The introduction of the MFRM approach, in particular, has considerably promoted research in rater variability (Bachman, Lynch, & Mason, 1995; Eckes, 2005, 2008; Engelhard, 1994; Kondo-Brown, 2002; Weigle, 1998). On the other hand, a lot of studies have been conducted by language assessment researchers on the factors that affect rater variability, such as the characteristics of the rater (Lumley & McNamara, 1995; Shi, 2001; Shohamy, Gordon, & Kraemer, 1992; Zhang & Elder, 2011), the test-taker (Haswell & Haswell, 1996; Hughes, Keeling, & Tuck, 1980), the tasks (Bachman et al., 1995; Upshur & Turner, 1999), and the rating scale and rating method in general (Barkaoui, 2007a, 2007b; Weigle, 1999).

From the early days, researchers in language assessment have tried to describe the rating process and how raters weight the various criteria, in an effort to better understand rater

variability (Chalhoub-Deville, 1995; Cumming, 1990; Erdosy, 2004; Freedman, 1979; Lumley, 2002; Milanovic, Saville, & Shen, 1996; Vaughan, 1991). A new direction in this connection is the classification of raters according to their weighting patterns (Eckes, 2008). However, the relationship between weighting patterns and rater variability has remained under-researched.

1.2 Research questions

Compared to other factors that affect rater variability, weighting patterns provide a direct indication of what raters value in the rating process and, as such, are expected to have a direct effect on the actual ratings. For this reason, the classification of raters according to their weighting patterns enables macroscopic descriptions of *types* of raters rather than idiosyncratic rating processes of individual raters. As this kind of study is still in its infancy, the relationship between the types of raters and the various aspects of rater variability has remained largely unexplored, a responsibility to be assumed in this study. The specific research questions are:

1. How successfully can raters be classified into types according to their weighting patterns?
2. How are types of raters different in the rating process?
3. To what degree are patterns of rater variability different across types of raters?

An English as a Foreign Language (EFL) speaking test for Chinese college students provides the context for this study.

1.3 Significance of the study

The meaning of the study lies both in its contribution to a better understanding of language assessment and in its implications for practice, especially in the design and development of performance tests and rating rubrics, and the selection and training of raters.

Rater variability is an issue that cannot be avoided in the assessment of language performance, which typically involves human judgment and decision-making in the rating process (McNamara, 1996; Engelhard, 2002). The following figure provides a schematic explanation of how rater variability arises.

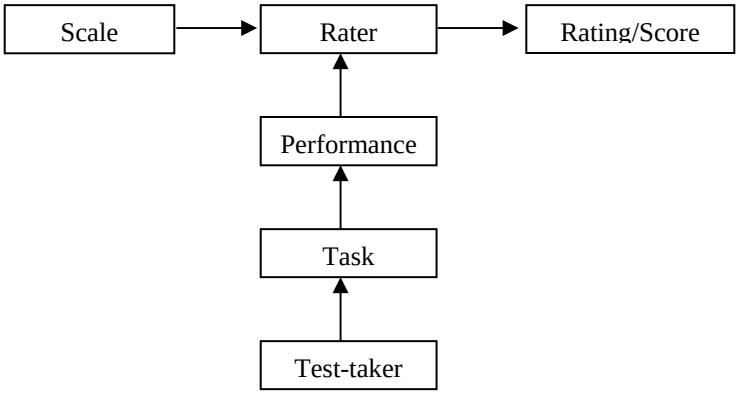

Figure 1.1. The flowchart of language performance assessment

Figure 1.1 is a flowchart of language performance assessment adapted from Kenyon (1992) and McNamara (1996). The key information here is that the path from performance to rating is mediated by the rater, and therefore the actual rating is inevitably subject to rater variability. A similar conceptualization of the rater as a mediating factor can be found in Bachman (2002). As the rater is an essential component of language performance assessment, the meaningfulness of

score interpretation is bound to be defected without accounting for the considerable variability in this process.

An essential part of inquiry into rater variability is the research into its contributing factors. Among these factors, the rating process and weighting patterns deserve special attention, as they are directly related to the final ratings given by the rater. In an effort to meet this need, this study will focus on the relationship between weighting patterns and rater variability, using relatively new approaches to quantify weighting patterns and classify raters (cf. Chalhoub-Deville, 1995; Eckes, 2008). It is hoped that the use of these different methods will further enrich the methodology for language assessment.

On the practical side, if rater types are found to make considerable difference in the process of rating and the actual ratings given by the raters, then this information can be used to guide test development and test use. For example, test developers may revise the rubric to emphasize certain aspects and deemphasize others, so that bias due to different weighting patterns is minimized. Rater trainers may suit training sessions for different types of raters to reduce the differences or, more radically, they may even design a screening test to bar those raters that refuse to be trained.

1.4 Limitations

Weighting patterns is only one way to capture the rating processes of different raters. The broader spectrum of rater variability studies include much more outcome and predictor variables than a single study can capture. The combination of these outcome and predictor variables would only be understood through a whole series of studies.

This limited choice of variables is further limited in theory and in practice. For example, the choice of criteria to be included in the description of weighting patterns is no more than the subjective decisions of the researcher, and so is the choice from a number of ways to classify the raters.

Nonnative-speaking raters are involved in this study, as these are the only raters available for the particular test under study. Therefore, no generalization to native-speaking raters is intended. The test involved is an EFL speaking test for Chinese college students, which limits the domain to second/foreign language assessment at the tertiary level.

Chapter 2 Review of Literature

2.1 Rater variability

2.1.1 A conceptual introduction

When a group of test-takers of various ability levels are rated on their performance, the ratings can deviate from the true distribution of ability in a limited number of patterns. To simplify exposition, suppose there are five test-takers, A, B, C, D, E, and their real ability can be calibrated on a ten-point scale as 1, 3, 5, 7, and 9 respectively (Figure 2.1a). For the time being, suppose the ability being measured is unidimensional in nature. Then there may be three patterns in which the ratings given by a rater may deviate from the true distribution (Figure 2.1, b-d).

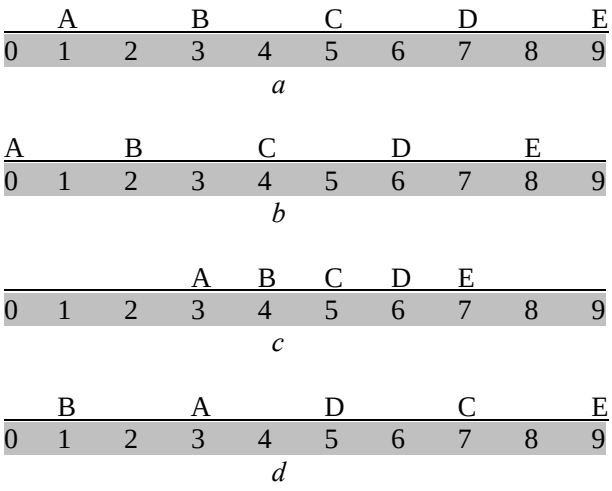

Figure 2.1. Graphical representation of three rater variability patterns

The first pattern of deviation is represented as Figure 2.1b, where all the test-takers are rated lower than their true ability levels. A rater that gives such ratings is considered severe. On

the contrary, when all the test-takers are rated higher than their true ability levels, the rater is considered lenient. In general terms, rater severity or rater leniency is the case of "consistently too high or too low" ratings relative to conceptual "true performance scores" (Saal et al., 1980). In practice, however, as "true performance scores" are unknown, severity is typically operationalized as mean ratings higher than the midpoint within the range of possible observed scores or the arbitrary origin on the scale in latent variable models, and leniency as mean ratings lower than the midpoint or origin.

The second pattern of deviation from the true ability distribution is illustrated by Figure 2.1c, where the ratings are clustered within a narrow space on the scale and variability of ratings is reduced from the true distribution. This situation is referred to as *restriction of range* in the literature. In its broad sense, the term entails ratings "clustered about any point on the rating continuum", though many researchers treat a special case of restriction of range, the so-called *central tendency*, as a separate pattern, where "ratings are clustered about the midpoint of the rating scale, reflecting raters' reluctance to use either of the extreme ends of the continuum" (Saal et al., 1980).

Figure 2.1d displays the third type of rater variability, where the ratings differ in sequence from the true ability levels. Here the positions of test-takers A and B on the scale are reverted, and so is the sequence between test-takers C and D. Such a pattern indicates a problem in rater agreement. When the sequencing of test-takers by the same rater varies on different occasions, intra-rater agreement is of concern; when the sequencing of test-takers varies across raters, inter-rater agreement becomes an issue.

The three patterns of deviation in Figure 2.1 provide a simplified vision of rater severity, restriction of range, and rater agreement. In practice the situation may be much more

complicated, with different combinations of these patterns. For example, when the five test-takers in Figure 2.1a are given scores 2, 3, 4, 5, and 6 respectively, both restriction of range and rater severity are of concern. If the five scores are 2, 4, 5, 3, and 6 respectively, then there is also reason for concern over rater agreement.

The idea of halo effect involves multidimensionality, but may be explicated with similar figures. To simplify presentation, only two criteria are included in the discussion, say, form and content. Suppose again the true standings of the same five test-takers in form is quantified as 1, 3, 5, 7, 9 on the ten-point scale, while their standings in content are 3, 1, 7, 5, 9, as Figure 2.2a shows.

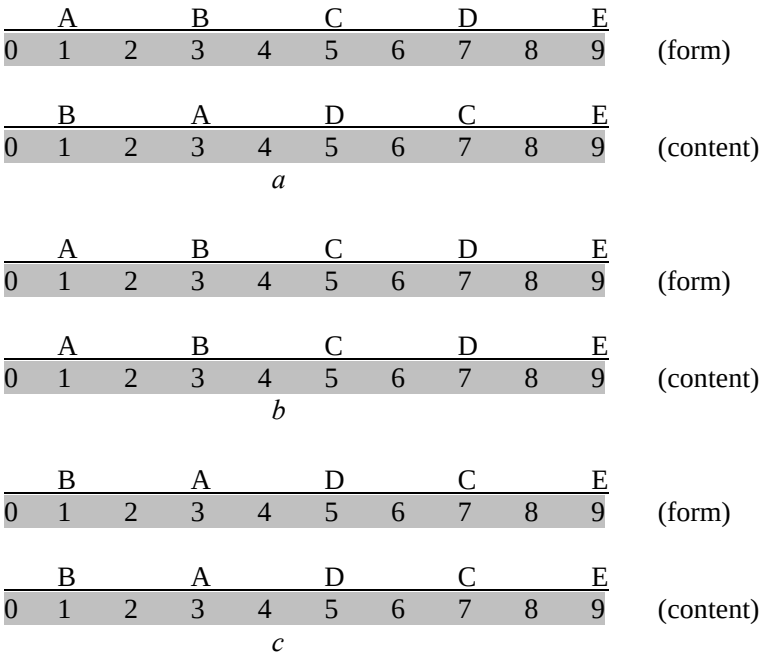

Figure 2.2. Graphical representation of halo effect

While the true ability distribution in form is distinguished from the distribution in content in Figure 2.2a, in actual ratings the two criteria may be indistinguishable from each other for

some raters, who may see form as the dominant criterion and distort content ratings to agree with form ratings (Figure 2.2b), or may be dominated by the content criterion and do the reverse (Figure 2.2c). Either pattern marks the rater's "failure to discriminate among conceptually distinct and potentially independent aspects of a ratee's behavior" (Saal et al., 1980).

2.1.2 Detecting rater variability

In the CTT framework (Gulliksen, 1950; Lord & Novick, 1968), factorial analysis of variance (ANOVA) proved to be the most systematic analysis of rater effects (Guilford, 1954) among a hodgepodge of various methods (Saal et al., 1980). With ANOVA, the rater main effect and its interaction with other factors can be tested for statistical significance to find out whether rater variability has contributed significantly to the rating. Table 2.1 is adapted from a list of different operational definitions from earlier studies compiled by Saal et al. (1980).

Table 2.1

Operational definition of some rater variability phenomena (after Saal et al., 1980)

Rater variability phenomena	Operational definition
Halo effect	Dimension intercorrelations Principal component or factor analysis Standard deviations Rater × Ratee interaction
Leniency or severity	Mean dimension ratings Rater main effect Skewness
Central tendency or range restriction	Mean dimension ratings Standard deviations Kurtosis Ratee main effect
Inter-rater reliability or agreement	Standard deviations Correlations Ratee main effect Rater × Ratee interaction

In Saal et al. (1980), halo effect is captured by high intercorrelations among the different subscores, i.e. if a subscore is given to each aspect of the test-taker's performance. This is a natural consequence of the patterns depicted in Figures 2.2b and 2.2c. If principle component or factor analysis is conducted on the correlation matrix, then halo effect will be indicated by fewer principal components or factors than the design. A single principal component or factor is not uncommon if the number of aspects is not large. A third possible indicator of halo effect is small variance among subscores given by the same rater. Suppose a rater gives five subscores to an essay, if small or no differences are found among the five subscores, then the rater may not discriminate among the five aspects. The fourth operational definition on the list of Saal et al. (1980) is based on a Rater × Ratee × Trait analysis of variance (ANOVA), proposed earlier by Guilford (1954). The indication of the halo effect is statistically significant Rater × Ratee interaction, especially when such interaction explains a large proportion of the rating variance. The underlying logic is that if a rater tends to overrate some ratees and underrate others, then the rater is subject to the failure to discriminate among different aspects of these ratees' behavior.

As discussed above, leniency may simply be operationalized as mean ratings higher than the midpoint, and severity as mean ratings lower than the midpoint. In a Rater × Ratee × Trait ANOVA this would be exhibited as a significant rater main effect, especially when the main effect explains a large proportion of the rating variance. Distribution of ratings provides the third way to gauge leniency and severity, with the former being represented by significant negative skewness and the latter by significant positive skewness.

In agreement with Figure 2.1c, restriction of range may be identified with small variance in the ratings. This extends naturally to the proximity of the mean dimension ratings to the

midpoint of the scale. Just as skewness is an indication of leniency or severity, kurtosis of the rating distribution may be used as a measure of range restriction, where a leptokurtic distribution indicates restriction of ratings to a small range, and a platykurtic distribution is a sign of widespread ratings. In terms of the Rater × Ratee × Trait ANOVA, the absence of a significant ratee main effect is an indication of strong central tendency.

Most operational definitions of inter-rater agreement involve some form of correlation between ratings given by several raters. Apart from correlations, the standard deviations of the ratings assigned to a particular ratee by several raters also reflect the proximity of these ratings. In terms of the Rater × Ratee × Trait ANOVA, a significant ratee main effect is an indication of inter-rater reliability, besides being an indication of the absence of range restriction. Moreover, a significant Rater × Ratee interaction indicates low inter-rater reliability, besides being an indication of strong halo effect.

The ANOVA approach based on CTT may be extended naturally into G-Theory, as analysis in G-Theory is based on the decomposition of variance components obtained from ANOVA. Instead of testing the statistical significance of main effects and interactions, however, G-Theory provides an estimation of the values and relative sizes of variance components for these effects and interactions (Cronbach et al., 1972; Shavelson & Webb, 1991; Brennan, 2001a). While all the main effects and interactions estimated by the ANOVA have their counterparts in a G-study, the generalizability or dependability coefficient estimated from a G-study is a direct indication of the degree of inter-rater consistency.

The use of the MTMM matrix in detecting rater variability was first proposed by Campbell and Fiske in their seminal paper on the MTMM paradigm, in which they laid down the principles for detecting halo effect by regarding each rater as representing a different method (Campbell &

Fiske, 1959). With each measure related to one trait and one method only, the relationship between any two measures may feature one of four logically possible combinations of traits and methods: same trait same method, or *monotrait-monomethod* (MTMM); same trait different methods, or *monotrait-heteromethod* (MTHM); different trait same method, or *heterotrait-monomethod* (HTMM); and different trait different methods, or *heterotrait-heteromethod* (HTHM). Applied to the detection of halo effect, Campbell and Fiske (1959) suggested that this exists when ratings of the same trait by different raters do not correlate higher than ratings of different traits by the same rater. When rater is identified with method, this is tantamount to saying that rater effects are detected when MTHM correlations are lower than HTMM correlations.

The comparison between observed MTHM and HTMM correlations is challenged on the basis that measurement error has not been taken into consideration. Lance, Dawson, Birkelbach, and Hoffman (2010), for example, mathematically proved that the relationships between observed MTHM and HTMM correlations are ambiguous, due to the attenuated correlations between latent traits. This problem is neatly addressed in the CFA approach to the MTMM matrix, which does separate measurement error from the true relationship. Still regarding each rater as representing a different method, rater effects can now be inferred from large method factor loadings in CFA models (Marsh & Grayson, 1995; Marsh & Hocevar, 1988).

So far, the MFRM approach has been the most versatile in detecting rater variability, whose complexity calls for some detail in explication. The MFRM model is an extension to the Rating Scale Model (RSM) proposed by Andrich (1978a, 1978b), intended for m ordered response categories scored with successive integers, such as 0, 1, …, m. According to a rewriting of Andrich's original model by Wright & Masters (1982), the probability of person n responding in

category x to item i is

$$P_{nix} = \frac{\exp \sum_{j=0}^{x} \left[\beta_n - \left(\delta_i + \tau_j\right)\right]}{\sum_{k=0}^{m} \exp \sum_{j=0}^{k} \left[\beta_n - \left(\delta_i + \tau_j\right)\right]} \qquad k = 1, 2, \ldots, m;\ x = 0, 1, \ldots, m \qquad (2.1a)$$

where $\tau_0 \equiv 0$, so that $\exp \sum_{j=0}^{0} \left[\beta_n - \left(\delta_i + \tau_j\right)\right] = 1$ (Wright & Masters, 1982, p. 49). The denominator in Equation (2.1a) is a normalizing factor, so that the probabilities of person n responding in all m categories to item i sum up to 1. The parameters to be estimated for this model are: the ability of person n denoted by β_n, the difficulty of item i denoted by δ_i, and m thresholds corresponding to the $m + 1$ rating categories, denoted by τ_1, τ_2, …, τ_m. The thresholds are defined as the points where the person has an equal probability (.50) of responding in two adjacent categories. The values of the thresholds are assumed to be the same across items (Wright & Masters, 1982, p. 48).

By the same token, the probability of person n responding in category $x - 1$ to item i is

$$P_{ni(x-1)} = \frac{\exp \sum_{j=0}^{x-1} \left[\beta_n - \left(\delta_i + \tau_j\right)\right]}{\sum_{k=0}^{m} \exp \sum_{j=0}^{k} \left[\beta_n - \left(\delta_i + \tau_j\right)\right]}. \qquad (2.1b)$$

It follows that the odds of person n responding in category x relative to the same person responding in category $x - 1$ is

$$\frac{P_{nix}}{P_{ni(x-1)}} = \frac{\exp\sum_{j=0}^{x}\left[\beta_n - \left(\delta_i + \tau_j\right)\right]}{\exp\sum_{j=0}^{x-1}\left[\beta_n - \left(\delta_i + \tau_j\right)\right]}, \tag{2.2}$$

as the denominators cross out. Taking the natural logarithm of both sides of Equation (2.2) and simplifying the equation produces a form that are frequently quoted in studies using MFRM models.

$$\ln\left(\frac{P_{nix}}{P_{ni(x-1)}}\right) = \beta_n - \delta_i - \tau_x. \tag{2.3a}$$

where τ_x denotes the difficulty of category *x* relative to category *x* – 1. The same parameters are estimated, but Equation (2.3a) facilitates an intuitive understanding, as everything is expressed in log-odds units, or logits. Stated in words, the log-odds of person *n* responding in category *x* relative to the same person responding in category *x* – 1 is the person's ability (β_n) minus the item difficulty (δ_i) and the difficulty of category *x* relative to category *x* – 1, denoted by τ_x.

In MFRM models, β_n and δ_i are called facets, different from the technical terms used in G-Theory, where one of the facets is given the special status of the object of measurement, usually the person (ability). Similar to G-Theory, however, new facets can be added to the model. This facilitates the inclusion of rater severity and domain or task difficulty in the model, which then takes the form

$$\ln\left(\frac{P_{nijkx}}{P_{nijk(x-1)}}\right) = \beta_n - \delta_i - R_j - D_k - \tau_x . \qquad (2.3b)$$

where the new terms R_j denotes the severity of Rater j, and D_k denotes the difficulty of Domain k. In terms of actual ratings with successive integers, Equation 2.3b expresses the natural logarithm of the probability of a person obtaining a score of x relative to $x - 1$ on item i after adjusting for rater severity and domain difficulty. If the person's ability is higher than the difficulty of item i after these adjustments, then the log-odds is larger than 1, and the probability of the person getting a score of x relative to $x - 1$ is larger than .50.

In MFRM models, ability levels of different persons, severities of different raters, and difficulties of different domains are estimated simultaneously and placed on the same linear scale and expressed in logits. This is analogous to equating test forms of various difficulties according to a common reference (Engelhard, 2002, p. 269). Placing estimates of different facets on the same scale provides a framework of reference for interpreting the values of estimated parameters. This is especially useful when fully crossed designs are impractical.

The detection of rater variability in the MFRM framework is facilitated by the information from the parameter estimates and the fit indices after fitting the model to the observed data. In essence, persons with higher abilities are expected to earn higher ratings than less able persons, and more difficult items or skills should result in lower ratings than easier ones. Inconsistency with these expected patterns indicates rater bias, which is captured in parameter estimates and fit statistics (Lunz & Linacre, 1998, p. 53).

The most informative fit indices in the MFRM framework are the Outfit and Infit statistics, short for outlier-sensitive and inlier-sensitive fit statistics. One essential element that goes into

the calculation of the Outfit mean-square for item *i* is the model-implied variance, or expected variance, of the rating x_{nij} given to person *n* on item *i* by rater *j*:

$$Var\left(x_{nij}\right) = \sum_{x=0}^{m} \left(x - E\left(x_{nij}\right)\right)^2 P_{nijx} \tag{2.4a}$$

where P_{nijx} is the probability of person *n* being rated on item *i* by rater *j* in category *x*, defined in the same way as in previous equations; and $E\left(x_{nij}\right)$is the expected value of x_{nij} , calculated as $E\left(x_{nij}\right) = \sum_{x=0}^{m} xP_{nijx}$. The other essential element for calculating the Outfit mean-square is the observed variance of the ratings on item *i*:

$$S^2\left(x_{nij}\right) = \sum_{n=1}^{N}\sum_{j=1}^{J} \frac{\left(x_{nij} - E\left(x_{nij}\right)\right)^2}{NJ} \tag{2.4b}$$

where *N* is the number of persons and *J* the number of ratings on item *i* per person. The Outfit mean-square, U_i, for item *i* is the ratio of these two variances:

$$U_i = \frac{S^2\left(x_{nij}\right)}{Var\left(x_{nij}\right)} = \sum_{n=1}^{N}\sum_{j=1}^{J} \frac{\left(x_{nij} - E\left(x_{nij}\right)\right)^2 / Var\left(x_{nij}\right)}{NJ} = \sum_{n=1}^{N}\sum_{j=1}^{J} \frac{z_{nij}^2}{NJ}, \tag{2.4c}$$

where $z_{nij}^2 \equiv \left(x_{nij} - E\left(x_{nij}\right)\right)^2 / Var\left(x_{nij}\right)$, the standardized residual squared. In contrast, the Infit mean-square, V_i, for item *i*, is the ratio of the sum of squares of rating residuals relative to the model-implied, or expected, sum of squares:

$$V_i = \frac{\sum_{n=1}^{N}\sum_{j=1}^{J} z_{nij}^2 Var\left(x_{nij}\right)}{\sum_{n=1}^{N}\sum_{j=1}^{J} Var\left(x_{nij}\right)} = \frac{\sum_{n=1}^{N}\sum_{j=1}^{J} \left(x_{nij} - E\left(x_{nij}\right)\right)^2}{\sum_{n=1}^{N}\sum_{j=1}^{J} Var\left(x_{nij}\right)}. \quad (2.5)$$

According to Wright and Masters (1982, pp. 99-101), the Outfit is sensitive to unexpectedly high or low difficulty (outliers) relative to the ability of person *n* and severity of rater *j*, hence the name Outfit, outlier-sensitive fit statistic. In contrast, the Infit carries with it different weightings for the squared residuals, so that extreme difficulties have less influence on the magnitude of the item fit statistic. As a result, it is sensitive to inliers, or an accumulation of small deviations that are less or more consistent than expected (Lunz & Linacre, 1998, p. 54), hence the name Infit, inlier-sensitive fit statistic. As the weightings are indicators of amount of information, the Infit is also called the information-weighted fit statistic.

The commonality between the Infit and Outfit statistics is that both can be expressed as the ratio of observed statistics relative to their expectation. The Outfit compares the observed variance of item *i* with the model-implied variance, while the Infit compares the sum of squared rating residuals with the expected sum of squares. Therefore, both have expectation of 1 (Lunz & Linacre, 1998, pp. 53-54).

This account of Infit and Outfit mean-squares for item *i* extends immediately to the same statistics for person *n* or rater *j*, though the summation in Equations 2.4c and 2.5 will be over *I* items and *J* raters for the calculation of person fit statistics, and over *N* persons and *I* items for the calculation of rater fit statistics. In any case, if the mean-square deviates considerably from the expected value of 1, this is indication of misfit. Values greater than 1 indicate larger

variability than expected, whereas values less than 1 are a sign of less variability than expected. As a rule of thumb, Linacre (1989) suggested that an Infit or Outfit value between .50 and 1.50 may indicate productive measurement, while Engelhard (2002) suggested a range of .60 to 1.50 for indication of acceptable fit. In practice, interpretation of these statistics depends on the facet involved and its substantive meaning.

The sample reliability of separation provides useful information about how well the elements within a facet are separated to reliably define the facet. This is calculated as

$$R = (SD^2 - MSE) / SD^2, \tag{2.6}$$

where SD^2 is the observed variance of element measures for a facet, on the logits scale, and *MSE* is the mean square calibration error, estimated as the mean of the calibration error variances for each element within a facet. Take the reliability of item separation for example,

$$MSE_I = \frac{\sum_{i=1}^{I} s_i^2}{I} \tag{2.7}$$

where I is the number of items, and s_i^2 the sample variance of item i. Calculation of the *MSE* for person and rater separation are exactly parallel to Equation 2.7. The sample reliability thus calculated indicates whether items, persons, and raters are sufficiently well separated in difficulty, ability, and severity, so that several statistically distinct levels can be defined to provide meaning to the facets (Wright & Masters, 1982).

In the detection of rater variability, both the parameters estimated and the fit indices provide useful information. The following is a summary of Engelhard's (1994) explanation of how rater severity, halo effect, central tendency and restriction of range are detected after fitting an MFRM model to the data. To simplify discussion, this account will start from an MFRM model with three facets: person, item, and rater. First of all, a high sample reliability of rater separation is evidence of systematic difference in rater severity, which establishes the foundation for comparing rater severities. On this basis, distribution of rater severity estimates enables the identification of extreme cases. As the severities of various raters are placed on the same scale through MFRM modeling, the mean and variance of severities provide the reference for interpreting individual rater severities. Similarly, the sample reliability of person separation and the distribution of person ability provide information about central tendency and restriction of range. While a low sample reliability of person separation indicates central tendency, a small variance of the person ability estimates is the result of restricted range. Thirdly, the Outfit and Infit statistics for raters provide information on the consistency of rater severity. As the expected value is 1 for both the Outfit and Infit mean-squares, a statistic larger than 1.5 indicates that the ratings given by a rater to some person-item pairs differ considerably from their expected values, interpretable as inconsistent ratings across person-item pairs, or intra-rater inconsistency. On the other hand, an Outfit or Infit mean-square less than 0.5 means that the rater's ratings of person-item pairs cluster within a narrow range around the expected value, which is an indication of restriction of range.

Halo effect is of no concern if a single domain is involved in the model. To discuss the detection of halo effect, the domain facet will be added to the MFRM model as a fourth facet (see Equation 2.3b). A most obvious consequence of halo effect would be a large number of

similar ratings in all or most domains, such as 333 or 334 on a three-domain item, which will be indicated by rater Outfit and Infit mean-squares less than 0.5. A low sample reliability of domain separation provides another indication that the domains are not sufficiently well separated from each other. A detailed examination of the difficulty estimates of different domains may reveal that some of the domains are well separated from each other, while others are not.

Following the fashion of Saal et al. (1980), this brief account is tabulated in Table 2.2 for easier reference.

Table 2.2

Rater variability phenomena defined in MFRM terms (after Engelhard, 1994, 2002)

Rater variability phenomena	Operational definition
Halo effect	Small (< .5) Outfit and Infit statistics on the rater facet Low sample reliability of domain separation Close clustering of domain difficulty values
Leniency or severity	High sample reliability of rater separation Extreme rater severity values
Central tendency or range restriction	Low sample reliability of person separation Small variance of the person ability estimates
Inter-rater reliability or agreement	Rater × person interaction Rater × item interaction Rater × domain interaction etc.
Intra-rater reliability or agreement	Large (> 1.5) Outfit and Infit statistics on the rater facet Differential facet functioning

2.1.3 Studies in language assessment

Studies on rater variability in language assessment have drawn heavily upon the psychometric methods outlined above, especially the MFRM approach. The study of Engelhard (1994), for example, though published on the *Journal of Educational Measurement*, exemplified the MFRM approach to the rater variability issue in the context of a writing assessment. The

work of McNamara (1996) has further promoted the MFRM approach in language performance assessment. On the basis of this approach, he proposed four different types of rater variability: 1) rater severity; 2) rater-item interactions and rater-candidate interactions; 3) rater consistency or random error; and 4) systematic variations among raters in the use of available mark range (McNamara, 1996, p. 145). In terms of Table 2.2, type 1 is identical with rater severity, types 2 and 3 are different sources of rater inconsistency, while type 4 concerns restriction of range.

In recent years, the use of the MFRM approach in detecting rater variability has been most popular in language assessment (Eckes, 2005; Elder, Barkhuizen, Knoch, & von Randow, 2007; Elder, Knoch, Barkhuizen, & von Randow, 2005; Knoch, Read, & von Randow, 2007; Kondo-Brown, 2002; Lumley & McNamara, 1995; Schaefer, 2008; Weigle, 1998; Wigglesworth, 1993). However, G-Theory has also been used by some researchers (Bachman et al., 1995; Schoonen, 2005). As the two different approaches produced similar results, they are considered complementary approaches to the detection of rater variability (Bachman et al., 1995).

Most studies in language assessment, however, do not simply aim at detecting rater variability, but typically explore the factors that contribute to the variability. An overview of these factors will be left for the following section.

2.2 Weighting patterns

2.2.1 Overview of factors that affect rater variability

Following common practice in language assessment (McNamara, 1996; Barkaoui, 2007b), the factors that affect rater variability will be broadly grouped into rater-relevant, ratee-relevant, and task-relevant factors.

Commonly studied rater-relevant factors include gender, native language, educational

background, professional background, and training. Haswell and Haswell (1996) claimed that a rater tends to be more severe to an essay written by a test-taker of the same gender and more lenient to an essay written by a test-taker of the opposite gender. However, their raters were asked to rate only two essays, one by a female student and the other by a male student. The small sample size of the essays being rated seriously threatened the internal validity of their conclusion.

Most studies on the effect of the rater's native language on rater severity have found no significant effect (Kim, 2009; Kobayashi & Rinnert, 1996; Shi, 2001; Zhang & Elders, 2011), but the study of Kobayashi (1992) suggested that raters of different native languages may vary in how strictly they rate certain aspects in analytic rating. Kobayashi (1992) also found differences in rater severity between raters of various educational backgrounds in interaction with their native languages. In rating clarity of meaning and organization in English writings, native English-speaking professors and graduate students were more lenient than their Japanese-speaking counterparts, but the English undergraduates were more severe than Japanese undergraduates.

Professional background is another factor that may affect inter-rater reliability and rater severity. In this connection, mixed results have been obtained from empirical studies. There seemed to be no difference in inter-rater reliability between EFL teachers and non-EFL teachers (Shohamy et al., 1992), or between native English speakers preparing for TESL career and nonnative-speaking EFL instructors (Connor-Linton, 1995). In the latter case, the two groups exhibited no significant difference in their mean scores either. In one study, ESL teachers and English teachers in the United States produced comparable mean scores when they rated compositions written by college students (Brown, 1991); in another, the English professors were found to be more lenient than the ESL professors (Song & Caruso, 1996). The length of

experience in teaching and holistic evaluation was also found to have an effect on rater severity, but again the situation is complicated. While raters with more years of experience tended to be more lenient in their holistic evaluation, this effect was not evident in analytic evaluation (Song & Caruso, 1996).

For most researchers, training has proven an effective way to improve rater consistency. Shohamy et al. (1992), for example, compared the inter-rater reliability of 10 trained and 10 untrained raters. Their results indicated improved inter-rater reliability through training. Unfortunately, however, some studies do not include a control group in the design (Sweedler-Brown, 1985; Weigle, 1998). The same problem is prevalent in studies on the effect of training on rater severity, which unanimously claim that raters rate more similarly after training, with reduced spread in rater severity estimates (Elder et al., 2005, 2007; Knoch et al., 2007; Lumley & McNamara, 1995; Weigle, 1994, 1998; Wigglesworth, 1993). Typically these studies estimate rater severity with MFRM analyses and compare the estimates before and after training. As with rater consistency, the lack of a control group in the design weakens the validity of statistical conclusion.

Far fewer studies have been conducted on ratee-relevant factors, and most have focused on contrast effect, i.e. the contrast between a test-taker and other test-takers rated before her. It is generally believed that a test-taker tends to be under-rated after more proficient test-takers, and over-rated after less proficient test-takers (Daly & Dickson-Markman, 1982; Hales & Tokar, 1975; Hughes et al., 1980). Broadly speaking, the performance of an interlocutor in speaking tests may also constitute contrast effect, but this has not been confirmed (Davis, 2009).

Discourse features related to a special group of test-takers is a non-negligible factor. For example, students from some eastern cultures such as Chinese and Japanese have been immersed

in inductive instead of deductive patterns most recommended in English writing. Kobayashi and Rinnert (1996) reported that native English teachers were significantly more severe than native Japanese teachers in evaluating inductive introduction.

As with raters, the gender and native language of the ratee have not been found to have a significant effect on the rater severity (Brown, 1991; Eckes, 2005). The study by Haswell and Haswell (1996), discussed above, seemed to suggest that a test-taker may expect a higher score from a rater of the opposite gender and a lower score from a rater of the same gender. However, the use of only one writing sample from each gender rendered the results unreliable.

Studies on task-relevant factors have focused on the characteristics of the tasks used for assessing language performance, the rubric used, and the scoring procedures. Typically these are related to inter-rater reliability. It is generally believed that standardizing the assessment procedures improves the inter-rater reliability, but how such procedures are standardized may make a difference. For example, Penny, Johnson, & Gordon (2000) devised a procedure that they called *rating augmentation* and reported improved inter-rater reliability through augmentation. The raters were asked to judge the same essay in a stepwise manner, first giving the essay a general grade, and then deciding whether the essay was better or worse than the benchmark paper within the same grade. Schoonen (2005) compared holistic and analytic scoring with five raters, and found that it was easier to achieve a desired generalizability of .80 with holistic scoring among his raters. Likewise, the study by Barkaoui (2007a) also produced higher inter-rater reliability with holistic scoring than with multiple-trait scoring.

Assessment procedures may also have an effect on rater severity. Upshur and Turner (1999) compared the effect of two methods of a speaking test: story-retell and audio-pal (oral letter to a coming exchange student). Most of their raters demonstrated bias, but some against the story-

retell format, while others against the audio-pal format. If the stakes of the assessment is also considered as a feature of the test task, then there is evidence that rater severity will be affected. Baker's (2010) study found that his four raters gave significantly lower scores when they were told that the essays were intended for certification purpose than when they were told that the essays were intended for research.

2.2.2 Rater conception

All the factors reviewed above may be called external in comparison to rater conception, or rater cognition, which involves the internal workings of the rater's mind in the rating process. Two recurrent themes can be identified in most discussions of rater conception: weighting patterns and rating process. An early implication of this dichotomy can be found in a review of direct writing assessment by Huot (1990): "Other than results that measure the importance of content and organization in rater judgment of writing quality, little is known about the way raters arrive at these decisions" (p. 258). Here "the importance of content and organization" in rater judgment is a specific example of how raters weight the different criteria in the rubric, while "the way raters arrive at these decisions" is what is meant by rating process.

An explicit presentation of the same dichotomy was attempted by Wolfe and his colleagues, who devised a model of *scorer cognition* (Wolfe & Ranney, 1996; Wolfe, 1997; Wolfe, Kao, & Ranney, 1998), graphically represented as Figure 2.4. Here, weighting patterns and rating process were phrased as the *framework of writing* and the *framework of scoring* respectively. The *framework of writing* is a "mental representation of the criteria contained in the scoring rubric", while the *framework of scoring* is a "mental representation of the process through which a text image is created, compared to the scoring criteria, and used as the basis for generating a

scoring decision" (Wolfe, 1997, p. 89). In cognitive terms, the framework of scoring is a script which specifies how a variety of possible mental procedures, called *processing actions*, are used to read the essay and evaluate it. Roughly in sequence, the rater takes in information from the text through the interpretation process and builds a text image; features of the text image are then considered in the evaluation process for their relative importance, and matched to features in the framework of writing; after a scoring decision is made, the rater needs to incorporate corrective feedback into their scoring activities during the justification process.

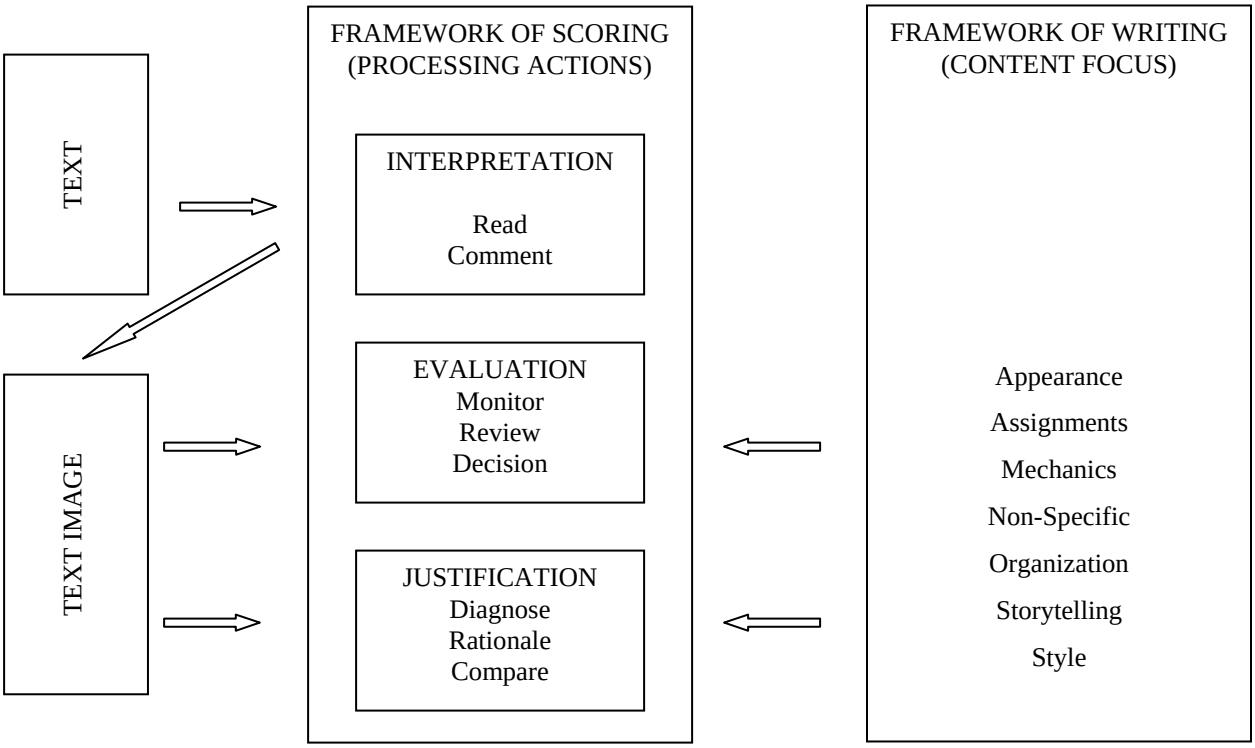

Figure 2.4. Model of scorer cognition (Wolfe, 1997, p. 89)

The framework of writing is empirically identified with the features of an essay upon which raters focus as they make scoring decisions, referred to as *content focus* in Figure 2.4. Examples of content foci are:

1) *Appearance:* considerations of the legibility or length of the essay,
2) *Assignment:* considerations of the extent to which the student complies with the writing prompt,
3) *Mechanics:* considerations of the student's ability to control spelling, punctuation, and grammar in writing,
4) *Non-Specific:* general considerations about the control or skill demonstrated by the writing,
5) *Organization:* considerations of the student's ability to control the structure and focus of the writing,
6) *Storytelling:* considerations of the student's ability to communicate ideas in writing, to develop these ideas, to use narrative elements, and to construct a story, and
7) *Style*: considerations of the student's ability to use words and sentences effectively to convey a personal voice. (Wolfe, 1997, p. 90)

The arrows in Figure 2.4 represent the basic relationship that the processing actions draw on both text and content focus in essay scoring. This is important for understanding the relationship between the processing actions and content focus, i.e., content focus provides the criteria necessary for the processing actions.

A similar framework was proposed by Cumming (1990) and revised in Cumming, Kantor, & Powers (2002), including 28 decision-making behaviors (Table 2.3). While the self-monitoring behaviors in the list of Cumming et al. (2002) look like what Wolfe (1997) called *processing actions*, the rhetorical, ideational, and language foci correspond closely to *content focus*.

Table 2.3

Decision-making behaviors while rating TOEFL writing tasks (Cumming et al., 2002)

Self-Monitoring Focus	Rhetorical and Ideational Focus	Language Focus
Interpretation Strategies		
Read or interpret prompt or task input or both	Discern rhetorical structure	Classify errors into types
Read or reread composition	Summarize ideas or propositions	Interpret or edit ambiguous or unclear phrases
Envision personal situation of the writer	Scan whole composition or observe layout	
Judgment Strategies		
Decide on macrostrategy for reading and rating; compare with other compositions; or summarize, distinguish, or tally judgments collectively	Assess reasoning, logic, or topic development	Assess quantity of total written production
Consider own personal response or biases	Assess task completion or relevance	Assess comprehensibility and fluency
Define or revise own criteria	Assess coherence and identify redundancies	Consider frequency and gravity of errors
Articulate general impression	Assess interest, originality, or creativity	Consider lexis
Articulate or revise scoring decision	Assess text organization, style, register, discourse functions, or genre	Consider syntax or morphology
	Consider use and understanding of source material	Consider spelling or punctuation
	Rate ideas or rhetoric	Rate language overall

Though Table 2.3 does not imply any sequence among specific behaviors, Cumming et al. (2002) did construct what they called a prototypical sequence of decision-making while rating compositions, including the following three stages:

1. Scan the composition for surface-level identification, such as length, format, paragraphing, script (typed or handwritten).
2. Engage in interpretation strategies, reading the essay while exerting certain judgment strategies,
    a. Classifying error types (lexis, syntax, morphology, spelling), leading to an assessment about the command of language,

b. Identifying comprehensibility, leading to an assessment of language use and rhetorical strategies,

c. Interpreting rhetorical strategies (in terms of relevance, rhetorical knowledge and performance, coherence, redundancies, topic development), leading to an assessment of content and organization, and

d. Envisioning the situation and personal viewpoint of the writer.

3. Articulate a scoring decision, while summarizing and reinterpreting judgments. (p. 74)

In brief, this starts with a scan of the composition for surface-level identification, continues with interpretation strategies, and ends with the articulation of a scoring decision. Some researchers have proposed three-stage models similar to this one (Erdosy, 2004; Lumley, 2002), while others have described variations within the interpretation and judgment stages (DeRemer, 1998; Milanovic et al., 1996; Smith, 2000).

2.2.3 Weighting patterns

Huot's (1990) comment above also reflects that much more research has been done on weighting patterns than on rating process. Historically, weighting patterns were first treated as being common among raters, but differences between raters were soon incorporated into the study. Some researchers today adhere to rich descriptions of the foci of raters, others have attempted to classify raters according to their macroscopic weighting patterns.

Many early studies on weighting patterns adopted a quantitative approach. Usually the scores given by the raters, either holistic or analytic, were treated as the dependent variable, while the scoring criteria constituted the independent variables. Under this setting, a natural tendency was to manipulate the levels of the independent variables through experimental design

and to examine their effect on the dependent variable. Analysis may be accomplished either through ANOVA (Freedman, 1979; Rafoth & Rubin, 1984) or regression (Breland & Jones, 1984).

Over time, more researchers have adopted the qualitative approach, inferring weighting patterns from raters' comments on and responses to essays. The comments may come from regular responses of writing teachers to student writings, or are elicited particularly for the study. Typically the comments are coded systematically and frequencies of different comments are compared to infer about the relevant weights attached to different scoring criteria, where a higher frequency is identified with a heavier weight (Gamaroff, 2000; Hamp-Lyons, 1991; Milanovic et al., 1996; Smith, 2000; Vaughan, 1991; Zamel, 1985).

Many researches in either the quantitative or qualitative tradition were built upon the implicit assumption that raters are a homogeneous group. Under such an assumption, ANOVA or regression analysis typically did not include any predictors that represent the systematic differences between raters. Qualitative analysis of comments did not address such differences either. More recent studies, however, started to treat systematic variability among raters. The commonest among these differences are the native language and professional background of the raters. In the first place, raters assessing language performance in their native language may focus on different aspects from raters assessing language performance in another language. So far, most studies seem to confirm such differences, but the patterns of focus vary across studies. For example, Japanese speaking raters of English essays in the study of Connor-Linton (1995) focused more on content, word choice, and grammar, while English speaking raters focused more on intersentential features of the discourse and specific intrasentential grammatical features. In contrast, Shi (2001) discovered that native English teachers attended more positively to

content and language, while native Chinese teachers attended more negatively to the organization and length of essays. Cumming et al. (2002) found that native-English composition assessors devoted more attention to rhetoric and ideas than the ESL/EFL assessors did. Furthermore, the native-English assessors seemed to divide their attention in a more balanced way to rhetoric and ideas and language, while the ESL/EFL assessors tended to give more weight to language than to rhetoric and ideas. A more recent study by Kim (2009) found that the judgments of the native English speakers were generally more detailed and elaborate in the areas of pronunciation, specific grammar use, and accuracy of the transferred information. Another study by Zhang and Elder (2011) revealed that nonnative English speakers made more comments on linguistic resources, native English speakers made more comments in other criteria, such as interaction, demeanor, compensation strategy, and other general comments. In general, then, the rater's native language does seem to affect weighting patterns, but no consensus on the patterns of the effects may be clearly described, as different criteria were involved in these studies.

In contrast, the effect of professional background of weighting patterns seems to be weak. Regardless of their experience or expertise in essay scoring, raters have been found to focus on similar features of an essay as they formulated scoring decisions (Huot, 1993; Song & Caruso, 1996; Wolfe & Ranney, 1996).

Continuing to address the differences between raters, recent attempts to describe weighting patterns reflect two tendencies. Some researchers address the issue by describing the general weighting patterns of individual raters, assuming idiosyncrasy among raters. Others try to classify raters according to common patterns, assuming the existence of typologies.

The study of Erdosy (2004) represents the individualized description of weighting patterns. This study is a detailed account of how four raters of various cultural background, native

language, and professional experience constructed scoring criteria without a scoring rubric while rating 60 TOEFL essays. In terms of weighting patterns, the researcher coded the decision-making behaviors of each rater elicited through a verbal protocol and calculated the proportional frequencies of these behaviors. The coding was based on an earlier version (Cumming, Kantor, & Powers, 2001) of the descriptive framework detailed in Table 2.3 above (Cumming et al., 2002). In spite of the comprehensive nature of the original framework, Erdosy (2004) attempted to grasp the general patterns of each rater by summing the frequencies of the two general categories: rhetorical-ideational and language. This simplification enabled a clear depiction of each rater's general weighing patterns. In pseudo names, rater Sam commented more frequently on language than on rhetoric and ideas, while rater Chris did just the opposite, focusing overwhelmingly on rhetorical-ideational qualities. In contrast, raters Alex and Jo were basically balanced in weighting language and content.

If Erdosy's (2004) individualized descriptions had been attempted on a large number of raters, he would have come up with a typology of raters, for there were only three logically possible general patterns with two categories of focus: emphasis on form, emphasis on content, and emphasis on both form and content. This simplified typology will be the focus of the present study, but before a detailed discussion about that, recent attempts at rater classifications will be reviewed here first.

Rater classification by weighting patterns is exemplified by two recent studies. Eckes (2008) explicitly asked 64 raters trained in TestDaF writing assessment to indicate on a four-point scale the importance they would attach to each of nine criteria: fluency, train of thought, structure, completeness, description, argumentation, syntax, vocabulary, and correctness. The data were subjected to a two-mode clustering technique which produced a joint classification of raters and

criteria, from which six rater types were identified. For example, one type of rater was identified as it was the only group who deemed vocabulary and syntax important. A technical problem that accompanied the Eckes (2008) attempt was that the numbers of raters classified into the types were highly unbalanced. Most saliently, only one rater was classified into one of the six types, which cast doubt on the validity of the classification results. As the total sample size was only 64, whether such results reflected true typologies or random errors remained unclear. That said, the Eckes (2008) study is still noteworthy as it pointed to a new direction: rater classification based on macroscopic weighting patterns.

The study of Schaefer (2008) was not designed to classify raters, but results from the study suggested that raters may be classified according to their bias patterns. The researcher described two severity patterns among a subgroup of raters who displayed significant interaction between rater severity and rating categories: some raters displayed severity bias toward content and organization, but lenient bias toward language use and mechanics, while others were lenient with content and organization, but severe with language use and mechanics. This seemed to suggest that content and organization was a different dimension from language use and mechanics.

According to the few clues available so far (Eckes, 2008; Ersody, 2004; Schaefer, 2008), a practical approach to rater classification on the basis of weighting patterns would be to focus on a small number of rating criteria, reduced to low dimensions. The dichotomous categorization of rating criteria into rhetorical-ideational and language foci (Cumming, 1990; Cumming et al., 2002; Erdosy, 2004) and the empirical distinction of content and organization from language use and mechanics (Schaefer, 2008) suggest that the dichotomous division into form and content may be helpful for an initial exploration into this issue. A second implication from these attempts is that a quantitative approach may be more appropriate for weighting patterns and rater

classification, as the discovery of macroscopic patterns depends on some sample size that extends far beyond what is manageable in a qualitative inquiry.

2.2.4 Weighting patterns and rater variability

Throughout this discussion, weighting patterns, and rater perception in general, have been assumed to be a factor that affect rater variability. Up to date, however, this relationship has been under-researched, and most discussions have focused on the relationship between rater agreement and weighting patterns, the two issues that are closely associated with the findings of Diederich, French, and Carlton (1961).

Many researchers have inferred a causal relationship between rater disagreement and weighting patterns from the findings of Diederich et al. (1961). The following comment is a case in point:

> [Diederich et al. (1961)] has proved two significant facts: (1) that we disagree widely in our holistic judgments of writing, and (2) that the basis of our disagreements seems to lie in the different weights which we attach to a few traits of writing. (Hirsch, 1977, p. 178)

On the other hand, weighting patterns may also have an effect on rater severity. To elaborate this in a simplified way, let there be two criteria called *form* and *content* in the scoring rubric, following the distinction between rhetorical-ideational and language foci (Cumming, 1990; Cumming et al., 2002; Erdosy, 2004) or the distinction of content and organization from language use and mechanics (Schaefer, 2008). Let three raters rate three essays according to these criteria. Suppose that in the rating process, Rater A emphasizes form exclusively, Rater B emphasizes content exclusively, and Rater C gives equal weights to form and content. Suppose further that the raters are not different from each other in any other ways. Also suppose that

Essay One excels in form but is poor in content, Essay Two excels in content but is poor in form, and Essay Three is balanced in form and content. From the perspective of the essays, Essay One would expect to get a high score from Rater A but a low score from Rater B, Essay Two would expect a low score from Rater A but a high score from Rater B, while Essay Three would expect to get equal scores from all three raters, and Rater C will assign equal scores to all three essays, as Table 2.4 shows.

Table 2.4

Combination of weighting patterns and essay profiles

	Good form, bad content	Good content, bad form	Balanced
Emphasis on form	High score	Low score	
Emphasis on content	Low score	High score	Equal scores across raters
Equal weights	Equal scores across essays		

In terms of weighting patterns, Rater A gives all weights to form, Rater B gives all weights to content, and Rater C gives equal weights to form and content. If these weights are normalized such that the total weight for each rater is 1, then the weight of any criterion for a single rater would be in the range of 0 and 1. The weighting patterns of the three raters can be tabulated as follows.

Table 2.5

Weighting patterns of three hypothetical raters

	Weight on form	Weight on content	Total weight
Rater A	1	0	1
Rater B	0	1	1
Rater C	.5	.5	1

As Table 2.5 shows, Rater A gives a weight of 1 to form and a weight of 0 to content, as he emphasizes form exclusively. Similar transformation from verbal description to numerical representation applies to Raters B and C. For Rater C, the weights are .5 on both form and content, as this rater gives equal weights to both dimensions. The total weight of each rater is 1.

Table 2.5 provides a hypothetical numerical example of weighting patterns. In practice, the weights on different criteria are much more various, and a rater seldom gives full weight to only one criterion and zero weight to other criteria (Chalhoub-Deville, 1995; Eckes, 2008). Nonetheless, the same idea applies and the weighting patterns of Raters A through C can be considered prototypical patterns when two criteria are involved.

If the broad distinction between form and meaning is carried through, the three prototypical weighting patterns may be labeled *form-oriented*, *content-oriented*, and *balanced*. *Form-oriented* raters are characterized by their dominant emphasis on form in the rating process, *content-oriented* raters make dominant emphasis on content, and *balanced* raters give approximately equal weights to form and content.

Returning to Table 2.4, it can be clearly seen that the ratings from the three types of raters are subject to not only rater severity, but also inter-rater agreement issues. While the patterns of rater severity may be the same among raters with the same weighing patterns, different weighting patterns tend to result in different patterns of rater severity. In turn, agreement between raters with the same weighting patterns may be higher than agreement between raters with different weighting patterns.

To go one step further, when the rater is required to analytically rate several tasks and/or aspects of the language performance, the various weighting patterns may lead to different severity patterns on the various tasks and/or aspects, or even alter the factor structure of the

assessment. Therefore, analysis of the factor structure of the actual ratings also provides information on halo effect, as this type of rater variability is by definition connected with the factor structure. As to the relationship between weighting patterns and restriction of range, it may be thus hypothesized. When a rater does not attach much importance to a certain aspect of the language performance, he may be insensitive to variation in this aspect, which results in ratings restricted in range. As restriction of range is often found to cause decreased correlation (Guilford, 1954; Gulliksen, 1950), it should also be somewhat connected to rater agreement and the factor structure of the assessment.

2.3 Methodological considerations

Before a study on the relationship between weighting patterns and rater variability can be designed, some methodological issues deserve preliminary considerations. However, these will be restricted to the discussion of weighting patterns. Methodological considerations on rater variability and factor structure are primarily statistical, and the above treatment has provided sufficient detail.

Up till now the term *weighting patterns* has been used in a broad sense to refer to the distributional pattern of attention, focus, importance or weight on different traits or characteristics of the language performance or on different criteria in the rating rubric imposed on or constructed by the rater. Description of the weighting patterns of a rater thus answers a broad list of questions like the following:

What does the rater focus on?

What does the rater value?

What is important for the rater?

What are the criteria of good language performance for the rater?

How does the rater weight the different criteria?

etc.

Approaches to such questions fall into two broad categories, to adopt the dichotomy of qualitative and quantitative inquiries. In this context, the two approaches differ mainly in how data are collected, as statistical procedures may be applied to data collected in either approach.

2.3.1 Qualitative approaches to weighting patterns

Most recent studies reviewed above have adopted a qualitative approach to weighting patterns. Typically, the researcher uses the verbal protocol to elicit responses from the rater, in the form of comments on the features of the language performance being assessed or justifications for the ratings given to such performance. Usually the raters are given the freedom to comment on anything they find noteworthy. Alternatively the verbal responses may come in the written mode. These responses may be concurrent, made in the rating process, or retrospective, after the rating process is complete. Rater responses are then transcribed, coded, and classified according to a certain framework for subsequent statistical analysis. To gain a clear overview of the rich variety of qualitative methods for describing weighting patterns, the studies of Diederich et al. (1961), Cumming (1990), Milanovic et al. (1996), and Lumley (2002) are compared in their mode (spoken vs. written) and timing (concurrent vs. retrospective) of response, as well as their descriptive framework (Table 2.6).

Table 2.6

Comparison of four qualitative studies on weighting patterns

Study	Response		Descriptive framework
	Mode	Timing	
Diederich et al. (1961)	written	concurrent	55 categories of comments under seven headings: ideas, style, organization, paragraphing, sentence structure, mechanics, verbal facility
Cumming (1990)	spoken	concurrent	28 decision-making behaviors under four kinds of focus: self-control, content, language, organization
Milanovic et al. (1996)	spoken (concurrent & retrospective) & written (retrospective)	concurrent (spoken) & retrospective (written & spoken)	11 composition elements which raters focused on: length, legibility, grammar, structure, communicative effectiveness, tone, vocabulary, spelling, content, task realization, punctuation
Lumley (2002)	spoken	concurrent	174 codes grouped into six categories: task fulfillment and appropriacy, conventions of presentation, cohesion and organization, grammatical control, general comments, additional comments

This brief comparison suggests that the descriptive framework is the most informative feature that distinguishes one study from another. Most studies adopted a hierarchical classification scheme, grouping the specific behaviors or processing actions observed into larger categories, which reflect the researcher's criteria of language performance. The most popular way to organize these categories seems to be the hierarchical units of a discourse. Diederich et al. (1961), for example, started from content to form, and from larger units of discourse to smaller units. Cumming (1990) followed a similar pattern, with content, organization, and language, while adding self control as a meta-cognitive component. Similarly, the more complex categories in Milanovic et al. (1996) may also be matched to the different levels of a discourse, with content, task realization, communicative effectiveness, and tone on higher levels, while the other elements on lower levels. As for Lumley (2002), the first four categories may also be sequenced from high to low levels, from task fulfillment and appropriacy, to cohesion and organization, to

conventions of presentation, and to grammatical control. While these four categories corresponded to the categories in the scoring rubrics, the other two were included more or less as *ad hoc* categories.

On the other hand, the descriptive frameworks also differ from each other in the number of specific behaviors or processing actions, which reflects the degree of detail in the description. Lumley's (2002) study, for example, included 174 codes, compared to 28 in the study of Cumming (1990). While detailed coding seems to promise accuracy, it is not recommendable from a practical perspective. For example, 26 of the behaviors identified by Cumming (1990) jointly accounted for less than 10% of the total frequency in the interview data. Similarly, Lumley (2002) also reported that 704 out of 781 comments (around 90%) related to the assessment of task fulfillment and appropriacy fell into seven subcategories.

This brief comparison also suggests that there is a common tendency among researchers to describe general patterns instead of the details of specific rating behaviors or processing actions. The low frequencies of most specific behaviors or actions in contrast to the high frequencies of a few justify such a practice. In conclusion, while it is good practice to go to some details in identifying specific rating behaviors or processing actions for accurate description, the general patterns remain to be the utmost goal of this endeavor.

In qualitative studies, weighting patterns are typically described as the frequency distribution patterns of the various criteria under consideration. A heavy weight is identified with a high frequency. For example, rater Sam in Erdosy's study (2004) was believed to give a heavier weight to language than on rhetoric and ideas, as he commented more frequently on the former than on the latter. In contrast, rater Chris focused overwhelmingly more on rhetorical-ideational qualities and could be described as a content-oriented rater.

2.3.2 Quantitative approaches to weighting patterns

Similar to qualitative inquiries, studies in quantitative approaches have also aimed at the general patterns instead of specific rating behaviors or processing actions. However, weighting patterns are typically calculated from quantitative data. Generally, weights fall into two broad categories: derived weights and self-perceived weights.

Early studies like Freedman (1979), Rafoth and Rubin (1984) and Breland and Jones (1984) did not explicitly ask their raters to weight the different criteria in terms of their importance. Instead, they quantified the various criteria and regressed the holistic scores from the raters on the criteria to obtain the weights, or used ANOVA to find out what criteria had a significant effect on the holistic score. Freedman (1979), for example, found three criteria with the most significant influence through multi-way ANOVA, while Breland and Jones (1984) reduced the list of influential criteria from 20 to eight with regression. Due to their indiscriminative treatment of raters, however, these researchers stopped one step ahead of quantifying the weights given to the criteria by individual raters and classifying raters accordingly.

Eckes's (2008) attempt illustrated the alternative way to obtain weights by eliciting explicit responses from the raters. His raters were asked to indicate the importance they would attach to nine criteria under consideration. On this basis, Eckes (2008) was not only able to describe the different weighting patterns of each rater, but was able to classify the raters according to their similarities and differences in weighting patterns.

The classification of raters according to weighting patterns reflects a new direction in the quantitative approach to weighting patterns. This new attempt is in perfect agreement with the common tendency among researchers to describe general patterns. Statistical methods for

classification have not been much used in language assessment, and most recent studies that do involve classification have been conducted for the purpose of diagnosis, using the Rule Space Methodology and similar methods based on the Q-matrix, or incidence matrix as it is referred to in standard textbooks of linear algebra (Buck & Tatsuoka, 1998; Jang, 2009a, 2009b; Lee & Sawaki, 2009). Cluster analysis used in Eckes (2008) is of course a standard statistical method for classification.

2.3.3 A note on speaking assessment

In the preceding discussion of rater variability and factors that affect rater variability, no distinction has been made between speaking and writing assessments. However, most empirical studies reviewed so far have involved writing assessment, and the two major models of rater conception (Cumming et al., 2002; Wolfe, 1997) were both based on writing assessment. The intention of this study, however, is to explore the relationship between weighting patterns and rater variability in speaking assessment, which has been underrepresented in the language assessment literature. This casts doubts on the applicability of the rating process models discussed above.

To clear such doubts, verbal protocols will be conducted in this study to provide some details of the rating process in speaking assessment, with a focus on the weighting patterns. Presumably, pronunciation and intonation are regarded as essential components in any scoring rubric for speaking assessment, and less weight is usually given to grammatical accuracy by many raters. Except for this, there is no reason why qualitative and quantitative methods summarized above should not be applicable to speaking assessment.

Chapter 3 Method

3.1 Overview

Investigation of the relationship between weighting patterns and rater variability involves a measure of weighting patterns for individual raters, classification of raters into types according to weighting patterns (Research question 1), and comparison of different types of raters in the various types of rater variability (Research question 3). A related issue is the differences between types of raters in the rating process (Research question 2).

To answer research question 1, the raters were presented with computer-generated profiles featuring strengths and weaknesses in various criteria under consideration for 120 hypothetical test-takers. The raters were asked to provide holistic ratings for each profile. These ratings were regressed on the values of the various criteria to derive the relative weights of the criteria, which were then subjected to cluster analysis to classify the raters.

For question 2, a small sample of raters from each type participated in a verbal protocol study. The raters were asked to rate the recordings of five real test-takers and justify their ratings in the process.

As for question 3, the real ratings of a sample of raters were subjected to a series of statistical analyses to detect and compare the various patterns of rater variability across the different types of raters.

3.2 Context

The context of the study was the Test for English Majors – Band 4, Oral Test (TEM4-Oral),

a national test administered to college EFL majors in China toward the end of their sophomore year (Writing Group of Syllabus for TEM4-Oral, 2008). The test is developed by the Institute of Oral English Studies of Nanjing University with the entrustment of the English Team of the National Foreign Languages Teaching Advisory Board of China. This is an annual event that takes place each May, with a steadily expanding population of test-takers that reached around 22,000 in 2010. The test is of high stakes, as test-takers who pass it receive a certificate of qualification, proficiency, or excellence according to how well they pass the test, which may play an important role in their job-hunting effort during the last year at college.

The TEM4-Oral is a computer- or tape-mediated speaking assessment comprising three tasks. The oral response of the test-taker to each task is recorded into a tape or as a computerized sound file, and the rater scores the recordings of the test-taker. Task I, retelling, requires the test-taker to retell a mini-story after listening twice to it. The scoring rubric takes the form of a checklist of points covered in the original story. The rater matches the retold story to the checklist and counts the points retrieved by the test-taker. Usually, 20 to 25 points are included in the checklist, and the percentage of correctly retrieved points is converted to a 100-point scale.

Task II is a talk based on a given topic. In this task, the test-taker is usually required to relate some real story/stories, and then comment on the significance of the story/stories. The scoring rubric of Task II includes three major criteria: topic relevance, richness in content, and organization.

Task III takes the form of a discussion. The test-takers pair up, and are given different roles to play, pro and con, in discussing a controversial issue. The scoring rubric of Task III also includes three major criteria: task relevance, role performance, and effective communication.

For Tasks II and III, the three major criteria are operationalized according to the specific

topic or situation involved. Though three criteria are involved in these tasks, the raters are required to give only a holistic/total score to each task. While the rater produces a holistic/total score for each of the three tasks, he is also required to give two more holistic ratings, one on pronunciation and intonation, and the other on grammar and vocabulary. These ratings are based on the test-taker's performance in all three tasks. The 100-point scale is adopted for all holistic ratings, but the raters are required to give scores with only a 0 or 5 in the ones place, such as 55, 60, 65, and 70. This practically reduces the number of possible scores to 21.

In sum, with the points of Task I converted to the 100-point scale, each rater gives five ratings in the 100-point scale: one on each of the three tasks, one on pronunciation and intonation, and one on grammar and vocabulary. To obtain more reliable ratings, all recordings are rated by two raters.

3.3 Participants

The participants of this study were 126 Chinese-speaking teachers of EFL majors who served as TEM4-Oral raters in 2010. Of these raters, 20 (16%) were male, while 106 (84%) were female. In terms of education, 119 (94%) had an MA degree, mostly in linguistics, applied linguistics, TESOL, English literature, and translation and interpretation, and 15 (12%) held a Ph.D. degree in similar areas to the MA degrees. At the time of the study, their average duration of experience as a college teacher of English was 10 years. Among these raters, 93 (74%) evaluated their own students as distributed similarly to the test-takers of TEM4-Oral, 16 (13%) believed their own students to be more proficient than the average test-takers, while 17 (13%) placed their own students at a lower level than the average test-takers. Except for two raters, all raters had taught English courses to freshmen or sophomores during the past three years. Among

the common courses, Basic English (integrated reading and writing) and Listening had the highest frequencies of response, 79 and 60 respectively, while Grammar was the least taught, with a frequency of only 12. The other courses, Reading (39), Integrated Skills (36), Pronunciation (27), Writing (27), and Speaking (26), had similar frequencies of response. In terms of training, more than half (77, or 61%) of the raters had no prior rating experience for TEM4-Oral, while 27 (21%) had taken part in rating only once before. Each of these raters worked consecutively for five days in one of two successive cohorts, as the catering capacity at the rating site was limited. Of the 126 raters, 62 worked in the first cohort, and 64 in the second cohort.

While all 126 raters were involved in procedures designed to address the first research question, a sample of 21 raters participated in the verbal protocols designed for the second research question. Each TEM4-Oral rater was expected to attend a one-day training session and work for five days. The raters worked for about six hours each day during normal working hours, during which each was expected to rate a set of 32 recordings, each lasting 10 minutes. Due to the intensity of workload for the raters, it was only possible to schedule a meeting with the raters after their day's work. To make the best use of time, the researcher had to monitor the pace of the raters with the clerical help from the rating site, and make appointments with potential interviewees shortly before they finished their day's work. This means that only a convenient sample of raters could be obtained. Also, each verbal protocol was limited to around one hour, and at most three raters could be interviewed each day. The number of raters included in the verbal protocols was 21, as the verbal protocols lasted seven days.

To conduct this research, an application for UCLA Institutional Review Board (IRB) approval was filed and approved. In addition, permission to access the participants and relevant

test data was granted in written form by the director of the TEM4-Oral development team. As required by UCLA IRB procedures, the raters were informed of the purpose and procedures of the study and given complete freedom to reject the researcher's invitation, or to withdraw anytime during the study. They were paid for participation according to local standards. Their identity was kept confidential and all files that contained identifiable information about the participants were encrypted and kept in a safe place accessible only to the researcher. All analyses were completed by the researcher alone, and relevant data files were kept inaccessible to any other parties.

3.4 Materials

3.4.1 The value judgment task

In order to collect data for classification purposes, the raters were presented with a value judgment task, which requested them to judge the EFL oral proficiency of hypothetical test-takers with computer-generated score profiles featuring strengths and weaknesses in five criteria: topic relevance, richness in content, organization, pronunciation and intonation, and grammar and vocabulary. These are aspects included in the rating rubric for Task II of TEM4-Oral.

The hypothetical score profiles were generated through R 2.9.1 (R Development Core Team, 2010) with the following codes:

```
write.table(round(matrix(runif(600,1,9),5,120),0),file="…/profile.csv", sep = ",")
```

where "profile.csv" is the name of the file generated, in Microsoft® Excel, 2003 CSV (Comma Separated Value) format and the "…" stands for the actual directory of the CSV file. The file

included 600 values in the range of 1 through 9, generated to conform to the random uniform distribution, rounded up to integers, and arranged in a matrix of five columns and 120 rows. The five columns were matched to the five criteria, each evaluated on a nine-point scale (1 through 9), and the 120 rows corresponded to 120 imaginary test-takers. The CSV file was imported into PASW17.0 (SPSS Inc., 2009), in which the 120 score profiles were transformed to 120 figures. Figure 3.1 is an example.

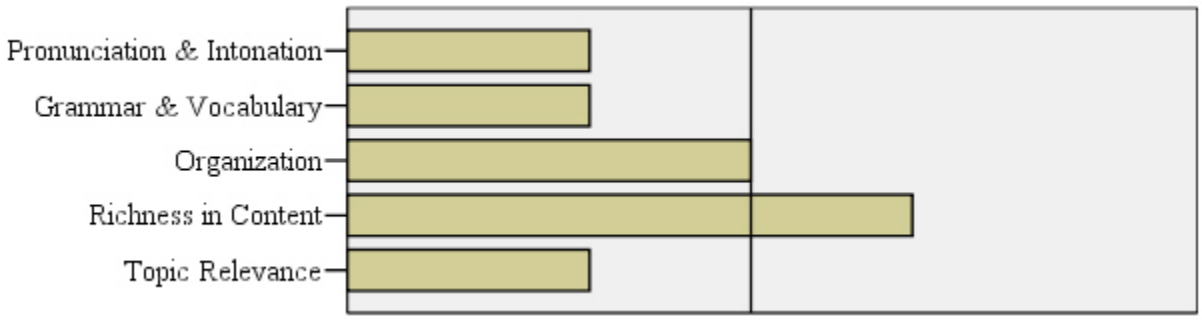

Figure 3.1. Graphical representation of a score profile

The five horizontal bars in Figure 3.1 stand for the five criteria, with a longer bar representing a higher score on each criterion. The vertical line in the middle of the bar chart represents the mean of all the imagined test-takers, which provides a reference for interpreting the length of the various bars. Accordingly, the score profile in Figure 3.1 represents a test-taker who is below average in pronunciation and intonation, grammar and vocabulary, and topic relevance, average in organization, and above average in content richness.

The 126 raters were be presented with the 120 figures and requested to judge the oral proficiency of the hypothetical test-takers on the 100-point scale that they would use for TEM4-Oral, i.e., with scores ending only in 0 or 5. As was discussed above, this practically reduced the number of possible scores to 21, coinciding with the 1-20 numerical scale widely used in value

judgment tasks (Anderson, 1982, p. 6).

The 120 figures were printed on 30 pages, with four charts on each page. The raters were provided with a scoring sheet to write down the holistic ratings. With a few exceptions, most of them completed the task in 30 to 40 minutes, which means 15 to 20 seconds for each profile.

3.4.2 The verbal protocols

Materials used for the verbal protocols included the segmented recordings of five test-takers and a summary question. As was mentioned above, time for each interview was limited to one hour, and so it was impractical to include too many test-takers. On the other hand, it was advisable to include more than one or two test-takers to guarantee a variety of profiles. As form and content were the two main foci in terms of weighting patterns, it was considered desirable to have test-takers strong in both form and content, strong in either aspect, and strong in neither aspect. This would be four different profiles, and adding a random profile to the beginning would provide an anchor against which ratings of the four profiles could be adjusted. Therefore, the recordings of five test-takers were used. The researcher prepared the recordings of 10 test-takers, two for each profile, and asked two raters to rate them according to TEM4-Oral rubrics, on the basis of which the recordings more typical of the intended profiles were selected.

To give raters time to talk about their ratings, the number of tasks included in the verbal protocols should also be limited. From the above description of the three tasks, the second task has the most typical form of a speaking test. Being an oral composition task, performance on this task is most comparable to essay writing. Besides, the three criteria associated with this task (topic relevance, richness in content, and organization) are often shared with scoring rubrics in writing assessment. In contrast, the retelling task is confounded with listening comprehension,

note-taking, and memory. Furthermore, the checklist used in the rating process demands a lot of efforts to complete, leaving little energy for the rater to pay attention to other aspects of the performance. On the other hand, performance in the discussion is dependent on the proficiency of both test-takers, as it is a co-constructed discourse. It is also a challenge for the rater to rate only one test-taker while listening to two simultaneously, as it is easy to take one for the other, especially when the two test-takers are of the same gender. For these reasons, only Task II was included in the verbal protocol study.

The recording of each test-taker on Task II was segmented into five parts according to preliminary trials with recordings from past years. Efforts were made to give a similar length to each part and yet to keep the beginning and ending sentences in each part complete. The final duration of most parts ranged from 20 to 30 seconds. It was deemed practical for the rater to recall the content of a segment of recording of that length and comment on salient features of that segment. Appendix F shows how the recordings of two test-takers were segmented, in which the double slashes signify the border between segments.

To capture possible fluctuations in the rating process, the raters were asked to rate all the segments they had heard after listening to each segment. In agreement with real rating practice, they were asked to give a holistic score to the talk, a score to pronunciation and intonation, and a score to grammar and vocabulary, all on the 100-point scale.

After rating each segment, the raters were asked to justify the ratings and comment on the salient features of the segment. After the whole process of rating and justification, they were asked a summary question: “What do you value most in real TEM4-Oral rating: pronunciation and intonation, grammar and vocabulary, organization, richness in content, or topic relevance? Why?”

3.5 Procedures

The study was conducted in three phases. The first phase centralized on the classification of raters, the second phase on the description of different types of raters with a sample of raters from each type, and the third phase on the discovery of differences between types of raters in rater variability.

For meaningful classification, the weighting patterns of all raters were measured first, by way of the value judgment task described above. This was completed toward the end of the training session. Regression analyses detailed in the following section were conducted to classify the raters.

To describe the different types of raters, a sample of 21 raters was selected for the verbal protocol study. As detailed in Section 3.3, this selection was not randomized due to the limited time available, as the verbal protocols could only be scheduled after the target raters had completed their day's work. Therefore, the raters were selected on a first-come-first-choice basis, in that the first rater who finished the day's work automatically became choice A, the first rater who finished one hour after choice A automatically became choice B, and choice C might be any rater who finished the day's work before the interview with Choice B ended. Of the 21 raters, 12 took part in the verbal protocols individually, six in pairs, and three as a group. The reason to include pairs and groups was to contrast the responses of different raters.

The verbal protocols were recorded with a digital voice recorder. The recordings were transcribed and coded for statistical analysis. The coding scheme will be described in the following section. The ratings produced during the verbal protocols were also analyzed to find out the differences between types of raters.

In the last phase, the scores of 33 raters in the real TEM4-Oral rating were obtained from the test developer. These scores were analyzed with a view to detecting and comparing rater variability across rater types.

3.6 Analyses

3.6.1 Relative weights

The weighting pattern of each rater consisted of the relative weights given to the five criteria under consideration. These are called *relative weights* because the total weight across all criteria was standardized to be unity through a procedure proposed by Hoffman (1960). The procedure starts with a multiple regression of the rater's judgment on the five computer-generated scores, which yielded the beta weights of the predictors. To correct for the effect of possible multicollinearity, Hoffman (1960, p. 121) proposed that the relative weight of each predictor should be calculated as

$$w_{oi} = \frac{\beta_{oi} r_{oi}}{R^2_{0.12...k}} \tag{3.1}$$

where

β_{oi} = the beta weight for the *i*th predictor;

r_{oi} = correlation with judgment of the *i*th predictor; and

$R^2_{0.12...k}$ = the squared multiple correlation coefficient reflecting the best linear combination of the *k* predictors in prediction of judgment.

3.6.2 Rater classification

Both hierarchical and *k*-means clustering were used to classify the raters according to the

relative weights they gave to the five criteria. Hierarchical clustering provided information on the most appropriate number of clusters, and *k*-means clustering provided the final cluster results with *k* set to the desired number of clusters.

3.6.3 Coding of verbal protocol data

The verbal protocols were transcribed verbatim by an undergraduate Chinese student, and then double-checked by the researcher. The transcript was then divided into coding units. In this context, a coding unit was defined as a continuous segment of discourse about the same rating criterion. For example, the following transcript was divided into four units, with numbers added:

> "1) I mean I did find him to have poor pronunciation and intonation, 2) but he made his points toward the end, 3) and I would raise his pronunciation score. 4) I think his pronunciation was okay if only he expressed his ideas clearly." (Appendix E)

The rater in this excerpt first commented on pronunciation and intonation, then shifted to content ("made his points"), but returned to pronunciation. Toward the end she mentioned both pronunciation and content ("expressed his ideas"). As there were three shifts, the excerpt was divided into four coding units.

As three scores—content, pronunciation and intonation, grammar and vocabulary—were given by the raters to the test-taker recordings in the verbal protocols, the division of the transcript and the identification of themes in each coding unit were limited to the three criteria. Both the researcher and a second coder, an experienced TEM4-Oral rater, coded all the units. Cohen's kappa (Cohen 1960) was calculated as a measure of intercoder reliability. The two coders then discussed the discrepancies case by case and reached an agreement on each unit.

After coding, the frequencies of comments on the various criteria were calculated, and the

different types of raters were then compared in the frequency and percentages of themes covered in the verbal protocols through MANOVA. The content of the verbal protocols was analyzed in terms of the following issues: self-perceived weights, relationship between criteria, and responses to negative features in content (digression) and pronunciation and intonation. Where applicable, the types of raters were compared in the frequencies of relevant responses. The self-perceived weights were derived from the raters' answer to the summary question toward the end of the verbal protocols (Section 3.4.2), the relationship between criteria was extracted from the raters' comments during the verbal protocols, while the responses to negative features in content and pronunciation and intonation were summarized according to the raters' comments on two of the test-taker profiles, one suspected of digression, and the other weak in pronunciation and intonation.

3.6.4 Rater variability

The primary purpose of this study was to explore the differences across rater types in rater variability. For this purpose the actual TEM4-Oral scores given by 33 raters were analyzed through MFRM, HLM (Hierarchical Linear Modeling), G-Theory, and CFA where applicable. The HLM analyses were conducted in lieu of factorial ANOVA in the CTT tradition because the test-takers were nested within the raters. As the interpretation of the results from most of these statistical procedures has been discussed in Chapter 2, only the use of CFA for detecting the halo effect will be justified here.

For the TEM4-Oral, the detection of the halo effect is most convenient by way of CFA models, which is clear from Figure 3.2.

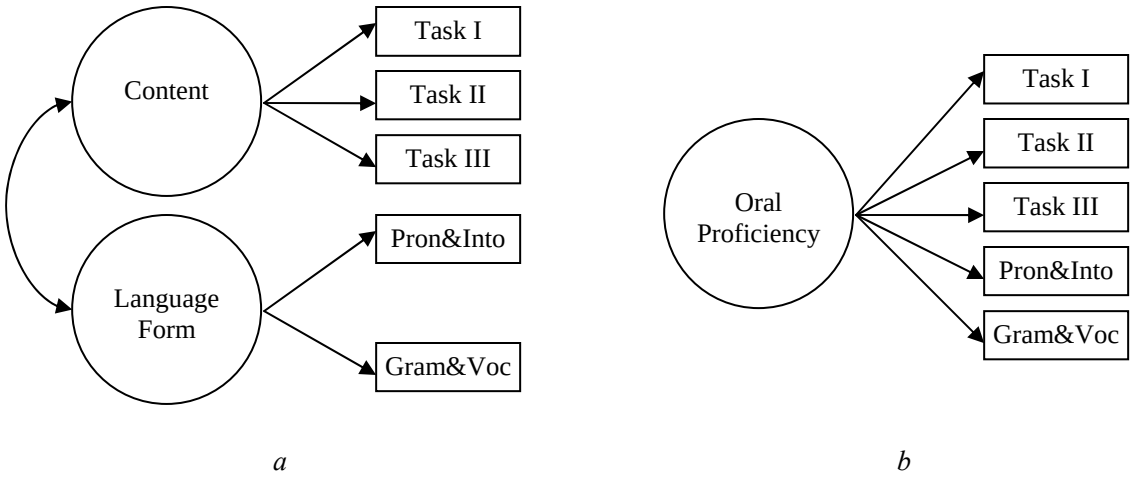

Figure 3.2. The factor structure of TEM4-Oral scores without and with halo (Note: Pron&Into = Pronunciation and Intonation; Gram&Voc = Grammar and Vocabulary)

According to the description in Section 3.2, the structure of the test can be represented by Figure 3.2a. The scores of the three tasks all contribute to a common factor related to content, while the score for pronunciation and intonation and the score for grammar and vocabulary are indicators of language form. The two factors may be correlated.

When there is halo effect, however, the factor structure of the scores will be altered into Figure 3.2b, with spuriously high correlations among the observed scores. This is so because model *b* can be obtained through fixing the correlation between content and language form in model *a* at unity, which spuriously increases the correlation between language form and the three tasks, or the correlation between content and the two form-related indicators. Statistically, model *b* is nested within model *a*, and a chi-square difference test can be conducted to find out which model fits the data better. For the above reason, the halo effect is evident if model *b* fits better than model *a*. This provides a basis for comparing the different types of raters.

Chapter 4 Weighting Patterns and Rater Types

This chapter reports the findings related to the first two research questions of the study:

1. How successfully can raters be classified into types according to their weighting patterns?
2. How are types of raters different in the rating process?

Results of statistical analyses related to research question 1 include 1) the holistic ratings given by the raters in the value judgment task, 2) the results of the multiple regression analyses, 3) the relative weights calculated from the regression weights, and 4) the results of cluster analyses. Results of statistical analyses related to research question 2 include 1) the intercoder reliability of the verbal protocol transcript coding and 2) the results of group comparison in the percentages of different themes covered by the raters in the verbal protocols. After the results of statistical analyses have been reported, the results of the qualitative analysis of the verbal protocol content will be reported (using pseudonyms of the raters) to determine how different types of raters perceived the weights they gave to different criteria and how they defined the relationship between criteria. The comparison ends with the different types of responses to digression and salient negative features of pronunciation and intonation. The verbal protocols were conducted in Chinese, but the relevant parts of the transcript are translated into English in this report. Table 4.1 provides an overview of the various analyses completed and the specific issues they address.

Table 4.1

Overview of analyses in answer to research questions 1 and 2

Issues	*Data*	*Analyses*
Research question 1		
Data preparation	Raters' holistic ratings from the value judgment task	Descriptive statistics Multiple regression analyses
Patterning of relative weights	Regression weights	Calculation of relative weights according to Hoffman's (1960) equation Descriptive statistics
Classification of raters	Relative weights	Cluster analyses Descriptive statistics MANOVA ANOVA
Research question 2		
Data preparation	Verbal protocol transcript coding	Calculation of intercoder reliability (Cohen's kappa)
Theme coverage	Frequency and percentages of themes covered in the verbal protocol	Descriptive statistics MANOVA ANOVA
Self-perceived weights	Verbal protocol transcript	Content analysis
Relationship between criteria	Verbal protocol transcript	Content analysis
Response to digression and salient negative features in pronunciation and intonation	Verbal protocol transcript Verbal protocol transcript coding	Content analysis Descriptive statistics

4.1 Research Question 1: Weighting patterns

4.1.1 Holistic ratings in the value judgment task

In the value judgment task, the raters were presented with 120 profiles of simulated scores on the five criteria used in TEM4-Oral: topic relevance, richness in content, organization, pronunciation and intonation, and grammar and vocabulary. On this basis each rater produced a holistic rating for each of the profiles. A total of 126 raters, 62 from the first cohort and 64 from the second, took part in this task. The minimum, maximum, and mean holistic ratings, as well as the standard deviation associated with each rater, are listed in Appendix A; here only the grand mean and standard deviation of the individual mean ratings and standard deviations are reported in Table 4.2.

Table 4.2

Grand mean and standard deviation of individual mean ratings and standard deviations

Cohort	*N*	Mean Ratings		Standard Deviations	
		M	*SD*	*M*	*SD*
1	62	61.39	5.86	10.74	2.01
2	64	60.76	3.79	10.80	1.96
Total	126	61.07	4.91	10.77	1.98

Table 4.2 suggests a rather wide range of mean ratings across individual raters, with the standard deviation of the mean ratings approaching 5. An examination of Appendix A shows that the minimum mean rating was 46.75, in stark contrast to the maximum of 75.63. The grand mean across both cohorts was 61.07. While the mean standard deviation was 10.77 across both cohorts, the range of standard deviations spanned a minimum of 5.52 and a maximum of 16.68 according to Appendix A. This provides some preliminary evidence of restriction of range among parts of the raters.

4.1.2 Multiple regression analyses

To prepare data for subsequent analyses, the holistic ratings given by each rater in the value judgment task were regressed on the five criteria. The relative weights were calculated from the unadjusted R^2 (Hoffman, 1960), and the value of this statistic from each regression model was reported in Appendix B. Thus, only a summary is given here. The mean R^2 was .78, with a standard deviation of .08. Of all the R^2 values, 81% fell in the range of plus/minus one standard deviation (.70 to .86). Six were lower than .65, with a minimum of .40. With the exception of these few values, the multiple regression analyses were considered successful in the sense that they guaranteed reliable results for the calculation of relative weights for the five criteria.

The specific beta weights associated with all five criteria are also reported in Appendix B, together with the correlation between each profile score and the holistic ratings. These provided the data for calculating the relative weights in accordance with Equation 3.1 (Hoffman, 1960) given in Section 3.6.1.

4.1.3 Relative weights

The relative weights derived from the beta weights are summarized in Table 4.3 below.

Table 4.3

Mean and standard deviation of relative weights

Cohort	*N*	*M(SD)*				
		Relevance	*Richness*	*Organization*	*Gra/Voc*	*Pro/Int*
1	62	.30(.15)	.25(.11)	.12(.07)	.17(.10)	.16(.14)
2	64	.27(.12)	.30(.14)	.11(.07)	.14(.08)	.17(.12)
Total	126	.28(.14)	.28(.13)	.12(.07)	.15(.09)	.17(.13)

Note: Relevance = Topic relevance; Richness = Content richness; Gra/Voc = Grammar and vocabulary; Pro/Int = Pronunciation and intonation

Two interesting tendencies can be seen in Table 4.3. First, it is clear that topic relevance and richness in content had much larger weights than grammar and vocabulary and pronunciation and intonation, and that organization was given the least weight. Second, the relative weights of organization and grammar and vocabulary tended to have smaller standard deviations than the other three weights. These tendencies were present in both cohorts as well as in the whole sample. In addition, the mean relative weights of the three content-related criteria averaged .23 for the whole sample, in contrast to .16 for the two form-related criteria. This suggests a general belief among the raters that content deserves more weight than form.

4.2 Research Question 1: Classification of raters with cluster analyses

4.2.1 Preliminary cohort-specific classification results

The relative weights derived from the multiple regression analyses were used as input data in the cluster analyses to classify the raters. In essence, the statistical method of cluster analysis was used to recognize weighting patterns from the relative weights and identify the rater types with the patterns recognized. As is explained in Section 2.2.4, low weights on form-related criteria and high weights on content-related criteria is a different weighting pattern from high weights on content-related criteria and low weights on content-related criteria, whereas balanced weights on all criteria constitute yet another weighting pattern. As there may be variations within content-related criteria, such as a low weight on topic relevance and a high weight on richness in content, or variations within form-oriented criteria, such as a low weight on pronunciation and intonation and a high weight on grammar and vocabulary, the number of weighting patterns may extend well beyond the three described in Section 2.2.4.

A common practice of cluster analysis is to determine the number of clusters through hierarchical clustering, and then conduct *k*-means clustering for final classification. *K*-means clustering is preferred because the cluster membership of cases can be changed in the process of this type of clustering for more appropriate classification results, while hierarchical clustering tends to trap cases in inappropriate clusters. However, the number of final clusters *k* needs to be determined before *k*-means clustering can be conducted, and hierarchical clustering is one way to achieve this (Lattin, Carroll, & Green, 2003; Mooi & Sarstedt, 2011; Punji & Stewart, 1983). The preliminary classification of raters in this study was based on hierarchical cluster analysis using Ward's method, applying squared Euclidean Distance as the similarity measure. This was done separately for the two cohorts of raters.

To determine the number of clusters in hierarchical clustering, the distances at which clusters are combined can be used as criteria. This information comes from either the agglomeration schedule or the dendrogram. An agglomeration schedule reports the clusters being combined at each stage of the clustering process as well as the distance between the two clusters, measured through an agglomeration coefficient. While the coefficients reported in the schedule become progressively larger from stage to stage, at certain points the coefficient may show a sudden, exceptionally larger increase. Together with substantive considerations, this sudden increase of distance between two clusters may indicate that the clustering process should be ended at this point and the number of clusters be determined. On the other hand, a dendrogram is a tree graph for displaying the clustering results, with branches joining the clusters being combined at each stage. The dendrogram is often scaled such that the length of a branch indicates the distance at which two clusters are joined. Therefore, the longest interval on the

scale over which the number of clusters does not change corresponds to the most possible solution.

To better illustrate the changes in coefficients over stages of the hierarchical cluster analysis, plots of the coefficients of the last 10 stages in the agglomeration schedules are presented in Figure 4.1 for cohort 1 and in Figure 4.3 for cohort 2. Dendrograms for cohorts 1 and 2 are presented in Figures 4.2 and 4.4 respectively.

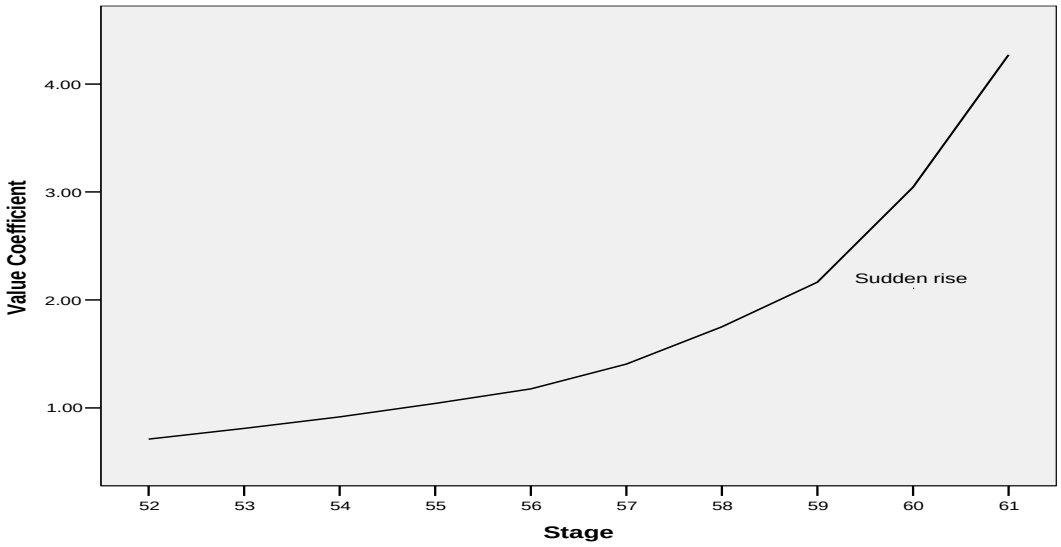

Figure 4.1. Agglomeration schedule of hierarchical cluster analysis for cohort 1

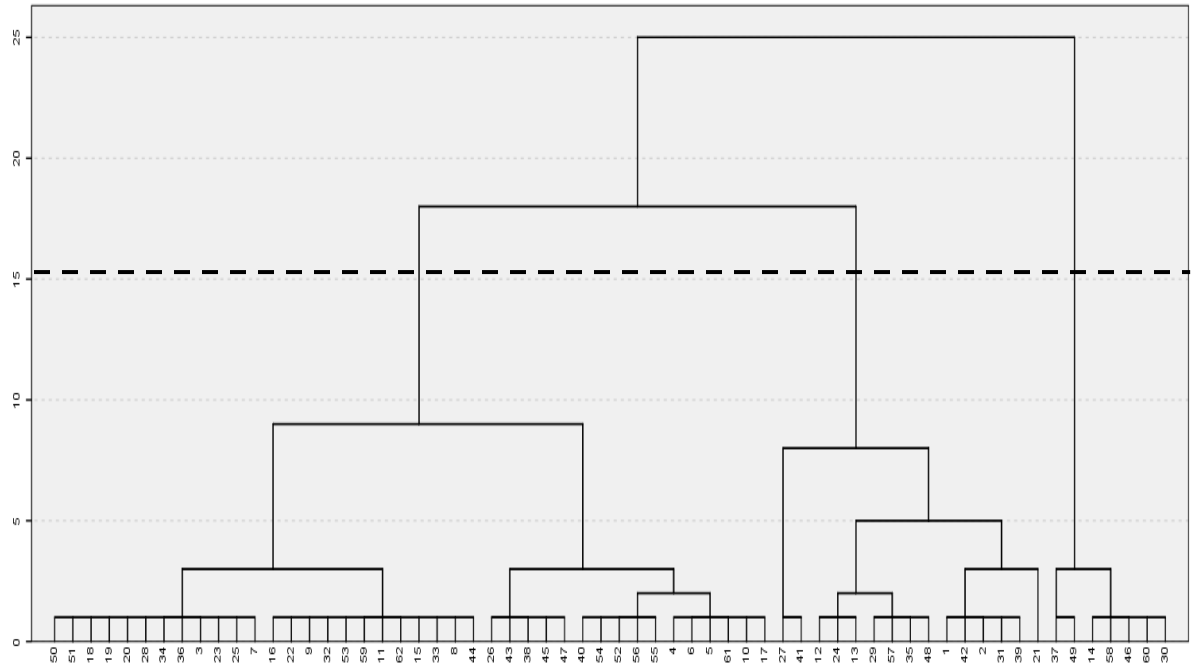

Figure 4.2. Dendrogram of hierarchical cluster analysis for cohort 1

For cohort 1, the agglomeration schedule (Figure 4.1) shows a sudden rise in the agglomeration coefficient from stage 59 to stage 60. This coefficient increases steadily by .23, .34, and .41 from stage 56 to stage 59, but by .88 from stage 59 to stage 60. Prior to stage 60, there were three clusters; so the sudden rise was in favor of the three-cluster solution. Similarly, the dendrogram (Figure 4.2) shows that cutting the tree into three branches with the dashed line would result in the widest range of rescaled distance (18 – 9 = 9) over which the number of clusters in the solution does not change (Everitt, Landau, Leese, & Stahl, 2011, p. 95; Lattin et al., 2003, p. 281), which also supported the three-cluster solution.

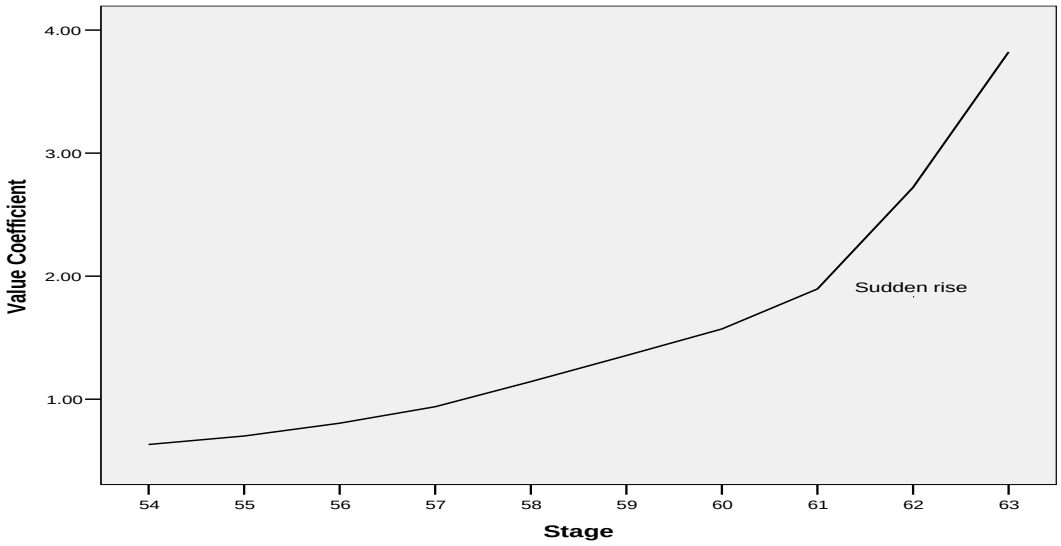

Figure 4.3. Agglomeration schedule of hierarchical cluster analysis for cohort 2

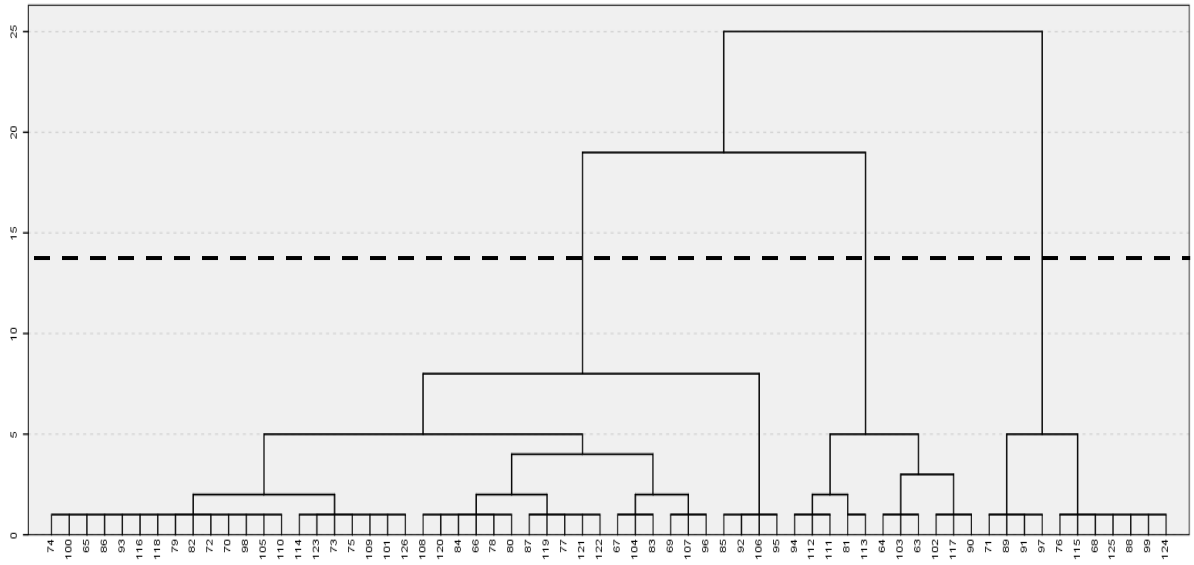

Figure 4.4. Dendrogram of hierarchical cluster analysis for cohort 2

For cohort 2, the agglomeration schedule (Figure 4.3) shows a sudden rise in the agglomeration coefficient from stage 61 to stage 62. This coefficient increases steadily by .21, .22, and .32 from stage 58 to stage 61, but by .83 from stage 61 to stage 62. Prior to stage

62, there were three clusters; so the sudden rise was in favor of the three-cluster solution. As in the case of cohort 1, the dendrogram (Figure 4.4) shows that cutting the tree into three branches with the dashed line results in the widest range of rescaled distance (19 – 8 = 11) over which the number of clusters in the solution does not change (Everitt et al., 2011; Lattin et al., 2003), which also supported the three-cluster solution.

The three-cluster solution corresponded well with the substantive theory discussed in the literature review, and was supported by both the MANOVA and univariate tests conducted to compare the three clusters in the means of the five relative weights. The descriptive statistics of the relative weights and the results of the univariate tests are presented in Tables 4.4 and 4.5, respectively.

Table 4.4

Mean and standard deviation of relative weights across clusters for cohort 1

Cluster	*N*	*M(SD)*				
		Relevance	Richness	Organization	Gra/Voc	Pro/Int
1	15	.18(.07)	.19(.09)	.07(.03)	.21(.14)	.35(.18)
2	40	.28(.07)	.28(.11)	.16(.06)	.17(.07)	.12(.05)
3	7	.65(.14)	.20(.09)	.06(.04)	.05(.03)	.05(.04)
Total	62	.30(.15)	.25(.11)	.12(.07)	.17(.10)	.16(.14)

Note: Relevance = Topic relevance; Richness = Content richness; Gra/Voc = Grammar and vocabulary; Pro/Int = Pronunciation and intonation

Table 4.5

ANOVA results for mean comparison between clusters for cohort 1

Source		*SS*	*df*	*MS*	*F*	*p*
Relevance	Between Groups	1.061	2	.530	84.175	.000
	Within Groups	.372	59	.006		
Richness	Between Groups	.098	2	.049	4.754	.012
	Within Groups	.611	59	.010		
Organization	Between Groups	.119	2	.060	21.841	.000
	Within Groups	.161	59	.003		
Gra/Voc	Between Groups	.120	2	.060	7.676	.001
	Within Groups	.461	59	.008		
Pro/Int	Between Groups	.708	2	.354	37.352	.000
	Within Groups	.559	59	.009		

Note: Relevance = Topic relevance; Richness = Content richness; Gra/Voc = Grammar and vocabulary; Pro/Int = Pronunciation and intonation

For cohort 1, the multivariate result was significant for cluster, Pillai's Trace = 1.381, $F(10, 112) = 24.992$, $p = .000$. Univariate tests indicated that the three clusters differed significantly in all five means of relative weights at the $p = .05$ level (Tables 4.4 and 4.5). The general pattern that can be recognized from Table 4.4 is that Cluster 1 had higher means in the relative weights of grammar and vocabulary and pronunciation and intonation but lower means in the relative weights of relevance and richness than the other two clusters, Cluster 3 had higher means in the relative weight of relevance but lower means in the relative weights of grammar and vocabulary and pronunciation and intonation than the other two clusters, while Cluster 2 lay in between in the relative weights of relevance, grammar and vocabulary, and pronunciation and intonation. For Cluster 2, the mean weights of richness and organization were also closer to the weights of the other three criteria. In terms of the conceptual classification discussed in the literature

review, Cluster 1 corresponded closely to the form-oriented type, Cluster 2 to the balanced type, and Cluster 3 to the content-oriented type.

The same correspondence between real data and conceptual taxonomies was established for cohort 2. The descriptive statistics of the relative weights and the results of the univariate tests for this cohort are presented in Tables 4.6 and 4.7, respectively.

Table 4.6

Mean and standard deviation of relative weights across clusters for cohort 2

Cluster	*N*	*M(SD)*				
		Relevance	Richness	Organization	Gra/Voc	Pro/Int
1	11	.12(.07)	.26(.11)	.05(.03)	.19(.13)	.38(.11)
2	42	.29(.10)	.26(.09)	.15(.06)	.16(.05)	.15(.06)
3	11	.34(.11)	.52(.12)	.05(.02)	.04(.02)	.04(.03)
Total	64	.27(.12)	.30(.14)	.11(.07)	.14(.08)	.17(.12)

Note: Relevance = Topic relevance; Richness = Content richness; Gra/Voc = Grammar and vocabulary; Pro/Int = Pronunciation and intonation

Table 4.7

ANOVA results for mean comparison between clusters for cohort 2

Source		*SS*	*df*	*MS*	*F*	*p*
Relevance	Between Groups	.319	2	.160	17.494	.000
	Within Groups	.557	61	.009		
Richness	Between Groups	.622	2	.311	30.255	.000
	Within Groups	.627	61	.010		
Organization	Between Groups	.134	2	.067	26.562	.000
	Within Groups	.154	61	.003		
Gra/Voc	Between Groups	.152	2	.076	16.807	.000
	Within Groups	.276	61	.005		
Pro/Int	Between Groups	.698	2	.349	75.527	.000
	Within Groups	.282	61	.005		

Note: Relevance = Topic relevance; Richness = Content richness; Gra/Voc = Grammar and vocabulary; Pro/Int = Pronunciation and intonation

The MANOVA conducted to compare the three clusters in the means of the five relative weights yielded a significant multivariate result, Pillai's Trace = 1.397, $F(10, 116) = 26.888$, $p = .000$. Univariate tests also indicated that the three clusters differed significantly in all five means of relative weights at the $p = .05$ level (Tables 4.6 and 4.7). The general pattern that can be recognized from Table 4.6 is that Cluster 1 had higher means in the relative weights of grammar and vocabulary and pronunciation and intonation but lower means in the relative weight of relevance than the other two clusters, Cluster 3 had higher means in the relative weights of relevance and richness but lower means in the relative weights of grammar and vocabulary and pronunciation and intonation than the other two clusters, while Cluster 2 lay in between in the relative weights of relevance, grammar and vocabulary and pronunciation and intonation. For Cluster 2, the mean weights of richness and organization were also closer to the weights of the other three criteria.

However, there were some differences in the specific patterning between the two cohorts. Most notably, for cohort 1, the mean relative weight in relevance distinguishes Cluster 3 from the other two clusters most effectively, but for cohort 2, the mean relative weight in richness did this job most effectively. On the whole, however, the same correspondence can be established between the clusters and the conceptual classification, with Cluster 1 matching the form-oriented type, Cluster 2 the balanced type, and Cluster 3 the content-oriented type.

According to the *Number* column of Tables 4.4 and 4.6, roughly two thirds of raters were classified as the balanced type, while only one third of raters were either form- or content-oriented.

4.2.2 Preliminary classification results for the whole sample

To correct for potential bias due to the small sample size of a single cohort, the raters of both cohorts were classified again using the same clustering method. The relative weights of both cohorts were consolidated into a large dataset upon which the cluster analysis was conducted. To help determine the number of clusters, plots of the coefficients of the last 10 stages in the agglomeration schedules and the dendrogram are presented in Figures 4.5 and 4.6 respectively.

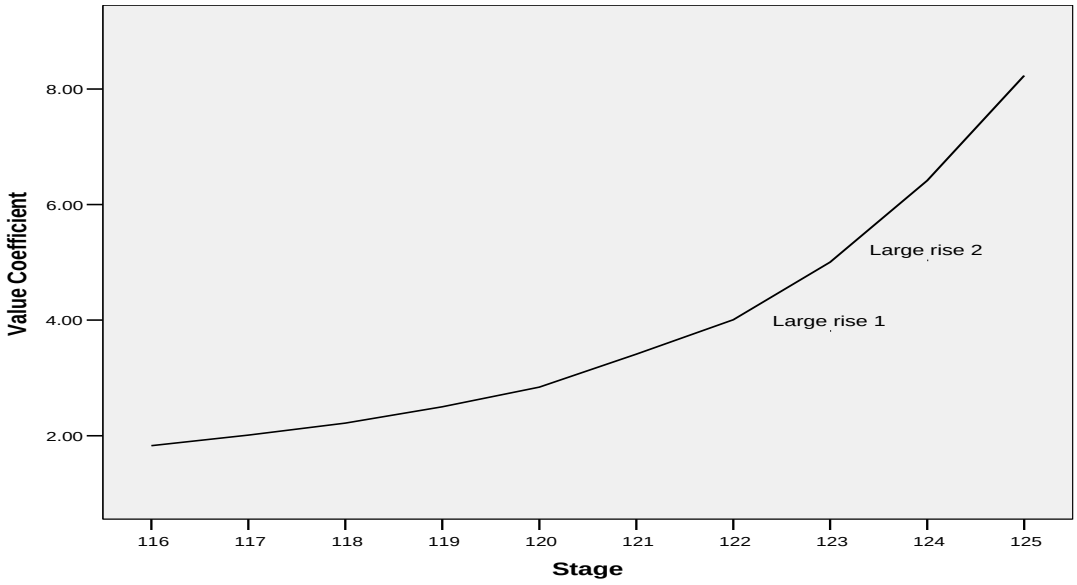

Figure 4.5. Agglomeration schedule of hierarchical cluster analysis for both cohorts

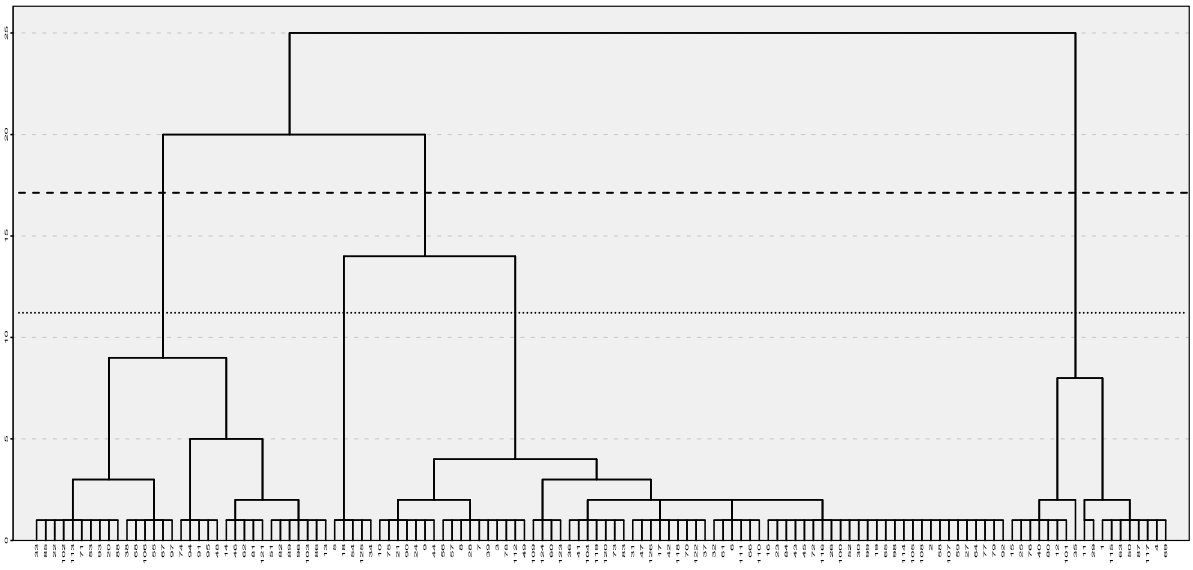

Figure 4.6. Dendrogram of hierarchical cluster analysis for both cohorts

This time, the four-cluster solution was found to compete with the three-cluster solution. The agglomeration schedule showed a sudden rise in the agglomeration coefficient from stage 122 to stage 123, followed by a steeper rise from stage 123 to stage 124. Though these were not clearly identifiable in Figure 4.5, in numerical terms the coefficient had been increasing steadily by .34, .57, and .59 from stage 119 to stage 122, but by 1.00 from stage 122 to stage 123, and by 1.41 from stage 123 to stage 124, which made it hard to judge whether the three- or four-cluster solution fit the data better.

The dendrogram (Figure 4.6), however, rendered a little more support to the three-cluster solution than to the four-cluster solution. For the three-cluster solution marked by the broken line, the range of rescaled distance over which the number of clusters in the solution does not change was 20 – 14 = 6. This distance was 14 – 9 = 5 for the four-cluster solution marked by the dotted line.

The correspondence between real data and conceptual taxonomies was retained for the pooled sample. The descriptive statistics of the relative weights and the results of the univariate tests for the whole sample are presented in Tables 4.8 and 4.9 respectively.

Table 4.8.

Mean and standard deviation of relative weights across preliminary clusters for both cohorts

Cluster	*N*	*M(SD)*				
		Relevance	Richness	Organization	Gra/Voc	Pro/Int
1	18	.15(.05)	.17(.06)	.06(.03)	.23(.15)	.39(.17)
2	75	.32(.13)	.23(.06)	.15(.06)	.16(.06)	.14(.07)
3	33	.27(.14)	.45(.11)	.08(.05)	.09(.06)	.11(.11)
Total	126	.28(.14)	.28(.13)	.12(.07)	.15(.09)	.17(.13)

Note: Relevance = Topic relevance; Richness = Content richness; Gra/Voc = Grammar and vocabulary; Pro/Int = Pronunciation and intonation

Table 4.9

ANOVA results for mean comparison between preliminary clusters for both cohorts

Source		*SS*	*df*	*MS*	*F*	*p*
Relevance	Between Groups	.448	2	.224	14.581	.000
	Within Groups	1.890	123	.015		
Richness	Between Groups	1.327	2	.664	112.562	.000
	Within Groups	.725	123	.006		
Organization	Between Groups	.174	2	.087	26.904	.000
	Within Groups	.397	123	.003		
Gra/Voc	Between Groups	.221	2	.111	16.955	.000
	Within Groups	.803	123	.007		
Pro/Int	Between Groups	1.059	2	.529	54.775	.000
	Within Groups	1.189	123	.010		

Note: Relevance = Topic relevance; Richness = Content richness; Gra/Voc = Grammar and vocabulary; Pro/Int = Pronunciation and intonation

The MANOVA conducted to compare the three clusters in the means of the five relative weights yielded a significant multivariate result, Pillai's Trace = 1.278, $F(10, 240) = 42.482$, $p = .000$. Univariate tests indicated that the three clusters differed significantly in all five means of relative weights at the $p = .05$ level (Tables 4.8 and 4.9).

The general pattern that can be recognized from Table 4.8 is that Cluster 1 had higher means in the two relative weights associated with form but lower means in the three relative weights associated with content, Cluster 3 had higher means in the relative weight of richness but lower means in the relative weights of grammar and vocabulary and pronunciation and intonation than the other two clusters, while Cluster 2 was closest to the grand mean in the four relative weights other than relevance, for which it had the highest mean among the three clusters. On the whole, the correspondence between the clusters and the conceptual classification was retained in the pooled sample, with Cluster 1 matching the form-oriented type, Cluster 2 the balanced type, and Cluster 3 the content-oriented type. As per this analysis, roughly 60% of the raters could be classified as balanced, 14% as form-oriented, and 26% as content-oriented.

4.2.3 Final classification results for the whole sample

Despite some discrepancies in the cluster membership of individual raters and the descriptive statistics of different clusters, the three-cluster solution worked as well in both cohort-specific and whole-sample analyses. Therefore, the final cluster membership of the raters was determined through a *k*-means cluster analysis on the same dataset of relative weights, where $k = 3$. The detailed classification results were given in Appendix C, but a summary of the results, including the descriptive statistics of the relative weights and the results of the univariate tests, is presented in Tables 4.10 and 4.11.

Table 4.10

Mean and standard deviation of relative weights across final clusters for both cohorts

Cluster	*N*	*M(SD)*				
		Relevance	Richness	Organization	Gra/Voc	Pro/Int
1	20	.52(.14)	.21(.10)	.07(.05)	.10(.07)	.10(.08)
2	82	.27(.06)	.31(.13)	.15(.06)	.16(.08)	.12(.06)
3	24	.15(.07)	.22(.10)	.07(.04)	.18(.11)	.39(.13)
Total	126	.28(.14)	.28(.13)	.12(.07)	.15(.09)	.17(.13)

Note: Relevance = Topic relevance; Richness = Content richness; Gra/Voc = Grammar and vocabulary; Pro/Int = Pronunciation and intonation

Table 4.11

ANOVA results for mean comparison between final clusters for both cohorts

Source		*SS*	*df*	*MS*	*F*	*p*
Relevance	Between Groups	1.545	2	.772	119.729	.000
	Within Groups	.793	123	.006		
Richness	Between Groups	.271	2	.135	9.347	.000
	Within Groups	1.781	123	.014		
Organization	Between Groups	.166	2	.083	25.198	.000
	Within Groups	.405	123	.003		
Gra/Voc	Between Groups	.081	2	.040	5.265	.000
	Within Groups	.943	123	.008		
Pro/Int	Between Groups	1.435	2	.717	108.481	.000
	Within Groups	.813	123	.007		

Note: Relevance = Topic relevance; Richness = Content richness; Gra/Voc = Grammar and vocabulary; Pro/Int = Pronunciation and intonation

The correspondence between real data and conceptual taxonomies remained unchanged in the results of the *k*-means cluster analysis. The MANOVA conducted to compare the three clusters in the means of the five relative weights yielded a significant multivariate result, Pillai's Trace = 1.232, $F(10, 240) = 38.463$, $p = .000$. Univariate tests indicated that the three clusters differed significantly in all five means of relative weights at the $p = .05$ level (Tables 4.10 and

4.11). Similar to the hierarchical clustering results, Cluster 1 corresponded to the content-oriented type, with the highest mean weight in relevance but lowest mean weights in grammar and vocabulary and pronunciation and intonation among the three clusters. Cluster 3, which corresponded to the form-oriented type, was the opposite: with the highest mean weights in pronunciation and intonation and grammar and vocabulary but lowest mean weight in relevance. Cluster 2 was balanced with means of relevance, pronunciation and intonation, and grammar and vocabulary running in between those of the other two types. However, it also had the highest means in richness of content and organization among the three types.

To recapitulate the analyses conducted up to this point, the raters' holistic ratings from the value judgment task were regressed on the five scoring criteria in the simulated score profiles. The resulting regression weights were then transformed into relative weights using Hoffman's (1960) equation. According to cluster analyses conducted on the relative weights, the raters were classified into three groups, which corresponded to the three types discussed in the literature review. As per the *Number* column in Table 4.10, roughly 65% of the raters were classified as balanced, 19% as form-oriented, and 16% as content-oriented. According to this result, 10 of the 21 raters included in the verbal protocols were form-oriented, nine were balanced, and only two were content-oriented.

4.3 Research Question 2: Characterization of rater types

4.3.1 Coverage of themes

Various aspects of the rating process could have been examined to characterize the three rater types. This study focused on comparing the rater types in the coverage of themes, self-perceived weights, and the conceptualization of relationship between criteria. Apart from these

general comparisons, the rater types were also contrasted in their responses to digression and salient negative performance in pronunciation and intonation.

The full transcripts of the verbal protocols were broken into 954 segments according to the principles set forth in Section 3.6.3 and two coders then coded the segments independently. The two coders agreed on 901, or 94.4%, of all the segments in terms of what themes were covered in each segment, and Cohen's kappa was estimated at .923. The two coders then discussed the discrepancies case by case and reached an agreement on each segment. The resultant distribution of raters' comments on the three themes—content, grammar, and pronunciation—is reported in Appendix D.

For comparative purposes, the percentages of comments on the three themes were calculated for each rater. These percentages were detailed in Appendix D. The reasons for comparing the different types of raters in terms of these percentages were twofold: 1) the number of comments varied considerably across raters, and 2) the percentages conveyed similar information to relative weights, which was the major concern of this study.

Table 4.12 reports the mean and standard deviation of these percentages for each type of rater.

Table 4.12

Mean and standard deviation of the percentages of comments

Type of rater	Content		Grammar and vocabulary		Pronunciation and intonation	
	M	*SD*	*M*	*SD*	*M*	*SD*
Form-oriented (n = 10)	.27	.07	.36	.05	.38	.04
Balanced (n = 9)	.35	.06	.33	.06	.33	.01
Content-oriented (n = 2)	.39	.05	.36	.05	.26	.06

The general patterns that can be identified from Table 4.12 are that, among the three types of raters, form-oriented raters had the highest percentage in pronunciation and intonation but the lowest percentage in content, whereas content-oriented raters had the highest percentage in content but the lowest percentage in pronunciation and intonation. Balanced raters lay in the middle in both these percentages. However, the three types of raters had similar percentages in grammar and vocabulary.

A MANOVA was conducted to compare the three types of raters in the means of these percentages, and the results of subsequent univariate tests are given in Table 4.13.

Table 4.13

ANOVA results for mean comparison between types of raters in the percentages of comments

Source		*SS*	*df*	*MS*	*F*	*p*
Percentage in content	Between Groups	.041	2	.021	4.694	.023
	Within Groups	.079	18	.004		
Percentage in grammar	Between Groups	.004	2	.002	.714	.503
	Within Groups	.056	18	.003		
Percentage in pronunciation	Between Groups	.029	2	.015	7.753	.004
	Within Groups	.034	18	.002		

The multivariate result indicated significant differences across the three types of raters: Pillai's Trace = .540, $F(4, 36) = 3.330$, $p = .02$. Univariate tests also indicated that the three types differed significantly in the percentages of comments on content and pronunciation and intonation at the $p = .05$ level, but not in the percentage of the comments on grammar and vocabulary (Table 4.13).

Prior to *post hoc* multiple comparisons, Levene's test of equality of error variances showed that the error variances were not significantly different, $F(2, 18) = .156$, $p = .857$ for content, $F(2, 18) = .103$, $p = .903$ for grammar and vocabulary, and $F(2, 18) = 1.983$, $p = .167$ in the case of pronunciation and intonation. Therefore, the Least Significance Difference (LSD) tests were conducted. The results showed that in the percentage of comments on content, the content-oriented and balanced raters had significantly higher means than the form-oriented raters ($p < .05$); in the percentage of comments on pronunciation and intonation, the form-oriented raters had a significantly higher mean than the other two types of raters, and the balanced raters had a significantly higher mean than the content-oriented raters; but no significant difference was found among the three types of raters in the percentage of comments on grammar and vocabulary.

This general pattern was similar to the pattern of relative weights across the three types of raters reported in the previous section. This is evidence that the form-oriented raters paid more attention to pronunciation and intonation, the content-oriented raters paid more attention to content, and the balanced raters divided their attention roughly equally.

4.3.2 Self-perceived weights

The summary question in the verbal protocols asked the raters to assess the relative importance they attached to the five criteria in the rating process. The transcript associated with this question was examined, and the frequencies of various response patterns of the three types of raters are reported here.

Of the 10 form-oriented raters, all reported that pronunciation and intonation was their top concern. While five of the 10 raters placed relevance or richness of content in the second place, four gave this place to grammar and vocabulary, and one reported that his sole consideration was pronunciation and intonation.

The balanced raters were more varied in terms of their top concerns. Six of them placed relevance and richness of content in the first place, two placed richness or relevance in the second place, following pronunciation and intonation, and one seemed to attach equal importance to pronunciation and intonation and richness of content.

The two content-oriented raters were similar to the six balanced raters who placed relevance and richness in the first place. While attaching most importance to content, both of them were aware of the strong impact of form-related criteria. Pronunciation and intonation, if "particularly good" or "bad" (Lee and Mia), would affect their overall judgment, but otherwise they were most concerned with content.

In sum, differences in self-perceived weights among the three types of raters are best characterized as general tendencies: The form-oriented raters all gave top priority to pronunciation and intonation, most of the balanced raters tended to place relevance and richness on a higher level of importance than pronunciation and intonation, and the content-oriented raters were more emphatic on content, but were still affected by pronunciation and intonation.

Such differences may be characterized as a continuum, with form and content at the two ends (Figure 4.7).

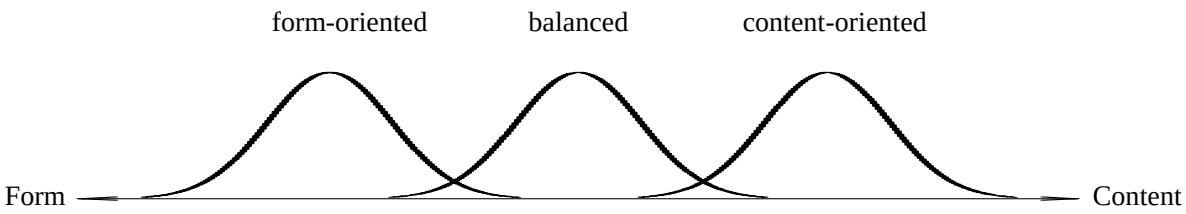

Figure 4.7. Three types of raters on the continuum of relative weights

At the form end of Figure 4.7, raters give full weights to form, whereas full weights go to content at the content end. Clearly, the form-oriented raters may be located closer to the form end, the content-oriented raters closer to the content end, and the balanced raters somewhere in between.

The continuum in Figure 4.7 is used to emphasize that all three types of raters were concerned with both form and content, and that their difference lay in whether they were concerned more with form or content or were balanced in their concern. Furthermore, a conceptual distribution can be imposed on each type of raters, such that variation within each type is allowed and that there is overlapping between two adjacent types to account for the borderline cases.

4.3.3 Relationship between criteria

Further description of the three types of raters was based on how they conceptualized the relationship between the form- and content-related criteria. The following summary was based on the parts of the verbal protocol transcripts concerning the raters' justification after rating each

segment of the recordings. Typical examples are extracted from the transcripts wherever applicable.

A peculiar behavior of the form-oriented raters was that they would base their assessment of content on form-related reasons. In rating the fourth segment of the third clip, Charlene decided that she would mark down the content by five points, and justified her decision by two reasons: 1) there was a lack of diversity in the sentence patterns used by the test-taker, and 2) there was some redundancy in the spoken discourse (Appendix E). While redundancy is certainly related with content, diversity in the sentence patterns is obviously an issue of grammar. Therefore, this is a clear case of justifying a reduction in the content score with a form-related reason. For another example, in rating the second segment of the second clip, Mina noticed that the test-taker was proceeding at a low rate of speech and decided that she would give a low content score. She reasoned that there would not be sufficient information if the test-taker spoke at such a rate (Appendix E). Likewise, Lou also penalized the test-taker in content for a low speech rate. Her typical comment was "her lack of fluency has obstructed the conveyance of message" (Appendix E), made on the first segment of the fifth clip:

> "But her pronunciation and intonation and her grammar and vocabulary were indeed not so good. In fact, her pronunciation and intonation was not bad in the first segment, but her fluency, out of whatever reasons, her lack of fluency has obstructed the conveyance of message, and so I gave her a failing score." (Appendix E)

A more common phenomenon, also evident with some balanced raters, was the anchoring of content scores in reference to pronunciation and intonation and grammar and vocabulary at the beginning of the rating process, when judgment of the content was still difficult. In rating the first segment of the second clip, May decided to give a tentative score to the content according to

the test-taker's performance in pronunciation and intonation and grammar and vocabulary. She explained that she was not able to make an overall judgment at this early stage of development. She also explained elsewhere (fifth segment of the fourth clip) that first impression on pronunciation and intonation was misleading (Appendix E). Similarly, Kim's tactic was to give the same score to all three subscales at the beginning, when she was not able to predict the content of the whole clip (first segment of clip two). For the form-oriented raters, this strategy came very handy, and Todd admitted that he paid attention only to the intonation at the beginning and tended to neglect the content under the influence of the intonation (first segment of clip two):

> "As to content, I think, maybe this was also affected by intonation, but anyway, his content did not have much, so to speak, strong attraction, not many merits, and I think this was mainly affected by his intonation, I listened only for his intonation at the beginning." (Appendix E)

Compared to the form-oriented raters, the balanced raters tended to distinguish content from form more clearly. After the fifth segment of the fourth clip, Hermione first commented on the grammatical errors, but immediately added that these did not affect the expression of the ideas:

> "There were some grammatical errors…. But these did not affect the expression of the ideas. I feel okay with it. Er, about this, about what she said, about what she had learnt, her expression was to the point, relevant." (Appendix E)

A brief summary of this excerpt is that grammar was bad, but content was okay. In other words, grammar and content were distinct criteria. Similarly, after giving two examples of grammatical

errors made by the test-taker, Josie decided that she would still give him a passing score in content, as she was still able to understand his ideas (Clip 2, Segment 4).

Another distinction was that the balanced raters sometimes based their assessment of form on content-related reasons, just the opposite of form-oriented raters. In commenting on the fifth segment of the second clip, Mona said that she found the pronunciation and intonation of the test-taker to be under average, but she decided to raise his score in pronunciation because he made his points clearly. She believed that "his pronunciation was okay if only he expressed his ideas clearly" (Appendix E). Elsewhere, she did the same thing for vocabulary (Clip 4, Segment 4). This response was not uncommon among balanced raters. Tisha, for example, commented that she would give a passing score to a test-taker if the test-taker conveyed sufficient message, even if he made a lot of errors in grammar (Clip 2, Segments 1 & 5). Josie, on the other hand, explained that inadequacy in content was the cause of inadequacy in form:

> "In the third segment, I think that, her poor language was caused by, er, a problem in the content, because she was recalling, she was recalling the situation in the college entrance exams…and so here, I found out that she did not so well in fluency, and I would give her a lower score in pronunciation and intonation."

However, the balanced raters also admitted that the assessment of content may be affected by pronunciation and intonation. In Hermione's words, "if the pronunciation of a student is really terrible, it will affect his content score" (Appendix E). Similarly, good pronunciation may help raise the content score, as Mona explained after the fifth segment of the fifth clip:

> "I will take away some points from her monologue. I don't know how unexpected is her 'unexpected', I mean, I will give her 50 or 55 for her content, I think. If I give her 55, it's because her pronunciation is not too bad." (Appendix E)

Likewise, the anchoring effect of pronunciation and intonation was also evident among the balanced raters. After the second segment of the second clip, Mona raised the content score a bit and explained why she did so:

> "I feel that sometimes, I mean the first segment, the influence of pronunciation and intonation was stronger, and I would judge that he was under average…I mean his pronunciation and intonation tended to affect his, the understanding of his idea. In the second segment I found out that he was, making his points…making some points in terms of content…and I can raise his score in content, I think." (Appendix E)

4.3.4 Typical contrasts

While the preceding sections reported findings from the comparisons in general aspects, this section focuses on the diverse responses of the different types of raters to digression and salient negative performance in pronunciation and intonation. The three types of raters were compared in the scores they gave to recordings with these features, and in the pattern of response to such features, summarized as frequencies from the verbal protocol transcripts.

A major rationale for distinguishing various types of raters according to their weighting patterns is the assumption that rater variability may result from the interaction between the raters' weighting patterns and the test-takers' characteristics. In accordance with Table 2.4, the form-oriented raters may be more sensitive to poor pronunciation and intonation and poor grammar and vocabulary, but less sensitive to poor content; whereas the content-oriented raters may have the opposite tendency, being more sensitive to poor content than to form-related features. In the same vein, the balanced raters may be equally sensitive to weaknesses in content and form.

Unfortunately only two participants in the verbal protocol were classified as content-oriented, which made the contrast between the form- and content-oriented raters less reliable. As a caution, interpretation of the contrastive study in this section should be based mainly on the difference between the form-oriented and the balanced raters.

The contrastive study reported here pertains to three clips of recordings used in the verbal protocols. These were chosen in accordance with the stereotypes depicted in Table 2.4. Clip 2 was chosen because the test-taker was heavily accented, and was given the lowest mean score in pronunciation and intonation by the two raters in the recording selection process (Section 3.4.2). The preceding sections have shown that most raters tended to be affected by salient features in pronunciation and intonation, and Clip 2 was chosen in this analysis with a view to finding out whether the types of raters differed in the degree to which they were affected. In contrast, Clip 5 was regarded as digressive, and the choice of this clip aimed at comparing the probabilities of a digression being discovered by different types of raters, which may be taken as a measure of sensitivity to content irrelevance. Clip 3 was chosen because it was most balanced in the three subscales according to the ratings obtained from the recording selection process.

The general responses of the three types of raters to the three types of test-takers can be deduced from the descriptive statistics in Table 4.14.

Table 4.14

Mean and standard deviation of scores given to three clips in the verbal protocols

		Form-oriented (n = 10)		Balanced (n = 9)		Content-oriented (n = 2)	
Clip	Subscale	*M*	*SD*	*M*	*SD*	*M*	*SD*
2	Content	60.00	5.27	59.44	7.68	57.50	3.54
	Pronunciation	60.00	5.27	57.22	7.55	52.50	3.54
	Grammar	60.00	4.71	57.78	6.18	50.00	.00
3	Content	75.00	3.33	73.89	8.58	70.00	.00
	Pronunciation	75.50	4.38	73.89	6.51	72.50	3.54
	Grammar	75.00	3.33	72.78	7.12	72.50	3.54
5	Content	65.50	6.85	62.22	9.39	57.50	3.54
	Pronunciation	67.00	7.53	66.11	3.33	62.50	10.61
	Grammar	67.50	6.77	66.11	4.17	62.50	3.54

As can be seen from Table 4.14, each type of rater gave similar mean scores to Clip 3 on all three subscales. The form-oriented and balanced raters gave Clip 2 similar mean scores on all three subscales, while the content-oriented raters gave this clip a higher mean score in content but lower mean scores in form-related criteria. In contrast, Clip 5 got a lower mean score in content but higher mean scores in form-related criteria from the balanced and content-oriented raters, whereas the form-oriented raters gave this clip similar mean scores on all three subscales. On the whole, Clip 2 fit the stereotype of a test-taker with salient problems in linguistic form, Clip 5 could be matched to the stereotype with salient problems in content, while Clip 3 served as an example of test-takers with a balanced profile.

On the basis of this profiling, the justifications provided in the verbal protocols were analyzed to compare the different types of raters in how they responded to salient problems in form or in content. For Clip 2, the raters' responses may be classified into three types according to how the raters evaluated content in spite of the heavy accent of the test-taker. The first type either made no justifications on content at all, or made vague comments such as "I could guess

what he means" (Mina) and "he has said something, and so, deserves a passing score" (Phoebe). In contrast, their comments on the two form-related subscales were much more specific. The second type made general comments on content, which reflected the criteria the raters had in mind. For example, "as to a failing case, let's have a look (at the rubric). First, the talk is digressive and totally irrelevant to the topic; second, it is totally illogical and confusing. As to this test-taker, I don't think he falls into this category in the rubric" (Josie). However, no specific content of the clip was mentioned in this type of response, for the mention of specific content is the feature of the third type. An example of specific content being mentioned is the following:

> "Further into the story, I sensed that his talk was roughly about a birthday, and then his family, that is, his parents celebrated his birthday for him, but I don't think he proved that he was able to 'learn something'. On this subscale, I gave him 65 in the beginning, as he was able to talk about an 'unexpected experience', but then he, he didn't 'learn something', and so I took 10 points off to make it 55, and it's a fail, on this subscale." (Kim)

Here "birthday", "family", and "his parents celebrated his birthday for him" were all specific content from the clip. Apart from that, the rater also covered the general criteria of "learn something" and "unexpected experience", which were requirements of the task.

Table 4.15 reports the number of raters from each type that were classified into each type of response.

Table 4.15

Number of raters by type of rater and type of response

Comment on content	Form-oriented	Balanced	Content-oriented
Lacking or vague	5	0	0
General only	1	4	1
Specific	4	5	1

In agreement with the results reported in Section 4.3.1, five of the 10 form-oriented raters made no comment on content, or only made vague comments, in stark contrast with the other two types of raters, of whom none belonged to this category. The overall picture here seemed to be that the form-oriented raters had a considerable chance of being influenced by salient negative features in pronunciation and intonation, such that they may neglect the content of the test-taker's talk to a certain degree. In contrast, the balanced and content-oriented raters tended to be less influenced by the accent of the test-taker, being able to direct their attention to the content.

The instruction of the task under discussion was to "talk about one unexpected experience you've had and what you've learned from it", and the rubric stated that a talk about the experience of another person was a case of digression. The content of Clip 5 was considered to be such a case for most of the talk was about how a severe disease developed in the speaker's aunt and then took away her life (Appendix F). Clearly this was the experience of another person rather than the speaker herself. Besides, dying from a severe disease after a long time of hospitalization could hardly be defined as "unexpected".

When the raters were divided according to whether they detected the digression in Clip 5, the three types of raters again displayed different tendencies. Table 4.16 lists the number of raters in each type who were able and unable to detect the above-mentioned digression.

Table 4.16

Number of raters in each type able and unable to detect digression in Clip 5

Digression detection	Form-oriented	Balanced	Content-oriented
Yes	3	7	2
No	7	2	0

As is evident in Table 4.16, seven out of 10 form-oriented raters did not regard the clip as digressive, whereas seven out of nine balanced raters and both content-oriented raters did. In general, therefore, form-oriented raters seemed to be less sensitive to topic irrelevance than the other two types of raters.

Relating Table 4.16 to Table 4.14 suggests that the detection of the digression in Clip 5 may be used to explain the different patterns of mean scores for the three subscales of this clip across the rater types. As the form-oriented raters did not tend to detect the digression, the mean score of content was comparable to the mean scores on the other two subscales. In the same vein, the balanced and content-oriented raters gave lower mean scores to content than to form most probably because they detected the digression. This reasoning was supported by the verbal protocols. In Section 4.3.3, for example, Mona, a balanced rater, commented that "I don't know how unexpected is her 'unexpected', I mean, I will give her 50 or 55 for her content, I think" (Appendix E). Similarly, Mia, a content-oriented rater, clearly announced that she would reduce the score of content for digression: "As for the score of the content, I will reduce it to 60, because her emphasis seems to have fallen on the story of another person, not of the speaker herself".

The tendency to reduce the content score for digression was shared by the three form-oriented raters who detected the digression. May, for example, commented at the end of the clip:

> "I think this test-taker had some problem in her topic. It should have been her own experience, and then its influence on herself. She did talk about what she witnessed, but, er, not much of her own experience. It was, after all, the story of another person. Therefore, there was a certain degree of digression, and I will bring the final score of content back down to 65, (but not lower), because her story was complete."

May suspected digression after hearing the first segment, and gave the score of 65. After the fourth segment, she raised the score to 70 for the test-taker began to talk about her own feeling, but decided to bring the final score back down to 65 at the end of the clip for the above reason. Similarly, Elaine and Charlene, the other form-oriented raters who detected the digression, also decided to reduce the content score on similar reasons.

In general, therefore, detection of digression led to reduced content scores, which explained the different patterns of mean scores for Clip 5 across the types of raters.

In comparison, it is hard to establish a relationship between different patterns of mean scores for Clip 2 and the raters' intensity of attention to content in the context of salient negative features in pronunciation and intonation. The balanced raters exhibited more attention to content than the form-oriented raters did, but there was little difference between the two types of raters in the pattern of mean scores, as mean scores were parallel across the three subscales for both types of raters. And though the mean score of content was higher than that of the form-related subscales for the two content-oriented raters, one of them gave the same scores to both content and pronunciation and intonation. It seemed that the impact of salient negative features of pronunciation and intonation was so strong that few raters were free from this impact.

This strong impact was even clearer when the transcript of Clip 2 (Appendix F) was compared to that of Clip 5 in terms of content. To accomplish this, some problems were cleared in advance that might hinder understanding of the transcripts as written texts. For Clip 2, these included the following:

- Repetition: e.g. "In, 2008, in, 2008…the exam, the examination…we just work hard…we just work hard…"
- Self-repair: "…was proaching me, was approaching me…and regretful, and grateful…"

- Hesitations and long pause: marked by “er” in the transcription. There were four such cases in the beginning and four toward the end.
- Mispronunciation of “moved” as “movid”, which happened twice.

Similar problems were shared by the transcript of Clip 5:

- Repetition: e.g. “…started to…started to, to…some very famous, some very famous doctors…”
- Self-repair: “decept, detect…she can’t work, she couldn’t work…her prain, her brain…”
- Hesitations and long pause: marked by “er” in the transcription. Only one case in the transcript.
- Mispronunciation of “leave” as “live”, which was repeated three times.

A condensed version of both clips was derived by correcting the mispronounced words and clearing the repetitions, self-repairs, and hesitations, but keeping the original wordings, including grammatical errors. The resultant text from Clip 2 ran like this:

> “In 2008, I am going to take the NCEE. At that time everyone was prepared for the examination, I am also. During that time, I am very nervous and forget everything. You know, we just work hard to pass through that examination. One day, when I returned, I’m very tired. I open the door, and see nobody at home. And inside the room was very dark. Just as I am going to turn on the light, I saw a light in the far way was approaching me. It’s a candle. I’m very surprised. And just as I’m wondering, someone was sing, “happy birthday to you.” I know, its’ my birthday, and my family prepared it for me. I am very moved. During that time, I’m just working hard, and forget everything, including my birthday. So, our family take a seat, and celebrate me for my birthday. That’s my

unexpected experience. I'm very glad that, at my working hard time, my family can bear my birthday in their mind. I am very moved, and grateful to them."

Here NCEE stands for the National College Entrance Examinations. And the "clean" version of Clip 5 was:

"When I was Junior 3, my aunt got a very strange and serious disease. All the doctors can't detect what kind of this disease exactly is, and my aunt started to lie on bed. She couldn't work any more. And her IQ was just like a child; she even can't count how much does seven plus eight. And she couldn't talk any more. We tried everything to save her, and even sent her to Shanghai to meet some very famous doctors, but still couldn't save her. And five months later my aunt's doctor said her brain can't work any more, and she can only rely on the breathing machine, so we decided to give up and my aunt died in her 50's. So after this, everyone of my family feel very upset, and we think we can't live without such an important members. But later I found that people we love will leave one day. The only thing we can do is accept the fact and live our own life better. This is the best way to memorize our loved ones, and I think even they have passed away they still don't want to see us feel sad about their death."

In terms of organization, both clips followed a temporal sequence in the development of the story, and ended the whole talk with a comment. The "clean" version of Clip 2 included 176 words, of which the first 150 were the story, and the last 26 were comments. In contrast, the condensed version of Clip 5 included 201 words, of which the first 155 were the story, and the last 46 were comments. Therefore, in terms of richness in content, the two clips were comparable in the narrative part, but Clip 5 featured a more detailed discussion of the lesson learned.

In terms of relevance to the topic of “one unexpected experience you’ve had and what you’ve learned from it”, the speaker in Clip 2 was very “surprised” at being greeted by a candle in the dark and a birthday song when he returned home in total negligence of his own birthday, which met the requirements for an “unexpected experience” that he himself had. For another, he was glad that his family kept his birthday in their mind, which was something that he “learned”, albeit not a subliminal “moral of the story” type. In contrast, the speaker in Clip 5 clearly missed the mark. As mentioned above, dying from a severe disease after a long time of hospitalization could hardly be defined as “unexpected”, and the experience happened to the speaker’s aunt rather than to herself. The comment in Clip 5, however, was more like a “moral of the story” lesson compared to the case of Clip 2.

In conclusion, when content was considered independently, there was no reason that Clip 2 should be inferior to Clip 5, and a lower content score for Clip 2 than for Clip 5, as Table 4.14 shows, was most likely the strong impact from the negative features of the pronunciation and intonation.

4.4 Summary

In an attempt to answer the first two research questions, this chapter has reported how and how well the raters were classified into three types: form-oriented, balanced, and content-oriented. The high R^2 values of the multiple regression analyses of the data from the value judgment task provided a strong basis for calculating the relative weights from the regression weights. As similar results were obtained from the hierarchical cluster analyses, and the results were found to agree well with conceptual categories of raters, the three-cluster solution of *k*-means clustering was adopted for final classification.

The meaningfulness of the three-cluster solution was examined through a series of comparisons of the different types of raters. Quantitatively, the three types of raters were found to have different patterns of attention, in terms of the percentages of comments on the themes of content, grammar, and pronunciation. The form-oriented raters made the highest percentage of comments on pronunciation and intonation, whereas the content-oriented raters made the highest percentage of comments on content, and the balanced raters seemed to distribute their attention equally. Qualitatively, the three types of raters exhibited different self-perceived weights and different beliefs in the relationship between form and content. All 10 form-oriented raters reported that pronunciation and intonation was their top concern, compared to only two (out of nine) balanced raters and neither of the two content-oriented raters. In contrast, both content-oriented raters reported that content-related criteria were their top concern, compared to six of the nine balanced raters and none of the form-oriented raters. While the form-oriented raters tended to base their assessment of content on form-related reasons, the balanced raters made a clearer distinction between content and form, and sometimes even based their assessment of form on content-related reasons.

The hypothesized interaction between the types of raters and the types of test-takers was basically confirmed. Salient negative features in pronunciation and intonation seemed to have a stronger impact on the form-oriented raters than on the other two types of raters, such that half of the form-oriented raters failed to comment on the content of a talk with a heavy accent. On the other hand, the balanced and content-oriented raters had a larger chance of detecting a digression in the talk than the form-oriented raters. Sensitivity to digression seemed to be the cause of lower mean scores on content than on the form-related subscales.

In spite of the differences, however, all three types of raters seemed to be subject to the strong impact of pronunciation and intonation, especially when salient positive or negative features were detected. Salient features in pronunciation and intonation tended to distract the raters from paying due attention to content. This was especially so in the rater's first impression. Adjustment to the first impression was contingent on careful consideration of the content. Probably out of these reasons, the different sensitivity to negative features in pronunciation and intonation did not lead to differences in the mean scores across the three subscales for the balanced raters.

On the whole, although the strong impact of salient features in pronunciation and intonation seemed to be shared by the three types of raters, the three-type classification has been found to be meaningful, thus providing a departure point in the search of relationship between weighting patterns and rater variability.

Chapter 5 Variability across Rater Types

With the first two research questions addressed in the preceding chapter, this chapter focuses solely on the third research question: To what degree are patterns of rater variability different across types of raters? In answer to this question, the three rater types—formed-oriented, balanced, and content-oriented—were compared in the four types of rater variability discussed in Chapter 2: severity, reliability, restriction of range, and halo effect. As the sparse dataset of test scores available from the test developer necessitated the linking of the test-taker groups in the statistical analyses, MFRM was used for the principal analyses for its ability to complete linking in the process of calibration. Wherever applicable, however, analyses based on HLM, G-Theory, and CFA were conducted for comparative purposes. Therefore, the MFRM results pertaining to the various patterns of rater variability will be reported first, followed by comparative results from the HLM, G-Theory and CFA analyses.

Table 5.1 provides a summary of the analyses conducted to detect each type of rater variability.

Table 5.1

Indicators of rater variability from different statistical models

Rater variability	MFRM	Alternative models	
		Indicators	Model
Severity	● Rater type severity ● Rater type × subscale interaction effects	● Main effect of the rater type ● Coefficients of the rater type × subscale interaction terms	HLM
Reliability	● Percentages of large standardized bias scores in the rater type × test-taker interactions and rater type × test-taker × subscale interactions	● Percentages of variance components ● G-coefficient ● Phi-coefficient	G-Theory
Range restriction	● Percentages of categories used ● Percentages of unexpected responses closer to the midpoint than the expected ratings	● Percentages of variance components	G-Theory
Halo effect	● Outfit and Infit statistics ● Reliability of subscale separation ● Comparison of models with subscale difficulty restricted and unrestricted ● Unexpected rater type × subscale interaction	● Comparison of single-factor model and two-factor model ● Factor loadings ● Inter-factor correlations	CFA

5.1 MFRM Analyses

5.1.1 Modeling process

The dataset of test scores available from the test developer was sparse in nature, with a considerable number of missing values. As each test-taker was rated on all five subscales, the test-taker was fully crossed with the subscale facet. As each rater gave scores to the same test-takers on all five subscales, the rater was also fully crossed with the subscale facet. However, each group of 32 test-takers was rated by a different pair of raters, and the test-taker was thus nested in the rater pair. Therefore, for the purpose of linking, it was necessary to obtain a sufficiently large subset of test scores within the administrative permission of the test developer such that all raters contained in the subset were properly linked. The final dataset of test-takers

used in this study comprised the five subscale scores of 3,894 test-takers, given by 33 different raters, seven of whom were classified as form-oriented in the analyses described in Chapter 4, 11 as balanced, and five as content-oriented. The remaining 10 raters were unclassified as they did not participate in the value judgment task.

The structure of the dataset can be graphically represented as Figure 5.1.

Test-taker Group	Subscales	Raters									
		1	2	3	4	5	6	7	8	9	…
1	A	*	*								…
1	B	*	*								…
1	C	*	*								…
1	D	*	*								…
1	E	*	*								…
2	A								*	*	…
2	B								*	*	…
2	C								*	*	…
2	D								*	*	…
2	E								*	*	…
3	A				*	*					…
3	B				*	*					…
3	C				*	*					…
3	D				*	*					…
3	E				*	*					…
.	.	.	.	.	.	.	.	.	.	.	…
.	.	.	.	.	.	.	.	.	.	.	…
.	.	.	.	.	.	.	.	.	.	.	…
i	A	*				*					…
i	B	*				*					…
i	C	*				*					…
i	D	*				*					…
i	E	*				*					…
.	.	.	.	.	.	.	.	.	.	.	…
.	.	.	.	.	.	.	.	.	.	.	…
.	.	.	.	.	.	.	.	.	.	.	…

Figure 5.1. Structure of the dataset of test scores

Each asterisk in Figure 5.1 stands for an observed data point, while the blank areas represent missing data. Each group includes 32 test-takers. Thus, for example, the first five rows signify that all 32 raters of Group 1 were rated on all five subscales by raters 1 and 2, but not by any other raters.

It is clear from the figure that there are a considerable number of missing values. The linking of test-taker groups and raters could only be achieved through anchor groups and raters. For example, raters 1 and 2 are linked through Group 1, while raters 1 and 5 through Group *i*, thus making it possible to link raters 2 and 5 through rater 1. In turn, Groups 1 and *i* are linked through rater 1. Similarly, Groups 3 and *i* are linked through rater 5. Since rater 4 is linked to rater 5 through Group 3, this completes the linking of raters 1, 2, 4, and 5, as well as test-taker groups 1, 3, and *i*. This explains how the 3,894 test-takers and 33 raters were linked in the MFRM analyses.

Prior to the MFRM analyses, the dataset was screened of all data points with the value zero, as the meaning of this value was unclear. In principle, zero values occurred only on the three content-related subscales, and there were two major reasons for their existence: failure of the test-taker to complete a certain task, or recording failure. As no validated record was available from the test developer, there was no way to find out the real reason behind each zero value. This screening resulted in the loss of 83 (or .31%) of the 26,385 data points.

The primary MFRM model adopted for the analyses was

$$\ln\left(\frac{P_{nijkx}}{P_{nijk(x-1)}}\right) = \beta_n - G_i - R_j - S_k - \tau_x . \qquad (5.1)$$

where β_n stands for the ability of test-taker n, G_i the severity of rater group i, R_j the severity of rater j, S_k the difficulty of subscale k, and τ_x the difficulty of response category x relative to category $x - 1$. The inclusion of the term G_i corresponds to the focus of this study on rater types rather than on individual raters. Together, the terms G_i and R_j account for rater severity. Thus Equation 5.1 states the probability of a test-taker with ability β_n obtaining a score of x relative to a score of $x - 1$ after adjusting for the severity of rater type i and individual rater j, as well as the difficulty of subscale k.

The five subscales were treated as parallel in this model. According to the design of TEM4-Oral, the first three subscales (retelling, talk, and discussion) should best be included as tasks or items, while the last two (pronunciation and intonation and grammar and vocabulary) as domains. However, these two scores were based on the test-taker's performance on all three tasks and there was no way to derive such scores for a single task. In other words, the domains were not crossed with the tasks. For this reason, it was impossible to treat the three tasks as one facet and the two domains as another, and the only possible solution was to treat all subscales as parallel.

A further complication arose in initial running of the model in Facets 3.58 (Linacre, 2005), as the raw data, in the form of percentage scores, did not fit the model. The number of unexpected responses with an absolute standardized residual greater than 2 was at a high of 3,219, accounting for 12.24% of the 26,302 valid responses used for estimation. This was much larger than the 5% recommended by Linacre (2010). To bypass this problem, the data were transformed into rating scales with 20 categories. This caused little difference in substantive interpretation for four of the subscales which, as mentioned in Chapter 3, were essentially 21-point rating scales, which were reduced to 20 categories after the deletion of the zero category.

The only exception, the retelling subscale, was essentially a 48-point partial credit scale, but was collapsed to the 20-point rating scale for agreement with the other scales.

This transformation successfully solved the problem of overall fit in the second run of the model. The number of unexpected responses with an absolute standardized residual greater than 2 dropped to 1,457, accounting for 5.54% of all the 26,302 valid responses used for estimation. This run, however, yielded extremely high Infit and Outfit values for the retelling task, 1.95 for Infit mean square, and 2.29 for Outfit mean square, as well as a near-zero value for discrimination (-.12), which indicated that retelling may be a totally different dimension from the other four subscales (Baghaei, 2008). This coincided with the different rating procedures of retelling, in essence a checklist of detailed points covered in the retold story, from the general impression type of rating shared by the other four subscales (Section 3.2). This substantive reasoning led to the drop of the retelling subscale in the final MFRM model.

After dropping the retelling subscale from the model, the valid responses totaled 21,044, of which 1,080 unexpected responses were identified with an absolute standardized residual greater than 2, accounting for 5.13% of all the valid responses. Of these, 185 had an absolute standardized residual greater than 3, accounting for .88% of all the valid responses. As these two percentages were close to the 5% and 1% recommended for overall fit (Linacre, 2010), the data could be said to fit the model sufficiently well, which justified the reporting of the following results.

5.1.2 General results

Figure 5.2 gives an overall picture of the results from the MFRM analysis.

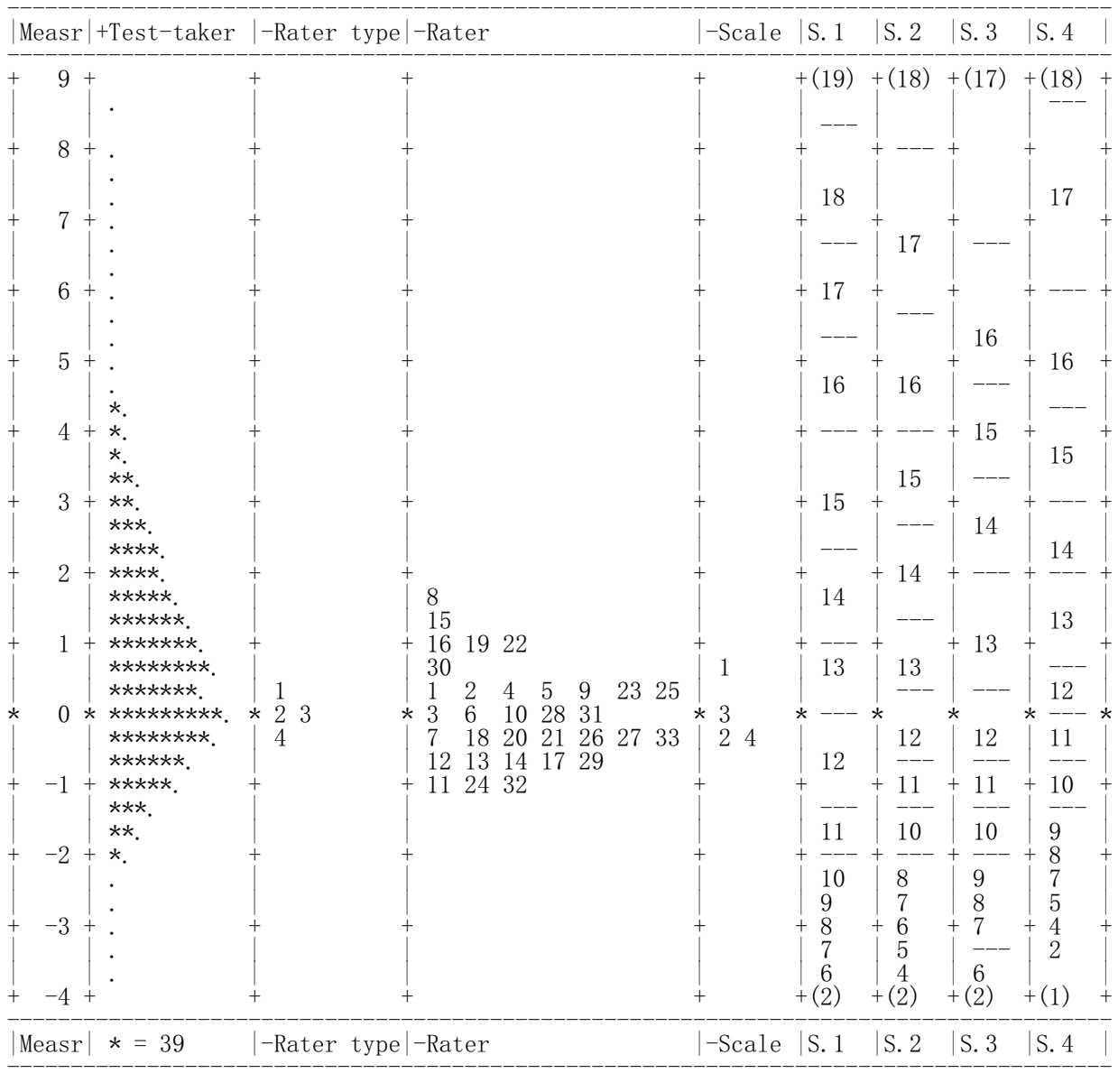

Figure 5.2. Variable map from Facets 3.58 analysis of TEM4-Oral (Note: Measr = Measure; Test-taker: each asterisk represents 39 test-takers, and a dot fewer than 39 test-takers. Rater type: 1 = Form-oriented; 2 = Balanced; 3 = Content-oriented; 4 = Unclassified. Scale: 1 = Talk; 2 = Discussion; 3 = Pronunciation and Intonation; 4 = Grammar and Vocabulary.)

In Figure 5.2, all the facets were calibrated on the same logit scale represented by the first column. The second column displays the distribution of test-taker ability along the logit scale, where each asterisk stands for 39 test-takers, and a dot for fewer than 39 test-takers. Test-takers with higher ability were placed at higher positions in the column. The third column shows the

severity of the rater types, with each number signifying a different group of raters. The group with higher severity is placed at a higher position in the column. The fourth column gives the severity for individual raters, represented by their numbers. The fifth column displays the difficulty of the four subscales, with the more difficult subscale toward the upper end. The last columns use horizontal bars to represent the transition points between score categories for each group of raters. The four rater groups were represented by S.1 through S.4.

According to the figure, the ability estimates of the test-takers were widely distributed along the logit scale, from a low of nearly -4 logits to a high of over 8 logits. The majority of the test-takers, however, were distributed between -2 logits and 4.5 logits. The distribution was negatively skewed. While the rater types were closely clustered around 0 logits with a small variation, the individual raters had a larger spread, covering a range from -1 logits to nearly 2 logits. The subscales were also somewhat separated from each other, especially subscale 1 (talk), which had a higher difficulty than the other three subscales. The transition points between categories were somewhat different across the rater types, but a more detailed examination will be left until a later section.

While Figure 5.2 facilitates an intuitive interpretation of the overall results, Table 5.2 presents the same results through summary statistics for the various facets.

Table 5.2

Summary statistics for the MFRM analysis of the TEM4-Oral data

Statistics	Test-takers	Rater types	Individual raters	Subscales
Measure	.83	.00	.00	.00
Model *S.E.*	.46	.01	.04	.01
χ^2	58689.6*	579.8*	9786.7*	2037.4*
df	3893	3	32	3
Separation index	3.52	11.06	16.66	21.50
Separation reliability	.93	.99	1.00	1.00

*$p < .01$.

The mean logits of the raters and the subscales were anchored to be zero in the analysis, but the mean logit of the test-taker ability was not restricted. This explains why the test-takers had a mean logit of .83. Elsewhere, the interpretation of the summary statistics across the four facets is straightforward. The χ^2 for each facet is marked with an asterisk, indicating statistical significance at the $p = .01$ level on the corresponding degrees of freedom. This is a sign that at least two elements of the facet had significantly different logit measures. The separation index for each facet indicates how many statistically distinct strata can be identified among the elements of the facet. The high separation reliability estimates are also signs that the elements of each facet can be reliably identified from each other.

For the major concern of this study, rater types, the χ^2 was estimated at 579.8, significant at the $p = .01$ level on 3 degrees of freedom. This indicates that at least two of the rater types had significantly different severities. The same inference can be made from the separation index, 11.06, which indicates that 11 statistically distinct strata could be identified among the rater types, much more than the number of rater types. The high separation reliability estimate, .99, is a sign that the rater types could be reliably identified from each other. On the basis of all this

information, the same inference can be made that there were significant differences in severity across the rater types.

5.1.3 Severity of rater types

As the χ^2 test revealed significant differences in severity across the rater types, pair-wise comparisons between rater types were conducted to find out which pairs of rater types exhibited significant differences using the following equation suggested by Linacre (2010):

$$t = (\text{Measure1} - \text{Measure2})/\text{sqrt}(\text{SE1}^2 + \text{SE2}^2). \quad (5.2)$$

This LSD (Least Square Difference) test statistic has $N - k$ degrees of freedom. In this case, N = 33 was the number of individual raters, and k = 4 the number of rater groups, including the unclassified group. The results of these *t*-tests, together with the relevant measures of rater type severity, and the Infit and Outfit mean squares, are reported in Table 5.3.

Table 5.3

Pairwise comparison of the rater types in severity

Variable	Measure	Model *S.E.*	Infit MnSq	Outfit MnSq	*t*	*df*	*p*
Comparison 1					6.26	29	.000
Form-oriented	.21	.02	1.08	1.09			
Balanced	.07	.01	1.11	1.13			
Comparison 2					9.55	29	.000
Form-oriented	.21	.02	1.08	1.09			
Content-oriented	-.06	.02	.94	.92			
Comparison 3					5.81	29	.000
Balanced	.07	.01	1.11	1.13			
Content-oriented	-.06	.02	.94	.92			

All Infit and Outfit mean squares were close to the expected value of 1, which indicates that all three rater types fit the model sufficiently well. Of the three rater types, the form-oriented type was found to be the most severe, with a measure of .21 logits; while the content-oriented type was the most lenient, with a measure of -.06 logits. The balanced type came in the middle, with a measure of .07 logits. The grand mean would be zero if the measure for the unclassified rater group, -.23 logits, was taken into consideration. The series of *t*-tests found out that all pairs of rater types were significantly different in severity at the $p = .01$ level. In other words, the form-oriented type was more severe than the other two types, and the balanced type was more severe than the content-oriented type. If the unclassified group of raters was left out of consideration, then the balanced type would have its severity close to the grand mean. Against this reference, the form-oriented type was biased toward the severe end, while the content-oriented type toward the lenient end.

5.1.4 Interaction effects

Interaction effects were the second concern of this analysis, which provided information to address the following two issues: the severity of the three rater types on specific subscales after accounting for their general severity, and the degree of biases of each rater type. The first issue was an extension of the concern over the severity of the rater types, while the second issue was directly related to the reliability of the rater types.

The rater type × subscale interaction provided information for the first issue. The MFRM analysis reported four interaction terms that had an absolute standardized value (*t*-score) greater than 2. The form-oriented type was found to be unexpectedly severe on pronunciation and intonation, but lenient on grammar and vocabulary; the content-oriented type was unexpectedly

severe on grammar and vocabulary, but lenient on discussion. In contrast, the balanced type was not subject to significant rater type × subscale interaction. The effects of such biases are more specifically displayed in Table 5.4, which lists all the statistically significant pairwise differences between the rater types in severity on each subscale.

Table 5.4

Significant pairwise differences between the rater types in severity on subscales

Variable	Measure	Model *S.E.*	Contrast	Joint *S.E.*	*t*	*df*	*p*
Discussion							
Form-oriented	-.26	.03	.15	.05	3.00	1,908	.003
Content-oriented	-.41	.04					
Balanced	-.22	.02	.20	.05	4.37	2,548	.000
Content-oriented	-.41	.04					
Pronunciation and intonation							
Form-oriented	.09	.03	.17	.04	4.42	2,867	.000
Balanced	-.08	.02					
Form-oriented	.09	.03	.16	.05	3.24	1,909	.001
Content-oriented	-.07	.04					
Grammar and vocabulary							
Form-oriented	-.32	.03	-.16	.04	-3.93	2,867	.000
Balanced	-.17	.02					
Form-oriented	-.32	.03	-.34	.05	-7.06	1,909	.000
Content-oriented	.02	.04					
Balanced	-.17	.02	-.19	.04	-4.27	2,550	.000
Content-oriented	.02	.04					

The first thing that should be noted here is that the three rater types did not differ significantly in severity on the talk subscale, but significant differences were found on the other three subscales. On the discussion subscale, the content-oriented raters were more lenient than the other two types. On the pronunciation and intonation subscale, the form-oriented raters were more severe than the other two types. On the grammar and vocabulary subscale, significant

difference was found for each pairwise comparison, with the form-oriented raters being more lenient than the other two types, and the content-oriented raters more severe than the other two types.

Figure 5.3 is a graphical presentation of the same information.

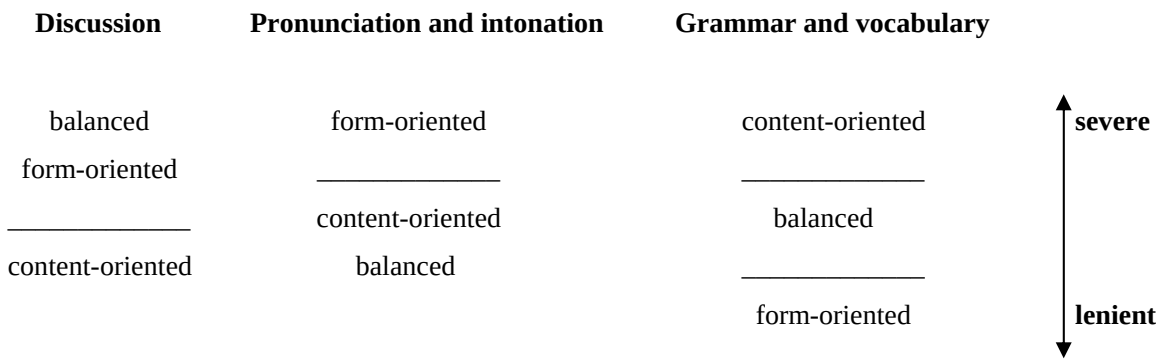

Figure 5.3. Differences between the rater types in severity on subscales

Figure 5.3 represents the different severities of the rater types in the same way as Figure 5.2, with the most severe rater type on the top and the most lenient at the bottom of each subscale. A horizontal line between two rater types signifies a statistically significant difference. According to its relative position in the figure, the form-oriented type was more severe than the balanced type on one subscale (pronunciation and intonation), but more lenient on another subscale (grammar and vocabulary); the form-oriented type was more severe than the content-oriented type on two subscales (discussion and pronunciation and intonation), but more lenient on another subscale (grammar and vocabulary); the balanced type was more severe than the content-oriented type on one subscale (discussion), but more lenient on another subscale (grammar and vocabulary). As the general severity of the form-oriented type was the highest among the three types, its extra severity on pronunciation and intonation deserves special

attention. Similarly, the extra leniency of the content-oriented type on discussion should also be noted, as this rater type had the lowest general severity.

For comparing the rater types in the degree of biases, summary statistics for the rater type × test-taker interaction and the rater type × test-taker × subscale interaction are reported in Table 5.5.

Table 5.5

Summary statistics for the interaction analysis

Statistics	Type of Interaction					
	Rater type × Test taker			Rater type × Test taker × Subscale		
	Form-oriented	Balanced	Content-oriented	Form-oriented	Balanced	Content-oriented
N combinations	1,114	1,755	797	4,455	7,018	3,185
% large *t* scores[a]	5.12	3.87	5.52	5.77	5.67	4.58
Minimum *t*	-4.39	-4.41	-4.13	-5.84	-6.83	-4.37
Maximum *t*	3.62	3.78	4.41	3.67	3.94	3.66

[a]Percentage of absolute *t* scores (standardized bias scores) equal to or greater than 2.

Due to the different numbers of raters in various types, the number of combinations varied considerably. Therefore, it is more reasonable to compare the rater types in the percentage of large *t* scores, i.e. standardized bias scores equal to or greater than 2. In this respect, little difference was found across the three rater types. This percentage ranged from 3.87 to 5.52 in the case of rater type × test-taker interaction, and from 4.58 to 5.77 in the case of the three-way interaction. These percentages may be considered low, which indicated that the three types of raters were sufficiently consistent in rating the different test-takers in general, and sufficiently consistent in rating the different test-takers on different subscales. The minimum and maximum *t* scores were also comparable among the three rater types, as the two bottom rows of Table 5.5

show. On the whole, therefore, the three rater types differed little in terms of rater type × test-taker interaction and three-way interaction and were sufficiently consistent as groups.

5.1.5 Restriction of range

According to Engelhard (1994, 2002), low sample reliability of person separation and small variance of the person ability estimates (see Table 2.2) can be indications of range restriction. For the whole sample, the reliability of test-taker separation was estimated at .93 (Table 5.2), while the standard deviation of test-taker ability estimates was 1.74. As this statistic is in logit unit, a figure greater than 1 can be interpreted as large standard deviation. Accordingly, restriction of range is of little concern for the whole sample of raters.

Differences across rater types in restriction of range may be discerned initially from the rightmost columns of Figure 5.2. As the unclassified rater group is not relevant for this purpose, only Columns S.1 through S.3 are of concern here. The test-taker ability estimates were negatively skewed, and so it is reasonable to examine the number of categories used by each of the three rater types that corresponded to the range of ability measure from -2 to 4. From Column S.1, it is clear that five categories (11 to 15) corresponded to this range for the form-oriented type. In contrast, Column S.2 shows that the balanced type used six categories (10 to 15) for this range of ability measure, while the content-oriented type covered this range with roughly five and a half categories (10 to halfway 15), according to Column S.3. This comparison reveals the general tendency that the form-oriented type was more subject to restriction of range, the balanced type affected least by it, while the content-oriented type was situated in between.

More specific information may be obtained from Table 5.6, which lists the frequencies of various categories used by the different rater types, together with relevant Rasch statistics. To

simplify presentation, the range of categories was limited from 8 to 17, as 99% of the category scores fell in this range.

Table 5.6

Frequencies of various categories used by different rater types

Cat. Score	Form-oriented			Balanced			Content-oriented		
	Count (%)	Ave. Meas.	Tran. Point	Count (%)	Ave. Meas.	Tran. Point	Count (%)	Ave. Meas.	Tran. Point
8	55 (1%)	-2.30	-3.25	110 (2%)	-1.69	-2.62	57 (2%)	-1.94	-2.94
9	67 (2%)	-2.34*	-2.98	192 (3%)	-1.52	-2.35	110 (3%)	-1.82	-2.51
10	285 (6%)	-1.54	-2.59	634 (9%)	-1.12	-1.96	269 (8%)	-1.15	-1.99
11	410 (9%)	-1.05	-2.01	689(10%)	-.66	-1.41	292 (9%)	-.65	-1.37
12	1,063(24%)	-.45	-1.23	1,504(21%)	-.12	-.72	716(22%)	-.02	-.65
13	1,026(23%)	.37	-.12	1,571(22%)	.64	.21	793(25%)	.97	.43
14	805(18%)	1.31	1.13	1,216(17%)	1.57	1.42	517(16%)	2.13	1.91
15	486(11%)	2.40	2.45	639 (9%)	2.57	2.72	251 (8%)	3.31	3.35
16	157 (4%)	3.50	3.86	293 (4%)	3.83	4.08	112 (4%)	4.42	4.71
17	60 (1%)	4.16	5.25	89 (1%)	4.72	5.70	39 (1%)	5.73	6.50

*Note: Cat. Score = Category Score; Ave. Meas. = Average Measure in logits; Tran. Point = logit measure for the expected score corresponding to the category score less .5 score points, regarded as the transition point between two adjacent category scores; * = reversed measure which does not increase with each higher category.*

The *Tran. Point* column in Table 5.6 gives the same information as Columns S.1 to S.3 in Figure 5.2. For the form-oriented type, for example, the transition point was -2.01 logits for category 11, indicating that a test-taker with this ability measure would be rated half way between categories 10 and 11 and rounded up to 11. Similarly, the transition point for category 10 was -1.96 logits for the balanced type and -1.99 logits for the content-oriented type, indicating that a test-taker with this ability measure would be rated half way between categories 9 and 10 and rounded up to 10. Toward the higher end, the transition point for category 16 was 3.86 logits for the form-oriented type, 4.08 logits for the balanced type, and 4.42 logits for the content-oriented type. Therefore, in the range between -2 and 4 logits, there were five categories (11 to 15) for the form-oriented type, but six categories (10 to 15) for the balanced type. The

correspondence between the categories and this range was not so neat for the content-oriented type, but roughly there were five and a half categories (10 to halfway 15). This is the same as the information contained in Columns S.1 through S.3 of Figure 5.2, but in numerical terms.

The percentages give similar information. The five most frequently used categories were 11 to 15 for the form-oriented type, together accounting for 85% of all the category scores given by this type of rater. In comparison, categories 11 to 15 accounted for only 79% of all the category scores for the balanced type, and only 80% for the content-oriented type. This indicates that the form-oriented type was subject to restriction of range to a greater extent than the other two types.

A special type of range restriction is central tendency. Myford and Wolfe (2004) suggested that fine-grained information on this tendency may be obtained from the table of unexpected responses with an absolute standardized residual equal to or greater than 2, as observed ratings subject to central tendency tend to be closer to the midpoint of the scale than the expected ratings. The fair-mean average of test-taker ability, estimated at 12.68 by Facets, was used for the midpoint of the scale, and all the unexpected responses reported for each rater type were compared to the expected ratings to see which was closer to the midpoint. The result of this comparison is reported in Table 5.7.

Table 5.7

Number and percentage of unexpected responses indicating central tendency

Rater type	*N* unexpected responses indicating central tendency	*N* combinations	%
Form-oriented	51	4,455	1.14
Balanced	102	7,018	1.45
Content-oriented	100	3,185	3.14

As Table 5.7 shows, the percentages of unexpected responses closer to the midpoint of the scale than the expected ratings were 1.14% for the form-oriented type, 1.45% for the balanced

type, and 3.14% for the content-oriented. These percentages were based on all the possible combinations of rater type, test-taker, and subscale, which are displayed in the third column of the table. This information indicates that the content-oriented type had a stronger degree of central tendency than the other two types of rater.

5.1.6 Halo effect

Engelhard (1994, 2002) suggested the following as signs of a halo effect (see Table 2.2): Small (< .5) Outfit and Infit statistics on the rater facet, low sample reliability of domain separation, and close clustering of domain difficulty values. As the concern of this study is over rater types, the same principles were applied to the rater type facet rather than the individual rater facet. In terms of the Outfit and Infit mean squares, all three types had both statistics close to 1, in the range from .92 to 1.13 (Table 5.3). The reliability of subscale separation was estimated at 1.00 (Table 5.2), and the subscale difficulty values were .51 for talk, -.20 for discussion, -.12 for pronunciation and intonation, and -.19 for grammar and vocabulary. As is visually evident in Figure 5.2, the talk subscale was set off from the other three subscales in difficulty, while the other three subscales tended to cluster more closely. This information seems to suggest that there may be halo effects across these three subscales, though the talk subscale tends to be independent from such effects.

Another test of the halo effects, suggested by Linacre (Myford & Wolfe, 2004), involves anchoring all subscales at the same difficulty (usually 0) before fitting the MFRM model. For this study, this test has the advantage of detecting rater types who are likely to be exhibiting halo, as such rater types will show better fit than rater types that are free from the halo effects. Another sign of halo, which can be deduced from model comparison, is that the fit of halo-prone

rater types will fit the restricted model as well as the original model, or even better. To facilitate this test, the Infit and Outfit mean squares resulting from the restricted model are reported in Table 5.8.

Table 5.8

Infit and Outfit mean squares across rater types with all subscale difficulties anchored at 0

Rater Type	Infit Mean Square	Outfit Mean Square
Form-oriented	1.07	1.09
Balanced	1.08	1.12
Content-oriented	.91	.92

Comparison of the Infit and Outfit mean squares to those reported in Table 5.3 shows that the greatest size of change resulting from the fixing of subscale difficulty values was .03 in Infit mean square, and .01 in Outfit mean square. This suggests that all three rater types were subject to halo effects to a certain extent, but that there was little difference across the three types in the degree of the halo effects.

A more fine-grained analysis helped identify the specific rater types and subscales affected by the halo effects. Myford and Wolfe (2004) suggested that the rater × trait interaction, or the rater type × subscale interaction in this case, may be used to detect the halo effects. Specifically, if a type of rater gives unexpectedly low scores on an easy subscale, or high scores on a difficult subscale, then this is halo. This is so because such unexpected scores are drawn closer to the rater type severity measure from the expected subscale difficulty value. Accordingly, among the rater type × subscale biases reported in Section 5.1.4, the following were indications of halo effects: a) the severe rating of the form-oriented type on pronunciation and intonation, and b) the severe rating of the content-oriented type on grammar and vocabulary.

To see how this was so, the expected values of the relevant rater type severity and subscale difficulty are tabulated together with the bias measures in Table 5.9.

Table 5.9

Rater type by subscale bias measures indicating halo effects

Expected values				Rater type × subscale measure	Bias		
Rater type	Measure	Subscale	Measure		Measure	Model *S.E.*	*t* score
Form-oriented	.21	Pronunciation	-.12	.09	.20	.03	6.64
Content-oriented	-.06	Grammar	-.19	.02	.21	.04	5.69

The rater type × subscale measure can be compared to the expected values of the relevant subscales to detect halo effects. In the case of the form-oriented type and the pronunciation and intonation subscale, Table 5.9 shows that the expected difficulty of the pronunciation and intonation subscale was -.12. As the severity of the form-oriented type on the pronunciation and intonation subscale was .09, the bias measure was .09 – (-.12) = .21 (reported as .20 in Table 5.9, which was the rounded value produced by Facets), which corresponded to a *t* score of 6.64, given the model S.E. of .03. Thus, the rating of the form-oriented type on pronunciation and intonation was significantly more severe than the expected value of -.12 at the $p = .05$ level. It deviated from the expected difficulty of this subscale toward the expected severity of the form-oriented type (.21) by .20 logits. Similarly, the severity of the content-oriented type on the grammar and vocabulary subscale deviated from the expected difficulty of the relevant subscale (-.19) toward the expected severity of this rater type (-.06) by .21 logits, to such an extent that it resulted in a greater logit value (.02) than the expected severity.

5.1.7 Summary

The MFRM analyses provided rich information that helped detect various patterns of variability concerning the three rater types. To begin with, when the balanced rater type was regarded as the reference type, the form-oriented rater type was found to be biased toward the severe end, while the content-oriented rater type toward the lenient end. In addition, the form-oriented rater type displayed extra severity in pronunciation and intonation, while the content-oriented rater type exhibited extra leniency in discussion. In terms of the rater type × test-taker interaction and the three-way interaction, however, the three rater types were comparable in the percentage of unexpected responses that had absolute standardized bias scores equal to or greater than 2, which is a sign of comparable degrees of consistency across rater types. The comparison of the frequencies of various categories used by different rater types indicated that the form-oriented type was subject to restriction of range to a greater extent than the other two types, though the content-oriented type was found to have a stronger degree of central tendency than the other two types according to the comparison of the percentages of unexpected responses closer to the midpoint of the scale than the expected ratings. Comparison of the Infit and Outfit statistics resulting from the restricted MFRM model with all the subscales fixed at the same difficulty to those from the original model suggested overall halo effects for all three rater types, while analysis of the rater type × subscale interaction effects pinpointed the specific halo effects: the form-oriented type on the pronunciation and intonation subscale, and the content-oriented type on the grammar and vocabulary subscale.

5.2 Alternative Analyses

5.2.1 Modeling process

For comparative purposes, analyses based on HLM, G-Theory, and CFA were conducted on the same dataset as was described in Section 5.1.1. As the retelling subscale was excluded from the MFRM analyses, it was also disregarded in these alternative analyses, so that comparison of results was more meaningful.

Of these analyses, HLM was performed in lieu of factorial ANOVA. The latter was recommended in earlier literature for the detection of rater variability (Guilford, 1954; Saal et al., 1980), but the former had much more to be recommended, as the structure of the dataset was obviously hierarchical, with test-takers nested within raters, and raters within rater types. With HLM, the interdependence among test-takers within raters and among raters within rater types could be accounted for (Raudenbush & Bryk, 2002).

Strictly speaking, test-takers were crossed with raters within each group of 32 test-takers, but few of the rater pairs who rated the same group of test-takers belonged to the same rater type, which complicated the interrelationship between the facets. For example, in Figure 5.1, raters 1 and 2 were crossed with test-takers in Group 1, but rater 1 may belong to the form-oriented type while rater 2 to the balanced type. Similarly, raters 8 and 9 were crossed with test-takers in Group 2, but rater 8 may belong to the form-oriented type while rater 9 to the content-oriented type. Therefore, the crossing of test-takers with raters was neither nested within rater types nor crossed with the facet.

An alternative was to disregard the crossing between the raters and the test-takers and to regard the test-takers as nested within the raters. This may lead to a certain degree of information loss on the test-taker and rater levels, but the raters would be neatly nested within rater types. As

the major concern was over rater types, this alternative was adopted in the final HLM model. On level one, the criterion variable was the subscale score (level-1), and the predictor variable was solely the unconditional mean of the rater (level-2). On level-2, the means-as-outcome of the raters was predicted by the rater type, subscale and their interaction. Therefore, the level-1 model was

$$Y_{ijk} = R_{0jk} + e_{ijk} \tag{5.3}$$

where Y_{ijk} stands for the score given to test-taker *i* by rater *j* on subscale *k*; R_{0jk} the intercept for rater *j* on subscale *k*, and e_{ijk} the error for test-taker *i* rated by rater *j* on subscale *k*. The level-2 model was

$$R_{0jk} = \gamma_{01}T_l + \gamma_{02}S_k + \gamma_{12}T_lS_k + \mu_{0jk} \tag{5.4}$$

where the intercept for rater *j* on subscale *k* is the sum of the effect of being in rater type *l* (i.e. $\gamma_{01}T_l$), the effect of subscale *k* (i.e. $\gamma_{02}S_k$), the interaction between rater type *l* and subscale *k* (i.e. $\gamma_{12}T_lS_k$), and the effect of everything else on which rater *j* differed from other raters (μ_{0jk}). The subscale scores were centered before the analysis, so no grand mean term was included in Equation 5.4.

While the HLM model provided information to facilitate inference about rater type severity and rater type × subscale interaction, G-studies were conducted to estimate the degree of consistency associated with each rater type. The only computer program for handling unbalanced

designs available at the time of the study, urGENOVA, was "not intended for situations in which there are large amounts of missing data" (Brennan, 2001b, p. 15), which was the case in this study (Figure 5.1). Therefore, a separate G-study was done on each test-taker group whose scores from both raters were available. There were 43 such groups, for each of which a rater × test-taker × subscale design was adopted. The results from all G-studies involving a rater of a certain type were then pooled for inference about the rater type. The number of G-studies involving each rater group was 18 for the form-oriented type, 31 for the balanced type, 13 for the content-oriented type, and 24 for the unclassified group. The different rater types were then compared in terms of the mean values of the relevant variance component percentages, and G- and phi-coefficients.

The factor structure of the TEM4-Oral was then examined through CFA separately for each rater type. The data associated with all raters of the same type were combined into a subset by stacking all the cases rated by one rater on those rated by another. In effect, the data structure displayed in Figure 5.1 was transformed into an $n \times 4$ matrix for each subset, where n is the number of cases, and 4 the number of subscores. For test-taker groups who had been rated by two raters, only the scores given by one of the raters were retained, so that the cases were independent from each other. In terms of Figure 5.1, for example, group 1 had been rated by raters 1 and 2, but only the scores given by one of these raters were retained, on a random basis. As a result of this selection and the case-wise deletion of missing data, the final number of cases included in the CFA was $n = 922$ for the form-oriented subset, $n = 1{,}465$ for the balanced subset, and $n = 669$ for the content-oriented subset. As each case had four subscores, the raw data were 922×4, $1{,}465 \times 4$, and 669×4 matrices respectively. For each rater type, two models were fitted to the subset, a single-factor model, and a two-factor model (Figure 5.4).

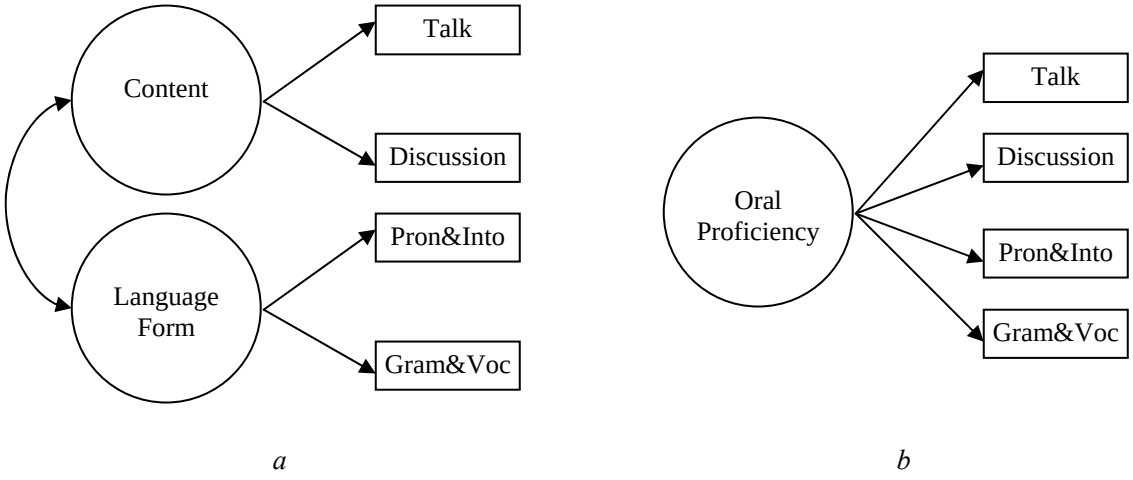

Figure 5.4. The factor structure of TEM4-Oral scores without and with halo (Note: Pron&Into = Pronunciation and Intonation; Gram&Voc = Grammar and Vocabulary)

Figure 5.4 is different from Figure 3.2 for the retelling subscale was excluded from the real analyses, but otherwise the two models remained the same. In the two-factor model (Fig. 5.4a), the subscales of talk and discussion were accounted for by the content factor while the subscales of pronunciation and intonation and grammar and vocabulary were accounted for by the language form factor, which was correlated with the content factor. In the single-factor model (Fig. 5.4b), all four subscales were accounted for by the same common factor (oral proficiency). As the single-factor model was nested in the two-factor model, a chi-square difference test was conducted to find out which model fit the subset of data better. This information was then used to infer about the halo effects.

The HLM analyses and the G-studies were performed in PASW Statistics 17.0 (SPSS Inc., 2009) using the Mixed Models and Variance Components procedures respectively, while the CFA were conducted in EQS 6.1 (Bentler & Wu, 2005).

5.2.2 Severity of rater types

Extending the ideas of Guilford (1954) and Saal et al. (1980), the main effect of the rater type can be examined to determine whether the three rater types were significantly different in severity. The general results from the HLM analysis based on Equations 5.3 and 5.4 are reported in Table 5.10.

Table 5.10

General results from the HLM analysis

Source	Numerator *df*	Denominator *df*	*F*	*p*
Rater type	3	29	.769	.521
Subscale	3	20999	181.526	.000
Rater type * Subscale	9	20999	7.943	.000

According to Table 5.10, the fixed effect of the rater type was not significant at even the $p = .05$ level, with $F(3, 29) = .769$, $p = .521$. The rater type × subscale interaction, however, was significant, with $F(9, 20999) = 7.943$, $p = .000$.

More specific information on the main and interaction effects is provided in Table 5.11, which reports the estimates of fixed effects from the HLM analysis.

Table 5.11

Estimates of fixed effects from the HLM analysis

Parameter	Estimate	*S.E.*	*df*	*t*	*p*
[Type=1]	.36	.22	31.996	1.684	.102
[Type=2]	.07	.17	31.986	.409	.685
[Type=3]	-.20	.26	31.990	-.775	.444
[Type=4]	.43	.18	31.979	2.362	.024
[Subscale=1]	-.88	.06	20999.000	-13.631	.000
[Subscale=2]	-.17	.06	20999.001	-2.608	.009
[Subscale=3]	.08	.06	20998.999	1.297	.195
[Subscale=4]	0[a]	0	.	.	.
[Type=1] * [Subscale=1]	.09	.10	20999.000	.930	.352
[Type=1] * [Subscale=2]	.12	.10	20999.000	1.156	.248
[Type=1] * [Subscale=3]	-.46	.10	20998.999	-4.543	.000
[Type=1] * [Subscale=4]	0[a]	0	.	.	.
[Type=2] * [Subscale=1]	.16	.09	20999.000	1.782	.075
[Type=2] * [Subscale=2]	.22	.09	20999.000	2.434	.015
[Type=2] * [Subscale=3]	-.17	.09	20998.999	-1.895	.058
[Type=2] * [Subscale=4]	0[a]	0	.	.	.
[Type=3] * [Subscale=1]	.43	.11	20999.000	3.797	.000
[Type=3] * [Subscale=2]	.55	.11	20999.000	4.946	.000
[Type=3] * [Subscale=3]	-.00	.11	20998.999	-.021	.984
[Type=3] * [Subscale=4]	0[a]	0	.	.	.
[Type=4] * [Subscale=1]	0[a]	0	.	.	.
[Type=4] * [Subscale=2]	0[a]	0	.	.	.
[Type=4] * [Subscale=3]	0[a]	0	.	.	.
[Type=4] * [Subscale=4]	0[a]	0	.	.	.

a. This parameter was set to zero.

b. Type: 1 = Form-oriented; 2 = Balanced; 3 = Content-oriented; 4 = Unclassified. Subscale: 1 = Talk; 2 = Discussion; 3 = Pronunciation and Intonation; 4 = Grammar and Vocabulary.

Though no significant differences were detected among the means of the three rater types, Table 5.11 shows that the form-oriented type gave the highest mean raw score (.36) and the content-oriented type the lowest mean raw score (-.20) among the three rater types. This was reverse to the order of severity estimated in the MFRM analyses.

5.2.3 Interaction effects

As Table 5.10 shows significant interaction effects, the specific differences across different combinations of rater type and subscale were examined. Table 5.11 gives the results of this

comparison from the perspective of the rater types. After accounting for the main effects of the subscales, the form-oriented type was severe on the subscale of pronunciation and intonation (Subscale 3); the balanced type was severe on the subscale of pronunciation and intonation and lenient on the subscales of talk and discussion (Subscales 1 and 2); while the content-oriented type was lenient on the subscales of talk and discussion.

While the rater type × subscale interaction terms in the HLM analyses provided further information about the severity of the rater types after accounting for their general severity level, the two-way and three-way interaction terms from the G-studies provided information for inferring the consistency of the rater types. The mean percentages of variance components yielded by the G-studies and the results from the D-studies, with the corresponding standard deviations, are reported in Table 5.12.

Table 5.12

Percentages of variance components and coefficients from G- and D-studies

Rater Type	Mean % of variance components (*SD*)							Mean D-study results (*SD*)	
	Rater	Test-taker	Subscale	Rater * Test-taker	Rater * Subscale	Test-taker * Subscale	Rater * Test-taker * Subscale	*G*	*phi*
Form-oriented	4.49 (5.67)	42.61 (12.63)	4.45 (4.80)	17.14 (8.88)	4.08 (3.61)	9.47 (5.96)	17.76 (4.25)	.77 (.10)	.72 (.11)
Balanced	7.54 (7.73)	40.31 (8.63)	3.26 (3.37)	16.82 (6.26)	3.85 (2.83)	9.16 (3.95)	19.07 (5.40)	.76 (.07)	.70 (.08)
Content-oriented	5.70 (7.99)	42.27 (12.73)	2.87 (2.75)	19.27 (9.67)	3.96 (3.09)	9.08 (5.06)	16.84 (3.23)	.75 (.11)	.70 (.14)

Here the mean D-study results were based on a fully-crossed design with two raters and four subscales (exclusive of the retelling subscale). According to Table 5.12, there was not much

difference between any two rater types in any of the statistics reported. The largest difference in the percentages was that between the form-oriented type (4.49%) and the balanced type (7.54%) in the rater facet, which was 3.05%. For all three rater types, interactions involving both the rater type and the test-taker facets were substantially large, in the range from 16.82% to 19.27%. In contrast, the smallest interaction effect was that between rater and subscale, at around 4% for each rater type. In agreement with the small differences across the rater types in the percentages of variance components, the mean G-coefficients and phi-coefficients were also similar across the rater types, with the largest difference at .02.

The standard deviations of the percentages and coefficients were generally large. For the main effects of and two-way interaction between the rater and subscale facets the measures of dispersion were comparable in size to the means. Percentages of variance components involving the test-taker facet had smaller standard deviations when compared to the means, but even the smallest coefficient of variation was as large as 19.17%, in the three-way interaction for the balanced raters. The coefficients of variation were generally smaller for the G- and phi-coefficients, which fell in the range between 8.90% (G-coefficient for the balanced raters) and 20.05% (phi-coefficient for the content-oriented raters). In sum, there was considerable variability in the percentages of variance components and the G- and phi-coefficients within each type of rater.

5.2.4 Restriction of range

In lieu of a test of rater main effect, the percentage of variance component of the test-taker facet can be used to infer about the restriction of range for each rater type (Section 2.1.2). Again,

as Table 5.12 shows, there was not much difference across rater types, the largest difference being 2.30%, between the form-oriented (42.61%) and the balanced types (40.31%).

5.2.5 Halo effect

Saal et al. (1980) suggested that the factor structure of the test can be examined to detect the existence of halo effects. It was reasoned in Section 3.6.4 that the structure of the TEM4-Oral can be represented as a two-factor model, and that if a single-factor model fits the data better, it is an indicator of halo effects.

Of the subsets of data used in the CFA, none conformed to the multivariate normal distribution, as Mardia's normalized multivariate kurtosis was estimated at 27.18 for the form-oriented subset, 33.35 for the balanced subset, and 62.89 for the content-oriented subset, well beyond the +3 to -3 range (Bentler, 2006). Therefore, the two models were fitted to the covariance matrices with the Maximum Likelihood estimator, but the Satorra-Bentler (S-B) scaled chi-square was used for the chi-square test (Satorra & Bentler, 1988, 1994), and the corresponding Satorra-Bentler scaled chi-square difference test (Satorra & Bentler, 2001) was conducted with the sbdiff.exe program (Crawford & Henry, 2003). The summary table of correlations, means, and standard deviations is given in Appendix G, and the results of the series of CFA modeling are reported in Table 5.13.

Table 5.13

Results of CFA modeling and Satorra-Bentler scaled chi-square difference tests

Model	CFI	RMSEA (90% CI)	*S-B* χ^2	*df*	*S-B* χ^2 difference	*df*	*p*
Form-oriented							
(*n* = 922)					32.107	1	.000
Single-factor	.966	.138 (.101, .178)	36.951	2			
Two-factor	.999	.035 (.000, .102)	2.096	1			
Balanced							
(*n* = 1,465)					6.227	1	.013
Single-factor	.994	.067 (.038, .100)	15.235	2			
Two-factor	.996	.076 (.038, .124)	9.530	1			
Content-oriented							
(*n* = 669)					20.950	1	.000
Single-factor	.980	.135 (.092, .183)	26.309	2			
Two-factor	.997	.071 (.015, .145)	4.393	1			

The general information from Table 5.13 is that the two-factor model fit the data significantly better than the single-factor model for all three rater types, as in each case, the Satorra-Bentler scaled chi-square difference test indicated that the two-factor model fit significantly better at the $p = .05$ level. A closer examination of the fit indices suggests that the single-factor model did not perform too much worse than the two-factor model for the balanced type, for which the difference between the two models in both the CFI and the RMSEA estimates was smaller than .01. Furthermore, when the single-factor model was fitted, the RMSEA estimates for the form-oriented and content-oriented types were greater than .10, clearly indicating model misfit, but this estimate was at a low of .067 for the balanced type, indicating adequate fit of the single-factor model. This suggests that the balanced type may be subject to the halo effect to a somewhat greater degree than the other two rater types.

The standardized factor loadings for the three subsets are reported in Table 5.14.

Table 5.14

Standardized Loadings of the two-factor model for the three subsets of data

	Form-oriented		Balanced		Content-oriented	
	content	language form	content	language form	content	language form
Talk	.731		.726		.778	
Discussion	.852		.799		.878	
Pronunciation		.829		.840		.923
Grammar		.933		.909		.944

As can be seen in Table 5.14, the standardized factor loadings ranged from .726 to .944, in the extreme high end. Together with the correlations between the two factors, these give further information about the halo effects. The inter-factor correlations were ϕ = .900 for the form-oriented subset, ϕ = .959 for the balanced subset, and ϕ = .904 for the content-oriented subset. All these were extremely high, and the value for the balanced subset was close to 1, in considerable agreement with the small differences between the single-factor model and the two-factor model in fit indices, as were reported in Table 5.13.

Combining information from Tables 5.13 and the inter-factor correlations, the form- and content-oriented types seemed more able to distinguish content-related criteria from form-related ones than the balanced type. This, however, does not necessarily mean that the balanced type was subject to the halo effect to a greater degree, as its loadings on the content factor (.726 and .799) were among the lowest, which suggests that this type of rater made a clear distinction between the talk and discussion subscales.

5.3 Summary

In answer to the third research question, this chapter examined the various patterns of rater variability associated with the three rater types through MFRM, HLM, G-Theory, and CFA. In

this context, MFRM analyses provided the most comprehensive results for comparing the three rater types, and were treated as the primary tool. Analyses based on the other statistical methods were conducted for alternative and supplementary sources of information.

According to the MFRM analyses, the form-oriented type was the most severe, the content-oriented type the most lenient, and the balanced type of moderate severity. In addition to their general severity, the form-oriented type was found to be unexpectedly severe on pronunciation and intonation, but lenient on grammar and vocabulary, whereas the content-oriented type was unexpectedly severe on grammar and vocabulary, but lenient on discussion. In terms of the rater type × test-taker and rater type × test-taker × subscale interactions, the three rater types were comparable in the percentage of standardized bias scores equal to or greater than 2, signifying similar degrees of rater reliability. Comparison in the frequency of various categories used by different rater types indicated that the form-oriented type had a stronger tendency of range restriction than the other two types, while the percentage of unexpected responses indicating central tendency was highest for the content-oriented type. All three rater types may be subject to the halo effect, as each rater type fit the restricted MFRM model with all subscale difficulties fixed at zero as well as the baseline model with all subscale difficulties free. More specifically, the severe ratings of the form-oriented type on pronunciation and intonation and the severe ratings of the content-oriented type on grammar and vocabulary were regarded as signs of the halo effect, as these ratings were deviated from the expected difficulty of the relevant subscale toward the expected severity of the rater type.

The alternative analyses, however, yielded some mixed results. In terms of the rater type severity, no rater type main effect was reported by the HLM analysis, indicating parallel severity among the rater types. Also, more significant rater type × subscale interactions were reported,

including severe ratings of the form-oriented and balanced types on pronunciation and intonation, and lenient ratings of the balanced and content-oriented types on talk and discussion. Here, only the severe rating on pronunciation and intonation on the part of the form-oriented type was the same as the MFRM result. Rater × test-taker and rater × test-taker × subscale interactions, however, were also reported to be comparable across rater types in the G-studies. Furthermore, G-coefficients and phi-coefficients from the D-studies were also similar across the rater types, leading to the same conclusion on rater reliability as that from the MFRM analyses. The percentage of variance component of the test-taker facet was estimated to be similar across rater types in the G-studies, indicating a similar level of restriction of range across rater types. In terms of the halo effect, the comparison of the CFA models generally supported the two-factor model over the single-factor model, indicating that none of the three rater types was severely subject to the halo effect. However, fit indices and factor loadings seemed to suggest that the form- and content-oriented types were more able to distinguish content-related criteria from form-related ones than the balanced type.

The discrepancies in results between the alternative analyses and the meanings of these results in relation to the characteristics of the three rater types will be discussed in the next chapter.

Chapter 6 Discussion

This chapter discusses the methodological issues related to the collection and analysis of data, as well as interpretation of the findings reported in Chapters 4 and 5. For each research question, the focuses will be on the justifications of using a certain method to collect and analyze data, and the consistency of the corresponding findings. The findings will be summarized again in relation to the research question, and interpretations and implications of the findings will be discussed. After a summary of the limitations of this study and the directions for future studies, a brief conclusion will end this chapter,

6.1 Research question 1: How successfully can raters be classified into types according to their weighting patterns?

6.1.1 Methodological considerations

As the first step in answer to this research question, the weighting patterns were derived from regression weights calculated from the holistic ratings given by 126 raters to 120 simulated profiles consisting of scores on five subscales: topic relevance, richness of content, organization, grammar and vocabulary, and pronunciation and intonation. The inclusion of these five subscales was based entirely on the rating rubric used for the talk subscale of the TEM4-Oral, as the task was rated on exactly these five criteria in real practice.

The simulated score profiles were presented to the raters in graphic forms instead of numeric figures. Two of the advantages of the graphic scale are its clear meaning and consistent interpretation (Aguinis, 2007). As Anderson (1982) suggested, the meaning of a graphic scale is

clear even for a child, for it corresponds to “an internalized length scale that develops from reaching movements and other activity in the child’s local space” (p. 7). On the other hand, the advantage of consistent interpretation is relative to numerical rating scales. It has been demonstrated that the meaning of scales may change with different numeric labels (Amoo & Friedman, 2001; Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991). Another advantage of the graphic presentation over the numeric version, specific to this study, is that raters may be tempted to calculate a summary score analytically from the given values instead of forming a general impression. Such calculation should be avoided as it differs considerably from the real rating process of TEM4-Oral.

Another issue related to the use of the value judgment task as an instrument for deriving relative weights from the raters is its indirect nature. As an alternative, the raters could have been asked to directly state the relative weights they gave to the five criteria. According to the empirical report on the most recent National Research Council (NRC) assessment of doctoral programs (Ostriker, Holland, Kuh, & Voytuk, 2011), the two methods may yield different but correlated results. In the NRC study, the final weights based on direct surveys were justifiable as they were calculated as the mean weights taken across all participants. In this study, however, asking the raters to directly give the relative weights might have suffered the following problems. First, individual weightings would have been subject to inconsistency across occasions, and consistent weights would have been contingent on repeated weightings, which was impractical within the limited time. Second, there is always discrepancy between self-proclaimed weights and the actual weights functioning in the rating process. In contrast, giving summary ratings on score profiles is more similar to the actual rating process, and the need for consistency may be addressed by presenting the raters with a large number of score profiles,

which proved practical in this study. For these reasons, relative weights based on regression were preferred in this study, and self-perceived weights were elicited in answer to the second research question to provide more meaningful interpretation of the classification results, which will be discussed in greater detail below.

The R^2 values from the regression analyses, 120 (or 95%) of which were estimated to be higher than .65, provided the empirical support for the consistency of the relative weights based on the regression analyses.

The second step in addressing the first research question was the classification of the raters through cluster analysis. Here the two cohorts were regarded as two independent samples, and the similar results of the hierarchical clustering on both samples provided the primary evidence for the stability of the three-cluster solution.

Mathematically, there could be a myriad of clustering patterns, and the selected solution would only be meaningful if a close correspondence could be established with substantive theories or facts. The meaningfulness of the three-cluster solution in this study was supported by its correspondence with the tripartite categorization of raters into the form-oriented, balanced, and content-oriented types, which is not only theoretically meaningful, but also concordant with the findings from the qualitative study conducted by Erdosy (2004).

Just as there are a variety of possible clustering patterns, so there are numerous ways to classify the raters of an EFL speaking test, many of which involve higher degrees of specificity than the tripartite scheme. To understand the complex variation among raters, a detailed account, or "thick" description typical of ethnography (Geetz, 1973; Ryle, 1968), is needed. The first purpose of this study, however, was to classify the raters in terms of their weighting patterns on a

macroscopic basis, and a tripartite classification based on the clustering results is a sufficient fulfillment of this aim.

6.1.2 Interpretation of the results

Two criteria were applied to the judgment of how successfully the raters were classified according to their weighting patterns: 1) the degree to which the different types of raters could be distinguished from each other according to their weighting patterns, and 2) the meaningfulness of this distinction.

Statistical information related to the first criterion included the mean and standard deviation of relative weights across the rater types and the associated comparisons by way of MANOVA, reported in Tables 4.10 and 4.11. In brief, significant differences were found in the mean weights on all five criteria across rater types. Most saliently, the form-oriented type exhibited overemphasis on pronunciation and intonation, with a mean weight of .39, whereas the content-oriented type gave an extremely high mean weight of .52 to topic relevance. These differences indicated a clear distinction between the rater types.

The meaning of this distinction arises from a more detailed examination of the weighting patterns associated with the rater types. In this connection, three general features may be observed. First, the three types did not differ in all criteria to the same degree. The largest range of mean weights was observed in the criterion of topic relevance, where the difference between the maximum weight (.52) and the minimum weight (.15) was .37. The other large range of mean weights was found for the criterion of pronunciation and intonation, where the difference between the maximum weight (.39) and the minimum weight (.10) was .29. For each of the other three criteria, however, the range of mean weights was either .09 or .08. The second observable

feature was that the balanced rater type did not give equal or near equal weights to all criteria. For this type of rater, the maximum mean weight was .31, and the minimum mean weight was .12, covering a range of .19. This deviated considerably from the expectation for a balanced distribution of weights, for which each criterion should have a mean weight of .20. As almost two thirds of the raters were classified as balanced, this deviation from the expectation should be taken as a reflection of the general implicit belief among these raters, and the balanced type should be described according to these empirical values. As the third observable feature, there was large variation across rater types in the range of mean weights within rater types. This range was .19 for the balanced type, .32 for the form-oriented type, and .45 for the content-oriented type. As the range was much smaller for the balanced type than for the other two types, it was still meaningful to call this type balanced, but in a relative rather than an absolute sense.

According to these three distributional features in the weighting patterns, the balanced type was characterized as a group of raters who attached greater importance to topic relevance and richness of content, but in doing so, remained moderately sensitive to organization, grammar and vocabulary, and pronunciation and intonation. Taking the balanced type as a reference, the content-oriented type is distinguished by its overemphasis on topic relevance, while the form-oriented type by its overemphasis on pronunciation and intonation. The overemphasis on any criterion is accompanied by a certain degree of downplaying the other criteria, resulting in enlarged unbalance in the mean weights within rater types.

In this interpretation, the rater classification according to the weighting patterns was considered successful. In an absolute sense, this is different from the theorization in Chapter 2, especially in the weighting pattern of the balanced type, but in contrast to the other two types, the balanced type could still be called balanced.

6.2 Research question 2: How are types of raters different in the rating process?

6.2.1 Methodological considerations

To describe the differences across rater types in the rating process, verbal protocols were elicited from a selection of raters. Somewhat against the expectation to include raters of diverse weighting patterns in the verbal protocol study, the resultant constitution of the rater sample was unbalanced, with more form-oriented raters (n = 10) and fewer content-oriented raters (n = 2) than wished. Generally speaking, the verbal protocols were of a qualitative nature, but as statistical analyses were conducted on the relevant frequencies and percentages, this lack of balance and the small number of raters of the content-oriented type posed a major threat to the generalizability of the quantitative results. Therefore, the quantitative comparisons of the rater types should be perceived with caution.

Another general note of caution concerns the limitation of the inter-type comparison to four aspects: coverage of themes, self-perceived weights, relationship between criteria, and response to digression and salient negative features in pronunciation and intonation. Most clearly, differences in these aspects were an underrepresentation of potential differences across rater types in the rating process. According to Wolfe's (1997) model of scorer cognition (Figure 2.4), for example, these aspects all fell in the category of content focus, while the processing actions were totally missing. Even in terms of content foci, the list could have been much longer and more fine-grained, as Cumming et al. (2002) showed (Table 2.3). The justifications for including these four aspects in the comparison were twofold: 1) the focus of study was on weighting patterns instead of the more general process of rating; and 2) the raters were classified according to their weighting patterns. In practice, the first three aspects served as the triangulated

validation of the different weighting patterns associated with the rater types, while the last aspect provided the link between weighting patterns and the raters' responses to two special features in test-taker performance: digression and salient negative features in pronunciation and intonation. The choice of these features in the comparison was closely related to the overemphasis of the content-oriented type on the former, and the form-oriented type on the latter.

Another area of underrepresentation was the inclusion of only the second task (talk based on a given topic) of the TEM4-Oral in the verbal protocol studies. As was explained in Chapter 3, the inclusion of only one task was due mainly to the lack of time for the verbal protocols, and the preference of the second task was that oral composition has the most typical form of a speaking test, that performance on such a task is most comparable to essay writing, and that the three criteria used for this task (topic relevance, richness in content, and organization) are often shared with scoring rubrics in writing assessment. In contrast, the first task (retelling) is confounded with listening comprehension, note-taking, and memory, and the checklist method for scoring is qualitatively different from the general practice of rating. Similarly, performance of a test-taker in the discussion is often confounded with that of the interlocutor, and identifying one test-taker from the other poses extra burden on the rater. The findings of the preliminary MFRM analysis, reported in Section 5.1.1, indicated that the retelling subscale reflected something different from the construct underlying the other four subscales. Moreover, the findings reported in Section 5.1.6 suggested that the talk subscale was more succinctly distinguished from the subscales of pronunciation and intonation and grammar and vocabulary, while the discussion subscale clustered closely with the two form-related subscales. These were both evidence in support of the use of the second task in the verbal protocols, as the tripartite criteria (content, pronunciation and intonation, and grammar and vocabulary) used in the verbal protocols made the most sense

when content was succinctly distinguished from the form-related criteria.

The number of participants varied across verbal protocols. Twelve of the raters took part individually, six in pairs, and three worked in a group. If pair and group work may introduce some interaction between participants or even peer pressure from each other, the analyses reported in Chapter 4 suggested that such effects were minor, as the raters of different types exhibited distinct features in paired or grouped verbal protocols. The beliefs of Lou and Josie, the former a form-oriented rater and the latter a balanced one, for example, contrasted starkly even though the two were working as a pair in the verbal protocol (Section 4.3.3).

With these considerations in the general design, the way the data were collected and analyzed for the four aspects will now be justified.

The comparison of the rater types in the frequencies and percentages of certain themes covered in the verbal protocol followed common practices in similar studies (Cumming, 1990; Cumming et al., 2002). The choice of the three specific themes—content, pronunciation and intonation, and grammar and vocabulary—was based on the actual rating scheme followed by the TEM4-Oral raters. This is certainly a limited reflection of the themes covered by the raters, but this helped focus the characterization of the three types of raters on the weighting patterns under consideration.

It was mentioned above in Section 6.1.1 that there is always discrepancy between self-perceived weights and the actual weights functioning in the rating process. To avoid this problem, self-perceived weights were elicited immediately after each verbal protocol to correspond better to the real rating process. A second means to assure this correspondence was that the verbal protocol was concurrent with the ongoing rating process, so that the performance

of the raters in the verbal protocol could be regarded as a natural extension of the real rating process.

Beliefs in the relationship between criteria, however, were inferences from the raters' justifications rather than their explicit proclamations. As this was not planned beforehand, no questions on these were asked of the raters in the verbal protocols. As a result, data were not available from every rater, and no frequencies or percentages were reported.

As was explained above, the choice of digression and salient negative features in pronunciation and intonation in the comparison was closely related to the overemphasis of the unbalanced rater types. The choice of negative features rather than positive ones as the focus of comparison was based on earlier findings concerning TEM4-Oral raters. According to verbal protocols conducted by Wang (2008), the majority of these raters adopted the subtractive scoring approach on the criteria of pronunciation and intonation and grammar (but not vocabulary), i.e. when they detected negative features they tended to subtract certain points from the anchor score formed through first impression. In other words, the majority of raters listened for negative features rather than positive ones in order to rate the test-takers on pronunciation and intonation and grammar. Whether they were able to deliver themselves from this inclination to give due attention to other features of the test-taker's performance, especially content-related features, thus becomes a touchstone for detecting overemphasis on pronunciation and intonation. By the same token, the negative content-related feature of digression was used as a touchstone for detecting overemphasis on topic relevance.

6.2.2 Interpretation of the results

As a summary of the findings related to the second research question was given at the end of Chapter 4, the interpretation of the results will be discussed in a holistic rather than aspect-specific manner in order to achieve a coherent understanding of the differences between the three types of raters in the rating process. Two themes will be covered in the following discussion: 1) the anchoring and masking effects of pronunciation and intonation; and 2) mitigation of the anchoring and masking effects of pronunciation and intonation. These themes provide a unified framework for interpreting the differences between the different types of raters in the rating process and the relationship between the weighting patterns and the actual ratings given by the raters.

6.2.2.1 Anchoring and masking effects

The most prevalent phenomenon among the three types of raters was the anchoring effect (Tversky & Kahneman, 1974), or the influence of the first impression based on pronunciation and intonation in this context. As was reported in Section 4.3.3, the raters based their judgment of the test-taker' overall performance on their initial judgment of pronunciation and intonation, and then made adjustments on the various subscales after hearing more of the recordings. Section 4.3.3 gave examples of form-oriented and balanced raters, but this practice was also common among content-oriented raters. Mia, a content-oriented rater, for example, clearly stated that she was not able to make a judgment on the content of Clip 2 at the beginning of the verbal protocol. Her way out was to give the same score to content as she gave to pronunciation and intonation, which remained unchanged through the whole process until the end of the recording, when she decided that the test-taker did not talk much about what he had learned from his experience, for

which reason she lowered the content score by one category. In itself, the case of Mia might be just an idiosyncratic one, but an *ad hoc* analysis of the scores given by the 21 raters during the verbal protocols sheds further light on the generality of the anchoring effect. As each rater gave a content score to each of the five segments of each of the four clips, the number of scores on content given by the 21 raters was 420, or 21 times 20. Similarly, there were 420 pronunciation and intonation scores. A comparison of the pronunciation and intonation scores with the corresponding content scores shows that 69 of the 420 differences were greater than one category. Of these 69 cases, only six were associated with the first segment, which is clear evidence of extensive anchoring effect.

A second note on Mia's case was that the final adjustment she made was only one category, on a 21-category scale. Again, this miniscule adjustment was prevalent among the raters participating in the verbal protocols. It was noted above that only 69, or 16%, of the 420 differences were greater than one category. In particular, only 18 of these 69 cases were associated with the final segment. What this implies is the lasting effect of their first impression, as the difference between pronunciation and intonation and content scores was enlarged only slightly from the beginning to the end of the verbal protocols.

When the anchoring effect of pronunciation and intonation persisted to the end, this became a masking effect, to borrow the term from psychology and physiology (Gelfand, 2010). Under the masking effect, some raters failed to give due attention to features other than pronunciation and intonation in the test-taker's performance, typically features of the content. This is clear from the frequencies reported in Table 4.15, which show that half of the form-oriented raters made no or only vague comments on the content of Clip 2, as the clip was marked by salient negative features in pronunciation and intonation. In contrast, all of the balanced and content-

oriented raters made either general or specific comments on content, despite the negative features. A similar case was Clip 5, whose pronunciation and intonation was unanimously deemed much better than Clip 2. For this clip, the majority of form-oriented raters failed to detect digression, while most balanced and both content-oriented raters clearly stated the digression (Table 4.16).

Taken together, the influences of pronunciation and intonation described above were twin to each other: 1) the first impression took shape at the beginning of the rating process (the anchoring effect), and 2) the first impression became the masking effect when it persisted to the end of the rating process.

The anchoring and masking effects of pronunciation and intonation were undesirable as they tended to bias the judgment of the rater toward a single criterion. A well-trained rater would be expected to be free from such effects and base his ratings on all relevant features of the test-taker. For this reason, the ability to mitigate these effects becomes an essential criterion for distinguishing the different types of raters.

6.2.2.2 Mitigation

For these raters, the mitigation of the anchoring and masking effects took several forms. The weakest form of mitigation was the distinction between different criteria. An example of this was provided in Section 4.3.3, where Hermione, a balanced rater, made a clear distinction between grammar and vocabulary and content in her justification. For another example, Mona, another balanced rater, made the following comment on the second segment of Clip 2: “I think he may have, because of his local accent, some problems in pronunciation and intonation, but he made his points, and I can raise his score in content, I think” (Appendix E). Obviously, what she

did was to make a clear distinction between pronunciation and intonation and content, which resulted in a higher rating for the content. In contrast, the form-oriented raters seemed to be more inclined toward confusing content with form. As was reported in Section 4.3.3, Charlene, a form-oriented rater, decided to mark down the content by one category on the grounds of redundancy, a content-related issue, and lack of variation in the sentence patterns, a form-related issue. Similarly, Lou, another form-oriented rater, penalized the test-taker in content for a low speech rate, on the basis that "her lack of fluency has obstructed the conveyance of message" (Section 4.3.3). For the content-oriented raters, the degree of distinction between form- and content-related criteria was not documented in the justification they made, but can be inferred from their response to salient negative features in pronunciation and intonation, in which case they were still able to direct their attention to content and made general or detailed comments on it (Section 4.3.4).

For the raters participating in the verbal protocols, the different degrees of distinction between content and form resulted in different degrees of variation in the mean scores on the three subscales used. For the three clips reported in Table 4.14, the form-oriented raters gave almost the same mean scores on the three subscales, while the mean scores for content varied moderately from the mean scores on the two form-related subscales for the balanced and content-oriented raters. The largest degree of variation was associated with the content-oriented raters. This is another indication that the balanced and content-oriented raters were able to make a clearer distinction between content and form than the form-oriented raters.

A stronger form of mitigation was the assessment of form on content-related reasons. A case in point is the belief of Mona that "his pronunciation was okay if only he expressed his ideas clearly" (Section 4.3.3). Similarly, Josie explained the inadequacy of Clip 1 in

pronunciation and intonation on the grounds of inadequacy in content (Section 4.3.3). This form of mitigation was not discovered in the verbal protocols involving the form-oriented raters.

Overemphasis on topic relevance may be regarded as the strongest form of mitigation, which was associated mostly with the content-oriented raters. A comparison of the mean content scores given to Clips 2 and 5 in the verbal protocols exemplifies the effect of this mitigation. As Table 4.14 shows, the form-oriented raters gave a higher mean content score to Clip 5 than to Clip 2 (65.50 > 60.00), and so did the balanced raters (62.22 > 59.44). In contrast, the content-oriented raters gave the same mean content score to both clips (57.50). It was demonstrated in Section 4.3.4 that there was no reason that Clip 2 should be inferior to Clip 5 when content was considered independently, and that a lower content score for Clip 2 than for Clip 5 was a reflection of the effects of negative features in pronunciation and intonation. In the same vein, the same mean content scores for both clips given by the content-oriented raters was a reflection of the mitigation, and emphasis on topic relevance was at work here.

The most general pattern of the different strengths of mitigation was the different percentages of comments on the three criteria made by the three types of raters in the verbal protocols, reported in Table 4.12. The form-oriented raters had a significantly higher mean than the other two types of raters in the percentage of comments on pronunciation and intonation, but a lower mean in the percentage of comments on content. The balanced raters had a higher mean than the content-oriented raters in the percentage of comments on pronunciation and intonation, but a lower but not statistically significant mean in the percentage of comments on content. This suggests that the form-oriented raters were most subject to the anchoring and masking effects of pronunciation and intonation, whereas the content-oriented raters displayed the strongest

mitigation of such effects, though the balanced raters were somewhat more similar to the content-oriented raters than to the form-oriented raters.

This result was in agreement with the self-perceived weights of the raters reported in Section 4.3.2: the form-oriented raters all gave top priority to pronunciation and intonation, most of the balanced raters tended to place relevance and richness on a higher level of importance than pronunciation and intonation, and the content-oriented raters were more focused on content. In terms of the opposition between subjection to and mitigation of the anchoring and masking effects of pronunciation, this again suggests most severe subjection on the part of the form-oriented raters and the strongest mitigation on the part of the content-oriented raters.

6.2.2.3 The subjection-mitigation continuum

In Figure 4.7, the three types of raters were placed on a continuum of relative weights with the form-oriented raters closer to the form end, the content-oriented raters closer to the content end, and the balanced raters in the middle. As the opposition between subjection and mitigation of the anchoring and masking effects was analogous to the opposition between the form and content ends, a figure isomorphic to Figure 4.7 could be used to represent the subjection-mitigation continuum (Figure 6.1).

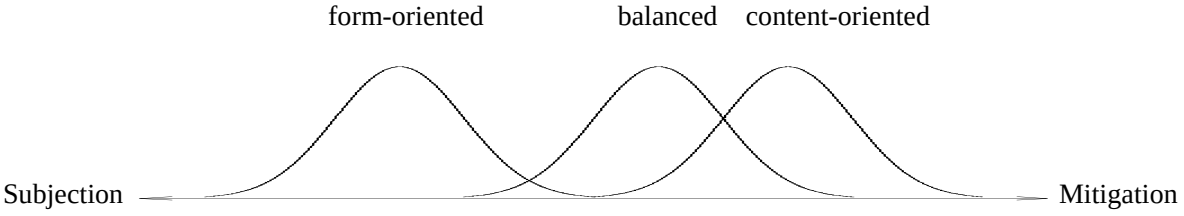

Figure 6.1. Three types of raters on the subjection-mitigation continuum

Figure 6.1 defines a continuum with subjection to the anchoring and masking effects at one end, and mitigation at the other. The three types of raters are placed on their positions according to findings from Chapter 4: the form-oriented raters closer to the subjection end, the content-oriented raters closer to the mitigation end, and the balanced raters somewhere in between, but closer to the content-oriented raters than to the form-oriented raters according to the above discussion.

The subjection-mitigation continuum graphically represented in Figure 6.1 provides a unified interpretation of the findings relevant to the second research question, summarized in Figure 6.2.

Form-oriented	**Balanced**	**Content-oriented**
First-priority criterion		
Pronunciation and intonation	Pronunciation and intonation or content	Content
Percentage of comments on pronunciation and intonation		
Largest among the three types of raters	Medium among the three types of raters	Smallest among the three types of raters
Percentage of comments on content		
Smallest among the three types of raters	Medium among the three types of raters	Largest among the three types of raters
Distinction between content-related criteria and form-related ones		
Confusion: assessment of content on form-related reasons	Clear distinction, with some degree of form assessment on content-related reasons	Clear distinction
Difference between mean score for content and mean scores on form-related criteria in the case of digression and negative features in pronunciation and intonation		
Negligible	Small	Moderate
Tendency to detect digression		
Small	Large	Large

Subjection ⟷ **Mitigation**

Figure 6.2. A detailed view of the subjection-mitigation continuum

In essence, the differences between the three types of raters in the rating process reflected the different degrees of subjection to the anchoring and masking effects of pronunciation and intonation and the different strength of mitigation of such effects. The form-oriented raters were subject to the anchoring and masking effects to the largest extent with their first priority on pronunciation and intonation, largest percentage of comments on pronunciation and intonation and smallest percentage of comments on content in the verbal protocols. They did not make a clear distinction between content-related criteria and form-related ones, but frequently based their assessment of content on form-related reasons. Their attention to content was weakest when salient negative features were present, and they tended to neglect digression due to their overemphasis on pronunciation and intonation. The content-oriented raters were the opposite, exhibiting the greatest degree of mitigation of the anchoring and masking effects. They gave first priority to content-related criteria in the rating process, making the largest percentage of comments on content and smallest percentage of comments on pronunciation and intonation. They tended to make the clearest distinction between content-related and form-related criteria, even when negative features were salient in pronunciation and intonation. Their degree of mitigation may even take the form of overemphasis on topic relevance, which made them sensitive to digression. The balanced raters were situated somewhere between the form- and content-oriented raters in most aspects. However, they were more similar to the content-oriented raters in terms of the percentage of comments on content, the distinction between form- and content-related criteria, and the tendency to detect digression. In emphasis, the balanced raters also attached greater importance to content than to language form, i.e., they were relatively rather than absolutely balanced, as was discussed in Section 6.1.2.

6.3 Research question 3: To what degree are patterns of rater variability different across types of raters?

6.3.1 Methodological considerations

As different statistical procedures were applied in address to the third research question, the key considerations in methodology were the explanation of the discrepancies between the results from different analyses, and the decision on which results to be interpreted in the following section. The discrepancies between the results fall into two broad categories: those due to the data structure, and those due to the different operational definition of a certain type of rater variability.

The dataset used in this study, whose structure was detailed in Section 5.1.1, was of an unbalanced nature, with a lot of missing values. For such a dataset, linking becomes an essential issue. In an MFRM model, linking is an integrated process (Linacre, 1989), provided that a sufficient number of anchor groups were included in the dataset. The HLM and G-theory models, however, were based on the assumption of random sampling, for which linking is an external process that needs to be modeled separately. From this perspective, the MFRM procedures should be regarded as the basic, and the HLM and G-theory analyses as alternatives. Discrepancies between the various analyses, therefore, should be resolved in favor of the MFRM results.

Theoretically, the MFRM models could be associated to the HLM and G-theory models, as all could be shown to be special cases of hierarchical generalized linear models (HGLM, Muckle & Karabatsos, 2009), and empirical findings generally supported the comparability between the MFRM and G-theory models in detecting rater effects (Bachman et al., 1995; MacMillan, 2000;

Sudweeks, Reeve, & Bradshaw, 2005). However, all these studies aimed at detecting variability among single raters, which is a different case from the focus on rater types in the present study.

The most notable discrepancy between the MFRM and HLM results lay in the ranking of the rater types in overall severity. The MFRM ranking, from severe to lenient, was form-oriented, balanced, and content-oriented, whereas the HLM analyses found no significant main effect for the rater type factor. As the MFRM and HLM were special cases of the HGLM, linking could be regarded as the principal reason underlying this difference. Therefore, the MFRM result will be treated as final as regards rater type severity in the following sections.

The second major discrepancy in the results was that between the MFRM and CFA results pertaining to the halo effect, which was due mainly to the different operational definition of the phenomenon. Myford and Wolfe (2004), for example, suggested that a large percentage of "identical ratings across traits" should be an indication of the halo effect, a method adopted in Engelhard (1994) and Eckes (2005). Alternatively, Myford and Wolfe (2004) cited a personal communication with Linacre and suggested anchoring all subscales at the same difficulty (usually 0) before fitting the MFRM model and comparing the fit indices of the subscale from this model with those from the unrestricted MFRM model, which was tried out in this study (Section 5.1.6). The logic of this alternative method is the detection of identical calibrations across traits, which is a latent estimate instead of observed ratings.

Obviously, this operational definition of the halo effect in MFRM is different from that proposed by CTT proponents. Saal et al. (1980), for example, suggested that subscale intercorrelation or factor analysis be used to detect the halo effect (Section 2.1.2). As subscales could be intercorrelated with each other and yet differ in difficulty, this method reflects a totally different conception of the halo effect. With reference to Figure 2.2, the MFRM and the CTT

procedures may be regarded as two sides of the same coin, and an eclectic treatment may prove more informative.

Engelhard (1994) warned against the failure of the CTT methods to distinguish between a true halo and an illusory halo by drawing on Murphy and Cleveland (1991). In brief, an illusory halo is the subscale intercorrelation intended by the test developers while a true halo is the unintended subscale intercorrelation. To avoid mistaking illusory halos for true ones, this study adopted a model comparison method by fitting both the intended two-factor model and the unintended single-factor model to the data. The results from the model comparisons and the results from the MFRM will be discussed as complementary in the following section.

Apart from the discrepancies in the results from the different statistical procedures, the exclusion of the retelling subscale from the analyses is another issue that deserves special attention. Both statistical and substantive reasons were given in Section 5.1.1, but these did not change the fact that important information may have been lost due to the exclusion of a part of the test from the analyses. As the MFRM model with the retelling subscale seriously violated the assumption of unidimensionality, the exclusion was considered a necessary evil, and the only hope may lie in the development of multidimensional MFRM models in the future, so that the dimension problem could be resolved.

Sampling is a more general problem in this study. Due to the administrative restrictions of the test developers, only a small subset of data was obtained. Furthermore, the selection of this subset was governed by the requirement to obtain sufficient interconnections among the test-taker groups, so that linking could be achieved in the MFRM analyses. Therefore, the dataset used in the study could not be regarded as a random sample from the whole database.

6.3.2 Interpretation of the results

With the above limitations in mind, the results from Chapter 5 will be discussed in the following sections. The results pertaining to each of the four types of rater variability will be reviewed briefly, followed by a discussion on the association between the results and the weighting patterns.

6.3.2.1 Severity

In terms of overall severity, the form-oriented raters were found to be most severe while the content-oriented raters to be most lenient. Furthermore, the form-oriented raters were also unexpectedly severe on the pronunciation and intonation subscale. In pairwise comparisons on this subscale, the form-oriented raters were found to be significantly more severe than the other two types of raters (Table 5.4). As reported in Chapter 4, the form-oriented raters attached the greatest importance to pronunciation and intonation. In the rating process, this was exhibited as the subjection to the anchoring and masking effects of pronunciation and intonation. According to earlier research on the TEM4-Oral raters, the majority of them tended to listen for negative rather than positive features in pronunciation and intonation and to reduce scores upon noticing negative features (Wang, 2008). It follows naturally that overemphasis on pronunciation and intonation leads to reduced ratings on this subscale. In addition, subjection to the anchoring and masking effects of pronunciation and intonation led the form-oriented raters to base the assessment of content on form-related reasons (Chapter 4). In turn, severity on the pronunciation and intonation subscale was generalized to overall severity, resulting in the highest severity of the form-oriented raters. By the same token, the content-oriented raters exhibited the strongest

degree of mitigation of the anchoring and masking effects, which contributed to their lowest overall severity.

Apart from their overall leniency, the content-oriented raters were excessively lenient on the discussion subscale. On this subscale, neither the form-oriented nor the balanced raters were found to be unexpectedly severe or lenient. In pairwise comparisons, the content-oriented raters were found to be significantly more lenient than the other two types of raters on this subscale (Table 5.4), which seems to suggest their strong ability to mitigate the anchoring and masking effects of pronunciation and intonation and to increase attention to the positive features in content. In a preceding study on the TEM4-Oral raters, Wang (2008) reported that these raters looked for negative features in topic relevance and organization, but positive features in richness, topic novelty and vividness. The tendency of the content-oriented raters and most of the balanced raters to notice topic irrelevance was confirmed in the verbal protocols reported in Chapter 4, while Table 4.14 suggests that the content-oriented raters had a stronger tendency to credit the merits in content of a test-taker who exhibited salient negative features in pronunciation and intonation. Wang (2008) also suggested that the negative features in topic relevance and organization were salient because they were uncommon, which may also have increased the probability of the content-oriented raters to notice the positive features in content when they made their judgment on the content-related subscale of discussion.

On the grammar and vocabulary subscale, the form-oriented raters were unexpectedly lenient, whereas the content-oriented raters were unexpectedly severe. Pairwise comparisons show that the form-oriented raters were most lenient on this subscale while the content-oriented raters were most severe and that all pairwise differences were statistically significant (Table 5.4). The stark contrast between form- and content-oriented raters seems to suggest the essential

importance of language form as a criterion in the rating process. Various criteria may be used in different rating rubrics, but criteria related to language form are never missed, even in "uncontrolled grading" when the raters were not given particular criteria (Diederich et al., 1961). However, the different severities on the grammar and vocabulary subscale, when combined with the different severities on the pronunciation and intonation subscale, seem to suggest that the different types of raters saw different subscales as the primary scale for language form. As the form-oriented raters overemphasized pronunciation and intonation, there is little doubt that pronunciation and intonation was their primary consideration related to language form. For the other two types of raters, however, grammar and vocabulary may be the most appropriate reflection of language form. In other words, pronunciation and intonation was regarded as language form itself by the form-oriented raters, but may be taken as a secondary criterion comparable to mechanics such as spelling and handwriting by the content-oriented and balanced raters. The form-oriented raters' treatment of pronunciation and intonation as the primary reflection of language form is clear from the verbal protocols analyzed in Chapter 4, especially in the findings that all of these raters regarded pronunciation and intonation as the primary concern. The downplaying of pronunciation and intonation by the content-oriented and balanced raters is also clear from the verbal protocols. Mona, a balanced rater, for example, regarded the negative features in pronunciation in Clip 2 as "local accent" (Appendix E). Coincidentally, in general discussion following the verbal protocols, two other balanced raters mentioned native-tinted accents in some Indian speakers of English as salient features that should be disregarded when assessing their English speaking ability

Thus, the differences across the three types of raters in their severities on the two form-related subscales may be a result of their choice of either pronunciation and intonation or

grammar and vocabulary as the primary scale for language form. As the form-oriented raters deemed pronunciation and intonation as the primary form-related scale, they were sensitive to negative features in this domain, resulting in the highest severity in pronunciation and intonation. Grammar and vocabulary, in comparison, was less essential and negative features in this domain had a weaker effect, hence the leniency of the form-oriented raters on this subscale. The content-oriented raters, in contrast, regarded grammar and vocabulary as the primary scale for language form, and became most severe on this subscale.

On the talk subscale, no significant differences in severity were found across the three types of raters. Most probably this resulted from the nature of the rating rubric for this subscale. The following is a translation from the original Chinese rubric for this subscale.

Overall requirements:

1. Must be an account of a specific event with a plot instead of a general discussion;
2. Must be the test-taker's own experience, related in the first person;
3. Must be an unexpected experience instead of an expected (as-planned) event;
4. Must cover experience/lessons/benefits/morals, etc. gained from the experience;
5. Must be an organized account of a complete event, with obvious gains.

Special cases:

1. Suspected recitation of prewritten content: rate as regular and annotate beside the rating;
2. Repetition of the story in Task I: rate in the 0-10 range;

3. Brief talk less than one minute: rate in the 0-40 range and annotate “insufficient content”;
4. Two or more events, either of which oversimplified or incomplete: rate in the 0-70 range;
5. General comments without concrete plots: rate in the 0-60 range;
6. Relating the experience of people other than the test-taker: rate in the 0-60 range;
7. Lessons from the experience missing: reduce 10 points.

The above rubric has two obvious features: 1) it is meticulously detailed; and 2) negative features are emphasized. In effect, the raters’ attention was restricted to a checklist of negative features in the test-taker’s talk. This, in turn, reduced the differences between the three types of raters and resulted in more comparable severities among them.

All in all, the highest overall severity and the unexpectedly high severity of the form-oriented raters on the pronunciation and intonation subscale can be attributed to the excessive sensitivity to negative features in pronunciation and intonation, or severe subjection to the anchoring and masking effects of pronunciation and intonation, which is directly related to the weighting patterns of the raters. For the content-oriented raters, downplaying the importance of pronunciation and intonation and increased attention to the positive features in content may have resulted in the lowest overall severity and the unexpectedly low severity on the discussion subscale. The leniency of the form-oriented raters and the severity of the content-oriented raters on the grammar and vocabulary subscale were also related to the different levels of importance attached to pronunciation and intonation. The focus of their attention on the negative features in pronunciation and intonation led the form-oriented raters to be less stringent on grammar and

vocabulary, while the tolerance of negative features in pronunciation and intonation may have shifted the attention of the content-oriented raters to the negative features in grammar and vocabulary in their effort to rate the formal features of the test-takers' performance. The balanced raters, in contrast, seemed to be relatively free from overemphasis on any of the criteria. Consequently, they ranked between the other two types of raters in overall severity, and were found to be unexpectedly severe or lenient on none of the subscales.

6.3.2.2 Reliability

Results from Chapter 5 suggested comparable overall reliabilities for the three types of raters. According to the MFRM results in Section 5.1.4, the three types of raters were comparable in the percentage of standardized bias scores equal to or greater than 2, in both the rater type × test-taker and rater type × test-taker × subscale interactions. The results of G- and D-studies reported in Section 5.2.3 confirmed this comparability, as rater × test-taker and rater × test-taker × subscale interactions had similar percentages of variance components across the three types of raters, and G- and phi-coefficients also had similar values. Furthermore, there was considerable variability within each type of rater in the percentages of variance components and the coefficients, which made it more difficult to differentiate the three rater types in reliability.

These results suggest that it is hard to predict differences in the overall reliability from the weighting patterns of the raters. According to general principles of test theory, standardization of the measurement procedure is one way to increase reliability (Kane, 1982; Lord & Novick, 1968). In the literature of writing assessment, proponents of holistic scoring have also argued for "instantaneous judgment" to avoid the influence of "tangential or irrelevant qualities" in the judgment and increase inter-rater reliability (Charney, 1984; McColly, 1970; Wolfe & Ranney,

1996). It seems that higher inter-rater reliability may be expected of raters who limit their attention to a single criterion rather then divide it among several criteria. This, however, was not the case with the TEM4-Oral raters. As Table 4.10 shows, the three types of raters may display different weighting patterns, but they still attached some importance to each and every criterion. The mean percentages of comments reported in Table 4.12 also show that attention of all three types of raters was divided among the different criteria. Therefore, no difference can be predicted in the overall reliability across the three types of raters.

6.3.2.3 Restriction of range

Like reliability, differences in the degree of range restriction are also hard to predict. Theoretically, the more criteria the raters pay attention to in the rating process, the greater variation in the ratings can be expected. As Tables 4.10 and 4.12 do not display differences among the three types of raters in this respect, it is impossible to predict the differences in the restriction of range across the three types of raters either. For this reason, empirical differences can only be explained on an *ad hoc* basis. In this connection, comparison in the frequency of various categories used by different rater types suggest that the form-oriented type had a stronger tendency of range restriction than the other two types, while the percentage of unexpected responses indicating central tendency was highest for the content-oriented type. This inference, however, should be taken with caution. For one thing, the largest difference across the three rater types in the percentage of the most frequently used categories was no more than 6% (Section 5.1.5), which may not be of much practical significance. For another, the percentage of variance accounted for by the test-taker facet was estimated to be similar across rater types in the G-studies (Section 5.2.4), which indicated similar levels of restriction of range across rater types.

6.3.2.4 Halo effect

In Section 6.3.1 a distinction was made between the two operational definitions of the halo effect: the traditional definition based on subscale intercorrelations, and the MFRM definition based on identical calibrations across subscales. According to the MFRM analysis, all three types of raters were subject to the halo effect to a significant degree, as each rater type fit the restricted MFRM model with all subscale difficulties fixed at zero as well as the baseline model with all subscale difficulties free. Specifically, the form-oriented raters tended to underrate pronunciation and intonation whereas the content-oriented raters tended to underrate grammar and vocabulary, so that these subscale severities were more similar to their overall severities than to the expected severities of the respective subscales. Therefore, the form- and content-oriented raters seemed to be more subject to the halo effect in the sense that they forced the severity of a certain subscale to be similar to their overall severity. The CFA results, however, indicated that the balanced raters were more subject to the halo effect, in that their ratings fit the single-factor model more adequately than the ratings of the other types of raters.

The tendency of the form- and content-oriented raters to force their severity on the form-related subscales to be similar to their overall severity may be attributed to the same reason that underlay their unexpected severity on these subscales. It was hypothesized above in Section 6.3.2.1 that both types of raters may regard language form as essential, but that the form-oriented raters may identify pronunciation and intonation with formal competence whereas the content-oriented raters may see grammar and vocabulary as a more important reflection of formal competence. As severity on formal competence was an essential contributor to overall severity, this resulted in more similar values between the overall severity and the severity on the subscale

regarded as the primary scale for language form, i.e. pronunciation and intonation for the form-oriented raters and grammar and vocabulary for the content-oriented raters.

In comparison, the balanced raters were less focused on any particular subscales, and were less affected by the inclination to bring a particular subscale to the same level as the overall severity. However, being less focused on any particular subscales may have produced the side effect of diffused, or distributed, attention. On the basis of visual research, Treisman (2006) suggested that distributed attention offers global and statistical properties of objects such as the frequencies, means, ranges, and variances, or the general layout of a scene, rather than individuated features. Extending the same principle to the lack of emphasis on the part of the balanced raters, these raters may in fact have rated the test-takers' performance more on the basis of global impression than individual traits. In other words, lack of a focus in the rating process may have led to less discrimination of and the consequent high intercorrelations among the four subscales under consideration, which resulted in the better fit of the single-factor model than in the cases of the form- and content-oriented raters.

6.3.3 Attributing rater type differences to weighting patterns

An important issue related to the interpretation of the results discussed in the preceding sections is that the severity of rater types should be interpreted as a separate facet from the severity of individual raters and that the actual ratings result from the combined severity of the rater type and the individual rater when all other conditions are held constant. As a distinct facet, the idiosyncratic severity of an individual rater may have a stronger impact on the actual ratings than the rater type severity. For example, the individual severity of rater 13, a form-oriented rater, was estimated at -.74. Added to the severity of the form-oriented type, .21, this resulted in

a combined severity of -.53. Similarly, the individual severity of rater 16, a balanced rater, was estimated at .93, which resulted in a combined severity of 1.00 when the severity of the balanced type, .07, was added. All other conditions held constant, a test-taker would get a lower score from this particular balanced rater than from this particular form-oriented rater. In contrast, the individual severity of another form-oriented rater, rater 4, was estimated at .30, which resulted in a combined severity of .51 when added to the rater type severity of .21. Similarly, the individual severity of another balance rater, rater 21, was estimated at -.48, which resulted in a combined severity of -.41 when added to the rater type severity of .07. All other conditions held constant, a test-taker would get a lower score from this particular form-oriented rater than from this particular balanced rater. These two examples illustrate that rater type severity and individual rater severity are two distinct facets in the MFRM model used, and should therefore be interpreted separately. In other words, stating that the form-oriented type was the most severe type is not equal to stating that every individual form-oriented rater was more severe than individual raters of the other two types. Furthermore, as the variable map in Figure 5.2 shows, the dispersion of severity among the individual raters was much larger than that among the rater types, which again suggests that the severity of an individual rater may have a stronger impact than the rater type severity.

The same caution applies to the interpretation of the results pertaining to the other types of rater variability. For example, although the single-factor model fit the ratings of the balanced raters better than it fit the ratings of the other types of raters, this does not mean that the same model fit the ratings of each and every balanced rater better than it did the ratings of each and every form- and content-oriented rater. In the final analysis, this notion is no different from the separation of between-group variation from within-group variation in an analysis of variance.

With this caution, the relation between weighting patterns and the different types of rater variability can be briefly summarized as follows. The form-oriented raters were most severe among the three types of raters in overall severity and on the pronunciation and intonation subscale. This was mainly attributed to their primary emphasis on pronunciation and intonation and the common tendency of the TEM4-Oral raters to attend to negative rather than positive features in this domain. In contrast, the content-oriented raters were most lenient in overall severity as well as on the discussion subscale, but most severe on the grammar and vocabulary subscale. The leniency was regarded as a result of the content-oriented raters' mitigated emphasis on pronunciation and intonation and increased attention on the positive features in content. The severity on the grammar and vocabulary subscale formed a contrast to the leniency of the form-oriented raters on this subscale, and was interpreted as a reflection of the different beliefs about what scale for language form was primary. It was reasoned that while the form-oriented raters regarded pronunciation and intonation as the primary scale for language form, the content-oriented raters may have downplayed the importance of pronunciation and intonation but emphasized grammar and vocabulary in assessing language form. For a unified interpretation, both types of raters were severe on the primary scale for language form. The balanced raters, with no particular focus in their weighting patterns, were situated between the form- and content-oriented in overall severity and were not found to be unexpectedly severe or lenient on any subscales. The subscale of talk was the only subscale that witnessed no significant difference in severity across the three types of raters. This lack of difference was attributed to the checklist-style of rubric used for this subscale, in contrast to the general-impression rubrics used for the other subscales under consideration.

While differential severity was the major association between weighting patterns and rater variability, no clear-cut relationship could be predicted in terms of reliability and restriction of range. Empirically, the three types of raters were found to be similar in overall reliability. More complicated results, however, were reported in restriction of range. According to frequency analysis, the form-oriented raters were associated with a greater degree of range restriction and the content-oriented raters with a greater degree of central tendency, but the differences across the rater types were of not much practical significance, and little difference was found in the G- and D-studies. To be on the safe side, no systematic relationship between weighting patterns and reliability or restriction of range was assumed.

While the form- and content-oriented raters were somewhat subject more to halo defined as equivalent severity on different subscales, the balanced raters were subject more to halo defined as high intercorrelations among subscales. The concepts of focused and distributed attentions were borrowed from vision research for an explanation of this contrast. From this perspective, the form- and content-oriented raters both had a focus in the rating process, i.e. pronunciation and intonation in the case of the form-oriented raters and topic relevance in the case of the content-oriented raters, whereas the balanced raters tended to distribute their attention to more criteria (Chapter 4). In the cases of the form- and content-oriented raters, it was on the subscale that reflected their respective focus that these raters tended to force their ratings to be similar to their overall severity. In the case of the balanced raters, it was reasoned that diffuse attention resulted in less discrimination of the subscales and the consequent high intercorrelations among them. This, in turn, led to the better fit of the single-factor model than in the cases of the form- and content-oriented raters. Despite the *ad hoc* nature of this explanation, the differences across the three types of raters were again attributed to their weighting patterns.

6.4 Implications

The implications of these findings, especially the association of the weighting patterns to the different types of rater variability, could be elaborated on from three perspectives, theoretical, practical, and research.

Theoretically, the validity of the concept of the weighting patterns has enriched the understanding of rater-related factors in language performance assessment. As the literature review in Chapter 2 shows, most of these factors have been studied empirically, both qualitatively and quantitatively, but description of the raters' weighting patterns has mostly been conducted on the basis of verbal protocols and is of a qualitative nature. While "thick" description of the weighting patterns is an inherited merit, this approach to weighting patterns is limited to casewise description, and a macroscopic view is hardly possible due to the difficulty in obtaining data from a large sample of raters. Therefore, the attempt of this study to elicit the raters' responses to score profiles through a value judgment task and to derive the weighting patterns from these responses serve as a viable answer to this practical difficulty. In essence, this macroscopic view of the weighting patterns serves as a cross-validation of the observations in previous studies about the differential foci of raters in the rating process, considerably increasing the generalizability of this concept. In terms of the rater cognition framework discussed in Chapter 2 (Figure 2.4 and Table 2.3), the macroscopic description of the weighting patterns provides one example of how the content focus can be quantified in a large-scale study, thus extending the application of such frameworks.

The association of weighting patterns with rater variability provides a new perspective in understanding the factors contributing to rater variability. Due to the exploratory nature of this

study, no causal relationship is assumed, but the association between the two phenomena provides at lease a rationale for future studies and a reference against which future findings can be interpreted.

In practice, understanding rater weighting patterns and devising ways to measure them have significance for language testers, especially in the development of rubrics and rater training. Concerning rubrics, the number of criteria to be included and the weighting of criteria are two considerations that may benefit from findings about rater weighting patterns. For one thing, the close fitting of the single-factor model to the ratings of the balanced raters suggests that simultaneous consideration of too many criteria may result in diffused attention to any particular criterion and change analytic scoring to holistic scoring, a problem that may be addressed by limiting the number of criteria in the rubric. For another, the anchoring and masking effects of pronunciation and intonation may be offset by giving this criterion a weaker weight than other criteria in the rubric (Adams & Frith, 1979; Jacobs, Zingraf, Wormuth, Hartfield, & Hughey, 1981). Rater training can also be tailored when the weighting patterns of the raters are known. The rater typology discussed in this study may provide a starting point along this line, for if the rater typology can be decided prior to training, different criteria may be emphasized for different types of raters so as to maximize homogeneity among the raters. Statistical adjustment of test scores by way of MFRM and other methods is useful, but the reduction of rater variability can never be neglected. For one thing, prevention is always more desirable than cure; for another, statistical adjustment may prove difficult, or even impossible, from time to time. A case in point is the impossibility of including the retelling subscale in the MFRM analyses of this study (Section 5. 1.1). Even after training, the weighting patterns of certain raters may still prove excessively deviant from the norm, in which case the measures of post-training weighting

patterns may provide the necessary information for the decision makers if additional training or even rater screening is desired (Ekbatani, 2008).

The quantification of weighting patterns also has its implications in research, especially when raters constitute one of the variables in the test design. For studies that examine the effect of raters as a random factor or facet, it is a necessary condition that the participating raters constitute a random sample. In determining the extent to which this condition is satisfied, researchers typically consider rater characteristics such as gender, age, level of education, general proficiency in the target language, and familiarity with the assessment, but weighing patterns have seldom been taken into consideration. In cases where the weighting patterns of the raters may have a significant effect, there is no reason not to incorporate this information in the design. For example, the weighting patterns of the raters may have a significant effect on the results of a G-study. The results from this study have not provided information on this, but the inter-rater reliability between two raters with similar weighting patterns may be expected to be higher than that between two raters with contrasting weighting patterns. More specifically, the agreement between two form-oriented raters may be higher than the agreement between a form-oriented rater and a content-oriented rater. This difference may be so large that the assumption of random raters is seriously violated.

6.5 Some unresolved issues

As in the case of many studies, this study has raised more questions than it has answered. To avoid being lost in the jungle of details, this section will be a general discussion about the generalizability of the findings from this study. The discussion will follow the Research Use Arguments (RUS) framework proposed by Bachman (2006, 2008). In the RUA framework,

generalizability is an umbrella term for the inferential links from an observation to a report, from a report to its interpretation, and from interpretation to its uses, including generalizations and decisions or actions informed by the interpretation. Three aspects of generalizability correspond to these links: consistency, meaningfulness and consequences. Consistency reflects the degree of agreement between the multiple reports of observation, meaningfulness is related to the interpretation of the observation reports, and consequences are the results from the generalizations and decisions or actions informed by the interpretations. In the light of the RUA framework, methodological weaknesses related to each research question detailed in previous sections are mainly issues of consistency. Similarly, the implications discussed in Section 6.4 are associated with decisions or actions that may be informed by the interpretation of findings from this study. Therefore, the issues of meaningfulness and generalization will be the foci of discussion here.

The inferential link from the findings of this study to their interpretation is restricted in breadth and depth. The narrowed inferential link resulted from the inadequate exploration of the primary hypothesis of this study, i.e. the interaction between the raters' weighting patterns and the test-taker profiles. To recapitulate the hypothesis presented in Table 2.4, there is a general tendency of a form-oriented rater to *over*rate a test-taker who is adequate in language form but inadequate in content but to *under*rate a test-taker who is adequate in content but inadequate in language form. A content-oriented rater would tend to rate the test-takers in a way opposite to the form-oriented rater, and a balanced rater would give balanced ratings across test-takers. A test-taker with balanced adequacy in language form and content could be expected to be relatively free from the biases related to the form- and content-oriented raters. Generally speaking, the test-taker profiles would be expected to interact with the raters' weighting patterns

to yield certain rating patterns. This interaction was explored in a limited manner in Section 4.3.4, where the special cases of digression and salient negative features in pronunciation and intonation were found to interact with the rater types. While this provides some initial support for the predicted interaction between test-taker profiles and raters' weighting patterns, the test-taker profiles were not included in the MFRM model used in this study, and no results provide information for testing this interaction.

Just as the exclusion of test-taker profiles from the MFRM model restricted the breadth of inferential link from the findings to their interpretation, the dependence on quantitative methods limited the depth of this inference, especially in the relationship between weighting patterns and rater variability. More specifically, the interpretation of the findings concerning this relationship is often of a hypothetical and *ad hoc* nature. The argument for the different primary scales for language form (Section 6.3.2.1) and the introduction of the distributed attention concept (Section 6.3.2.4) are cases in point. While these may be the processes underlying the rating process, additional qualitative studies would be needed so as to provide in-depth descriptions of these processes. For example, to understand the workings of the two types of halo effects differentiated in the above discussion, new verbal protocols could be designed to probe into the tendency of the different types of raters toward a specific type of halo. Such a study could also provide insight into the mechanisms of distributed attention, if this really underlies the rating process, and these mechanisms could be described in detail. Hopefully, such a study could also provide information about the relationship between the two types of halo. The "thick" description available from such qualitative studies would no doubt deepen and enrich the interpretation of findings from this study.

The generalization of the interpretation discussed in this chapter is another important consideration according to the RUA framework. This issue could be examined in terms of four aspects: units, treatments, outcomes, and settings (Shadish, Cook, & Campbell, 2002).

For a simplified description, the units of this study were nonnative-speaking college teachers of English in China. There are at least the following issues with this group of participants: differences between native- and nonnative-speaking raters, and differences between EFL teachers and teachers of other academic subjects. Whether results from this particular group of participants generalize to native-speaking raters and raters with different professional backgrounds is an issue that calls for further research. Moreover, the native language and the professional background of the participants are only two of the issues related to rater characteristics reviewed in Chapter 2.

For this study, the analogy to treatment in an experimental design is the classification of raters into three types—form-oriented, balanced, and content-oriented—according to cluster analysis of their weighting patterns. There are certainly other ways to classify the raters, and different ways to decide how many types the raters should be classified into. With different classifying schemes, the classification results may vary, and future studies could certainly be designed to find this out. Also, the inclusion of only two content-oriented raters in the verbal protocol considerably restricted the generalizability of the interpretation about this type of rater.

More fundamentally, the classification of raters was based on the weighting patterns derived from the value judgment task, which included computer-generated score profiles as the input, instead of real score profiles. In reality, however, when the raters are confronted with real test-takers, a different classification scheme may describe the typology of raters more accurately. Therefore, a value judgment task with real score profiles may provide more authentic input for

deriving the weighting patterns. To find this out requires a series of comparative studies, which may be implemented in two steps: 1) both computer-generated and real score profiles can be included in a single value judgment task but analyzed separately, and the classification results based on different types of score profiles can be compared; and 2) if the classification results differ significantly, subsequent verbal protocols and statistical analyses similar to the present study can be planned accordingly. Potentially, the classification results may be different in the number of rater types or in the constitution of the different rater types if the same number of rater types applies. Either the same or different classification results are obtained, these studies will enrich the knowledge of rater typology based on weighting patterns.

The outcomes of this study are two-fold: a description of the rating process based on verbal protocols, and a comparison of different types of raters in terms of the different patterns of rater variability. With respect to the current conception of rater variability, there is not much to be added.

There are multiple limitations in the settings in which language is assessed. Broadly, this study focused on second language speaking assessment. Second language writing assessment may well be different, in which case the anchoring and masking effects of pronunciation and intonation, for example, would be totally irrelevant. More specifically, the research context of TEM4-Oral carries with it certain peculiarities, such as the characteristics of the test-takers, the structure and tasks of the test, and the rating rubric. Of special importance is the rubric of the second task (talk), according to which the five criteria—pronunciation and intonation, grammar and vocabulary, organization, richness of content, and topic relevance—were included in the value judgment task designed to measure the weighting patterns of the raters. With different tasks and different rubrics, the criteria may vary, and the effect of such variations on the

classification results and the association of weighting patterns to rater variability may be hard to predict. Therefore, the settings of this study are another important aspect that limits the generalizability of the interpretation.

The issues of meaningfulness and generalization discussed so far, together with the methodological considerations detailed in earlier sections of this chapter, rightly echo the statement at the beginning of this section—this study has raised more questions than it has answered. This, however, could be viewed as the main value of the present study, as it is a study with "methodological rich points", points at which "researchers learn that their assumptions about the way research works and the conceptual tools they have for doing research are inadequate to understand the worlds they are researching" (Hornberger, 2006, p. 222).

6.6 Conclusion

With its limitations in perspective and possible further studies in prospect, the present study could be said to have served its intended aims, which were essentially of an exploratory nature. In review, these aims were the measurement of the weighting patterns of the raters in a large-scale EFL speaking test, the classification of these raters according to their weighting patterns, the characterization of the different types of raters in the rating process, and the association of the rater types with the different patterns of rater variability.

The raters of TEM4-Oral were classified into three types and named form-oriented, content-oriented, and balanced respectively according to their weighting patterns derived from a value judgment task. In the verbal protocols, the three types of raters were found to differ in theme coverage, self-perceived weights, degree of distinction between form- and content-related criteria, and response to digression and salient negative features in pronunciation and intonation.

In association with rater variability, the different types of raters were distinguished in overall severity and severity on three subscales, but no clear-cut relationship was found in terms of reliability and restriction of range, and mixed results were reported in terms of halo effect.

The series of findings tend to support the meaningfulness of the concept of weighting patterns in second language speaking assessment and suggest their potential effects on the rating process and rating results. The results of this study also strongly suggest that weighting pattern, as a rater characteristic, should be included as a systematic factor in a model of language performance assessment. In practice, this is an issue that should be addressed in test development and use, especially in the development of rubrics and the training of raters.

Appendices

Appendix A

Summary of holistic ratings in the value judgment task (N = 120)

Rater code	Min.	Max.	*M*	*SD*	R^2
101	20	85	63.13	10.04	0.69
102	20	80	54.38	10.52	0.80
103	20	85	46.75	12.78	0.58
104	40	95	71.00	9.38	0.66
105	25	90	57.79	16.68	0.88
106	35	95	61.29	13.76	0.71
107	40	95	64.92	14.19	0.76
108	35	85	63.79	8.43	0.79
109	20	80	52.58	12.14	0.77
110	40	85	64.29	9.06	0.76
111	30	90	62.29	11.37	0.74
112	30	85	62.25	10.77	0.63
113	30	90	58.13	12.60	0.76
114	40	90	58.88	12.19	0.74
115	30	90	70.42	11.48	0.77
116	30	85	67.92	9.80	0.87
117	40	85	65.54	7.93	0.80
118	30	90	57.63	12.98	0.84
119	40	90	66.04	9.46	0.76
120	50	90	69.42	8.65	0.70
121	45	85	65.33	8.32	0.77
122	40	90	62.29	10.47	0.78
123	40	95	70.75	11.41	0.77
124	40	90	65.42	10.99	0.64
125	30	85	62.46	10.13	0.78
126	25	80	53.71	12.14	0.89
127	30	80	56.54	10.29	0.84
128	30	85	61.50	12.31	0.81
129	30	85	62.79	12.23	0.79
130	30	85	62.83	11.59	0.93
131	20	85	53.92	12.74	0.77
132	20	70	46.92	12.06	0.87
133	50	85	72.00	5.52	0.66
134	35	90	67.17	11.77	0.40
135	15	80	49.17	14.79	0.68
136	35	85	59.92	9.85	0.81
137	30	80	58.42	9.19	0.80

138	40	95	67.58	9.03	0.75
139	35	80	55.08	9.50	0.73
140	40	80	58.17	9.33	0.73
141	35	80	61.21	9.17	0.73
142	35	85	63.38	9.38	0.73
143	30	90	61.29	10.64	0.75
144	30	85	62.96	9.25	0.79
145	30	85	63.71	10.70	0.84
146	40	80	62.08	10.24	0.78
147	25	80	53.21	13.15	0.82
148	45	95	75.63	11.29	0.74
149	25	85	55.88	11.11	0.74
150	30	85	61.29	8.78	0.77
151	30	85	55.58	12.54	0.83
152	35	85	57.13	8.71	0.78
153	30	75	58.17	8.93	0.82
154	30	95	66.54	13.47	0.83
155	30	85	61.71	10.40	0.80
156	30	95	68.08	12.76	0.83
157	40	85	61.58	8.30	0.84
158	30	85	62.54	12.28	0.85
159	50	85	66.42	7.62	0.77
160	20	90	58.79	12.70	0.78
161	40	85	59.71	8.41	0.79
162	40	80	58.79	8.43	0.83
201	40	85	59.79	8.39	0.76
202	30	85	58.58	9.35	0.80
203	30	80	59.33	9.44	0.84
204	30	80	57.33	10.41	0.90
205	30	80	58.92	9.81	0.76
206	35	85	58.83	10.14	0.82
207	20	75	55.63	10.48	0.78
208	40	85	58.88	9.52	0.79
209	40	95	62.96	12.59	0.79
210	30	90	61.96	11.80	0.54
211	40	90	66.38	8.28	0.66
212	20	90	61.17	12.62	0.71
213	20	85	62.00	12.66	0.82
214	20	90	64.46	13.95	0.76
215	40	85	63.04	8.03	0.86
216	20	90	55.33	16.15	0.68
217	35	90	63.54	10.30	0.88
218	20	95	73.29	12.15	0.79
219	20	85	57.79	12.73	0.86
220	40	85	63.46	9.89	0.72
221	20	85	55.83	12.76	0.80
222	30	85	53.58	12.25	0.70

223	55	90	69.00	6.94	0.67
224	35	80	59.67	9.76	0.82
225	30	80	58.42	10.19	0.78
226	40	85	61.38	8.70	0.76
227	50	85	70.29	7.31	0.58
228	40	80	59.38	7.32	0.82
229	20	85	54.75	13.90	0.81
230	20	85	58.17	8.91	0.86
231	40	85	65.83	8.41	0.82
232	20	90	58.71	14.24	0.70
233	30	80	58.29	10.82	0.81
234	40	90	62.42	10.41	0.85
235	30	80	58.33	8.61	0.76
236	40	90	64.13	10.41	0.78
237	30	90	60.92	11.45	0.85
238	20	80	52.46	12.52	0.79
239	20	85	60.88	10.99	0.82
240	20	95	59.88	13.91	0.86
241	30	90	61.25	12.49	0.81
242	40	85	61.67	9.01	0.80
243	40	80	60.42	7.38	0.92
244	35	85	59.58	11.33	0.74
245	30	85	58.92	10.25	0.85
246	30	80	58.38	10.56	0.91
247	10	80	54.58	14.50	0.85
248	30	90	63.63	12.82	0.79
249	30	85	61.88	12.20	0.76
250	30	85	59.88	9.95	0.79
251	35	90	63.29	10.16	0.85
252	30	85	60.75	11.59	0.91
253	30	90	59.08	11.54	0.80
254	40	90	64.54	10.71	0.80
255	35	85	62.46	9.87	0.83
256	30	90	65.96	10.95	0.75
257	40	90	65.54	11.66	0.83
258	20	85	57.92	10.34	0.85
259	30	85	60.88	10.24	0.84
260	30	85	63.13	9.15	0.82
261	30	90	59.58	13.85	0.71
262	35	80	58.79	9.26	0.82
263	30	80	58.13	10.25	0.87
264	35	95	63.17	12.38	0.76

Appendix B

Correlation between profile scores and holistic rating and beta weights from regression

Rater	Correlation with rating					Beta weights from regression					R^2
Code	Rel	Con	Org	Gra	Pro	Rel	Con	Org	Gra	Pro	
101	.44	.36	.24	.32	.47	.30	.37	.23	.33	.56	.69
102	.57	.42	.40	.39	.30	.40	.40	.35	.37	.40	.80
103	.42	.34	.36	.36	.28	.27	.35	.33	.35	.38	.58
104	.34	.27	.30	.29	.53	.19	.31	.31	.30	.62	.66
105	.91	.42	.20	.06	.01	.84	.25	.07	.02	.03	.88
106	.60	.39	.39	.33	.23	.45	.35	.33	.30	.32	.71
107	.50	.40	.43	.42	.27	.32	.40	.39	.40	.39	.76
108	.53	.41	.44	.44	.25	.35	.40	.39	.41	.36	.79
109	.53	.53	.36	.37	.21	.34	.52	.33	.35	.32	.77
110	.58	.50	.39	.39	.09	.40	.46	.33	.36	.20	.76
111	.42	.22	.15	.22	.66	.32	.24	.16	.25	.72	.74
112	.37	.26	.25	.42	.44	.24	.28	.24	.43	.53	.63
113	.63	.57	.37	.22	.12	.45	.52	.33	.19	.22	.76
114	.71	.48	.25	.19	.20	.58	.40	.18	.17	.27	.74
115	.45	.41	.28	.51	.33	.29	.42	.24	.51	.43	.77
116	.55	.42	.38	.46	.34	.38	.41	.34	.45	.45	.87
117	.65	.49	.42	.30	.15	.48	.44	.36	.27	.25	.80
118	.86	.35	.33	.16	.09	.77	.21	.20	.12	.13	.84
119	.59	.46	.32	.34	.29	.43	.42	.28	.33	.39	.76
120	.44	.54	.36	.41	.09	.24	.54	.33	.39	.21	.70
121	.52	.52	.40	.39	.16	.33	.50	.37	.37	.28	.77
122	.49	.63	.21	.24	.31	.30	.63	.22	.24	.42	.78
123	.53	.39	.35	.43	.31	.38	.38	.30	.42	.41	.77
124	.49	.42	.40	.37	.16	.32	.40	.36	.34	.27	.64
125	.44	.43	.16	.48	.41	.30	.43	.14	.49	.50	.78
126	.65	.43	.42	.36	.32	.47	.40	.37	.34	.42	.89
127	.58	.46	.43	.33	.32	.39	.45	.41	.31	.43	.84
128	.54	.41	.45	.42	.27	.35	.41	.41	.40	.38	.81
129	.35	.18	.18	.21	.73	.25	.23	.21	.24	.79	.79
130	.62	.46	.35	.39	.38	.45	.44	.31	.39	.48	.93
131	.63	.51	.32	.29	.24	.46	.47	.27	.27	.33	.77
132	.64	.46	.34	.45	.25	.48	.42	.27	.43	.35	.87
133	.46	.55	.23	.27	.27	.29	.54	.22	.27	.37	.66
134	.54	.34	.04	.19	.15	.48	.25	-.03	.19	.19	.40
135	.43	.39	.21	.62	.07	.30	.35	.14	.61	.17	.68
136	.53	.46	.20	.30	.48	.38	.45	.19	.31	.57	.81
137	.60	.53	.19	.36	.30	.45	.48	.15	.36	.39	.80
138	.42	.48	.21	.39	.40	.26	.49	.21	.41	.50	.75
139	.52	.38	.42	.39	.28	.35	.38	.39	.37	.39	.73

140	.36	.31	.27	.51	.42	.22	.33	.26	.52	.52	.73
141	.59	.43	.22	.22	.41	.45	.39	.19	.23	.48	.73
142	.59	.51	.27	.24	.30	.42	.48	.24	.23	.39	.73
143	.51	.35	.25	.43	.41	.38	.34	.21	.44	.50	.75
144	.54	.52	.45	.34	.17	.34	.51	.42	.31	.29	.79
145	.60	.35	.47	.42	.28	.44	.33	.41	.39	.39	.84
146	.73	.49	.35	.25	.06	.59	.40	.27	.22	.14	.78
147	.64	.52	.38	.27	.22	.47	.48	.33	.25	.32	.82
148	.57	.69	.13	.16	.14	.40	.64	.11	.16	.22	.74
149	.47	.44	.46	.46	.12	.28	.44	.42	.42	.24	.74
150	.45	.38	.22	.26	.55	.30	.40	.23	.28	.64	.77
151	.60	.66	.26	.25	.16	.42	.62	.23	.23	.27	.83
152	.58	.40	.41	.28	.35	.41	.39	.39	.26	.45	.78
153	.50	.58	.31	.38	.28	.30	.58	.29	.37	.39	.82
154	.80	.37	.28	.19	.28	.69	.27	.19	.17	.33	.83
155	.51	.56	.17	.26	.43	.34	.55	.17	.27	.52	.80
156	.51	.43	.38	.53	.24	.34	.42	.33	.51	.36	.83
157	.57	.44	.41	.51	.14	.40	.41	.34	.48	.26	.84
158	.58	.44	.41	.38	.32	.40	.42	.37	.37	.43	.85
159	.53	.43	.36	.41	.31	.36	.42	.32	.39	.41	.77
160	.61	.28	.30	.42	.36	.50	.24	.23	.41	.44	.78
161	.63	.42	.31	.46	.19	.49	.36	.22	.44	.28	.79
162	.70	.56	.29	.27	.17	.55	.48	.22	.25	.26	.83
201	.42	.38	.19	.41	.48	.28	.40	.19	.43	.58	.76
202	.56	.43	.43	.35	.29	.38	.42	.40	.33	.40	.80
203	.60	.48	.34	.32	.35	.42	.46	.31	.31	.45	.84
204	.66	.47	.44	.38	.24	.49	.43	.38	.35	.35	.90
205	.37	.55	.12	.16	.51	.20	.58	.16	.19	.60	.76
206	.45	.55	.24	.41	.37	.26	.56	.23	.42	.49	.82
207	.28	.36	.15	.26	.65	.13	.42	.20	.30	.74	.78
208	.57	.50	.30	.36	.30	.40	.48	.27	.35	.40	.79
209	.56	.59	.43	.26	.12	.36	.57	.40	.22	.23	.79
210	.48	.26	.36	.37	.23	.35	.24	.31	.35	.31	.54
211	.50	.36	.32	.26	.39	.36	.36	.31	.26	.48	.66
212	.50	.71	.18	.12	.11	.32	.68	.18	.11	.21	.71
213	.60	.49	.37	.42	.19	.42	.45	.32	.40	.30	.82
214	.47	.35	.15	.54	.37	.35	.33	.10	.55	.46	.76
215	.59	.48	.44	.36	.28	.40	.47	.40	.33	.39	.86
216	.43	.34	.48	.42	.22	.26	.35	.45	.39	.33	.68
217	.59	.48	.43	.40	.28	.40	.46	.38	.38	.39	.88
218	.40	.33	.30	.49	.45	.24	.36	.28	.50	.56	.79
219	.75	.60	.27	.24	.09	.59	.50	.20	.21	.18	.86
220	.57	.63	.20	.19	.17	.41	.59	.17	.18	.26	.72
221	.51	.35	.41	.38	.39	.35	.36	.39	.37	.49	.80
222	.51	.38	.35	.43	.27	.35	.36	.31	.42	.37	.70

223	.48	.55	.22	.27	.28	.31	.54	.21	.27	.38	.67
224	.65	.64	.23	.15	.23	.48	.59	.20	.14	.32	.82
225	.45	.41	.25	.23	.54	.30	.43	.26	.25	.63	.78
226	.39	.54	.35	.41	.24	.18	.57	.35	.40	.36	.76
227	.52	.56	.12	.19	.17	.38	.52	.10	.19	.25	.58
228	.52	.50	.44	.43	.18	.32	.49	.40	.41	.30	.82
229	.53	.70	.25	.21	.19	.32	.68	.25	.20	.30	.81
230	.56	.49	.40	.40	.29	.37	.48	.37	.38	.41	.86
231	.48	.56	.32	.38	.30	.28	.57	.31	.37	.42	.82
232	.48	.73	.20	.08	.04	.29	.70	.20	.07	.14	.70
233	.56	.73	.23	.20	.09	.36	.69	.21	.19	.19	.81
234	.68	.66	.29	.21	.08	.50	.58	.24	.18	.17	.85
235	.24	.57	.09	.36	.44	.06	.62	.14	.39	.55	.76
236	.56	.47	.34	.40	.26	.39	.45	.30	.39	.37	.78
237	.58	.44	.32	.37	.39	.42	.42	.28	.37	.49	.85
238	.59	.40	.44	.35	.28	.42	.37	.40	.32	.38	.79
239	.22	.40	.15	.56	.46	.06	.46	.16	.59	.57	.82
240	.57	.64	.35	.36	.11	.37	.61	.32	.33	.23	.86
241	.67	.60	.27	.21	.18	.51	.53	.22	.20	.27	.81
242	.55	.47	.37	.27	.38	.37	.47	.36	.26	.48	.80
243	.62	.49	.40	.40	.30	.44	.47	.35	.38	.41	.92
244	.37	.51	.21	.34	.43	.19	.54	.23	.35	.53	.74
245	.57	.46	.40	.38	.32	.39	.45	.37	.36	.43	.85
246	.63	.48	.39	.42	.29	.45	.45	.33	.40	.40	.91
247	.71	.35	.23	.31	.41	.60	.28	.16	.31	.47	.85
248	.65	.41	.36	.39	.23	.50	.35	.29	.36	.32	.79
249	.63	.41	.41	.35	.20	.47	.37	.35	.32	.30	.76
250	.46	.35	.45	.48	.29	.29	.37	.42	.46	.41	.79
251	.59	.64	.36	.34	.10	.39	.60	.32	.31	.22	.85
252	.61	.49	.37	.42	.30	.43	.47	.33	.40	.41	.91
253	.50	.36	.22	.35	.52	.36	.36	.20	.37	.61	.80
254	.58	.35	.40	.40	.33	.42	.33	.35	.39	.43	.80
255	.44	.46	.15	.20	.60	.28	.48	.18	.23	.69	.83
256	.59	.51	.30	.28	.28	.42	.48	.26	.27	.37	.75
257	.55	.47	.33	.34	.38	.37	.47	.31	.34	.49	.83
258	.50	.48	.33	.32	.45	.32	.49	.33	.32	.55	.85
259	.73	.65	.18	.17	.09	.58	.55	.11	.15	.17	.84
260	.55	.55	.26	.35	.31	.38	.53	.24	.35	.42	.82
261	.63	.37	.24	.37	.28	.52	.31	.17	.36	.35	.71
262	.68	.33	.18	.36	.40	.59	.26	.10	.36	.46	.82
263	.77	.44	.26	.19	.34	.64	.36	.19	.18	.40	.87
264	.62	.47	.35	.25	.29	.45	.44	.31	.23	.38	.76

Note: Rel = Topic relevance; Con = Content richness; Org = Organization; Gra = Grammar and vocabulary; Pro = Pronunciation and intonation

Appendix C

Relative weights and results of k-means cluster analysis

Rater	Relative Weights					Cluster
Code	Rel	Con	Org	Gra	Pro	Membership
101	.19	.19	.08	.15	.38	3
102	.28	.21	.18	.18	.15	2
103	.20	.20	.20	.22	.18	2
104	.10	.13	.14	.13	.49	3
105	.86	.12	.02	.00	.00	1
106	.39	.19	.18	.14	.10	1
107	.21	.21	.22	.22	.14	2
108	.24	.20	.22	.23	.11	2
109	.23	.36	.16	.17	.09	2
110	.31	.31	.17	.19	.03	2
111	.18	.07	.03	.08	.64	3
112	.14	.12	.09	.28	.37	3
113	.37	.39	.16	.05	.04	2
114	.56	.26	.06	.05	.07	1
115	.17	.22	.09	.34	.18	2
116	.24	.20	.15	.24	.18	2
117	.40	.27	.19	.10	.05	2
118	.80	.09	.08	.02	.01	1
119	.34	.25	.12	.15	.15	2
120	.15	.42	.17	.23	.03	2
121	.22	.34	.20	.19	.06	2
122	.19	.51	.06	.07	.17	2
123	.26	.19	.14	.24	.17	2
124	.25	.26	.23	.20	.07	2
125	.17	.24	.03	.30	.26	3
126	.34	.20	.17	.14	.15	2
127	.27	.24	.21	.12	.16	2
128	.23	.21	.23	.21	.13	2
129	.11	.05	.05	.06	.73	3
130	.30	.22	.12	.16	.20	2
131	.37	.31	.11	.10	.10	2
132	.35	.22	.11	.22	.10	2
133	.21	.45	.08	.11	.15	2
134	.64	.21	.00	.09	.07	1
135	.19	.20	.04	.55	.02	2
136	.25	.26	.05	.11	.33	3
137	.34	.32	.04	.16	.14	2

138	.14	.31	.06	.21	.27	3
139	.25	.20	.22	.20	.15	2
140	.11	.14	.10	.36	.30	3
141	.37	.23	.06	.07	.27	1
142	.34	.34	.09	.08	.16	2
143	.25	.15	.07	.25	.27	3
144	.24	.33	.24	.13	.06	2
145	.31	.13	.23	.20	.13	2
146	.55	.25	.12	.07	.01	1
147	.37	.31	.16	.08	.09	2
148	.31	.60	.02	.03	.04	2
149	.17	.26	.26	.26	.04	2
150	.18	.20	.07	.10	.46	3
151	.31	.50	.07	.07	.05	2
152	.30	.20	.20	.09	.20	2
153	.19	.40	.11	.17	.13	2
154	.66	.12	.06	.04	.11	1
155	.22	.38	.04	.09	.28	2
156	.21	.22	.15	.32	.10	2
157	.27	.22	.17	.29	.04	2
158	.28	.22	.18	.17	.16	2
159	.25	.23	.15	.21	.16	2
160	.40	.09	.09	.22	.21	1
161	.40	.19	.09	.26	.07	1
162	.46	.33	.08	.08	.05	1
201	.15	.20	.05	.23	.37	3
202	.26	.23	.22	.15	.15	2
203	.30	.26	.13	.12	.19	2
204	.36	.22	.18	.15	.09	2
205	.10	.42	.02	.04	.41	3
206	.14	.38	.07	.21	.22	2
207	.05	.19	.04	.10	.62	3
208	.28	.30	.10	.16	.15	2
209	.25	.42	.22	.07	.03	2
210	.31	.12	.20	.23	.13	2
211	.27	.20	.15	.10	.28	3
212	.22	.68	.04	.02	.03	2
213	.31	.27	.15	.21	.07	2
214	.22	.15	.02	.39	.22	3
215	.27	.26	.20	.14	.12	2
216	.16	.18	.31	.24	.10	2
217	.27	.25	.19	.17	.12	2
218	.12	.15	.11	.30	.32	3

219	.51	.35	.06	.06	.02	1
220	.33	.52	.05	.05	.06	2
221	.22	.16	.20	.18	.24	2
222	.25	.19	.15	.26	.14	2
223	.22	.44	.07	.11	.16	2
224	.38	.46	.05	.02	.09	2
225	.17	.23	.08	.07	.44	3
226	.10	.41	.16	.22	.11	2
227	.34	.50	.02	.06	.08	2
228	.21	.30	.22	.21	.06	2
229	.21	.59	.08	.05	.07	2
230	.24	.28	.17	.17	.14	2
231	.16	.39	.12	.17	.15	2
232	.20	.73	.06	.01	.01	2
233	.25	.62	.06	.05	.02	2
234	.40	.45	.08	.05	.02	2
235	.02	.46	.02	.18	.32	3
236	.28	.27	.13	.20	.12	2
237	.29	.22	.10	.16	.23	2
238	.31	.19	.22	.14	.14	2
239	.02	.23	.03	.41	.32	3
240	.24	.46	.13	.14	.03	2
241	.42	.39	.07	.05	.06	1
242	.25	.27	.17	.09	.22	2
243	.30	.25	.15	.16	.13	2
244	.10	.37	.07	.16	.31	3
245	.26	.24	.17	.16	.16	2
246	.31	.23	.14	.19	.13	2
247	.50	.12	.04	.11	.23	1
248	.41	.18	.13	.18	.09	1
249	.39	.20	.18	.15	.08	1
250	.17	.16	.24	.28	.15	2
251	.27	.45	.13	.12	.03	2
252	.29	.25	.13	.18	.14	2
253	.23	.16	.06	.16	.40	3
254	.31	.15	.18	.20	.18	2
255	.15	.27	.03	.05	.50	3
256	.33	.33	.10	.10	.14	2
257	.25	.27	.12	.14	.23	2
258	.19	.28	.13	.12	.29	3
259	.51	.42	.02	.03	.02	1
260	.25	.36	.08	.15	.16	2
261	.46	.16	.06	.19	.14	1

262	.49	.10	.02	.16	.23	1
263	.57	.18	.06	.04	.16	1
264	.37	.27	.14	.07	.14	2

Note: Rel = Topic relevance; Con = Content richness; Org = Organization; Gra = Grammar and vocabulary; Pro = Pronunciation and intonation

Appendix D

Distribution of comments on the three domains

Rater	Type	Number of comments			Percentage of comments		
		Content	Grammar	Pronunciation	Content	Grammar	Pronunciation
Mina	1	8	16	16	.20	.40	.40
Charlene	1	19	18	22	.32	.31	.37
Hermione	2	20	25	26	.28	.35	.37
Mona	2	25	10	20	.45	.18	.36
Lee	3	17	13	10	.43	.33	.25
Mia	3	16	18	12	.35	.39	.26
James	1	4	15	13	.13	.47	.41
Tisha	2	7	9	9	.28	.36	.36
Leanna	2	8	10	10	.29	.36	.36
King	2	23	25	22	.33	.36	.31
May	1	27	27	24	.35	.35	.31
Lou	1	7	10	10	.26	.37	.37
Josie	2	10	11	9	.33	.37	.30
Elaine	1	27	26	23	.36	.34	.30
Phoebe	1	6	10	12	.21	.36	.43
Kim	1	13	15	14	.31	.36	.33
Tel	2	10	7	8	.40	.28	.32
June	1	3	3	5	.27	.27	.45
Todd	1	12	15	18	.27	.33	.40
Hilda	2	22	21	20	.35	.33	.32
Sean	2	17	15	11	.40	.35	.26

Note: Type 1 = Form-oriented; Type 2 = Balanced; Type 3 = Content-oriented.

Appendix E

Excerpts of the verbal protocol transcript

(Translated from Chinese. Numbers denote corresponding sections in the body of the text.)

4.3.2 Self-perceived weights

Form-oriented raters:

Charlene: Pronunciation comes first. Yes. And then topic relevance comes second. Third is grammar.

Mina: In fact, I am probably more concerned with pronunciation and intonation. And then content, but when it comes to grammar and vocabulary, er, I am not so concerned.

James: I am most concerned with pronunciation and intonation….If ever pronunciation and intonation is good, scores go to 75 or 80….if ever pronunciation and intonation is poor, then scores will be 60 or 65.

May: To start with, I will give most attention to pronunciation and intonation….Second, second comes grammar and vocabulary.

Lou: Pronunciation and intonation, grammar and vocabulary, content in sequence of importance.

Kim: I think pronunciation and intonation comes first, vocabulary second, richness of content third, and then organization. Topic relevance is the last.

Todd: I think pronunciation and intonation, I am sure, I am most concerned with. And then comes grammar and vocabulary.

Phoebe: I have been most concerned with pronunciation and intonation….after this rating experience, I think I will remind my students of topic relevance.

June: My first impression goes with pronunciation and intonation ….then comes content, I mean, topic relevance, it won't do if it is off topic. And then, if everything is relevant to the topic, I will listen for the content, most probably to find out how he wraps the whole thing up, and whether he is able to, to end the story with a moral, or just give a plain ending to the story. That is, I may make a final judgment on the content at the end.

Elaine: First comes pronunciation and intonation…. And then comes content…. Grammar is not so important, in comparison.

Balanced raters:

Hermione: Er, I am most concerned with, er, I think it is, pronunciation on the one hand, I figure I care much about pronunciation, er, on the other hand, er, content, I believe it must be relevant to the topic, otherwise he doesn't even have to talk and a really low score will be given. Then it comes to his logic and richness in content.

Mona: I am, most concerned with content…. You have to give me real stuff…. We will, I mean, encourage students to speak with a good flow… but I am most concerned with content when it comes to rating.

Tisha: Er, I seem to be more concerned with pronunciation and grammar…. Can't control it. I feel, er, sometimes I think content, and development deserves more attention, but I will still be distracted uncontrollably…. Pronunciation, grammar, er, yes, distracted by pronunciation and intonation, and grammar.

Leanna: I will probably, consider the whole thing first. If a student did a good job in retelling, I will wait, with patience, to the end, and then decide how well he did in pronunciation, in grammar, etc.

King: Er, I am concerned with, er, pronunciation on the one hand, and content on the other…. Er, content is the stuff, pronunciation, I mean, there should be a lot of stuff…. If pronunciation is good, then the stuff has a beautiful appearance, and it makes it even better. As to grammar and vocabulary, I think we can place lower requirements in speaking, not as high as in writing.

Josie: For me, content comes first, and then grammar and vocabulary, then organization, then pronunciation and intonation. Anyway pronunciation and intonation won't be the most important.

Sean: I mean I am most concerned with content and topic relevance, next come organization and grammar and vocabulary, the whole thing should adhere, and last comes pronunciation and intonation.

Tel: That is to say pronunciation and intonation only gives you the first impression, while the key lies in content.

Hilda: In terms of the rubrics, I am more oriented toward the content and the specific stuffs, I mean the richness and liveliness of content and topic relevance…. And then, on this basis, if there is a rich variety of vocabulary, and the structure is complete, I mean, logical, the score will be even higher…. It is then that I consider pronunciation and intonation, but I am also affected by first impression.

Content-oriented raters:

Lee: I am probably most concerned with content. Besides, pronunciation should be clear…. If pronunciation is clear I will at least keep the pronunciation score at a certain level. And then, if the pronunciation is particularly good, even if the content is average, I may raise the score a bit.

Mia: My first priority goes to content…. Unless there is anything particular in other aspects… particularly bad… otherwise… I will base my judgment mainly on content.

4.3.3 Relationship between criteria

Charlene, form-oriented, Clip 3, Segment 4:

Cai: This ends Segment 4.

Charlene: I have reduced the content score by five points.

Cai: Content score reduced by five points?

Charlene: Yes. And then grammar and vocabulary remains unchanged.

Cai: So why did you reduce the content score by five points?

Charlene: I think that in content, er, for one thing the sentence patterns are not varied, and for another there is redundancy. She doesn't have to say so much.

Mina, form-oriented, Clip 2, Segment 2:

Mina: Seems he is speaking at quite a low rate. And then his ideas are rather…I have a feeling that within three minutes he will have…difficulty in bringing his story to an end.

Cai: Really? You think he speaks slowly?

Mina: He is a bit slow, and so his information won't be sufficient. Probably he will get a score of around 60.

Lou, form-oriented, Clip 5, Segment 1:

Lou: But her pronunciation and intonation and her grammar and vocabulary were indeed not so good. In fact, her pronunciation and intonation was not so bad in the first segment, but her fluency, out of whatever reasons, her lack of fluency has obstructed the conveyance of message, and so I gave her a failing score. She did quite well in the second segment, but in the third, the fourth, and the fifth segments, the more she spoke, the more language problems, such as the confusion of long and short vowels, poor enunciation of the dental fricative, and the "ei" sound, she had some problem in this vowel as well, and so on many occasions, this has, so to speak, obstructed the conveyance of message.

May, form-oriented, Clip 2, Segment 1:

May: For this, the first segment, the content of the monologue cannot, er, be easily evaluated on the whole. So I, er, for the time being, made the judgment according to his performance in the other criteria.

May, form-oriented, Clip 4, Segment 5:

May: Hmm, the final score for the monologue is 75.
Cai: Hmm.
May: Five points higher now, er, because I now find out that the first impression she gave me was actually a problem of pronunciation and intonation, and that has been somewhat misleading.

Kim, form-oriented, Clip 2, Segment 1:

Kim: In this beginning, as I am not pretty sure what he will talk about later, I can only evaluate his pronunciation, which is just adequate for me to understand what he says, and so I give a passing score to each of the subscales, exactly the passing score.

Todd, form-oriented, Clip 2, Segment 1:

Todd: As to content, I think, maybe this was also affected by intonation, but anyway, his content did not have much, so to speak, strong attraction, not many merits, and I think this was mainly affected by his intonation, I listened only for his intonation at the beginning.

Hermione, balanced, Clip 4, Segment 5:

Hermione: "as I can." There are some grammatical errors.
Cai: Hmm.
Hermione: But these did not affect the expression of the ideas. I feel okay with it. Er, about this, about what she said, about what she had learnt, her expression was to the point, relevant.

Hermione, balanced, general comment:

Hermione: If the pronunciation of a student is really terrible, it will affect his content score.
Cai: Hmm, you mean…
Hermione: Because…
Cai: You mean pronunciation and content are in fact…
Hermione: closely related.
Cai: Related.
Hermione: It will affect…
Cai: Hmm.
Hermione: It will affect me, at least. I think at least I won't, I, er, will be affected when I give the content score.

Josie, balanced, Clip 2, Segment 4:

Josie: Er, when it came to the fourth segment, that is, he made more serious errors, like "no gravity", like, "celebrate me", and the like.
Cai: Hmm.
Josie: But, on the whole I think I was still able to understand what he meant, and that's why I still gave him, a passing score, for after all I was able to understand what he wanted to say in his communication, though of course he made a lot of errors.

Josie, balanced, Clip 1, Segments 3 & 4:

Josie: And when it came to the third and fourth segments, I was most concerned over, in the third segment, I think that, her poor language was caused by, er, a problem in the content, because she was recalling, she was recalling the situation in the college entrance exams. So here, I think this actually made the task harder for her, she had to consider, before, what she had said before, and so here, I found out that she did not so well in fluency, and I would give her a lower score in pronunciation and intonation.

Mona, balanced, Clip 2, Segment 5:

Mona: I mean I did find him to have poor pronunciation and intonation, but he made his points toward the end, and I would raise his pronunciation score. I think his pronunciation was okay if only he expressed his ideas clearly.

Mona, balanced, Clip 4, Segment 4:

Mona: Hmm, I mean, I have raised her vocabulary score.
Cai: Hmm.
Mona: I mean, she, I think she is able to convey her ideas well enough.
Cai: Hmm.
Mona: It came out that I felt confused about what she wanted to say in the beginning, that's why I gave her a low score in vocabulary.

Mona, balanced, Clip 5, Segment 5:

Mona: I will take away some points from her monologue. I don't know how unexpected is her "unexpected", I mean, I will give her 50 or 55 for her content, I think. If I give her 55, it's because her pronunciation is not too bad.

Mona, balanced, Clip 2, Segment 2:

Mona: I, I think after this segment, I can give him a higher score.
Cai: In what?
Mona: In content, mainly.
Cai: Content?
Mona: Because he has said something, I mean, at the very beginning he gave me an impression of poor spoken English, I mean poor pronunciation and intonation.
Cai: Hmm.
Mona: It may have an influence. I feel that sometimes, I mean the first segment, the influence of pronunciation and intonation was stronger, and I would judge that he was under average.
Cai: You mean you base your judgment on the pronunciation and intonation in the first segment?
Mona: Yes, I mean his pronunciation and intonation tended to affect his, the understanding of his idea. In the second segment I found out that he was, making his points…
Cai: In terms of content?
Mona: Making some points in terms of content.
Cai: Hmm.
Mona: I think he may have, because of his local accent, some problems in pronunciation and intonation, but he made his points, and I can raise his score in content, I think.

Tisha, balanced, Clip 2, Segments 1 & 5:

Tisha: But if such a student did an especially good job in the first segment, and had rich content in the last segment, I might give him a score of 60 for grammar.

Appendix F

Transcript of talk by two test-takers

Clip 2:

In 2008, in 2008, I am going to take the NCEE. At that time everyone was prepared for the exam, the examination, I am also. During that time, I am very nervous and forget everything. You know, we just work hard, for, er, for, er, we just work hard to pass through, er, that, er, that examination. // One day, when I am, when I returned, I'm very tired. I open the door, and saw, and see nobody at home. And inside, and inside the room was very dark. Just as I, I am, I am going to, turn on the light, I saw a light in the far way was proaching me, was approaching me. It's a candle. // I'm very surprised. I'm very surprised. And, just I'm, as I, just as I'm wondering, someone was sing, "happy birthday to you." I know, its' my birthday, and my family prepared it for me. I am very movid. // During that time, during that time, I'm working, I'm just working hard, and forget everything, including my birthday. So, our family take a seat, and celebrate me for my birthday. // That's, that is my unexpected experience. I'm very glad that, at my hard time, er, not hard time, er, working hard time, er, my family can bear my birthday in their mind, I'm very, er, I'm very, you know, I am very movid, and, and regretful, and grateful to them.

Clip 5:

When I was Junior 3, my aunt got a very strange and serious disease. All the doctors can't decept, detect what kind of disease, what kind of this disease exactly is, and my aunt started to, er, started to, to, lie on bed. She can't work, she couldn't work any more. // And her IQ was just like a child; she even can't count how much does seven, seven plus eight. And she couldn't talk any more. We tried everything to save her, and even sent her to Shanghai to meet some very famous, some very famous doctors, but still, still couldn't save her. // And five months later my, my aunt's doctor said she, her prain, her brain can't work any more, and she can only live, rely, by rely on the breathing machine, so we decided to give up and, my aunt, my aunt died at her 50's, in her 50's.// So after this, all, everyone of my family feel very upset, and we can't, we can't, we think we can't live without such an important members. But later I found that people we love will live, will, will live, live one day. // We, the only thing we can do is accept the fact and live our own life better. This is the best way to memorize the, our loved one, our loved ones, and I think they, even they were, has passed away, even they have passed away they still don't, they don't want to see us feel, feel sad about their death.

Note: // marks the border between two segments in the verbal protocols.

Appendix G

Summary of correlations, means, and standard deviations of data subsets used in the CFA

	Talk	Discussion	Pronunciation	Grammar	*M*	*SD*
Form-oriented						
Talk	1				12.29	2.10
Discussion	.623	1			13.04	1.53
Pronunciation	.559	.628	1		12.70	1.81
Grammar	.608	.717	.773	1	13.08	1.82
Balanced						
Talk	1				12.08	2.30
Discussion	.580	1			12.82	1.65
Pronunciation	.563	.658	1		12.71	1.95
Grammar	.644	.689	.763	1	12.75	1.83
Content-oriented						
Talk	1				11.99	2.14
Discussion	.683	1			12.83	1.71
Pronunciation	.631	.741	1		12.56	1.86
Grammar	.677	.742	.872		12.46	1.96

References

Adams, M. L. & Frith, J. R. (Eds.). (1979). *Testing kit: French and Spanish*. Washington D. C.: Foreign Service Institute.

Aguinis, H. (2007). *Performance management*. New Jersey: Pearson Prentice Hall.

Amoo, T. & Friedman, H. H. (2001). Do numeric values influence subjects' responses to rating scales? *Journal of International Marketing and Marketing Research*, *26*, 41-46.

Anderson, N. H. (1982). *Methods of information integration theory.* New York: Academic Press.

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Andrich, D. (1978b). Application of a psychometric rating model to ordered categories, which are scored with successive integers. *Applied Psychological Measurement, 2*(4), 581-594.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*(4), 453-476.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1-34.

Bachman, L. F. (2006). Generalizability: A journey into the nature of empirical research in applied linguistics. In M. Chalhoub-Deville, C. A. Chapelle, & P. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 165-207). Dordrecht, Netherlands: John Benjamins.

Bachman, L. F. (2008). Generalizability and research use arguments. In K. Ercikan & W-M. Roth (Eds.), *Generalizing from educational research: Beyond qualitative and quantitative polarization* (pp. 127-148). New York: Taylor & Francis.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12,* 238-257.

Bachman, L. F. & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions, 22*(1), 1145-1146.

Baker, B. A. (2010). Playing with the stakes: A consideration of an aspect of the social context of a gatekeeping writing assessment. *Assessing Writing, 15*(3), 133-153.

Barkaoui, K. (2007a). Rating scale impact on EFL essay marking: A mixed method study. *Assessing Writing, 12,* 86-107.

Barkaoui, K. (2007b). Participants, texts, and processes in ESL/EFL essay tests: A narrative review of the literature. *The Canadian Modern Language Review, 64*(1), 99-134.

Bentler, P. M. (2006). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software, Inc.

Bentler, P. M. & Wu, E. J. C. (2005). *EQS 6.1 - Structural equation modeling software for Windows*. Encino, CA: Multivariate Software, Inc.

Breland, H. M. & Jones, R. J. (1984). Perceptions of writing skills. *Written Communication, 1*(1), 101-119.

Brennan, R. L. (2001a). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. L. (2001b). *Manual for urGENOVA Version 2.1* (Iowa Testing Programs Occasional Papers 49). University of Iowa.

Brown, J. D. (2004). Performance assessment: Existing literature and directions for research. *Second Language Studies, 22*(2), 91-139.

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly, 25*, 587-603.

Buck, G. & Tatsuoka, K. (1998). Application of rule-space methodology to language testing: Examining attributes of a free response listening test. *Language Testing, 15,* 118-142.

Campbell, D. T. & Fiske, D. W. (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.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18,* 65-81.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Connor-Linton, J. (1995). Crosscultural comparison of writing standards: American ESL and Japanese EFL. *World Englishes, 14*, 99-115.

Crawford, J. R. & Henry, J. D. (2003). The Depression Anxiety Stress Scales: Normative data and latent structure in a large non-clinical sample. *British Journal of Clinical Psychology, 42*, 111-131.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7,* 31-51.

Cumming, A., Kantor, R., & Powers, D. E. (2001). *Scoring TOEFL® essays and TOEFL® 2000 prototype writing tasks: An investigation into raters' decision-making and development of a preliminary framework* (TOEFL Monograph Series, Report No. 22). Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision-making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*(1), 67-96.

Daly, J. A. & Dickson-Markman, F. (1982). Contrast effects in evaluating essays. *Journal of Educational Measurement, 19*(4), 309-316.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*(3), 367-396.

DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, 5*(1), 7-29.

Diederich, P. J., French, J., & Carlton, S. (1961). *Factors in judgments of writing ability* (Educational Testing Service Research Bulletin No. 61–15). Princeton, NJ: Educational Testing Service.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197-221.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25,* 155-185.

Ekbatani, G. (2008). *Measurement and evaluation in post-secondary ESL*. New York: Taylor & Francis.

Elder, C., Barkhuizen, G., Knoch U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37-64.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly, 2*(3), 175-196.

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

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for ALL students: Development, implementation, and analysis* (pp. 261-287). Mahwah, NJ: Erlbaum.

Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions* (TOEFL® Research Report No. RR-70). Princeton, NJ: Educational Testing Service.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Chichester, UK: John Wiley & Sons, Ltd.

Freedman, S. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology, 71*(3), 328-388.

Gamaroff, R. (2000). Rater reliability in language assessment: The bug of all bears. *System, 28*, 31-53.

Geertz, C. (1973). *The interpretation of cultures*. New York: Basic Books.

Gelfand, S. A. (2010). *Hearing: An introduction to psychological and physiological acoustics* (5th ed.). London: Informa Healthcare.

Guilford, J. P. (1954). *Psychometric methods (2nd ed.).* New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hales, L. W. & Tokar, E. (1975). The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay question. *Journal of Educational Measurement, 12,* 115-117.

Hamp-Lyons, L. (1991). Reconstructing "Academic writing proficiency". In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 127-153). Norwood, NJ: Ablex.

Haswell, R. H. & Haswell, J. T. (1996). Gender bias and critique of student writing. *Assessing Writing, 3*(1), 31-83.

Hirsch, E. D., Jr. (1977). *The philosophy of composition*. Chicago: University of Chicago Press.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin, 57*, 116-131.

Hornberger, N. H. (2006). Negotiating methodological rich points in applied linguistics research: An ethnographer's view. In M. Chalhoub-Deville, C. A. Chapelle, & P. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 221-240). Dordrecht, Netherlands: John Benjamins.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1980). The influence of context position and scoring method on essay scoring. *Journal of Educational Measurement, 17*, 131-135.

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

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In B. Huot & M. Williamson (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206-236). Cresskill, NJ: Hampton Press.

Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfield, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach.* Rowley, Mass: Newbury House.

Jang, E. (2009a). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing, 26,* 31-73.

Jang, E. (2009b). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly, 6*(3), 210-238.

Jöreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In R. C. Atkinson, D. H. Krantz, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2, pp. 1-56). San Francisco: W. H. Freeman.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*(2), 125-160.

Kenyon, D. M. (1992). *Introductory remarks at symposium on development and use of rating scales in language testing*. Fourteenth Language Testing Research Colloquium, Vancouver, February 27-March 1.

Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26,* 187-217.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26-43.

Kobayashi, H. & Rinnert, C. (1996). Factors affecting composition evaluation in an EFL context: Cultural rhetorical pattern and readers' background. *Language Learning, 46*, 397-437.

Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly, 26*, 81-112.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(3), 1-31.

Lance, C. E., Dawson, B., Birkelbach, D., & Hoffman, B. J. (2010). Method effects, measurement error, and substantive conclusions. *Organizational Research Methods, 13*(3), 435-455.

Lattin, J., Carroll, D. J., & Green, P. E. (2003). *Analyzing multivariate data*. Pacific Grove, CA: Duxbury Press.

Lee, Y. & Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly, 6*(3), 239-263.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (2005) *Facets: Rasch-model computer programs* (Version 3.58.0). Chicago: winsteps.com.

Linacre, J. M. (2010). *A user's guide to Facets: Rasch-model computer programs.* Chicago: winsteps.com.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246-276.

Lumley, T. & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.

Lunz, M. E. & Linacre, J. M. (1998). Measurement designs using multifacet Rasch modeling. In G. A. Marcoulides (Ed.), *Modern methods for business research: Methodology for business and management* (pp. 47-77). Mahwah, New Jersey: Lawence Erlbaum.

MacMillan, P. D. (2000). Classical, Generalizability, and Multifaceted Rasch detection of interrater variability in large, sparse data sets. *The Journal of Experimental Education 68*(2), 167-190.

Marsh, H. W. & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.). *Structural equation modeling: Concepts, issues, and applications* (pp. 177-198). Thousand Oaks, CA: Sage.

Marsh, H. W. & Hocevar, D. (1988). A new, more powerful approach to multitrait-multimethod analyses: Application of second order confirmatory factor analysis. *Journal of Applied Psychology, 73,* 107-117.

McColly, W. (1970). What does educational research say about the judging of writing ability? *Journal of Educational Research, 64,* 147-156.

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

Milanovic, M., Saville, N., & Shen, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment* (pp. 92-114). Cambridge, UK: Cambridge University Press.

Mooi, E. & Sarstedt, M. (2011). *A concise guide to market research*. Berlin, Heidelberg: Springer-Verlag.

Muckle, T. J. & Karabatsos, G. (2009). Hierarchical generalized linear models for the analysis of judge ratings. *Journal of Educational Measurement, 46*(2), 198-219.

Murphy, K. R. & Cleveland, J. N. (1991). *Performance appraisal: An organizational perspective*. Boston: Allyn & Bacon.

Myford, C. M. & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 460-517). Maple Grove, MN: JAM Press.

Ostriker, J. P., Holland, P. W., Kuh, C. V., & Voytuk, J, A. (Eds.) (2011). *A data-based assessment of research-doctorate programs in the United States*. Washington, D.C.: The National Academies Press.

Penny, J., Johnson, R. L., & Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric. *Assessing Writing, 7*(2), 143-164.

Punji, G. & Stewart, D. W. (1983). Cluster analysis in marketing research: review and suggestions for application. *Journal of Marketing Research, 20*(2), 134–148.

R Development Core Team (2010). *R: A language and environment for statistical computing, reference index version 2.9.1.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rafoth, B. A. & Rubin, D. L. (1984). The impact of content and mechanics on judgments of writing quality. *Written Communication, 1*(4), 446-458.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks: Sage Publications.

Ryle, G. (1968). *The thinking of thoughts* (University Lectures 18). The University of Saskatchewan.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413-428.

Satorra, A. & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *ASA 1988 Proceedings of the Business and Economic Statistics Section* (308-313). Alexandria, VA: American Statistical Association.

Satorra, A. & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.

Satorra, A. & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507-514.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 26,* 465-493.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22*, 1-30.

Schwarz, N., Knäuper, B., Hippler, H-J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly 55*(4), 570-582.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental & quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing 18,* 303-325.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*(1), 27-33.

Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. In G. Brindley (Ed.), *Studies in immigrant English language assessment, Vol. 1* (pp. 159-189). Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Song, C. B. & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing, 5*, 163-182.

SPSS Inc. (2009). *PASW Statistics* (Version 17.0). Chicago, IL: SPSS, Inc.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*, 239-261.

Sweedler-Brown, C. O. (1985). The influence of training and experience on holistic essay evaluations. *The English Journal, 74*(5), 49-55.

Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition, 14,* 411-443.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Upshur, J. A. & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing, 16*, 82-111.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-25). Norwood, NJ: Ablex.

Wang, H. (2008). Decision making while scoring EFL tape-mediated speaking test performance. *CELEA Journal, 31*(4), 16-28.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11,* 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*, 263-287.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145-178.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10,* 305-335.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4*(1), 83-106.

Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication, 15*(4), 465-492.

Wolfe, E. W. & Ranney, M. (1996). Expertise in essay scoring. In D. C. Edelson & E.A. Domeshek (Eds.), *Proceedings of ICLS 96* (pp. 545-550). Charlottesville, VA: Association for the Advancement of Computing in Education.

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

Writing Group of Syllabus for TEM4-Oral. (2008). *Syllabus for TEM4-Oral (Revised ed.).* Shanghai: Shanghai Foreign Language Education Press.

Zamel, V. (1985). Responding to student writing. *TESOL Quarterly, 19*(1), 79-102.

Zhang, Y. & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing, 28*(1), 31-50.