Test Interactiveness of Holistic Scoring and Profile Scoring in EFL Classroom Writing Assessment:

A Different Perspective to Compare the Effectiveness of Two Grading Approaches

总评法与分项法的测试交互性

——大学英语课堂写作测试研究

张红霞 著

Zhang Hongxia

中国科学技术大学出版社

University of Science and Technology of China Press

Test Interactiveness of Holistic Scoring and Profile Scoring in EFL Classroom Writing Assessment:

A Different Perspective to Compare the Effectiveness of Two Grading Approaches

总评法与分项法的测试交互性

——大学英语课堂写作测试研究

张红霞 著
Zhang Hongxia

中国科学技术大学出版社
University of Science and Technology of China Press

图书在版编目(CIP)数据

总评法与分项法的测试交互性:大学英语课堂写作测试研究=Test interactiveness of holistic scoring and profile scoring in EFL classroom writing assessment:A different perspective to compare the effectiveness of two grading approaches/张红霞.—合肥:中国科学技术大学出版社,2006.12

ISBN 7-312-01783-5

Ⅰ.总… Ⅱ.张… Ⅲ.英语—写作—语言教学—教学研究—英文 Ⅳ.H315

中国版本图书馆 CIP 数据核字(2006)第 162812 号

出版发行 中国科学技术大学出版社
(安徽省合肥市金寨路 96 号,230026)
印　　刷 中国科学技术大学印刷厂
经　　销 全国新华书店
开　　本 700×1000　1/16
印　　张 13.25
字　　数 322 千
版　　次 2006 年 12 月第 1 版
印　　次 2006 年 12 月第 1 次印刷
印　　数 1—1000 册
定　　价 29.00 元

DEDICATION

This work is dedicated to the memory of my father
whose weak whispers of care have faded into the air

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE
INTRODUCTION

1.1 Context of the Problem

Teaching writing usually involves some direct measures of assessment, that is, direct assessment of writing performance. Although it may not be essential for instructors to read, comment upon, or grade every piece of writing that students produce for a course, a variety of assessment options are available for responding to writing assignments. Writing instructors can use an evaluation checklist to respond to students' writing, or can use a pass/fail system (hurdle) or a scale (ladder) (McNamara, 2000) to assign a grade for the overall writing or various aspects of writing. Instructors' comments can be fed back in written or verbal format. When grades are of necessity, there are a number of options available as well for instructors to make selections appropriate to their instructional purposes and courses objectives. Amongst the options, the holistic approach is the best known and most widely practiced in China, and the next best known might be the analytic or profile approach. The primary trait scoring and alternative assessments such as paper-based and electronic portfolios are much less known.

Which scoring approach to use, the holistic or the profile approach, has long been a question of concern for writing instructors and researchers. Both approaches have their own advantages and disadvantages. In holistic scoring, "a rapid overall rating" is made of writing (Shaw, 2002: 11). It is based on the overall impression of a continuous discourse according to its general properties (Elliot, Plata, & Zelhart, 1990: 17; Shaw, 2002: 11). It can save the trouble of getting into details when details are unnecessary. "A major advantage of holistic over analytic scoring is that each writing sample can be evaluated quickly by more than one rater for the same cost that would be required for just one rater to do the scoring using several analytic criteria" (Nakamura, 2004: 45). Put it simply, holistic scoring is generally more practical and efficient. Another advantage of this scoring approach is that rating holistically is closer to real-life reading behaviors and thus a more authentic and natural process than reading analytically (White, 1995). However, the holistic approach, though widely accepted in the field of writing assessment, has not been used on clear theoretical grounds (Huot, 1990a). Moreover, with only one assessment item of overall quality, holistic grading is potentially less reliable (Croker, 1999: 9). There are other aspects of holistic scoring that pose threats to its reliability and validity. For example, in actual practice, raters sometimes apply their idiosyncratic ratings in quality judgment (Nakamura, 2004: 45) rather

than observe general properties. Research has evidenced that holistic raters tend to be biased toward syntactic features or content while neglecting the overall rhetorical situation within which texts are produced (Huot, 1990a).

In the analytic scoring, evaluation is broken up into important dimensions or components (Shaw, 2002; Goldsby, 2004), and a set of grades are assigned to allow for uneven or jagged profiles (Hamp-Lyons, 1992; Shaw, 2002). Usually students within a single class are performing at different levels of writing achievement across different dimensions. For example, for one student, the grammar can be strong but the rhetorical organization and the vocabulary weak. For another student, though the overall meaning of writing is clear and well developed, the sentences are awkwardly structured. For a third student, the writing has a clear focus and good organization but the writing conventions such as spelling and punctuation can be poor. This is especially true with L2[①] students because different aspects of their writing ability tend to develop at different rates and stages (Johns, 1991; Hamp-Lyons, 1992; Tilbrook, 1996; Shaw, 2002). In these circumstances of uneven profile, it may be less useful to judge only the overall quality of the writing. Besides capturing various features of writing products, analytic scoring can also recognize different parts of the writing process such as planning and drafting (Johns, 1991, cited in Tilbrook, 1996). Moreover, producing a set of grades, or profile reporting, in relation to each important aspect of writing competence as identified in learning objectives can be more constructive and conducive to development of writing ability, and simultaneously contributory to assessment validity. It provides clear and rich diagnostic information about specific strengths and weaknesses in student writing, which is particularly useful and valuable for L2 students (Hamp-Lyons, 1992; Shaw, 2002; Goldsby, 2004). Judgments on various aspects of writing such as focus, development, organization, style and mechanics can help underdeveloped and developing students tell between higher- from lower-order concerns and thus can revision and refashion their writing purposefully and effectively, rather than focusing only on surface features of writing. However, with one single overall score generated from holistic scoring, students are often confused about what to address in rereading their own work. Undeniably, using analytic scoring does have some disadvantages. The main problem is that it is time-consuming in the processes of development, implementation, and scoring (Goldsby, 2004).

It is evident that each of the two grading approaches has its own distinctive strengths and shortcomings. They vary in practicality of application, quality of measurement, and their contributory relationship with teaching and learning. It seems paradoxical that the holistic scoring is more practical but tends to be less reliable or valid whereas the analytic grading is more reliable and valid but clearly less practical (Croker, 1999: 9). In the meanwhile, both

① The terms "L2" and "FL," "ESL" and "EFL" are used interchangeably in most cases in this paper, except that they are specifically referred to in the review of related literature.

approaches share some validity and reliability difficulties in implementation that pertain to rating scales (Fulcher, 1987; Matthews, 1990; Upshur & Turner, 1995). For example, both approaches involve subjective judgment (Douglas & Selinker, 1992; Bachman & Palmer, 1996; Croker, 1999), and it is not always clear that, even when using identical scales and arriving at like quality ratings, raters base their judgment on similar reasoning (Douglas & Selinker, 1992). On the other hand, some research since the 1960s has suggested that both scoring methods are valid and reliable, and even cost-effective means of assessing writing (Newbold, 1990: 4), and that evidence gleaned from using the holistic and analytic scoring instruments can inform test developers and writing instructors of examinees' proficiency levels (Shaw, 2002).

The above findings and contemplations of the two scoring approaches are sometimes in contradiction with each other. It seems difficult to draw any consistent or definite conclusions about them in comparison. In writing assessment, as in other fields of language assessment, there are many compelling problems which are inconclusive in answer. Among these problems are many aspects of direct writing assessment, especially grading issues (Huot, 1990a, 1990b & 1993). As of today, some of the issues remain largely unchanged. Although "a number of different criteria have been proposed over the years for rating different sorts of language performance assessments," however, issues such as "which of those already existing rating scales function well for different purposes and in different situations" and "how do analytic and holistic scoring methods compare in terms of effectiveness for different purposes and different situations" have not yet been clearly or conclusively answered (Brown, 2004: 122). It is not always possible to assert whether holistic rating is more or less useful than analytic rating as the practical consequences of any such differences which might occur between them are not always significant (Shaw, 2002: 11—12). It seems that there is no single best scoring system for all purposes (Perlman *et al.*, 1994).

Since there is no irrefutable or systematic information as to why particular scoring approaches worked or failed for specific instructional and assessment programs, "the choice of scoring method is not always easy" (Nakamura, 2004: 45). The relevant common sense is that whether to use the holistic approach or use the analytic scoring procedures with particular groups of students is dictated by the purpose of assessment and contextual needs. Different scoring approaches may serve different assessing purposes. Sometimes an overall impression can suffice for purposes such as placement and screening testing. Some other times, a profile of grades might be needed to provide writers sufficient information to do revisions. However, it is always easier said than done.

Though "the choice about the kind of rating scales to use is not always clear-cut. A useful approach to making a decision is to appeal to the Bachman and Palmer (1996) framework of test usefulness" (Weigle, 2002: 120). Bachman and Palmer hold that "the most important consideration in designing and developing a language test is the use for which

it is intended, so that the most important quality of a test is its usefulness" (1996: 17). Test usefulness has six qualities: reliability, construct validity, authenticity, interactiveness, impact, and practicality. "While these qualities are all important, it must be emphasized that it is virtually impossible to maximize all of them" (Bachman, personal communication, cited in Weigle, 2002: 48). Therefore, the choice of scoring procedures for writing assessment does not attempt to maximize each of the six qualities but should be based on deciding which qualities are most relevant in a given testing situation and determining the best possible combination of or the most appropriate balance among the six qualities (Bachman & Palmer, 1996; Weigle, 2002: 120).

A closer look at the six qualities of test usefulness reveals that not all the qualities have been researched or applied properly. Reliability and practicality have long been attended to in both research and practice since the prevalence of indirect writing assessment. From 1950s on, the validity problems in indirect writing assessment came to be widely recognized (Wiggins, 1992; Hamp-Lyons, 1992; Jones, 2001; Clark & Bamberg, 2003). Indirect assessment was not so much valid as it had been claimed to be. Assessment can be reliable without being valid (Croker, 1999), and validity is the more important determinant of assessment quality (Mabry, 1999). Since then, more and more attention has been given to validity. Assessing the quality of a test in relation to both validity and reliability has become the traditional standard practice (Jones, 2001: 2), and their importance in test design and test validation has been widely understood in the language testing world (Shaw & Jordan, 2002: 12). With the arising of process writing trend in rhetoric and composition in 1970s and 1980s (Emig, 1971; Hayes & Flower, 1980, etc.) and increasing attention to the communicative use of language (Canale & Swain, 1980; Bachman, 1990, Bachman & Palmer, 1996, etc.), using direct writing assessment and achieving high authenticity in assessment have become widely accepted. Proper authenticity of language assessment, or the extent to which assessment simulates real-world use of target ability, can ensure, in part, that the assessment actually measures what it intends to measure, and is hence more valid (Wiggins, 1992; Clark & Bamberg, 2003). Another test quality that has also been recently accentuated is impact, which relates to the washback effects of language tests on classroom teaching and learning and the effects of language tests on society at large (Bachman & Palmer, 1996: 29—30). Both authenticity and impact have been extensively used as quality indexes in validating UCLES① ESOL② tests (e. g., Saville, 2001; Shaw & Jordan. 2002; Taylor & Saville, 2002; Weir, 2002; Taylor, 2004; Bridges & Shaw, 2004). Among the six qualities of test usefulness, test interactiveness is the least researched or applied in test validation. It was first explicitly proposed as an essential test quality in Bachman and Palmer'

① University of Cambridge Local Examinations Syndicate.

② English for speakers of other languages.

s model of test usefulness (1996).

After recommending using the above model of test usefulness in decisions on rating procedures, Weigle further suggests that classroom writing assessment tend to be more concerned with construct validity, authenticity, interactiveness, and impact, than with reliability and practicality (2002: 175), which have driven the standardized testing of indirect writing and even the direct writing on one-to-two topics into shape and wide practice. According to Bachman and Palmer (1996), an interactional framework of language use should be considered in the assessment of language ability, which is particularly important in assessing writing ability. Writing is both product and process, and largely independent work. Authentic or performance-based writing assessment (Kim, 2002) should therefore take into account test takers' topical knowledge, language knowledge, affective schemata, and strategic competence. With all these considerations, we can have opportunities to more accurately evaluate and better interpret our student writers' performance in writing, and thus can make informed decisions about instruction and finally enhance learning.

In specific relation to holistic and analytic scoring procedures in writing assessment, Weigle offers comparative analyses in tabular form in terms of the test usefulness model (2002: 121). However, among the six quality aspects of the model, she makes no comparison with respect to test interactiveness between the two scoring methods, but leaves there the mark "n/a" for both approaches, which means "no account" or "not applicable." She only suggests that the interaction between the test taker and the test may be influenced by the rating scale if he/she knows how the writing will be evaluated.

It might be reasonable to think that the analytic grading approach would be more positive in promoting ESL writing test interactiveness than the holistic grading simply because the former provides more information for ESL students to write and revise. It could be so. However, what would be of persuasion and value is empirical evidence (Weigle, 2002; Nakamura, 2004). Whether using the analytic approach will produce more positive effects and how big the effect sizes will be, as compared with the traditional holistic approach, and whether it is worth for us to spend so much time in designing and using analytic grading specific for our students and assessment situation, are all questions that merit empirical explorations and answers.

A final issue that should be noted is that the majority of previous empirical studies on writing instruction and assessment have been based on statistical significance testing. It has been very extensively used in verifying or rejecting null hypothesis, and in analyzing and interpreting differneces in research outcomes, whether in the wider educational measurement circles or in language testing (e. g., in research on various aspects of writing instruction and assessment: Reid, 1992; Hedgcock & Lefkowitz, 1992; Sweedler-Brown, 1993; Tarone et *al.*, 1993; Devine, Railey, & Boshoff, 1993; Song & Caruso, 1996; Uzawa, 1996; Polio, Fleck, & Leder, 1998; Ferris & Roberts, 2001; Hirose, 2003, etc.). However, statistical

significance indicates the quality of research rather than the importance of comparative effect. More recent work has revealed that, to better meet the needs of modern science, such procedures are most often useful when used as an adjunct to other results such as effect sizes rather than as a stand-alone result (Robinson & Wainer, 2001). Actually, this understanding has been embodied rather impressively in the wider educational research (Madden, Stevens, & Slavin, 1986; Stevens, Slavin, & Farnish, 1989; Stevens & Durkin, 1992; Bramlett, 1994; Stevens & Slavin, 1995, etc.), but more slowly and much less commonly in the field of applied linguistics or language testing (Spencer, 1991 & 1999; Russell & Haney, 1997; Breland, Kubota, & Bonner, 1999; Thompson *et al.*, 2004a & 2004b; Breland, Lee, & Muraki, 2004; Wolfe & Manalo, 2004). To overcome the limitations of using only significance testing approach and to examine more accurately the effectiveness of intervention and the educational significance of any resultant effects, the interpretation of the outcomes in terms of statistical significance in this study would be coupled with effect size analyses.

Effect size (ES) analysis is a measure of the difference in outcome between intervention groups. It is a way of expressing the difference between two groups. In particular, if the groups have been systematically treated differently in an experiment, the effect size indicates how effective the experimental treatment is. It uses the idea of "standard deviation" to conceptualize the difference between the two groups. Effect size analysis emphasizes the most important aspect of an intervention, the magnitude of effect, rather than its statistical significance, which tends to conflate sample size and effect size. The underlying reason is that significance tests mainly depend on two things: the size of a sample and the size of effect. It would not be difficult to obtain a "significant" result if the effect was very big (despite having only a small sample) or if the sample was big (despite having only a tiny effect size). As a result, it may be dangerous to draw valid conclusions in either case, especially when the sample is small. On the other hand, effect size provides a standardized, scale-free measure of the relative importance of treatment effect, which is often accompanied with an estimate of its likely margin for error, or "confidence interval." For these reasons, effect size analysis is a helpful and important tool in reporting and interpreting effectiveness of treatment.

To sum up, studies examining comparatively different effects of holistic scoring and analytic scoring on learning in terms of test interactiveness are scant but imperative. Methodological imbalance in previous writing studies is another motivation that has led to the present writing assessment research.

1.2 Purpose of the Study

Mainly motivated by Weigle's suggestion that there may be a potential link between the use of rating scales and test interactiveness and that it is a question that should be empirically

addressed and answered (2002: 121), the present study attempted to look at a familiar question, the differences between holistic scoring and analytic scoring, but from a different perspective. It sought to investigate comparatively the effects of the holistic and the analytic grading approaches on the four components of test interactiveness, that is, language knowledge, topical knowledge, affective involvement, and strategic competence. It aimed to examine, in terms of learning gains, whether the two grading approaches would produce different results, or whether one of the scoring measures of writing assessment, the analytic profile and the holistic rubric, exerts more positive effects on our EFL students' interaction with writing prompts. It further investigated the sizes of the effects to provide evidence for their educational significance.

The present study was an intervention experiment, establishing different treatment status for two subject groups: one receiving the traditional holistic scoring as the control group, and the other receiving analytic scoring as the treatment group. Although the intervention instruction was relatively short, it was hoped that the current study would provide empirical information on test interactiveness of the profile approach and the holistic approach and thus enable us to have a complete picture of the usefulness of these testing instruments. It was also hoped that this study would generate useful information for decision-making in classroom writing assessment in order to provide a more supportive learning environment for our EFL student writers.

1.3 Research Questions

The paucity of research on test interactiveness of common measures of grading writing has left us an incomplete picture of the nature of the grading approaches and their influences on classroom learning. Therefore, a systematic, empirical examination of their comparative influences on learning would make much sense both in research and in practice.

The present research was situated in a college-level EFL learning context in China. Built upon the components of test interactiveness conceptualized in Bachman and Palmer (1996), it was to evaluate the efficacy of two different grading approaches, analytic vs. holistic, for second-year college English student writers. The umbrella research question asked in the present study is as follows:

> Does one of the two scoring measures of classroom writing assessment, the analytic profile and the holistic approach, have more positive effects upon test interactiveness, or EFL writers' interaction with writing prompts, than the other? If yes, what will be the sizes of the effects?

In relation to Bachman and Palmer's formulation of test interactiveness, the above cover question was broken down to four strands of secondary studies:

(1) Is there a significant difference in "language knowledge" between students who

receive analytic grading and those who receive traditional holistic grading? How big is the size of effect of the treatment?

(2) Is there a significant difference in "topical knowledge" between students who receive analytic grading and those who receive traditional holistic grading? How big is the size of effect of the treatment?

(3) Is there a significant difference in "affective involvement" in writing between students who receive analytic grading and those who receive traditional holistic grading? How big is the size of effect of the treatment?

(4) Is there a significant difference in "strategic competence" in writing between students who receive analytic grading and those who receive traditional holistic grading? How big is the size of effect of the treatment?

The answers that the above four secondary questions sought to find were actually evidence of two types, the observable or product-based evidence and the unobservable or process-based evidence. The first two questions were to assess students' writing proficiency as demonstrated in their writing outcome products, while the last two were to examine what was inside the student writers' minds while they were writing.

1.4 Significance and Rationale of the Study

Because of the prevalence of holistic scoring in our college English writing assessment, the questions of whether or not this grading approach facilitates learning and whether it, compared with the analytic grading method, exerts more positive effects on students' writing performance is of great importance. The present study is significant mainly in two ways. First, it will enable a better understanding of the importance of test interactiveness as an essential test quality as well as of the utility of the holistic and the analytic scoring approaches for a given assessment purpose in the given assessment situation. Second, it will also provide pedagogical insights for classroom writing instructors about effective ways to produce learning gains.

Since research on test interactiveness is sorely lacking, the findings of the study will provide precious empirical evidence of test interactiveness as an important test quality. It will also help to fill up the gap left thus far and present a full picture of test usefulness of the common measures of writing assessment under discussion. Whichever of the measures is to be found more positive and useful for learning, the research findings will have particularly important and practical implications for the development and validation of classroom assessment measures.

This study will also have benefits for writing pedagogy and writing learning. Hitherto, there have been no definite or systematic answers to the question in writing assessment about whether to adopt a holistic approach or an analytic approach. The choice should be dictated by