The CATESOL Journal

ANTHONY BECKER
Northern Arizona University, Flagstaff

Examining Rubrics Used to Measure Writing Performance in U.S. Intensive English Programs

A scoring rubric acts as a useful guide for evaluating the quality of students' written responses. In second language writing, scoring rubrics can be used to measure a variety of discourse and linguistic features. However, certain advantages and disadvantages are associated with particular rubrics (see Hamp-Lyons, 2003; Weigle, 2002). Therefore, numerous factors (e.g., purpose or resources) need to be considered when deciding which type of scoring rubric to use. This study describes the types and features of scoring rubrics that are used to measure English as a second language (ESL) students' writing in Intensive English Programs (IEPs) at multiple universities throughout the US. Forty-three IEP directors completed a questionnaire and interview that addressed the relevance/role of writing in their programs and the types/features of rubrics they use. The findings highlight some of the decision-making behaviors of IEP directors in their choices of scoring rubrics.

Introduction

Assessment of an examinee's performance on performance-based tasks (e.g., constructed-response essay) is often based on the judgment of experts, teachers, or trained raters. For classroom-based achievement tests, teachers are usually the primary judges of performance-based assessments (PBAs), while for proficiency tests, trained raters are often the primary judges of PBAs. The measurement of [writing] PBAs generally requires the assignment of a score, which is assumed to reflect the underlying construct or ability to be measured, relative to descriptors included in scoring rubrics. The rubrics that are commonly used to score writing PBAs include three main types: (a) analytic rubrics, (b) holistic rubrics, and (c) primary trait rubrics (Cumming, 1997; East & Young, 2007).

All three scoring types have certain advantages and disadvantages associated with their use. For example, although analytic scoring may improve reliability among measurements, the scoring of one individual trait can influence how another trait is scored (Xi & Mollaun, 2006). For holistic scoring,

one advantage is that there is an emphasis on positive aspects of an examinee's performance. However, holistic scoring typically offers little diagnostic information to identify test takers' strengths and weaknesses (Weigle, 2002). Finally, primary-trait scoring, while specific to a task, tends to be time consuming to develop. The consideration of advantages and disadvantages associated with using a particular scoring rubric are especially important when measuring L2 writing at Intensive English Programs (IEPs).

The success of [international] students at U.S. universities is largely dependent upon their ability to write (Rosenfeld, Leung, & Oltman, 2001). Therefore, an important objective for any IEP should be to ensure that students' scores reflect the writing ability of students and/or their readiness for mainstream university courses. To ensure this, numerous aspects of the rubrics must be considered. Aspects such as the features to include in the rubric, as well as the type and purpose of the scoring rubric, are just a few important considerations when developing and/or selecting a scoring rubric for measuring L2 writing. Without serious consideration of these aspects, it could be difficult for stakeholders (e.g., IEP administrators) to justify their decisions (e.g., advancing a student to another level).

Review of Relevant Literature

When attempting to assess what a test taker can actually do, PBAs are often used instead of objective testing. This is largely because of the perceived authenticity of PBAs and their potential for improving instructional practices (Bachman, 2002; Crooks, 1988). PBAs (such as writing a complaint letter, lab report, or research paper) offer the opportunity to closely mimic real-life situations, which can help to strengthen inferences made from a test task to the target language use domain—the situation in which a test taker will use the language outside of a test (Chapelle, Enright, & Jamieson, 2008). In an IEP, such assessments can be useful for indicating how well L2 students might perform on similar tasks found in mainstream university courses.

Scoring rubrics are typically used to score PBAs. The use of a scoring rubric is important for assessing writing performance because it "represents, implicitly or explicitly, the theoretical basis upon which [a] test is founded" (Weigle, 2002, p. 109). In the case of writing achievement, a scoring rubric can be used to indicate how well a student has achieved mastery of aspects of L2 writing (which might include conventions, structure, vocabulary, etc.) that were taught in a course or program. The three most common types of scoring rubrics (i.e., analytic, holistic, and primary-trait) used to score PBAs for writing are discussed below.

Types of Scoring Rubrics

Analytic Scoring. Analytic scoring includes individual traits, or components, of written expression. The analysis of several individual traits has prompted some researchers to label such scoring *multiple-trait scoring* (Hamp-Lyons, 2003). An analytic scoring rubric typically includes several writing components, such as accuracy, cohesion, content, organization, register, and appro-

priateness of language conventions (see Weigle, 2002), with each component being scored separately. Analytic scoring allows the rater(s) to focus on various aspects of an individual's writing and score some traits higher than others.

In L2 writing assessment, the first analytic scoring rubric to appear was the ESL Composition Profile (Jacobs, Zingraf, Wormuth, Hartfiel, & Hughey, 1981). This rubric, which provided the first conceptualization of scoring separate components for writing, consisted of five major analytic dimensions (i.e., development, organization, vocabulary, language use, and mechanics) designed to measure the writing of ESL students at North American universities. Today, the use of analytic rubrics to score writing continues to be prominent, as is evidenced by their use in the International English Language Testing System (IELTS), the Michigan Writing Assessment Scoring Guide (Hamp-Lyons, 1991), and earlier versions of Criterion (see Lee, Gentile, & Kantor, 2008).

As shown in Table 1, analytic scoring has several advantages and disadvantages. One of the greatest advantages is that the reliability of scoring is typically improved when raters use analytic rubrics. For instance, the results of several studies have found that analytic scoring provides the greatest chance for reliability between and within raters (Al-Fallay, 2000; East & Young, 2007; Knoch, 2009; Nakamura, 2004). This is largely believed to be true because analytic scoring enables raters to focus on and apply one scoring criteria at a time. In contrast, one major disadvantage of using analytic scoring is that the rating on one scale may influence the rating on another scale, referred to as the halo effect (Myford & Wolfe, 2003). This rater effect reflects the tendency of raters to let one trait influence evaluation of another trait. For instance, if a rater gives an essay a score of 1 for language use, that same rater may be influenced by this poor score, leading him or her to give another score of 1 for content.

Table 1
Advantages and Disadvantages of Analytic Scoring

Type	*Advantages*	*Disadvantages*
Analytic	Categories are not collapsed into one inflated score; can train raters easily (Cohen, 1994) Generalization to different writing tasks is possible (Weigle, 2002) Reliability is improved (Huot, 1996; Knoch, 2009) Can help to identify writers' strengths and weaknesses; provides diagnostic information (Bacha, 2001; Carr, 2000)	Rating on one scale may influence rating on another; scales may not be informative for respondents (Myford & Wolfe, 2003) Development can be time consuming and expensive (Hamp-Lyons, 2003; Weigle, 2002) Writing subskills cannot be separable (White, 1984) Raters may judge the scales holistically to match holistic impressions (Nakamura, 2004)

Because of its utility in providing diagnostic information, analytic scoring is often used in diagnostic testing. As analytic rubrics provide separate categories for writing components, they can help to identify the specific strengths and weaknesses of writers. Furthermore, analytic scoring is commonly used in classroom-based achievement tests, as this type of scoring aids in providing more directed feedback to students and teachers (Brown & Hudson, 2002).

Holistic Scoring. Holistic scoring takes the entire written response into account to assign an overall score for the performance. Instead of scoring writing components individually, these components are integrated into one impressionistic score. Holistic scoring generally places an emphasis on what is done well and not on what is lacking or deficient (White, 1985). For several well-known language tests, such as the Cambridge ESOL Exam and the Internet-based Test of English as a Foreign Language (TOEFL iBT), holistic rubrics are used to score examinees' written responses.

As shown in Table 2, several advantages and disadvantages are associated with using holistic scoring. The most widely recognized advantage of holistic scoring is its practicality (Weigle, 2002). Holistic scoring scales are relatively short and do not encompass several categories of criteria for which individual scores must be derived. As a result, holistic scoring is often considered a popular choice for rating PBAs. On the other hand, a major disadvantage of using holistic scoring is that it does not provide sufficient diagnostic information about examinees' writing. As Cohen (1994) notes, holistic rubrics provide only a composite score, which does not provide specific evidence of where and how much additional writing instruction is needed.

Table 2
Advantages and Disadvantages of Holistic Scoring

Type	*Advantages*	*Disadvantages*
Holistic	Emphasis is on what writers do well and not on deficiencies (Cohen, 1994) Validity is greater because it reflects authentic, personal reaction of reader (White, 1984) Scores are determined quickly (Weigle, 2002)	Scores do not provide diagnostic information; reliability is reduced (Song & Caruso, 1996) Scores can depend more upon the rater than upon text qualities (Hamp-Lyons, 2003) Information for deciding what to target next is insufficient (Nelson & Van Meter, 2007)

Holistic scoring is typically used for measuring written performance in large-scale testing situations. Large-scale tests (such as aptitude and placement tests) typically involve a large concentration of examinees taking the test at a given time. Therefore, because of its efficacy, holistic scoring is often used to make quicker, more efficient scoring decisions in these testing situations.

Primary Trait Scoring. The least common scoring type, primary-trait scoring, involves a decision about a single aspect that is central to the success of a writing task. For this type of scoring, the scoring rubric is developed in regards to a single feature of writing that is determined to be essential to a particular writing task. For example, a specific writing task might ask students to express their feelings from a particular point of view. The primary trait being scored for this task could include *use of dialogue*, *point of view*, or *tense aspect*, as these traits are considered necessary for successful completion of this particular writing task (Freedman, 1991).

As shown in Table 3, there are several advantages and disadvantages to using primary-trait scoring. The major advantage of primary-trait scoring is that it allows attention to be given to one writing trait at a time (Cohen, 1994). While this is similar to analytic scoring, primary-trait scoring differs from analytic scoring in that only one trait is focused on for the entire task. In contrast, the major disadvantage with primary-trait scoring is that it tends to be very time consuming. As Lloyd-Jones (1977) illustrates, with primary-trait scoring a scoring guide must be developed for every writing task, which can take "an average of 60 to 80 hours per task" (p. 38).

Table 3
Advantages and Disadvantages of Primary-Trait Scoring

Type	*Advantages*	*Disadvantages*
Primary-Trait	Attention is given to one writing aspect at a time (Cohen, 1994) Scale fits specific task at hand (White, 1985)	Scales are not integrative (Cohen, 1994) Development is labor intensive (Weigle, 2002)

It is uncommon to see primary-trait scoring being used in most testing situations. As Shaw and Weir (2007) indicate, due to "the lack of generalizability and the requirement to produce detailed rating protocols for each task, the primary trait approach is regarded as time-consuming and expensive to implement" (p. 149). As a result, primary-trait scoring is usually reserved for research situations or situations in which information is desired concerning learners' mastery of specific writing skills.

While understanding how the types of scoring rubrics are classified may be helpful when making decisions about which rubric to use, the purposes for using these rubrics are an equally important consideration (North & Schneider, 1998). This is particularly true for many IEPs, since these programs are often concerned with assessing students' proficiency (for placement and exit purposes) and achievement (for program advancement purposes).

Two common uses of scoring rubrics at IEPs are discussed below.

Uses of Scoring Rubrics

When choosing (or developing) scoring rubrics to measure L2 writing, the distinction between measuring proficiency and achievement is essential for deciding what should be included in the rubrics. In terms of measuring proficiency, the relevant knowledge, skills, and abilities that are considered most important for future learning need to be included in the scoring rubric. However, these underlying features of proficiency are not always concrete and are often difficult to define. Therefore, the scoring rubric represents the features that are part of the intended construct (Weigle, 2002). In contrast, the relevant content of a curriculum/syllabus/textbook are sought for inclusion in a scoring rubric designed to measure achievement. While language may be the medium through which this content is learned, the content represents what the scoring rubrics are designed to measure.

Rubrics to Measure Proficiency. Proficiency tests target the command one has of a target language at some point. Douglas and Chapelle (1993) assert that the results of such assessment represent the degree to which language learners have reached some level of language ability. In L2 writing assessment, an individual's writing ability is often based on the criteria included in the scoring rubrics. These criteria, as North and Schneider (1998) argue, should be grounded in some theory of language learning (e.g., Bachman & Palmer's [1996] model). However, the notion of writing ability is troubling to define. No general consensus exists for defining the construct of writing ability, and numerous models exist for understanding the nature of writing ability (see Weigle, 2002).

The lack of a clear model when defining L2 writing ability makes it particularly difficult to choose or design scoring rubrics to measure the writing proficiency of students. With numerous aspects of L2 writing to consider, IEPs may struggle to include the relevant features of writing that align with their curricular goals and objectives. This could ultimately affect the validity of such L2 writing tests, including the scoring validity (i.e., validity concerned with the reliability of test scores) and the construct validity (i.e., validity concerned with the construct being measured).

Rubrics to Measure Achievement. Achievement tests target the degree of achievement by individuals on a range of criteria related to a specific curriculum. Achievement assessments largely provide information about students' progress or readiness for subsequent levels of instruction (Hughes, 2002). In most IEPs, writing achievement is relevant to a course of study within a specific course syllabus or curriculum and is related to the outcomes and content of the local curriculum. However, it is sometimes challenging to ensure that the criteria used to describe examinee performance are clearly related to the goals and objectives of a given course (Brown & Hudson, 2002).

IEPs must be certain that the features they include in their scoring rubrics reflect what students have been taught in their IEP courses. It is important that rubrics for measuring writing proficiency are not substituted when choosing or developing scoring rubrics for classroom-based writing achievement assessments, as the criteria in these rubrics are often general and could fall short of measuring the content that IEP students have been taught.

In summary, PBAs are often used to measure IEP students' writing ability and achievement. To score such assessments, three different types of scoring rubrics can be used: analytic, holistic, and primary-trait. Each rubric type can be used to measure a variety of discourse and language features; however, numerous factors influence the inclusion (or exclusion) of these features. This study proposes to describe the types of scoring rubrics that are used to measure performance-based writing assessments in IEPs at multiple universities throughout the US and to understand the decisions that are made about rubrics, as well as the discourse/linguistic features that are deemed relevant in L2 writing assessment. This study seeks to address the following research questions:

1. What kinds of scoring rubrics are IEPs at selected U.S. universities using to measure writing performance?
2. What discourse/linguistic features are included in the scoring rubrics used by IEPs at universities throughout the US?
3. Does the inclusion of some features appear to be related to the type of test design (e.g., achievement or proficiency)?
4. How important are these features perceived to be for successful academic writing?

Method

Participants

Approximately 82 IEP directors at universities throughout the US were contacted for participation in this study. Of these, 43 IEP directors agreed to participate and voluntarily completed a questionnaire (discussed below).[1] The directors who responded to the questionnaire represented university programs from different geographical regions of the US: Midwest (n = 15), Northeast (n = 9), South (n = 6), Southwest (n = 8), and West (n = 5). At the time of the study, they reported being in charge of their respective IEPs for an average of 5.65 years (SD = 2.81). In addition, at the time of the study, they reported that the average number of second language students enrolled in their programs was 58.34 (SD = 17.66).

Materials

A 20-item questionnaire was developed for this study. The first part was intended to gather general information about the program. The second part addressed the relevance/role of writing in the program's curriculum. The final part was related to specific types and features of the rubrics used to measure writing in the program. The discourse and linguistic features included in the questionnaire were selected from an analysis of six different scoring rubrics that were provided by several IEP directors specifically for the present study.

In addition to the questionnaire, an interview was developed. The 15-minute interview consisted of 10 questions that were meant to probe for additional information related to IEP directors' responses to the questionnaire. The first two questions were related to the decisions made by the IEPs about their writ-

ing rubric(s). The remaining eight questions addressed the features that were included in the IEPs' writing rubric(s).

Design and Procedure

The web addresses of 82 IEP programs were collected through an online program directory (www.opendoors.iienetwork.org) provided by the Institute of International Education and the Teachers of English to Speakers of Other Languages (TESOL) Listserv. A participation-request email was sent to each of the 82 IEP directors. The email provided a brief explanation of the purpose of the study and a request for their participation in the study. If the directors agreed to participate, they were advised to complete the questionnaire, which was attached as a Microsoft Word document, and to return it (as an attachment) to the primary researcher. Respondents were also asked to provide a copy of their IEP's writing rubric(s) as an additional attachment.

Once the questionnaires and rubrics were collected and analyzed, all IEP directors were contacted for participation in a follow-up interview. The researcher called those directors (N = 18) who agreed to the interview, with all correspondence taking place via Skype (www.skype.com). The audio for the interviews was recorded using a recording option provided by the Skype software. In addition, the researcher wrote down segments of the directors' responses to the interview questions.

Data Analysis

The frequencies for all of the questionnaire responses were summed to provide total counts (which are reported in the Results section). Qualitative information from the interviews was transcribed verbatim by the researcher. Initially, the researcher read through the transcriptions to get a sense of the overall data. Afterward, the researcher followed Creswell's (2002) process of data analysis and interpretation. This framework for data analysis took place in five procedural steps:

1. The researcher carefully read through each interview and wrote ideas that came to mind in the margins.
2. While doing a second reading, the researcher bracketed words and phrases that were deemed important (i.e., words/phrases that relate to decisions about scoring scales). In addition, key words or phrases were written near the bracketed segments in the margins to provide codes to describe the bracketed materials.
3. Once the bracketing and coding were complete, the researcher made a list of the code words and began clustering similar or redundant codes.
4. Afterward, the researcher returned to the transcribed data and identified specific quotes that supported the codes.
5. The researcher then collapsed the list of codes into several themes formed from the initial codes. From the coding and the themes, a narrative description of the data was constructed.

Because of time constraints, an additional rater could not be consulted for coding decisions; as a result, interrater reliabilities (or rater agreement) could not be reported.

Statistical Analysis. A chi-square test of independence was performed to determine the relationship between the type of test design and the discourse/linguistic features included in the scoring rubrics. For each of the discourse/linguistic features (i.e., accuracy, content, coherence, complexity, grammar, language use, organization, vocabulary, and structure/syntax), each school's test-design type (e.g., achievement or proficiency) was coded: Achievement tests received a code [1] and the proficiency tests received a different code [2]. The chi-square test of independence for each feature was performed using the Statistical Package for the Social Sciences (SPSS), version 17.0.

Results

The following is a summary of the findings from the questionnaires and the IEP director interviews used in the present study. The results are reported as they pertain to the four research questions outlined earlier.

RQ1: What kinds of scoring rubrics are IEPs at selected U.S. universities using to measure writing performance?

Several items from the questionnaire and interview were used to address this research question. Table 4 presents the types of rubrics and test designs that the IEP directors indicated using to measure writing performance in their programs. The table indicates that no other scoring rubrics were reported to be used by the IEP directors for measuring writing performance, other than analytic and holistic scoring rubrics. The number of holistic rubrics (n = 26) is almost twice that of the number of analytic rubrics (n = 15). Interestingly, 2 directors indicated that they do not use any scoring rubrics for measuring writing performance. Instead, they indicated that their instructors make pass/fail judgments about students' writing performances. The table also indicates that writing proficiency was the construct of interest for most IEPs. In fact, the number of scoring rubrics designed to measure writing proficiency (n = 25) is almost twice that of the number of scoring rubrics designed to measure writing

Table 4
A Comparison of Rubric Types and Test Designs

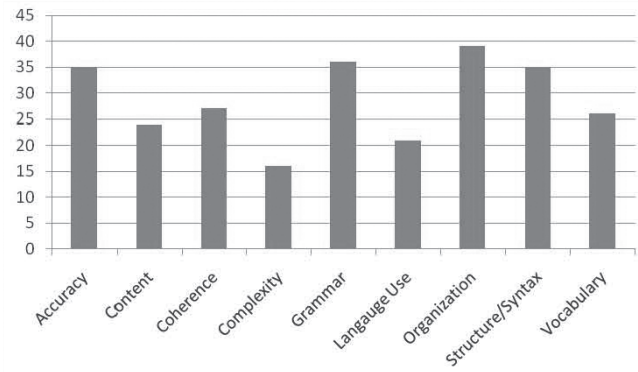
	Scoring rubric type				
Test design	Analytic	Holistic	Other	Total	%
Achievement	10	6	0	16	37
Proficiency	5	20	0	25	63
Other	0	0	0	0	0
Total	15	26	0	41	--
%	35	65	0	--	100

achievement (n =16). Finally, analytic scoring rubrics (n = 10) appeared to be used more than holistic scoring rubrics (n = 6) to measure writing achievement, while holistic scoring rubrics (n = 20) were used much more frequently than analytic scoring rubrics (n = 5) to measure writing proficiency.

RQ2: What discourse/linguistic features are included in the scoring rubrics used by IEPs at universities throughout the US?

Figure 1 illustrates the frequency of the specific discourse/linguistic features indicated by the IEP directors. As shown in the figure, organization (n = 39) and grammar (n = 36) were the most frequent features included in the scoring rubrics, followed by accuracy (n = 35) and structure/syntax (n = 35). The IEP directors reported that language use (n = 21) and complexity (n = 17) were the least frequent features included in the scoring rubrics. It is worth noting that 3 IEP directors also indicated that *fluency* was a feature included in their scoring rubrics. In addition, 2 directors stated that thesis statement and topic sentence formation were features included in their rubrics.

Figure 1
A Chart Comparing the Frequency of Discourse/Linguistic Categories



RQ3: Does the inclusion of some features appear to be related to the type of test design (e.g., achievement or proficiency)?

To determine if there were a relationship between the test design and the different discourse/linguistic features, a chi-square test of independence was performed. The hypothesis (and accompanying null hypothesis) for each of the discourse/linguistic features was as follows:

$H_{1\text{-}9}$: There is a sig. difference between the achievement and proficiency test designs.
H_0: There is no sig. difference between the achievement and proficiency test designs.

Table 5 presents the results of the nine chi-square analyses that were conducted to test these hypotheses. As indicated in the table, there were significant

chi-square values only for content [χ^2 (1, N = 23) = 5.26, p = .02] and language use [χ^2 (1, N = 21) = 3.86, p = .04]. Therefore, since these values are significant, the null hypothesis for these two features would be rejected and the experimental hypothesis for these two features would be accepted. In other words, for content and language use, there is a significant difference between the number of writing achievement and writing proficiency tests that include the measure of these two features.

Table 5
Summary of Chi-Square Analyses for Discourse and Linguistic Features

Feature	*Chi-square statistic (χ^2)*	*p*	*Significant*
Accuracy	.71	.40	No
Content	5.26	.02	Yes
Coherence	3.52	.06	No
Complexity	.25	.62	No
Grammar	.44	.51	No
Language use	3.86	.04	Yes
Organization	.23	.63	No
Structure/syntax	.71	.40	No
Vocabulary	1.96	.16	No

Note. *Alpha levels were set at .05 for all chi-square analyses.*

RQ4: How important are these features perceived to be for successful academic writing?

Table 6 indicates the number of IEP directors who considered the various discourse and linguistic features to be important and/or not important for

Table 6
Summary of Agreement Among IEP Directors (*N* = 18)

Feature	*Important*	*Not important*	*Agreement of importance (%)*
Accuracy	17	1	94
Content	12	6	67
Coherence	6	12	33
Complexity	7	11	39
Grammar	13	5	72
Language use	10	8	56
Organization	18	0	100
Structure/syntax	9	9	50
Vocabulary	18	0	100

successful academic writing. From the table, all of the directors agreed that *organization* and *vocabulary* were important for successful academic writing, while almost all of the directors agreed that *accuracy* was important. Approximately half of the directors agreed that *language use* and *structure/syntax* were important for successful academic writing, while around one-third of directors agreed that *coherence* and *complexity* were important in academic writing.

Discussion

This study proposed to describe the types of scoring rubrics that are used to measure performance-based writing assessments in IEPs at multiple U.S. universities and to understand the decisions that are made about rubrics, as well as the discourse/linguistic features that are deemed relevant in L2 writing assessment. These issues are discussed as they relate to the findings of this study.

Preference for a Scoring Rubric

The results from the questionnaires make it clear that holistic scoring rubrics are preferred over analytic rubrics, especially for measuring the writing proficiency of IEP students. This finding seems to align with the findings from much of the writing assessment literature (Cohen, 1994; Weigle, 2002). However, the reasons for choosing the holistic rubrics over the analytic rubrics were not consistent among IEP directors. For instance, 1 IEP director indicated that

> We prefer to use a holistic approach because instructors, uh or teachers, seem to be much less intimidated when they only have to look for one score. You see, I think they feel more comfortable knowing that there is really just one score for them.

This implies that ESL instructors are seemingly overwhelmed when having to score several criteria separately. According to this IEP director, having to derive a single score for a writing performance helps alleviate some of the anxiety that raters might potentially encounter.

Another motivation for choosing holistic scoring rubrics is related to efficiency. As another director said,

> Well, holistic [scoring] is so much easier to do and the teachers don't have to spend much time on their scoring. In our program, we have nearly 300 [ESL] students, with multiple tests. Using this [holistic] scoring, we can do things more quickly.

It is evident from this IEP director's response that holistic scoring appears to be less time consuming than other scoring approaches. This same director commented that she had some familiarity with multiple-trait scoring (i.e., analytic scoring), and it was her impression that this type of scoring approach required additional time to train teachers and have them score essays.

In contrast, several IEP directors indicated the utility of using analytic scoring rubrics instead of holistic rubrics. Specifically, 3 directors commented that the analytic rubrics appear to be much more informative about their students' writing and that these rubrics helped to guide teachers' feedback about students' weaknesses. One of the directors indicated that

> These analytic rubrics take a little bit more time to look at, but, but they give a lot more information. When you look at them, they have specific points about each student's writing. So, the teacher, the teacher can see exactly where a student is strong, uh, and where a student is weak.

These various responses make it apparent that some IEP directors have carefully considered the use of which rubrics to use for scoring students' writing performance. These considerations seem to center around the time it takes to score the essays and the information that the rubrics provide about what students do well, as well as where they need help.

Features of a Scoring Rubric

The scoring rubrics that IEP directors commented on included a variety of discourse and linguistic features. While the types of features (e.g., subject-verb-object ordering for structure/syntax) were not specified by most directors, the general categories mentioned earlier did give some indication of the features that were thought to be most important. For instance, organization and grammar were found to be the most frequent features included in the rubrics, followed by accuracy and structure/syntax. When asked why these features were included, most directors commented that they seemed to be appropriate aspects of L2 writing.

However, when asked which features were most important for successful academic writing, the perspectives of some IEP directors appeared to differ. For instance, 1 director suggested that

> Having a firm grasp of the content that is learned is vital. If a student can't master, or at least learn, the materials in their courses, they won't be able to do well. Also, I guess organization is pretty important, too. It seems that composition courses want students to have a good idea of how to organize their papers.

Interestingly, this same IEP director did not indicate that organization was a feature included in the scoring rubrics used by his IEP. Perhaps this director assumed that organization was something learned and practiced in class but not readily assessed in a writing test. Meanwhile, another IEP director determined that accuracy and vocabulary were features that were most integral to successful writing. This director said,

> In university comp[osition] courses, instructors want language to be accurate, as well as the information that you present. They don't want writing to

> be erroneous, [be]cause that will influence how they read the paper. Also, they want sources and facts to be accurate. … I imagine that composition instructors require the use of rich vocabulary too. You need to express yourself using concepts related to the topic you're writing about.

Based on this response, the IEP director developed a clear notion that accuracy and vocabulary are fundamental aspects of successful writing. However, it is also interesting to note that this same director did not indicate that accuracy is a measure of writing included in her IEP's scoring rubrics. When this same director was asked how the inclusion of features was decided for the scoring rubrics, she responded that

> When looking for a textbook, you know, to use at our program, the staff honestly didn't put much thought into the rubrics. But, after we got the textbooks, well, we looked through the existing rubrics that [the publisher] provided and we decided to include what we thought was important. In the end, we decided that certain grammar points were important for students, but we were more concerned with making sure they could organize their writing and didn't have too many mistakes, or errors, in their actual essays.

This director's response suggests that IEP staff actually corroborated on what to include in their rubrics, but only after they had decided which textbooks they would use to shape their curriculum.

It is worth noting that this IEP director discussed the adaptation of an already existing scoring rubric. This director was certainly not alone in this approach, as 27 IEP directors (of the 43 who responded) indicated they borrowed and/or adapted scoring criteria, descriptors, and scoring schemes (e.g., the number of scoring bands) from already existing scoring rubrics. Of this group, approximately 14 directors reported borrowing and/or adapting information from the TOEFL iBT scoring rubric. Several of these directors commented that they believed the TOEFL rubrics are likely to be reliable instruments, and so they thought the rubrics were appropriate to use for scoring writing.

Surprisingly, only 5 IEP directors indicated that their programs had developed their own scoring rubrics for writing. The directors from these five programs indicated that they were in close contact with assessment experts (either through professional association or because these experts were faculty members at the same university) who could oversee their efforts in developing their own scoring rubrics. However, these same IEP directors also indicated that these efforts were largely time consuming and required a substantial investment of time and money for their respective programs. The efforts of these IEPs highlight the capability of teachers and directors to get involved in the development of scoring rubrics that are directly relevant to their interests (Turner & Upshur, 2002).

Implications

The findings of this descriptive study have implications for research and for pedagogical practices. In terms of research, in demonstrating the different

views of IEP directors about the use of analytic and holistic rubrics, these findings are compatible with previous research (e.g., Bacha, 2001; Carr, 2000; East & Young, 2007) that highlights the perceived advantages and disadvantages of both scoring types. In this way, the findings reflect the necessity for IEPs to carefully consider the scoring methods being used, as a particular rubric can offer certain benefits that another rubric cannot (e.g., providing detailed feedback about writing).

These findings also underscore the value of understanding the features of writing that are perceived to be important for successful academic writing. The IEP directors in this study viewed several particular features of writing (e.g., accuracy, organization, and vocabulary) to be important for successful academic writing. This is along the same lines as Rosenfeld et al.'s (2001) report that command of these three aspects is necessary for writing at both the undergraduate and graduate levels in U.S. universities. However, since these features of writing can be viewed differently, the ways in which these three features (as well as other writing features) are defined in scoring rubrics needs to be closely examined. Furthermore, to validate the IEP directors' perceptions about what is important for successful writing, an analysis of the actual features included in students' writing might prove useful.

The findings of this study also have practical implications for educators and IEPs. Graduate programs (e.g., in TESOL or rhetoric/composition) and IEPs that work with writing teachers may find it worthwhile to provide a course and/or training in assessment or evaluation. Typically, very little attention is devoted to assessment. As Weigle (2007) indicates,

> Courses on teaching writing often devote only a limited amount of time to the discussion of assessment. Moreover, teachers often feel that assessment is a necessary evil rather than a central aspect of teaching that has the potential to be beneficial to both teacher and students. As a result, teachers sometimes avoid learning about assessment, or, worse, delay thinking about how they will assess the students until they are forced to do so. (p. 194)

If teachers do not receive adequate assessment training, it is difficult to expect them to make justified decisions about how to effectively assess their students' writing. With this in mind, graduate programs and IEPs might consider implementing assessment courses and/or training components that highlight important considerations for assessment (e.g., reliability and validity) and the test-development process, including how to develop and score writing tasks.

Finally, there is an obvious need for IEP administrators and teachers to make informed decisions about their assessment practices. As the results of this study indicate, decisions made about assessment are often the result of practicality. However, assessment decisions must also be based on theory, as what is practical is not always what is best for our students and teachers. Therefore, when developing and using assessment procedures, both theory and practicality should be carefully considered.

Conclusion

While there is a large body of research comparing the potential advantages and disadvantages of using certain types of scoring rubrics, there is very little research about the decisions behind choosing these rubrics. These decisions are important to understand, as the use of certain rubrics appears closely related to the construct being measured (e.g., achievement or proficiency) and the measures of writing deemed to be important. The findings of this descriptive study help to highlight some of the decision making of IEP directors in their choices of scoring rubrics used to measure their students' writing performance. One can hope that these findings will help to raise awareness of the numerous decisions that must be made, resulting in more informed decisions that can lead to more useful assessments of students' writing.

Author

Anthony Becker is a doctoral candidate at Northern Arizona University in Flagstaff in the Applied Linguistics Program, where he works as a graduate assistant in the Program of Intensive English. His research interests are second language assessment, cognitive aspects of writing, and teacher training and development.

Note

[1]Forty IEP directors responded to an individual email sent out by the primary researcher. Three additional directors responded to a posting on the Teachers of English to Speakers of Other Languages (TESOL) Listerv.

References

Al-Fallay, I. (2000). Examining the analytic marking method: Developing and using an analytic scoring schema. *Language & Translation, 12*, 1-22.

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System, 29*, 371-383.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.

Bachman, L. F. (2002). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice, 21*, 5-18.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge, England: Cambridge University Press.

Carr, N. (2000). A comparison of the effects of analytic and holistic rating scale types in the context of composition tests. *Issues in Applied Linguistics, 11*, 207-241.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the test of English as a foreign language*. Mahwah, NJ: Lawrence Erlbaum.

Cohen, A. (1994). Assessing written expression. In *Assessing language ability in the classroom* (pp. 303-357). Boston: Heinle & Heinle.

Creswell, J. W. (2002). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Upper Saddle River, NJ: Pearson Education.

Crooks, T. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 38*, 438-481.

Cumming, A. (1997). The testing of writing in a second language. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education*: *Language testing and assessment* (Vol. 7, pp. 131-139). Norwell, MA: Kluwer Academic.

Douglas, D., & Chapelle, C. (1993). Foundations and directions for a new decade of language testing. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 1-24). Alexandria, VA: TESOL.

East, M., & Young, D. (2007). Scoring L2 writing samples: Exploring the relative effectiveness of two different diagnostic methods. *New Zealand Studies in Applied Linguistics, 13*, 1-21.

Freedman, S. W. (1991). Evaluating writing: Linking large-scale testing and classroom assessment. *Center for the Study of Writing* (Occasional Paper, 27). Berkeley: University of California.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241–276). Norwood, NJ: Ablex.

Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 162-189). Cambridge, England: Cambridge University Press.

Hughes, A. (2002). *Testing for language teachers*. Cambridge, England: Cambridge University Press.

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication, 47*, 549-566.

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

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*, 275-304.

Lee, Y-W., Gentile, C., & Kantor, R. (2008). *TOEFL CBT essays: Scores from humans and E-rater* (TOEFL Research Rep. No. RR-81). Princeton, NJ: Educational Testing Service.

Lloyd-Jones, R. (1977). Primary trait scoring. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing* (pp. 33-69). New York: National Council of Teachers of English.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-faceted rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386-422.

Nakamura, Y. (2004). *A comparison of holistic and analytic scoring methods in the assessment of writing*. Retrieved February 9, 2009, from http://jalt.org/pansig/2004/HTML/Nakamura.htm

Nelson, N. W., & Van Meter, A. M. (2007). Measuring written language ability in narrative samples. *Reading & Writing Quarterly, 23*, 287-309.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*, 217-263.

Rosenfeld, M., Leung, S., & Oltman, P. (2001). *The reading, writing, speaking,*

and listening tasks important for academic success at the undergraduate and graduate levels (TOEFL Monograph No. MS-21). Princeton, NJ: Educational Testing Service.

Shaw, S. D., & Weir, C. J. (2007). *Examing writing: Research and practice in assessing second language writing.* Cambridge, England: Cambridge University Press.

Song, 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.

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from students samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly, 36*, 49-70.

Weigle, S. C. (2002). *Assessing writing*. New York: Cambridge University Press.

Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing, 16*, 194-209.

White, E. M. (1984). Holisticism. *College Composition and Communication, 35*, 400-409.

White, E. M. (1985). *Teaching and assessing writing*. San Francisco: Jossey-Bass.

Xi, X., & Mollaun, P. (2006). *Investigating the utility of analytic scoring for the TOEFL academic speaking test (TAST)* (TOEFL iBT Monograph No. 01). Princeton, NJ: Educational Testing Service.