The Conundrum of Classroom Writing Assessment

Drs. Joseph F. Engemann and Tiffany Gallagher

Brock University

Abstract

This article highlights the need for reliable, valid, and fair assessment of writing performance by classroom teachers to balance large-scale literacy assessment initiatives that have taken root in most jurisdictions across Canada. While these initiatives have established a focus on writing instruction and assessment, there is a notion that large-scale assessment may not, by itself, provide an accurate reflection of students' writing performance. It is important, therefore, for teachers to use effective assessment instruments, such as rubrics, to reliably, validly, and fairly determine how well their students write. The grading of written compositions through the use of rubrics can be effectively implemented by teachers with sufficient training whilst providing detailed student feedback. Classroom assessment of writing must therefore be coupled with professional development and teacher collaboration, which should lead to an improvement in student writing proficiency.

Dr. Joseph Engemann is a member of the Department of Graduate and Undergraduate Studies in Education at Brock University. He specializes in research on classroom assessment and evaluation, teacher induction and mentoring, and aboriginal science education.

Dr. Tiffany Gallagher is a member of the Department of Pre-service Education at Brock University. Her current research interests include literacy assessment, reading and writing strategy instruction, teachers with learning disabilities, and the role of the in-school resource teacher.

Long regarded as one of the “three Rs,” along with reading and arithmetic, writing is considered to be one of the essential academic skills that students acquire as an intended function of schooling (Barakett & Cleghorn, 2000). In keeping with this commonly accepted notion, there is a current culture of accountability in education, which emphasizes standards and evaluation of these basic academic skills (Hunter, Jones, & Randhawa, 1996). In all Canadian provinces and territories except Prince Edward Island and Nunavut, large-scale writing assessment is conducted either annually or over a two- or three-year period as a way of measuring the writing proficiency of students (Airasian, Engemann, & Gallagher, 2007). While claiming to address the need for public accountability, it has been argued that large-scale assessment is woefully inadequate at providing the detailed diagnostic evidence and formative recommendations needed to improve classroom pedagogy and student achievement (Froese-Germain, 1999; Simner, 2000).

According to the *Standards for the Assessment of Reading and Writing* (IRA/NCTE Joint Task Force on Assessment, 1994), valid assessment of writing should involve multiple perspectives and sources of data. Generally, large-scale writing assessment relies on a ‘one-off’ writing sample that is used as the reflection of students’ writing proficiency (IRA/NCTE Joint Task Force on Assessment, 1994). This single writing opportunity places pressure on classroom teachers to coach their students for these assessments (Mabry, 1999). Some teachers adjust their instructional practices and calibrate their writing assessment tools to approximate the protocols of large-scale assessments (Hunter et al., 1996). Other teachers find themselves grading students’ writing samples and providing them with constructive feedback specific to the criteria of the large-scale assessments (Earl, 1999; Skwarchuk, 2004).

Undoubtedly, teachers are eager to undertake practices that will provide them with the most accurate and efficient way of assessing their students’ writing. The focus of this article, therefore, is to provide an overview of classroom writing assessment with a focus on the use of rubrics. The need for teachers to ensure that writing assessment is reliable, valid, and fair will be emphasized.

An Overview of Classroom Writing Assessment

Writing may be considered one of the most difficult proficiencies that students are expected to master in elementary school (Gunning, 2002). Furthermore, classroom teachers typically find assessment of students' writing performance onerous, time-consuming, and subject to the vagaries of imprecision and subjectivity (Culham, 2003). As a result, there is a resurgence of interest in models of instruction that assist teachers in defining the components of good writing and that provide a structure for assessment that connects with identified characteristics (Culham, 2003; Spandel & Hicks, 2002). Currently, many teachers are searching for more reliable and valid methods of assessment of writing performance that they can employ as part of their regular classroom practice.

Over the past four decades, teachers have tended to assess writing with holistic grading methods (Hunter et al., 1996). Holistic grading involves making an assessment of the quality of a complete written composition against a prepared scale or rubric (Hunter et al., 1996). It involves assigning a global grade (percentage, rating, letter, etc.) as a measure of the students' level of writing performance. Holistic grading for writing is appropriate when the purpose of assessment is to obtain a broad perspective about the writing proficiency of a student or when the written composition cannot be assessed according to distinct criteria (Moskal, 2000). It is preferred by many teachers as a quick and efficient way of assessing students' writing. By contrast, analytic grading involves the breaking down of a written composition into components, each of which is assessed separately and then amalgamated with the scores from other components to derive an overall grade. Analytic grading can provide a more comprehensive outline of the strengths and weaknesses of students' writing performance than holistic grading, but it is detail-oriented and thus more time-consuming.

Reliable, Valid, and Fair Writing Assessment

Concerns about the reliability and validity of writing assessment are ever-present in the minds of teachers. For example,

some educators regard holistic grading as lacking uniform precision since there is a requirement to globally judge students' writing compositions. Indeed, without precise assessment tools, teachers may assess written compositions subjectively and inconsistently. Yet, there is evidence that holistic scoring measures can be highly reliable when graders are extensively trained in the application of the measures. In fact, in these cases, there can be strong correlations between holistic and total analytic scores (Hunter et al., 1996). Regardless of the type of method used to grade writing, if there is a lack of alignment between writing instruction and assessment, teachers may be led to assess components that were not included as an instructional focus. Teachers need to look long and hard at assessment measures to ensure that the outcomes described reasonably reflect their instructional objectives. Effective writing assessment protocols must be consistent, accurate, and reasonable or, in other words, reliable, valid, and fair.

Reliability

Reliability refers to the degree of stability or consistency of assessment among multiple raters at one point in time or a single rater at different points in time (Airasian et al., 2007). It also refers to the accuracy of an assessment tool itself (Isaac & Michael, 1997). The goal in assessment is to establish a consensus in the application of an assessment tool. In writing assessment, this may be accomplished through the use of a tool that has clearly articulated criteria and the implementation of in-service training sessions to establish high levels of inter-rater reliability (Stuhlmann, Daniel, Dellinger, Denny, & Powers, 1999). Ideally, these sessions should encourage teachers to hone their skills in consistently applying assessment tools.

Validity

Assessment tools must be used in instructionally valid ways. Validity is the extent to which assessment data are appropriate for making a decision about students and instruction (Airasian et al., 2007). An assessment tool must have content validity; that is, it must contain criteria and descriptors that accurately describe the students' products or performances. In classroom settings, an important

consideration is instructional validity, which is the degree to which the assessment measures the classroom instruction (Santrock, Woloshyn, Gallagher, DiPetra, & Marini, 2004). This implies that assessment tools should measure what was taught and whether students have had adequate opportunity to learn the skills and obtain the knowledge that is being assessed. At times, grading students' written compositions requires evaluative judgment and there is a temptation to subjectively utilize assessment tools (Aiken, 2000). To reduce teachers' misuse and misinterpretation of writing assessment measures, they must be provided with assessment standards that are easily identifiable and can be applied to every student in accordance with instructional objectives (Banks, 2005).

Fairness

In general, fair and equitable assessment is the ethical responsibility of all teachers. In Canada, the criteria for determining fairness and equitability in student assessment have been outlined as principles and guidelines in *Principles for Fair Student Assessment Practices for Education in Canada* (Joint Advisory Committee, 1993). This document was produced through a collaborative effort among many recognized organizations within the Canadian educational community and has been supported by the Canadian Teachers Federation, the Canadian School Boards Association, and the Canadian Society for the Study of Education. This framework outlines principles for classroom assessment that guide teachers in developing and choosing methods for assessment, collecting assessment information, judging and scoring student performance, and summarizing, interpreting, and reporting assessment findings (Joint Advisory Committee, 1993). An extrapolation of this framework into the assessment of written compositions means that teachers must adopt reasonable standards of writing assessment while, at the same time, maintain respect for students' individuality in writing characteristics such as tone or voice. In addition, the IRA/NCTE Joint Task Force on Assessment (1994) states that equity should be pursued in writing assessment, a goal that is unlikely to be achieved through the use of a single assessment. Instead, several varied assessment pieces are

required to provide a meaningful measure of the writing capabilities of students.

The Birth of the Rubric

One of the purposes of assessment is to encourage students' skill development by providing evaluative feedback; in writing assessment, rubrics are particularly effective for this purpose (Linn & Miller, 2005). A rubric is a set of criteria for different levels of performance (Airasian, 2005; McMillan, 2004). Rubrics, like the grading strategies that use them, may be identified as either holistic or analytic. A holistic rubric focuses on grading in a more general and overall fashion as the written composition is assessed in its entirety (Linn & Miller, 2005). By contrast, analytic rubrics focus on particular characteristics of the writing and pinpoint strengths and weaknesses of a composition based on specifically delineated criteria (Linn & Miller, 2005). Within an analytic rubric, the criteria represent the characteristics that help the rater identify the goals of instruction, while the descriptors identify the degree to which the student has attained these goals (McMillan, 2004). The descriptors provide illustrative detail about the criteria. The number of characteristics within a rubric can vary and the objective description of these criteria is often difficult for the rubric author to document (Linn & Miller, 2005).

To choose or construct an appropriate rubric, educators should consider the assessment purpose (Linn & Miller, 2005), measurement technique(s), and the implications of the measurement tool on evaluation (McMillan, 2004). High-quality rubrics have at least four characteristics: (1) content/coverage of the features that are used to measure the quality of the performance or product; (2) clarity/detail of the definitions, indicators, and samples of student work; (3) practicality of the rubric in terms of its usefulness for instruction and assessment; and (4) quality/fairness, which considers the degree of inter-rater reliability and fairness to all students (Arter & McTighe, 2001). In addition to these characteristics, a rubric should clearly connect instruction with assessment through its evaluative criteria (Payne, 2003).

As noted above, rubrics may be used for scoring or instruction

(Andrade, 2005). Scoring rubrics are used strictly to assign grades, whereas, instructional rubrics are used during the instructional process. Teachers use instructional rubrics to clarify learning goals and focus their teaching. In this fashion, an instructional rubric is provided to students along with an assignment and then used for formative feedback and summative evaluation. A teacher's formative feedback is often specific to rubric criteria and allows students to target aspects of their performance that require improvement. In essence, by providing students with assessment criteria, teachers make learning targets clear to students, which should improve their writing performance (Arter & McTighe, 2001).

Contemporary Rubric Use

Over the past decade, analytic trait-based rubrics have grown in popularity among educators in North America (e.g., *6 + 1 TRAIT* model, Culham, 2003; *Write Traits*®, Spandel & Hicks, 2002). These rubrics provide classroom teachers with specific measures to assess the six componential writing traits: ideas, organization, voice, word choice, sentence fluency, and conventions. Each trait is individually assessed on a 5-point scale allowing teachers the facility to provide students with specific feedback on separate traits in terms of strengths and weaknesses. In general, this kind of formative feedback is especially effective at improving students' learning (Black & Wiliam, 1998). Most particularly, analytic trait-based rubrics assist struggling writers to examine all aspects of good writing, not just those related to their specific need for growth and development (Schirmer & Bailey, 2000). More research needs to be conducted, however, to determine the veracity of the claims about the effectiveness and the reliability of analytic trait-based rubrics that are used to assess writing performance.

Reliability of Writing Rubrics

Contemporary forms of writing assessment raise the concern about whether different graders can reliably assess writing samples using complex scoring criteria. Consequently, an integral aspect of establishing reliability is to ensure that graders understand the criteria

that they are using to assess student performances and products, and that they come to a consensus as to how the criteria are to be applied. This process is one of negotiation and involves determining inter-rater reliability among graders. Inter-rater reliability is a measure of the degree of correlation of responses among graders who are assessing a single product or performance. It is an essential characteristic of any assessment that will be used for placement or promotion and for the differentiation of instruction among learners (Aiken, 2000). For writing assessment, an inter-rater reliability score is represented by an intraclass correlation coefficient, a generalized inter-rater measure derived from several graders scoring the compositions of a number of students (McGraw & Wong, 1996). The process of establishing inter-rater reliability among teachers requires an investment of time through professional development.

As an illustration, 10 literacy resource teachers participated in two consecutive 3-hour sessions aimed at discussing elementary writing assessment and evaluation (Gallagher, Mangat, Engemann, & Castle, 2005). Three scoring rubrics (*Write Traits*®, Spandel & Hicks, 2002; *6 + 1 TRAIT* model, Culham, 2003; *Wechsler Individual Achievement Test [WIAT]*, The Psychological Corporation, 1991) were used to establish an intraclass correlation coefficient for the grading of writing samples. A comparison was made among the teachers' total grades for each of the papers. Intraclass correlation coefficients were calculated for the *Write Traits*® rubric, $r = 0.94, p < .05$; for the *6 + 1 TRAIT* rubric, $r = 0.94, p < .05$; and for the *WIAT* rubric, $r = 0.96, p < .05$. In the end, the teachers were able to reliably use each of the three rubrics to score a series of writing samples. However, it should be noted that this result came after a substantial amount of discussion and negotiation among the raters about the meaning of the descriptors and qualifiers in the rubrics.

As they graded the student writing samples, the resource teachers openly provided comments about the efficacy of the rubrics (Gallagher et al., 2005). Initially, the resource teachers found it difficult to interpret the exact meaning of some of the adjectives used in the rubric descriptors. Through the process of discussing concerns, contrasting rubrics, debating grading decisions, and supporting judgments, the resource teachers struck a consensus on the application

of the scoring rubrics. Across the two training sessions, cohesion developed among the resource teachers and they perceived that their discussion contributed to the establishment of consistency in their grading. Consensus-building around the use of rubrics would appear to lessen the subjectivity that permeates the assessment of writing. This happens as raters negotiate the meaning of the descriptors within a rubric.

An important mechanism for developing consensus among teachers is through collaborative professional development. Engemann, Gallagher, and Castle (2005) found that most teachers eagerly seek to improve their instructional practice and assessment of student work either through attendance at workshops or by informal collaboration, mentoring, and dialogue with colleagues. Effective professional development has the benefit of helping teachers gain consistency in their assessment of student writing through the development of similar pedagogical practice. Rubrics are more reliably employed when teachers are trained to apply them consistently during assessment. Dialogue and debate are essential to broadening teachers' understandings of the elements of writing and the components of writing assessment.

Future Classroom Writing Assessment

Teachers can improve the reliability in their use of rubrics for grading written compositions by enlisting the support of colleagues. For example, teachers may ask other teachers to grade one or more compositions from students in their classrooms and compare the grading. Alternatively, teachers can receive collaborative professional development in the application of grading criteria. This should lead to higher rates of agreement among different graders with the accompanying likelihood that the rubric will be more reliably applied in assessment (IRA/NCTE Joint Task Force on Assessment, 1994). This is especially important for the use of analytic rubrics. As an illustration, primary-level teachers who were trained to interpret the scoring criteria on an analytic writing rubric, graded written compositions more consistently than untrained teachers (Stuhlmann et al., 1999). The teachers who received training in the use of a writing

assessment rubric tended to be more uniform in their interpretation of criteria within the rubric. However, teachers should not lose sight of the fact that reliable assessment is not necessarily valid assessment. The assessment task must be meaningful to draw worthy conclusions. Furthermore, writing assessment should be ongoing and reflective of the classroom writing instruction over a period of time.

Conclusion

The need for an improvement in student writing proficiency is evident to some who teach in institutions of post-secondary education. Many students are graduating from high school with a less than adequate level of writing proficiency. Teachers, teacher educators, and literacy consultants must assist each other in providing opportunities for all educators to improve the quality of writing instruction and assessment through collaboration and meaningful professional development.

References

- Aiken, L. R. (2000). *Psychological testing and assessment*. Boston: Allyn & Bacon.
- Airasian, P. (2005). *Classroom assessment –concepts and applications* (5th ed.). Boston: McGraw-Hill.
- Airasian, P. W., Engemann, J. F., & Gallagher, T. L. (2007). *Classroom Assessment: Concepts and Applications – 1st Canadian Edition*. Toronto, Canada: McGraw-Hill Ryerson.
- Andrade, H. G. (2005). *Teaching with rubrics*. *College Teaching*, 53(1), 27-30.
- Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom*. Thousand Oaks, CA: Corwin Press.
- Banks, S. (2005). *Classroom assessment – Issues and practices*. Boston: Pearson Education, Inc.
- Barakett, J., & Cleghorn, A. (2000). *Sociology of education*.

- Scarborough, Canada: Prentice-Hall Canada, Inc.
- Black, T., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2), 139-48.
- Culham, R. (2003). *6 + 1 traits of writing*. New York: Scholastic.
- Earl, L. M. (1999). Assessment and accountability in education: Improvement or surveillance? *Education Canada*, 39(3), 4-6.
- Engemann, J. F., Gallagher, T. L., & Castle, J. (2005). *Professional development perspectives of teachers and school administrators*. Presented at the Annual Meeting of the Canadian Society for the Study of Education (CSSE). London, Ontario, Canada.
- Froese-Germain, B. (1999). *Standardized testing: Undermining equity in education*. Report prepared for the National Issues in Education Initiative. Ottawa, Canada: Canadian Teachers' Federation.
- Gallagher, T. L., Mangat, M. A., Engemann, J. F., & Castle, J. (2005). *Establishing reliability between graders using writing rubrics*. Presented at the Annual Meeting of the Canadian Society for the Study of Education (CSSE). London, Ontario, Canada.
- Gunning, T. G. (2002). *Assessing and correcting reading and writing difficulties*. Boston: Allyn & Bacon.
- Hunter, D., Jones, R., & Randhawa, B. (1996). The use of holistic versus analytic scoring for large-scale assessment of writing. *The Canadian Journal of Program Evaluation*, 11(2), 61-85.
- Isaac, S., & Michael, W.B. (1997). *Handbook in research and evaluation* (3rd ed.). San Diego, CA: EdITS.
- IRA/NCTE Joint Task Force on Assessment (1994). *Standards for the assessment of reading and writing*.
- Joint Advisory Committee (1993). *Principles for fair student assessment practices for education in Canada*. Retrieved March 13, 2006, from University of Alberta, Centre for Research in Applied Measurement and Evaluation web site: <http://www.education.ualberta.ca/educ/psych/crame/research.htm>
- Linn, R., & Miller, M. D. (2005). *Measurement and assessment in teaching*. Columbus, OH: Pearson.

- Mabry, L. (1999). Writing to the rubric. *Phi Delta Kappan*, 80(9), 670-73.
- McGraw, K., & Wong, S. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30-46.
- McMillan, J. (2004). *Classroom assessment – principles and practice for effective instruction*. Boston: Pearson Education Inc.
- Moskal, B. M. (2000). Scoring rubrics part I: What and when. *Practical Assessment, Research, & Evaluation*, 7(3), 1-7.
- Payne, D. (2003). *Applied educational assessment*. Toronto, Canada: Wadworth Thompson.
- The Psychological Corporation (1991). *Wechsler Individual Achievement Test*. San Antonio, TX: Harcourt Brace & Co.
- Santrock, J., Woloshyn, V. E., Gallagher, T. L., Di Petta, T., & Marini, Z., (2004). *Educational Psychology-1st Canadian Edition*. Toronto, Canada: McGraw-Hill Ryerson.
- Schirmer, B., & Bailey, J. (2000). Writing assessment rubric – An instructional approach with struggling writers. *Teaching Exceptional Children*, 33(1), 52-58.
- Simner, M. L. (2000). A joint position statement by the Canadian Psychological Association of School Psychologists on the Canadian Press Coverage of the Province-wide Achievement Test Results. Retrieved from: http://cpa.ca/documents/joint_position.html.
- Skwarchuk, S.-L. (2004). Teachers' attitudes toward government-mandated provincial testing in Manitoba. *Alberta Journal of Educational Research*, 50(3), 252-82.
- Spandel, V., & Hicks, J. (2002). *Write traits classroom kits*. Wilmington, MA: Houghton-Mifflin Co.
- Stuhlmann, J., Daniel, C., Dellinger, A., Denny, R., & Powers, T. (1999). A generalizability study of the effects of training on teachers' abilities to rate children's writing using a rubric. *Journal of Reading Psychology*, 20(2), 107-27.