The Canadian Journal of Program Evaluation Vol. 11 No. 2 Pages 61–85
ISSN 0834-1516 Copyright © 1996 Canadian Evaluation Society

THE USE OF HOLISTIC VERSUS ANALYTIC SCORING FOR LARGE-SCALE ASSESSMENT OF WRITING

Darryl M. Hunter
Saskatchewan Department of Education
Regina, Saskatchewan

Richard M. Jones
Education Quality and Accountability Office
Toronto, Ontario

Bikkar S. Randhawa
University of Saskatchewan
Saskatoon, Saskatchewan

Abstract: This article discusses a variety of issues associated with the use of holistic and analytic scoring methods in large-scale student assessments. The article reviews the education literature on the subject, describes the results of a recent Saskatchewan study comparing the results of the two scoring approaches, and discusses some of the study's implications for future large-scale assessments.

Résumé: L'article présente une variété de problématiques liées aux méthodes d'évaluation analytiques et holistiques de la performance des élèves lors d'enquêtes à grande échelle. Cet article résume la littérature éducationnelle sur le sujet, décrit les résultats d'une récente étude en Saskatchewan qui compare les résultats des deux approches d'évaluation, et discute de certaines implications de l'étude sur de futures enquêtes à grande échelle.

Synonymous with red ink and course success for students, the bane of educators at the end of an instructional day, and often the single most expensive aspect of large-scale assessment programs, the marking of writing is a key evaluative activity in education. At one time, the scoring of essays and open-response items was overwhelmingly associated with language arts and the humanities; student evaluation in other disciplines was seen as more amenable to "objective" measurement through multiple-choice items. Yet increasingly, teachers in mathematics and the natural sciences too are exhorted to use more "authentic" evaluation techniques in their classrooms, including written responses. Likewise, large-scale assessment programs are attempting to capture and categorize more complex thinking and problem-solving skills through extended open-response items.

Because of the general retreat from tests composed entirely of multiple-choice questions, varying methods for scoring student writing have come into prominence during the past decade. Educators are exploring alternative marking techniques, asking which is the most efficient and appropriate. Issues of fairness and due process have also come to the fore. Consistency in evaluating student writing becomes critical when those evaluations influence "high-stakes" judgements involving placement or promotion of students. This concern with fairness is exemplified by the fact that virtually every educational institution, in its policy statements, sets out appeal procedures for those who would challenge the marks assigned.

Concern over due process has also led to codification of fair procedures for conducting the large-scale assessments that are multiplying across Canada as pressures for public accountability mount. Provincial achievement testing at the Grade 12 level, such as Alberta's Diploma Testing Program or Quebec's exit examinations, are examples of high-stakes testing programs that determine students' final marks. Simultaneously, large-scale assessment programs such as that of the Council of Ministers of Education, Canada's national School Achievement Indicators Program, British Columbia's Learning Assessment Program, and Nova Scotia's portfolio assessments are all used to gauge systemic rather than individual student performance for "low-stakes" public accountability purposes. Portfolio assessments in particular have stimulated concerns about reliability in scoring (LeMahieu, Gitomer, & Eresh, 1995).

As attention focuses on student outcomes, a debate is beginning to be heard between the proponents of various scoring techniques, the most visible polarity being between proponents of holistic methods and those championing an analytic approach to evaluating student writing. With *holistic scoring*, student work is rated as a complete unit against a prepared scale or rubric. The scorer reads the student response and a global score is awarded, which may be in the form of a percentage, letter grade, or rating number denoting the level of achievement. *Analytic scoring*, on the other hand, involves evaluating student work by breaking it down into its constituent elements or attributes and assigning a proportion of the available marks to each. The scorer reads the student response, assigns scores to each element or attribute being assessed, and then sums, averages, or proportionally weights the scores on the various dimensions to derive an overall score. Holistic scoring is less time consuming—the procedure emphasizes very rapid reading and rating—and is therefore less expensive; analytic scoring has the advantage of providing more detailed diagnostic information about student strengths and weaknesses in various skill areas.

Analytic methods appear to be more amenable to content-driven courses. In contrast, provincial curricula are increasingly moving away from an emphasis on discrete skills and isolated fragments of content to the cultivation of integrated higher-order skills in reasoning, problem solving, and divergent thinking. Holistic scoring is touted as suitable for evaluating these open-ended and higher-order skills. As such, the method is associated with the recent interest in performance-based and "authentic" assessments. Hence, holistic approaches may become embroiled in the same debates (Lewington, 1996) that surround these emerging views of assessment.

Despite concerns about fairness and consistency in scoring, there is little empirical research that systematically compares the two approaches to scoring writing. The purposes of this article, then, are to summarize the literature on holistic versus analytic scoring; describe the results of a recent study that compares the results of the two scoring approaches; and discuss this study's implications for future large-scale assessments and research.

LITERATURE REVIEW

Continuum of Scoring Methods

One reviewer recently identified some 19 different definitions of holistic procedures used in educational literature. In fact, one can discern a continuum of scoring practices (see Figure 1) along which five distinct methods can be situated. At one end of the axis lies general impression scoring as a holistic method, and at the opposite end lies atomistic scoring as an analytic scoring method (Goulden, 1989). The five methods range from general impression scoring's wide-angle assessment of the whole, without clearly delineated criteria, to atomistic scoring's microscopic assessment of constituent parts, without considering the overall effectiveness of the entire writing. As one moves across the spectrum, progressively greater constraints are placed on the marker's latitude in judgement: general

Figure 1
Continuum of Scoring Methods

Holistic Approaches			Analytic Approaches	
<——^————————^————————|^——————————^————— —^———————>				
General Impression Scoring	Holistic Scoring	Primary Trait Scoring	Analytic Scoring	Atomistic Scoring

impression scoring depends on the idiosyncratic judgements of scorers, and holistic, primary trait, and analytic procedures depend on increasingly prescriptive scoring scales, culminating with attempts to eliminate the marker's subjective response altogether in atomistic scoring.

General impression scoring, the earliest manifestation of holistic scoring procedures, was developed in the 1960s by an American organization, the Educational Testing Service, in pursuit of a valid and economical way of directly measuring students' abilities to write rather than continuing to rely on indirect measures such as multiple-choice tests. As described by its principal researcher, Paul Diederich, the method asked raters to read a paper and assign it a single score on the basis of their overall impression of its quality. Criteria were chosen by the individual marker and were not articulated until after the scoring session. Diederich's initial experiments without guides or controls demonstrated how erratic the scoring of essays could be (Diederich, 1974; White, 1985). General impression scoring came to be seen as marking without criteria; the unreliability demonstrated by these early experiments has subsequently fuelled the objections of many critics of holistic approaches, who tend to associate the method with other fads of the 1960s such as transcendental meditation.

Holistic scoring, as it has evolved from general impression scoring and been refined over the past two decades, can still be defined as an evaluative method that considers the overall quality of the product. To bring greater rigor and consistency, however, researchers and assessors added a series of controls: training or orientation exercises to develop a shared interpretive outlook on the criteria among all markers, prior to scoring; a scoring guide or rubric that describes the performance criteria to define each point on the scale; ongoing checks or audits of the quality of scoring throughout a marking session; independent double or even quadruple scoring of student writing; and detailed record keeping to monitor the consistency of scoring on an ongoing basis (White, 1985). Whereas general impression scoring involves individual raters applying their personal criteria, holistic scoring relies on a formal, group-centered process to internalize prepared criteria. What follows is a focused and deliberate effort by that group of raters to consistently apply the criteria throughout a scoring session. The scoring rubric, anchor, or sample papers that exemplify the ratings on its scale, and training papers that bring the scoring team to a common understanding of the performance levels, are the method's characteristic devices (Daiker & Grogan, 1991).

Primary trait scoring is a variant of holistic scoring in that it considers the effectiveness of the overall work and proceeds from a framework of prepared and shared criteria. However, whereas holistic scoring employs a rubric of more general characteristics that distinguish performance levels straddling a variety of writing genres and modes, primary trait scoring defines with greater precision and exclusiveness the criteria to be used in scoring. Raters focus on the specific rhetorical attributes or traits in a given piece of writing. A method developed through the work of the National Assessment of Educational Progress (NAEP) in the United States using rhetorical theory and discourse models, primary trait scoring is based on the twin premises that all writing is performed for a specific audience and that writing, if successful, will have the desired effect on that audience (Lloyd-Jones, 1977). Raters may be asked to make precise measures of students' skill in argumentation, or their ability to project themselves imaginatively into a prescribed situation, or their effectiveness in using statistical or documentary information to justify an assertion. The rating is always made by considering the writer's overall effectiveness in manipulating various discourse elements to achieve the desired effect.

Analytic scoring proceeds from a different set of assumptions than general impression, holistic, and primary trait scoring. If holistic scorers assume that the whole produces the parts, the analytic scorer says that the observable and measurable skills or attributes produce the whole. Raters first score the individual elements and then combine that series of judgements to produce a composite score. The analytic rater will judge a student's ability along a series of dimensions, such as organization, content, mechanics, and diction, and then calculate a composite score. It is "the recording and tabulating of subscores which separates analytic from holistic scoring" (Goulden, 1989, pp. 4–5).

Atomistic scoring is even more granular in its examination and analysis of writing components, and reflects a preoccupation with discrete bits of knowledge. It involves counting or noting the presence or absence of elements in the writing. A rater may count the number of spelling errors, the number of grammatical faults, the number of causes correctly described for World War I, or the number of positive and negative effects the North American Free Trade Agreement has had on the Canadian economy, or whether the writer has or has not formulated a clear thesis statement, before awarding or subtracting points based on his or her count. Atomistic methods equate scoring with taking inventory, and assume that a paper's overall merit is reflected in the presence of isolated attributes or the absence of errors. Its characteristic device for attaining objectivity is the answer key, the outline listing of correct responses or mechanical faults with point allocations for each.

Reliability

The increasing constraints placed on raters' judgements in the pursuit of greater scoring objectivity have led many to believe that reliability of scoring increases as one moves from general impression to atomistic methods. Yet a series of studies have demonstrated that high inter-rater and intra-rater reliability coefficients can be achieved for holistic scoring, even though White (1985) and Smith (1993) show that the method's proponents commonly inflate their coefficients and do not have standard ways of reporting reliability. Holistic scoring's popularity began to grow when Godshalk, Seinford, and Coffman's (1966) study of College Entrance Board assessments revealed high reliability of scores. By the mid-1970s, assessors were reporting that holistic scoring techniques consistently yielded inter-rater reliability coefficients in the .7 to .8 range and above if scorers were given special training sessions prior to their work (Myers, 1980). Cooper (1977) emphatically reiterated this point two years later, saying that when raters were from similar backgrounds and were trained to use a holistic scoring scale, they could achieve nearly perfect agreement in choosing the better of a pair of essays; scoring reliabilities in the .85 to .94 range on their summed scores were possible using multiple pieces of a student's writing.

Despite this research, there have been few empirical studies that systematically compare more than one scoring method. Most studies feature very small numbers of raters, such as Baurer's (1982), study which found acceptable inter-rater and intra-rater reliability for nine markers' scores on writing samples using three different scoring methods: analytic, primary trait, and holistic. Veal and Hudson (1983) compared holistic, analytic, atomistic, and primary trait scoring methods using the papers of 10th-grade students in Georgia. They found acceptable reliability (.69 to .76 range) for all four methods when employed concurrently. Quellmalz (1982) concluded that most holistic and analytic rating scales can demonstrate high inter-rater reliability. But she also warned that inter-rater agreement within a scoring session is insufficient for demonstrating scale reliability. Parallel to the problem of test-retest reliability, a scoring rubric must be stable and demonstrate that criteria can be applied consistently with a new set of raters to both a new set of papers and a set of previously scored papers.

If consistency between scorers and scoring sessions was a preoccupation of the 1970s and 1980s, the advent of the "authentic assessment" movement in the 1990s, with its focus on context, emphasized consistency of evaluation between large-scale and classroom situations. In addition to the traditional psychometric concerns of inter-rater and intra-rater consistency, here reliability also means congruence between third-party testing and classroom testing. In fact, Moss (1994, p. 7) reconceives the issue of reliability by emphasizing the processes involved in reaching a common understanding of student work:

> A hermeneutic approach to assessment would involve holistic, integrative interpretations of collected performances that seek to understand the whole in light of its parts, that privilege readers who are most knowledgeable about the context in which assessment occurs, and that ground those interpretations not only in the textual and contextual evidence available but also in a rational debate among the community of interpreters.

In Moss's view, reliability is not only a metric that describes how well people can agree about performance on a prespecified scale, but also includes the agreement that arises through exchange of ideas about a body of evidence and the shared interpretative tradition that arises from that exchange.

Validity

If one defines writing as a contextualized, communicative act and scoring as a method of rating overall communicative effectiveness, then one might logically expect that the validity of scores will increase as one moves across the continuum from the quantitative focus of atomistic scoring to the qualitative focus of general impression scoring. Indeed, White (1985) argues that the measurement of writing as a unit of expression rather than as a series of isolated skills conforms to more contemporary views of how to teach writing, as well as to postmodern literary criticism. Just as humans are more than the sum of their socialized behaviors, so too is an essay or speech more than a composite of purportedly "neatly sequential and comfortably segmented" parts (p. 30). Holistic scoring, he implicitly argues, has more face validity because of the lack of consensus about what the essential writing subskills are, and because of the raters' jurisdiction to go beyond punctuation and grammar and consider the unexpected ways in which writers demonstrate their idiosyncratic qualities in their written work.

Yet a number of scholars have questioned the validity of holistic methods. Charney (1984) argues that holistic scoring as a qualitative method has not received scrutiny of the kind leveled against the quantitative methods it purports to replace. She points to controversies among assessors about procedures for choosing essay topics and selecting criteria to judge student work, to raise doubts about holistic methods' predictive validity. She also questions their face validity by pointing to a series of studies suggesting that raters may be reacting to cosmetic attributes of student writing, such as neatness of handwriting, number of minor spelling errors, and other mechanical traits rather than those substantive features detailed by the scoring rubric.

Another sympathetic critic, Huot (1990a), believes that even though the issue of reliability has been settled through three decades of research, assessors' preoccupation with consistency has "caused the profession to assume, confuse, and otherwise neglect the validity" (p. 204) of these methods. He contends that concurrent validity has not been proven despite a number of early studies that correlated holistic raters' judgement of overall quality with syntactical features such as usage and mechanics, and later studies that demonstrated a strong relationship between holistic ratings and features such as content and organization. Of more concern, Huot concurs with Charney in speculating that the speed of reading and the rubric might cause raters to limit their focus in an effort to attain agreement, thereby distorting their judgement of writing ability: "a personal stake in reading might be reduced to a set of negotiated principles, and then a true rating of writing quality could be sacrificed for a reliable one" (p. 211). He states that a "major weakness is that the work done on the influences of holistic rating procedures is quite limited" (p. 208).

The pioneering attempt to investigate the relative effects of essay variables, reader variables, and environmental variables on holistic scores was Freedman's (1981) study of college composition ratings in the San Francisco Bay area. Essay characteristics easily accounted for most of the variance in scores, but the effect of trainers was also significant. Correlations in the range between .61 and .76 were found between holistic ratings and the analytic categories of voice, development, organization, sentence structure, and word choice. Freedman concluded that the only additional information the analytic scale yielded over the holistic score was a usage score.

Freedman was also among the first to attempt to identify the specific analytic elements upon which raters base their holistic judgements. In an earlier 1979 study, she rewrote student essays to manipulate content, organization, sentence structure, and mechanics before holistically rating the various pieces of writing. Although mechanics and sentence structure were associated with ratings, Freedman found that content scores and organizational factors best predicted the overall scores. One critic has noted that the rewritten essays reflected qualitative extremes rather than a natural range of student performance, and the experiment design was considered faulty (Charney, 1984). Moreover, it is highly unlikely that one could revise only one element of an essay without substantially altering others.

Another approach has been through protocol analysis. By interviewing 300 scorers before, during, and after a marking session, Braungart-Bloom (1986) investigated the relationship between holistic raters' perceptions of writing quality and the characteristics of that writing that contributed to its quality. Trained holistic readers perceived six general groupings of attributes: content, organization, sentence construction, usage, mechanical characteristics, and textual characteristics. They were able to identify the salient characteristics at the extremes of a six-point scale, but were less able to identify patterns of relationships among essay characteristics at the middle points, even following training and reading of 90,000 essays from Grade 9 students in New Jersey. This corroborated an earlier, more limited, and less controlled study conducted by Marsh and Ireland (1984) with 139 Australian student writings, using both holistic and analytic ratings, which found that "the predicted ability of teachers to differentiate among the components of writing effectiveness was so weak as to be of little practical value" (p. 18).

Another Australian scholar has also raised caveats about the predictive validity of holistic methods. After reviewing but not documenting the scoring of candidates' essays in the Western Australia Tertiary Admission English Examination for the 1969–79 period, and after analyzing the retrospective comments of raters for one series of exams, Hay (1982) suggested that skill in synthesis was a crucial factor in assigning ratings. The scores seemed to reflect psychological variables such as fatigue or self-confidence on the part of the marker, rather than any particular set of skills demonstrated by the writer. Experienced markers were unable to completely agree in their formulation of what determines their individual general impressions. He deduced that the *context* in which a marker arrives at his or her general impression is very significant. The refusal of Victoria State teachers to conduct the 1976 matriculation exams, and their recommendation of a lottery in place of them, is noted in passing.

Most systematic in its comparison of scoring methods is Veal and Hudson's (1983) study of several direct and indirect measures with 10th-grade Georgia student writings. Holistic, primary trait, analytic, and atomistic methods as well as a series of "objective" tests were administered simultaneously in the large-scale assessment. A .64 correlation between holistic ratings and total analytic ratings, including content, style, and mechanics, was found. Minor correlations ranging from .25 to .31 were found between holistic ratings and atomistic counts of punctuation, capitalization, awkward construction, agreement, and word-choice attributes. Of the indirect measures, the Iowa Test of Basic Skills demonstrated the strongest correlation with holistic procedures, ranging from .52 to .65 on its various language skill sections. The four different scoring methods were employed concurrently with different sets of papers or students; apparently, an identical set of papers was not scored consecutively using the different methods.

In recent reviews of holistic scoring by adherents of authentic assessment approaches, some have asserted that our constricted notions of validity must be expanded, like our conceptions of reliability. The social purposes of writing and assessing must be considered, including the effects of assessment on the educational system in which it occurs (Messick, 1989). Proponents of holistic scoring contend (Camp, 1993) that assessment must support instruction and foster development of higher-order thinking and problem-solving skills to demonstrate "systemic validity." Moreover, direct assessments of writing should both measure product quality and show an awareness of the process leading up to the product's creation. Valid assessment procedures will make explicit, and help both teachers and learners to understand, the characteristics of good performances and products. Just as writing is a social construction, so too is the assignment of a mark. Scoring procedures must be seen as part of a "contextualized experience" that is consistent with school reform and restructuring.

Both Pula and Huot (Huot, 1993; Pula & Huot, 1993) have twice investigated this notion of ecological validity by exploring the relationships between scorers' personal background, scoring experience, professional training, and work experience and their marking performance. They affirm that holistic raters tend to base their judgements primarily on content and organizational features of writing, and that holistic procedures may in fact promote a fluent and receptive reading of student writing, rather than throttling it for the sake of consistency. Raters bring their personal and professional pasts to a scoring session to create an immediate discourse community. The attendant socialization creates an environment conducive to a rich, personal, and consensual reading of student work.

Implicit in most criticisms of holistic scoring is the claim that the methodology is ungrounded in theory—that it is essentially a praxis that has evolved out of practical, administrative concerns over economy and consistency of scoring rather than any sound body of psychological theory. Although White (1985) has linked holisticism to current preoccupations with composition process and literary criticism, and others contend that holistic methods must be embedded in a theory of writing (Camp, 1993), Huot (1990b) likens our current understanding of scoring methodologies to our knowledge of chemistry in the age of alchemy. No attempt has been made to base holistic methods in a coherent body of principles such as those found in Gestalt theory.

Indeed, psychometricians might be well advised to revisit the propositions of Wertheimer, Kohler, and Koffka for insights into holistic scoring and its perceptual bases. Gestalt theory, it will be recalled, rejected the concepts of association of elements as the basis of perception and introspective analysis as the key to primary, original experience. The German psychologists, centered at the University of Berlin, attacked the behaviorists' idea of learning as building connections between stimuli and responses. In the Gestalt approach, there is an emphasis on innate organizing processes that give us, not isolated sensations, but patterns as a primary characteristic of perception. The overall configuration is seen as the basic unit of perception, and the whole determines the character and behavior of the parts, instead of the other way around. As the Gestaltists pointed out, a piano melody is the same in one key as in another, even though the individual notes all change. Moreover, the qualities of wholes—like the liltingness or the plaintiveness of a melody—do not reside in the individual notes.

A recently conducted large-scale assessment program in Saskatchewan explored the relationship between the quality of whole writings and selected individual analytic writing elements, by comparing holistic and analytic scoring approaches that were employed consecutively with an identical set of papers. The intent in the low-stakes assessment was to determine the degree to which holistically derived scores could be predicted from analytic approaches, individually and in aggregation, and to examine the relationship between the various elements of writing in producing the global score. Is the quality of a whole piece of writing greater than the summed or average quality of its parts?

SASKATCHEWAN'S PROVINCIAL LEARNING ASSESSMENT IN LANGUAGE ARTS

Samples

Saskatchewan's 1994 Provincial Learning Assessment in Language Arts was designed to provide reliable information to the Saskatchewan Department of Education and to the general public about Grade 5, 8, and 11 students' reading and writing skills and strategies. In May 1994, writing instruments were administered to a total random sample of approximately 1,600 students, about one-third of whom were at each of the three grade levels.

Procedure

The assessment form allowed for prewriting, drafting, and postwriting activity. In addition, all students were asked to submit a sample of what they considered to be their best piece of writing from regular classroom work completed within the preceding three months. Students were allowed three hours in three separate sessions, with some latitude in granting additional time if necessary.

Scoring

Students' work was scored using five performance levels in a holistic scale, and subsequently in five analytic dimensions. Analytic categories included content, organization, syntax, vocabulary, and mechanics, with criteria that were virtually identical to those used in the 1994 Council of Ministers of Education, Canada's (CMEC's) national School Achievement Indicators Program (SAIP). Holistic criteria incorporated the analytic elements and other features of Diederich's (1974) scale. A score of "0" was given to pieces of writing for which there was insufficient information, in the opinion of the scorer, to provide a rating. Final drafts on writing assessment forms and classroom writing samples were scored holistically using identical criteria and anchor papers chosen by scoring leaders to exemplify the criteria in the scoring rubric. Raters provided only the individual analytic ratings; a total analytic score was subsequently derived by obtaining an average of these component ratings.

Once holistic scoring was completed for the grade-level pool of final drafts for both the writing assessment forms and classroom samples, the final drafts on the assessment forms were marked again along the five analytic dimensions in five separate but consecutive scoring sessions. All scoring practices were organized according to the principles and procedures suggested by White (1985) and Myers (1980). Saskatchewan raters were asked to read quickly, to not think about a paper too much, and to score their first impression, making certain it matched the exemplar papers offered; raters were advised that their first judgement would likely be genuine and accurate, and that overdeliberation would likely allow tangential or irrelevant criteria to bias their scores.

Three procedures, in particular, were used to enhance scoring consistency with such a rapid reading. The first was systematic training of all raters before each scoring session. All raters were mailed preparatory information and asked to become familiar with criteria, performance levels, and scoring procedures before the five-day scoring session. Before each scoring session, criteria were discussed by all team members for each of the five performance levels. Anchor papers at each performance level were read, discussed, and provided to all teams of raters for reference during the scoring session. The training papers were scored and discussed to ensure that all team members had adopted a shared perspective of the criteria before "live" scoring began. Retraining occurred after all breaks. During the scoring session, scoring leaders provided ongoing rereading and monitoring of scores. Inconsistent or "uncalibrated" raters were corrected on an ongoing basis.

Approximately 10% of papers were randomly double-scored by each grade-level leader to ensure consistency. In addition, scoring leaders chose approximately 10 papers at selected points during each session for "internal audits," and required each team member to score this identical paper. To prevent previously assigned scores from prejudicing raters' judgements, raters used labels to cover all scores assigned. Inconsistent or uncalibrated raters were recalibrated immediately. Scoring sessions were interrupted and retraining of the entire team was conducted as necessary.

The third control involved independent quality checks of scoring. During each four-hour scoring session, an external auditor presented eight photocopied "calibration papers" at 30- to 50-minute intervals for simultaneous reading and scoring by all team raters. If all scores were adjacent, the scoring session proceeded. If one or more scores were not adjacent, the scoring session was interrupted and scoring discrepancies were discussed. Reorientation using training papers was conducted if necessary. Discrepant papers were adjudicated by the scoring leader. The scoring leader also reviewed all papers scored by the inconsistent rater(s) in the interval since the last calibration paper and adjudicated where necessary. For the scoring session overall, inter-rater reliability was .82. This was calculated by dividing the number of papers where perfect scorer agreement occurred, before third party adjudication, by 1,583, the total number of papers scored.

Results

For the three grades involved, correlations between the classroom sample writing scores, which students considered to be their optimal performance in the regular classroom setting during the previous three months, and their performance in a test situation ranged from .27 to .34 (see Table 1). It should be noted that not all students had all the data available; as a result, a pair-wise missing cases procedure in correlational analysis was employed. These low correlations in Table 1 are a very modest confirmation that the assessment gathered information that accurately represented students' realized classroom capabilities in teacher-assigned situations. As a measure of concurrent validity, the correlation coefficients suggest that the learning assessment was only marginally able to capture student's writing quality in optimal circumstances.

Table 1
Correlations and Descriptive Statistics for Different Scores on Writing Tests and Classroom Samples by Grade

Grade/	Variable	Holistic Score: Test (1)	Analytic Score: Test (2)	Holistic Score: Classroom Sample (3)	*N*	Mean	S.D.
5	1	1.00	.70	.27	568	2.75	.81
	2		1.00	.27	568	2.81	.64
	3			1.00	538	2.88	.95
8	1	1.00	.67	.30	516	2.69	.86
	2		1.00	.29	516	2.84	.63
	3		1.00		473	2.85	.97
11	1	1.00	.70	.27	500	2.94	.84
	2		1.00	.34	501	2.96	.68
	3			1.00	422	3.09	.87

An examination of the correlation coefficients in Table 2 indicates that holistic scores assigned to the essay correlate substantially, *r*= *.69*, with the total analytic scores. This correlation provides a coefficient of determination or r^2 of .48, which is almost the same as that obtained from the multiple regression analysis with the holistic score as the dependant variable and the five analytic component scores as predictors (see Table 3). This means that the common variance between holistic scores and total analytic scores is 48%, that is, 48% of the variance in holistic scores is explained by total analytic scores, and vice versa. The predictive power of the total analytic score is as

Table 2
Correlation Matrix of Holistic Scores, Analytic Components, and Analytic Total: Scores, Means, and Standard Deviations (*N* = 1,585)

Variable	1	2	3	4	5	6	Mean	S.D.
1. Holistic score							2.79	.84
2. Content	.54						2.79	.89
3. Organization	.53	.53					2.70	.91
4. Syntax	.51	.46	.48				2.96	.80
5. Vocabulary	.47	.45	.40	.48			2.96	.72
6. Mechanics	.55	.45	.47	.52	.47		2.93	.89
7. Analytic total	.69	.77	.77	.77	.71	.77	2.86	.65
8. Adjusted analytic*	n/a	.61	.61	.62	.57	.61	—	—

* Each correlation coefficient in the row represents a value between a specific component and the analytic total score, excluding that specific component to show the effect of removal or moderation on the correlation coefficient in the row above.

Table 3
ANOVA for Multiple Regression of Holistic Scores as the Dependent Variable and Five Analytic Components Scores as the Predictors

Source	df	SS	MS	F-ratio
Regression	5	533.68601	106.73720	288.98*
Residual	1576	582.09986	.36935	

Note. $R^2 = .48$.
* $p < .05$.

strong as that of the five components entered into the regression simultaneously (see Table 3). The latter also implies that for this study, it was not necessary to perform the multiple regression because simple regression of holistic scores is just as effective.

Correlation coefficients of holistic scores with the five analytic score components range from .47 with vocabulary to .55 with mechanics (see Table 2). In other words, the coefficients of determination for these analytic components range from .223 to .303. Thus, these simple correlations account for common variance in the range of 22.3% to 30.3%. However, the correlation coefficients of total analytic scores with the five component scores range from .71 to .77. That is, the redundant variance is in the range of 51.0% to 59.9%. These correlations are spuriously high because each component score is part of the total analytic score. If the total analytic scores are adjusted by removing a specific component score before using them for the calculation of correlation coefficients of the component scores with the adjusted total analytic scores, the values will be lower and perhaps closer to those found between the holistic score and the total analytic ratings. Corrected correlations are displayed in Table 2, row 8. These range from .57 to .62, as expected. These values are substantially lower than those obtained when all the components contributed to the analytic total score.

Cross-tabulations were also generated as a second means of comparing the results of the two scoring methods. The data presented in Table 4 are the result of asking the following question: Will the scores for individual students be the same if only holistic scores are used rather than analytic scores? As mentioned earlier, a single analytic score was assigned each paper by averaging the ratings for the five analytic components. For the purposes of Table 4, the analytic average score was truncated to the lowest whole number. Truncating the average analytic scores is necessary to equate them with

Table 4
Cross-tabulations of Holistic and Total Analytic Ratings

		Analytic Rating						
		0	1	2	3	4	5	Row total
	5			1	11	7	1	20 1.3%
	4			27	176	50	2	255 16.1%
Holistic Rating	3	1	5	332	426	19	1	784 49.5%
	2		48	303	76			427 27.0%
	1		44	40	2			86 5.4%
	0	4	4	3				11 .7%
	Column Total	5	101	706	691	76	4	1583
	%	.3	6.4	44.6	43.7	4.8	.3	100.0

Note. **The figure in each cell is the total number of papers rated by each scoring procedure.**

a holistic rating, which assigns the individual student writing entirely to one performance level or another without a decimal point. Truncating, rather than rounding to the nearest whole number, was deemed appropriate because students had not demonstrated that they were operating at the next-highest performance level on the five-point scale.

Table 4 demonstrates cross-tabulations between the ratings derived analytically and holistically. The number of papers receiving each holistic rating is provided along the *Y*-axis and the corresponding ratings awarded the same papers analytically along the *X*-axis. For example, 27 papers received a holistic rating of 4 and an analytical rating of 2; 176 papers received a holistic rating of 4 and an analytic score of 3; 50 papers received identical holistic and analytic ratings of 4; and so forth. With regard to the interchangeability of holistic and total analytic scores on the total pool of 1,583 papers scored in the Saskatchewan session:

- 52.3% of the papers were given identical ratings holistically and analytically.
- 96.8% of the papers were within plus or minus one rating.

DISCUSSION

These results indicate that the quality of a whole piece of writing can be predicted quite well by averaging five analytic component scores that many writing specialists have identified as essential writing skills. A comparison of the means of holistic and total analytic scores shows a significant difference ($p < .05$) in favor of analytic scoring. In other words, the total analytic scores on the average are higher than the holistic scores, albeit the difference is practically insignificant: .07. At the same time, the correlational studies in this project show that the total analytic score accounts for approximately half of the variance in the holistic ratings. These results may draw into question the claims of Gestaltists who say that the whole is greater than the sum of its parts. This study found that the whole writing is greater, but not necessarily of cumulatively better quality, than its constituent parts.

That the various analytic scales do not account for all the variance may be explained by those attributes of a written response that they did not measure: awareness of audience, stylistic flair, economy of language, and experimentation with phraseology, which were not independently gauged by the analytic rubrics. In that sense, the holistic approach may have enabled scorers to consider a broader range of writing qualities than was permitted by the five individual analytic scales. That there was no real difference in correlational values between the analytic component scores may suggest that for this large pool of writing, raters were able to reliably and accurately employ the five analytic scales when considering the quality of the whole writing. One cannot say that raters favored or attended to one or more analytic features when generating their holistic scores, as the research design did not ask raters what their foci were when assigning a summative holistic score, but rather asked them to rate each element independently on an analytic scale. No specific element of writing was more effective in predicting overall writing quality.

This study may contradict Hay's (1982) assertion that synthesis, as conceived in Bloom's taxonomy, is the operative skill demonstrated by experienced raters in a holistic approach. If determining overall writing quality was the sole result of synthesizing individual elements, nearly all of the variance in the holistic score should result from the interplay of the analytic subcomponents. That it does not may affirm White's (1985) assertion that holistic approaches enable markers to consider, in a consistent fashion, other qualities of writing than those specified in an analytic rubric. Rather than being skilled in synthesis, raters may need to be adept in evaluation—through integration of both specified and unspecified written elements or stimuli into a total configural estimation of whole writing quality. Training for a holistic scoring session is the process of negating or at least modulating previous, personal, idiosyncratic, evaluative experience, so that raters can consistently perceive that whole.

On the other hand, Hay's caveats about holistic scoring are confirmed by this study. Looking at individual papers, we notice that just over half of the papers received identical ratings holistically and analytically through an average of the component scores. Although nearly all were within one scale point of the holistic rating, suggesting that the two approaches yielded approximately similar results, they were not exactly the same results. Moroever, because such a restricted scale was employed, one that would inherently narrow the range of choice for raters, a difference of one scale point would constitute a considerable zone of variation in outcomes if the individual student scores were to be used for rank-ordering students.

In general, these findings support the results of both Sarah Freedman's (1981) study and Veal and Hudson's (1983) research, which demonstrated the strong correlations between holistic ratings and total analytic or analytic subcomponent scores. If a large-scale summative assessment of writing is needed for public accountability purposes, the Saskatchewan study recommends a holistically rated sample as a measure of overall writing competence, for reasons of its face validity, its reliability, and its cost-effectiveness. A second analytic rating, whereby a subsample of the holistically rated papers are also analytically rated, is suggested as a means of providing detailed diagnostic feedback such as that obtained in the classroom, and as a means of reducing error in measurement.

CONCLUSION

One can conclude that the holistic scores generated in this study were quite robust. The correlational and multiple regression studies suggest a substantial level of agreement between the results of the two methods. Considering the low-stakes purpose of this public accountability exercise, which was to generate a global profile of student literacy achievement as an indicator of overall systemic performance, the global scoring approach was effective in reliably and accurately gauging the quality of student writing on the assessment form. As a technique for efficiently categorizing and rating one aspect of provincial student literacy, holistic scoring was appropriate for suggesting strengths of and needed improvements to provincial programs and for reporting to the general public about key student outcomes. Coupling holistic with analytic ratings provided detailed diagnostic information with a more general profile of writing quality.

However, the cross-tabulation statistics show that perfect agreement on ratings given papers scored both ways occurred in only about one-half of the cases. Although 97% of the total analytic scores on the papers were within one scale point of the holistic scores and vice versa, this amounts to up to a 20% fluctuation in marks in relation to a five-point scale. Unfortunately, it is unknown whether the holistic or total analytic score, or neither of these—given the correlation with the classroom sample in this study—provides the "true" measure of student performance. Test developers and policy makers must carefully consider these implications, particularly where decisions such as Grade 12 exit or certification examinations are involved. In high-stakes situations, all papers—and not just a sample—should be double-scored, for three reasons. First, multiple holistic scores of student work will enable individual student performances to be independently rated by at least two markers to ensure that the mark is accurate and has been fairly assigned. Furthermore, averaging two ratings provides a better estimate of the true performance of the student because the error of measurement is reduced owing to replication. And third, multiple scoring generates marks carried to at least one decimal point to allow for more precise rank-ordering of students for high-school graduation, university entrance, or educational placement and promotion. Thus, double- or multiple-scoring of all papers, which will compromise holistic scoring's cost-effectiveness, is necessary for both precise and fair decision making when the destiny of individual students is at issue.

As the authentic assessment movement reverberates in Canada, holistic scoring techniques will be applied in many situations beyond writing assessments. Reading, the sciences, and even mathematics are some areas under consideration. The marking session for the reading component of the 1994 national SAIP has opened important avenues of research into the use of holistic methods for entire test forms encompassing extended-response, short-open-response, and objective items (Council of Ministers of Education, Canada, 1995). Preliminary scoring of individual items to train raters in identifying the "break points" (or key indication points that demonstrate the performance level at which the student is operating) in an assessment form, prior to assigning a summary global rating, holds promise in extending application of holistic approaches to subject matter and types of test questions beyond the bounds of essay answers.

At the same time, the proliferation of large-scale assessments across Canada may draw into question the systemic or curricular validity of writing scoring scales and methods. Are the performance descriptions found in "public accountability" scoring scales, whether of the holistic or analytic variety, a reasonable approximation of those skills sought in classrooms and provincial curricula? Do the "outcomes" described in a scoring rubric reflect both the instructional objectives found in teachers' day books and the criteria commonly used to evaluate a student's piece of classroom writing? Elements such as originality, flavor, style, or voice may be desired qualities in writing, but are they the subject of direct instruction in classrooms and sufficient basis for differentiating students in "high-stakes" testing situations? Because the scale defines student competence in written production, the content validity of scoring rubrics should be as rigorously scrutinized as test items or specifications. To prevent meandering judgements between assessments and to provide raters and audiences with concrete and operational tools, criteria should be explicit and detailed. For that reason, primary trait scoring should be increasingly considered as a holistic method, even though there is an inherent tension between precision in criteria and the applicability of a global approach to determining writing quality.

Moreover, research attention might be paid to the means of communicating holistic scores for public accountability purposes, to reinforce ecological validity. Little consideration has been given to how the results of a holistic scoring session should be packaged to fulfil the purposes of a large-scale assessment program with those audiences that have an interest in the results. Little is known about the degree of credibility parents and professionals place in holistically derived results. Because professionals must use the scores to adjust their instructional practices, and ultimately to calibrate their classroom evaluation practices with the profiles generated by a large-scale assessment program, their perceptions of report usefulness are important. The holistic performance descriptors can be relatively long; extended exemplars of student writing are necessary illustrations of holistic scoring criteria. Assessors often do not include this descriptive documentation in reports to either parents or educators, so that reports are less meaningful to an assessment's audiences. This may undermine subsequent instruction in specific skills and contravene current definitions of "systemic validity."

And finally, some scholarly effort might be made to place holistic scoring methods within the theoretical framework of Gestalt psychology. Köhler's principle of proximity and Ehrenfel's discussion of *Gestaltenqualität* may suggest ways in which assessors can more fruitfully assist markers to maintain high intra-rater reliability. Koffka's notion of figure and ground, and of the environmental field, could illuminate the interrelationship between a scoring rubric and an anchor paper, as well as ways to construct effective, grounded writing prompts. Likewise, Wertheimer's experiments with and seminars on the *Einstellung* effect, the mindset that predisposes a person to one mechanistic kind of mental act, may extend our understanding of how raters can be effectively and efficiently trained to develop the interpretive community that adopts and applies a shared outlook on the scoring criteria. Koffka's law of pregnance is related to the dynamic tension that raters repeatedly articulate during a holistic scoring session, between writing elements and the total configural quality of a piece of writing: as such, it may illuminate the art of crafting effective scoring rubrics. Wertheimer's propositions on productive thinking can guide scoring leaders as they assist raters to assimilate and apply scoring criteria (see, for example, Koffka, 1935). In brief, Gestalt theory can inform our understanding of scoring methods, which can be defined as extended exercises in controlled perception.

A recent review of large-scale testing practices in Canada concluded that, in an era of budgetary restraint, it may be premature to mourn the passing of the multiple-choice test (Traub, 1994). To forestall that eventuality, assessors will likely turn to holistic approaches because of their cost-effectiveness. However, assessors must be vigilant and not sacrifice the controls, which have enhanced their reliability, on the altar of budget restrictions. Saskatchewan's experience with holistic scoring suggests that both molar and molecular approaches can be economically employed and investigated in tandem with open-response tasks in a low-stakes testing situation. However, the results suggest that evaluators must be cautious when employing holistic methods in situations where high-stakes decisions will be made based on the results.

REFERENCES

Baurer, B.A. (1982). *A study of the reliabilities and cost efficiencies of three methods of assessment for writing ability*. Unpublished doctoral diss., University of Illinois, Urbana-Champaign.

Braungart-Bloom, D.S. (1986). *Assessing holistic raters' perceptions of writing qualities: An examination of a hierarchial framework following pre-post training and live readings*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Camp, R. (1993). Changing the model for the direct assessment of writing. In M.M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 45–78). Cresskill, NJ: Hampton Press.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English*, 18(1), 65–81.

Cooper, C.R. (1977). Holistic evaluation of writing. In C.R. Cooper & L. Odell, (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 3–32). Urbana, IL: National Council of Teachers of English.

Council of Ministers of Education, Canada. (1995). *1994 reading and writing assessment, school achievement indicators program* (Technical Report). Toronto: Author.

Daiker, D., & Grogan, N. (1991). Selecting and using sample papers in holistic evaluation. *Journal of Advanced Composition, 11*(1), 160–171.

Diederich, P. B. (1974). *Measuring growth in English*. Urbana, IL: National Council of Teachers of English.

Freedman, S.W. (1979). How characteristics of students' essays influence teachers' evaluation. *Journal of Educational Psychology, 71*, 328–338.

Freedman, S.W. (1981). Influences on evaluators of expository essays: Beyond the text. *Research in the Teaching of English, 15*(3), 245–255.

Godshalk, F.I., Seinford, F., & Coffman, W.E. (1966). *The measurement of writing ability*. New York: College Entrance Examination Board.

Goulden, N.R. (1989). *Theoretical and empirical comparisons of holistic and analytic scoring of written and spoken discourse*. Paper presented at the annual meeting of the Speech Communication Association, San Francisco.

Hay, J. (1982). General impression marking: Some caveats. *English in Australia, 59*, 50–57.

Huot, B. (1990a). Reliability, validity and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*(2), 201–213.

Huot, B. (1990b). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*(2), 237–263.

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

Koffka, K. (1935). *Principles of Gestalt psychology*. London: Routledge & Kegan Paul.

LeMahieu, P.G., Gitomer, D.H., & Eresh, J.T. (1995). Portfolios in large-scale assessment: Difficult but not impossible. *Educational Measurement: Issues and Practice, 14*(3), 11–28.

Lewington, J. (1996, May 10). Lab sole source of science-test answers. *Globe and Mail*, p. 6.

Lloyd-Jones, R. (1977). Primary trait scoring. In C.R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 33–68). Urbana, IL: National Council of Teachers of English.

Marsh, H.W., & Ireland, R. (1984). *Multidimensional evaluations of writing effectiveness* (ERIC Research/Technical Report).

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: Macmillan.

Moss, P.A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5–12.

Myers, M. (1980). *A procedure for writing assessment and holistic scoring*. Urbana, IL: ERIC Clearinghouse and the National Council of Teachers of English.

Pula, J.J., & Huot, B.A. (1993). A model of background influences on holistic raters. In M.M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp.237–265). Cresskill, NJ: Hampton Press.

Quellmalz, E.S. (1982). *Designing writing assessments: Balancing fairness, utility and cost* (CSE Report No. 188). Los Angeles: Centre for the Study of Evaluation, University of California.

Smith, W.L. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement technique. In M.M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 142–205). Cresskill, NJ: Hampton Press.

Traub, R. (1994). *Standardized testing in Canada*. Toronto: Canadian Education Association.

Veal, L.R., & Hudson, S.A. (1983). Direct and indirect measures for large-scale evaluation of writing. *Research in the Teaching of English, 17*(3), 290–296.

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