DOCUMENT RESUME

ED 349 830 FL 020 649

AUTHOR Hamp-Lyons, Liz
TITLE Holistic Writing Assessment of LEP Students.
PUB DATE Aug 92
NOTE 52p.; In: Focus on Evaluation and Measurement. Volumes 1 and 2 Proceedings of the National Research Symposium on Limited English Proficient Student Issues (2nd, Washington, DC, September 4-6, 1991); see FL 020 630.
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Elementary Secondary Education; English (Second Language); *Holistic Approach; *Limited English Speaking; *Scoring; *Student Evaluation; *Writing Evaluation

ABSTRACT

This paper argues for a direct, holistic assessment of writing of limited-English-proficient (LEP) students. Holistic writing assessment is the term used for tests that evaluate writing wholly through the production of writing. A holistic writing assessment has at least the following five characteristics: each individual taking the assessment must actually write at least one piece of continuous text of 100 words or more; the reader is provided a prompt and is given considerable room in which to respond to the prompt; every text is read by at least two or more reader-judges who have been through training for scoring of writing in that context. The judgments made by readers are tied in some way to some common yardstick, such as a set of sample essays, or one of several rating scales and the readers' responses to the writing are expressed as a number or numbers of some kind, instead of or in addition to written or verbal comments. Further information is provided on the use of holistic writing assessment and on scoring methods for holistic writing assessment methods. Responses to the paper by Denise McKeon and Joy Kreeft Peyton are appended. (VWL)

Holistic Writing Assessment of LEP Students

Liz Hamp-Lyons
University of Colorado, Denver

Introduction

I thought I was here to campaign for the death of standardized testing, but it turns out that I'm here to say "I told you so." -- not to my physically-present audience, for I am among the converted, but to federal and state bureaucrats who have been antagonistic to or simply afraid of alternatives to standardized testing in general and to direct writing assessment in particular. I only hope that some of those people will read this book, and that this and the many excellent papers from the Symposium will not stay among the converted.

The irony of alternative assessment is that such a term should be needed. We have come full circle to the assessments of the turn of the century, writing prime among them. Is there a connection between the US's role as the multiple choice test capital of the world and an increasing anxiety abut declining educational standards? I think so. Is there a connection between declining literacy and the rise in social ills? I think there is. President Bush's little booklet, AMERICA 2000: An Education Strategy says:

- For too many of our children, the family that should be their protector, advocate and moral anchor is itself in a state of deterioration.
- For too many of our children, such a family never existed.
- For too many of our children, the neighborhood is a place of menace, the street a place of violence.
- Too many of our children start school unready to meet the challenges of learning.
- Too many of our children arrive at school hungry, unwashed and frightened.
- And other modern plagues touch our children: drug use and alcohol abuse, random violence, adolescent pregnancy, AIDS and the rest.

But few of these problems are amenable to solution by government alone, and none by schools alone. Schools are not and cannot

be parents, police, hospitals, welfare agencies or drug treatment centers. They cannot replace the missing elements in communities and families. Schools can contribute to the easing of these conditions. They can sometimes house additional services. They can welcome tutors, mentors and caring adults. But they cannot do it alone. (p.10-11)

But this, it seems to me, is missing the point. Of course schools can't do these things alone; but neither can they achieve the AMERICA 2000 goal of universal literacy alone. Each requires the commitment of federal dollars. But AMERICA 2000 misses the point by a wide margin: It lays the blame for social ills at the doors of families and communities as though there were no record of the sociopolitical changes that have been primarily responsible for the increasing unemployment, poverty, exclusion and alienation lying behind these social ills. It blames "adult misbehavior" without acknowledging that not all the adults who've been misbehaving are in the children's homes or communities -- some of them are in high office, possessing the strings to the purses that contain the children's future opportunities. It lays the blame on the symptoms and not on the disease. And AMERICA 2000 goes on to propose curing the symptoms without attending to the disease.

AMERICA 2000 proposes that universal literacy is a more achievable goal than a nurturing family, a safe neighborhood and enough to eat. Happily, most of us will still be around in the year 2000 to assess the predictive validity of this proposal. My paper, then, is offered not as a claim that reformed practices in the assessment of writing will achieve the goals of AMERICA 2000, but as a range of options for improving writing evaluation as one very small practical contribution to one small part of the problem, within what I hope the National Education Goals Panel will swiftly realize must be a wholistic approach to problem-identification and solution-delivery to "make this land all it should be." (AMERICA 2000, cover page)

Holistic Writing Assessment

Definition "Holistic" writing assessment is the term used for tests which test writing wholly through the production of writing. While holistic writing assessments vary from national assessments such as the National Assessment of Educational Progress (NAEP) to teacher-made tests applied within a school building or even just one classroom, and from elementary school through college and graduate education, they all have certain things in common. A holistic writing assessment has at least the following five characteristics: First, each individual taking the assessment must actually, physically write at least one piece of continuous text of 100 words or longer and may write several pieces and/or considerably longer pieces. Second, while

the writer is provided with a set of instructions and a text, picture, or other "prompt" material, she or he is given considerable room within which to create a response to the prompt. Third, every text is read by at least one, usually two or more, human reader-judges who have been through training for the scoring of writing in that context. Fourth, the judgments made by readers are tied in some way, tightly or loosely, to some common yardstick, such as a set of sample essays, a description of expected performance at certain levels, or one or several rating scales. Fifth, the readers' responses to the writing are expressed as a number or numbers of some kind, instead of or in addition to written or verbal comments; scores on the test are recorded and can be retrieved for review by higher or external authority as needed. It should be clear from the above that a writing test is a **performance** test.

Contrasts "Objective" tests are tests in which discrete elements such as the ability to recognize correct English word order, sentence structure rules such as tense maintenance, and vocabulary items dominate. Objective tests call on recognition skills not production skills: test takers select from a narrow set of choices created by the testers. While these skills may be related to proficient writing, as statistical studies have shown, most of us do not accept that they can represent what proficient writers do. The second kind, "analytic" tests require the test taker to write continuous prose, but instead of evaluating the text they use various count measures, such as mean number of words, word length, sentence length, number of errors per sentence, t-unit length, proportion of simple to complex structures, etc., which are claimed to be highly correlated with writing quality. Analytic assessment of writing does not involve the application of discourse-level measures of writing quality. As with objective tests, an increasingly large number of people, including teachers and researchers, do not accept that analytic measures can represent writing ability. The people who argue FOR holistic writing assessment ground their arguments in construct validity. They believe writing **must** be assessed with a performance sample.

Why assess writing with a performance sample? We live in a society that makes greater demands on the competencies of its members than at any time since the Industrial Revolution, and yet makes it easier than ever before for these members to exist at the fringes of that society in ways that are minimally functional, functional only because of the accommodation of the society to ever lower levels of functioning. I live in a city where more than half the Hispanic population do not complete high school, where 29 percent of the population as a whole and 9 percent of the college population are black. No longer, it seems, does the definition of a civilized society include education for all. What has this to do with writing assessment? Everything.

I am convinced that the methods of testing that have been prevalent in the last half-century bear some responsibility both for the declining educational and literacy standards in this country, and for the changing attitudes to education. "Education" has been reduced to that which can be tested in multiple-choice format, and which can be compressed into an item answerable in 60 seconds or less (since standardized tests depend in large measure on the number of items for their reliability). Teachers find themselves test-driven away from significant educational goals and toward limited sets of assessable knowledge. Children find themselves repeating similar problems again and again, in modes containing extremely low intrinsic motivation, because these are the forms used and areas covered on the test. "Education" no longer means the drawing out of talents, interests and capacities that its Latin origin suggests. An education no longer implies preparation for life and citizenship, for social and moral responsibility. Take a field visit to the pond, to carry out an experiment on specific gravity, or to observe the mating rituals of the crested grebe? Stop and write a poem about the clarity, the smells, the sounds of the day? Freewrite about the scariness of having a plane crash just blocks away from school? Learn to mix clay, to shape and bake it, to feel the simple beauty of it under your fingers, the satisfaction of making? Listen to stories of the lives of your grandparents, your neighbors? Read stories of the ordinary people who inhabit the land, who have made it what it is, the Polish, Greek and Asian early immigrants , the more recent Russian and Vietnamese immigrants, the Native Americans, the descendants of slaves, the Chicanos and Chicanas? Go out into the community and confront social issues, consider resolutions and begin action? Why? It won't be on the test. In my city, where the school-age population is more than half Hispanic American, Cinco de Mayo passed in my son's school with no celebration, no mention. His entire first grade year passed without a field trip.

There are two arguments levelled against holistic writing assessment, or performance testing of any kind. They are, that it is too expensive, and that the results are unreliable. In terms of expense, writing tests are not that much more expensive than standardized tests, since their higher cost for scoring is counterbalanced by the higher development cost for standardized tests. The development and use of writing tests also requires the involvement of skilled people in values clarification, test design, and scoring, bringing benefits in teacher skill development that must also be laid against the cost of direct writing assessment. Writing tests **are** more expensive, and they **do** demand the involvement of a large number of skilled people. But the evidence suggests to me and many others that our views of the cost/benefit of different forms of testing must be redefined to encompass not only test design and administration costs but also human costs and the practical economic consequences of each lost productive citizen. Human costs are not merely figurative, they

are real. Teachers have always known this, but its truth has only recently been understood by business and industry, and it is this new understanding by corporate interests that lies behind the AMERICA 2000 initiatives.

The second argument, of unreliability, has been a diff cult one for proponents of direct writing assessment to counter, in part because reliability is poorly understood. People are used to standardized tests. Test taking, and judgments about answers, go on invisibly, and the judgment process is automated. Questions are rarely raised about what goes on behind the scenes, and it is easy to forget, with standardized tests, that they too are subjective. The items are developed and selected by human judges; they are answered by human beings whose experiences and judgments may be different from those of the test designers; the "correct" responses are decided by human judges, as are the distracting "incorrect" responses. Standardized tests too, then, are not objective, but the scoring method obscures that fact, and people feel confident that they can depend on the scores to be "accurate." Standardized tests are "sold" to us because they are reliable: But this reliability means only that, once someone has decided what the answer will be, a clerical system ensures that only that answer is credited, giving 100 percent scoring reliability. No writing test can compete with that. And yet, scoring reliability is only one side of the issue. A test must not only test something consistently; it must also test the right thing. In this respect standardized tests are more difficult to pin down than performance tests are. Standardized tests claim to test large collections of skills with names like "language proficiency," which in fact has yet to be satisfactorily defined, or smaller sets of skills such as "grammatical competence," but can test it only by sampling a very small subset of the elements that together make up a language user's range of grammatical knowledge. Because they test only a very small subset of the possible microcomponents that make up any one of these larger skill/ability sets, the possibility of a "miss," of testing an element not known by this particular test taker, or of a "false hit," of testing an element this test taker is more familiar with than most others, is quite large. These decisions about test content are made by a small number of test designers, and they are made with a mix of expert judgment and individual variation that is much like decisions made by readers of writing samples. In fact, training for essay readers is highly developed and frequently written about and researched; the same is not true of training for item writers. But because on standardized tests the human judgment processes occur before the individual takes the test and not after, it seems less responsible for the individual results. This is clearly not true.

Educational testers call what testing does to teaching, good or bad, "washback" or "backwash," and it is true there are few empirical studies of it. But look at this country, and you see a giant laboratory,

where the Method has been to construct an educational values system around standardized tests; where the Subjects have been America's school-age population; and the Results are before our eyes daily, on the streets and in the newspapers. Crime; drug abuse and drug pushing; teen pregnancy; gang violence; child abuse; spouse battering and family abandonment; homelessness; poverty. The highest neonatal mortality rate of any First World country. School dropout rates and illiteracy. College dropout rates and unemployment. Can we lay all this at the door of standardized testing? No, of course not. There are other well-documented sociopolitical factors which are in large part responsible. But I submit to you that the decreased attention to literacy in our schools, triggered by the decreased value placed on literacy by our school bureaucracies as represented by their mandatory testing policies, has led directly to decreased literacy at school exit and has been one factor in the rising numbers of semi-functional members of society. And this is a tragedy, not only a criminal waste of human resources, but a deprivation of joy, of growth, of self-knowledge, of opportunities for families to learn and love together. This tragedy cannot be measured. It is not limited to LEP students: It is a rot that has spread right through our education system and so through the society. Last night I walked past the Baptist Church just two blocks from this elegant hotel, where at 11 p.m. were twenty to thirty women and children crowded huddled onto the steps and in knots on the sidewalk. At 6 a.m. today I walked past the Department of Justice and read the words above the door: "Justice is the Greatest Purpose of Men on this Earth" and where I saw five or six men sleeping huddled on the warm air gratings of the building's narrow gardens. I passed the National Archives where I read the legend "The Heritage of the Past is the Seed of all our Futures." And I thought -- yes, and we are living it.

What part can alternative assessment, and holistic writing assessment in particular, play in providing a seed of hope for a more just future for our LEP, our minority, our poor and indeed all our children's futures? I believe it can play a part both through the message it sends to teachers, parents, and learners about what the society values, and through the concrete effects it has in necessitating a kind of "teaching to the test" which is congruent with the needs of the society and the individual future citizen.

In my view then **any** writing test is better than a standardized test. Later in this paper I make the specific argument that there is a form of holistic writing assessment that is ideally suited to LEP contexts. But before I do that, I want to describe the common writing assessment options currently in use. It is convenient to think of five components of a writing test: the writer, the task, the scoring method, the readers, and score reporting. While there is much that could be said on the subjects of writers, tasks, and readers (see Hamp-Lyons, ed. 1991), in this paper I focus on the scoring method

and score reporting, because I consider them to be particularly critical in the design of appropriate writing assessments for LEP students and for the evaluation of LEP education programs.

Scoring Methods for Holistic Writing Assessment

There is some confusion about the terms used in writing assessment, particularly the term "holistic assessment," and I believe it will be fruitful to establish and maintain a clear distinction between the terms "holistic methods of writing assessment" and "holistic scoring." There are several reasons for this confusion: One has been the desire by those in writing assessment to contrast all methods of evaluating writing through the judgment of actual samples of student writing with the objective and analytic methods almost universally used at the end of the 1970s, and still all too common today. The second reason is undoubtedly that direct writing assessment is still a very young field and there few people whose primary research interest lies within it, so that growth is both slow and somewhat haphazard. Although writing was almost universally assessed holistically in the early decades of the century, before the psychometric revolution of the 1930s, it was more of a "cottage industry," with few publications existing in the area. Once standardized tests were developed by and for the large government agencies—especially the Army and the intelligence agencies—research into writing assessment almost disappeared for a generation, and only concern about declining literacy levels in this nation brought it back. But the main reason for the confusion over terms is the difficulty of making clear to non-experts what a writing test is. To many people a writing test is simply the collection of writing, any writing, from students and then the making of impressionistic judgments about the quality of the results. Because the phrase "holistic scoring" has become the best-known one associated with writing assessment, it is not surprising that holistic assessment of writing and holistic scoring have become synonymous in the minds of many teachers and administrators. Add to this the failure of the writing assessment specialists to agree on terminology (a consequence of the youth of the field, referred to above), and the problem is difficult to eradicate. The distinction between holistic scoring and holistic methods of writing assessment is an important one. In a classic paper, Charles Cooper (1977) defined holistic evaluation as:

> any procedure which stops short of enumerating linguistic, rhetorical, or informational features of a piece of writing. Some holistic procedures may specify a number of particular features and even require that each feature be scored separately, but the reader is never required to stop and count or tally incidents of the feature. (p. 4)

This is the definition of holistic assessment used in this paper. "Holistic scoring," "primary trait scoring" and "multiple trait scoring," are all holistic methods for making judgments about writing, as is portfolio assessment with which I close my exploration.

Holistic Scoring

Holistic scoring seems to have been established independently in two similar forms in Britain and the United States, by Wiseman and his colleagues in England and known at that time as the "Devon method" (Wiseman, 1949), and by Educational Testing Service in the United States, best known through the work of Godshalk, Swineford, and Coffman (1966). In holistic scoring (or rather, in focused holistic scoring, the usual method currently) written texts are collected from test takers, usually responding to a quite general question or "prompt" within a limited time frame of 30 to 50 minutes. These are submitted to readers for scoring; readers usually meet together for training and scoring, although in many local holistic scorings readers take essays away to score them. Training is generally fairly limited, typically a session of two to four hours, and generally proceeds by referring immediately to essays and the writing standards they illustrate. There is a scale of some kind, most often running from 1 to 6 (with 6 usually being high), "benchmark" essays are used to show what an essay at each score level looks like. Readers read practice essays and try to match the "expert" scores previously assigned to those essays. The theoretical foundation upon which holistic scoring rests is that readers make judgments of texts as a whole: that they are unable to separate out facets or parts of the essay and identify them. While proponents of holistic scoring argue that holistic scoring "reinforces the vision of reading and writing as intensely individual activities involving the full self" (White, 1985, p33) and that any other approach is "reductive," ultimately agreement on scoring standards is typically reached by each reader adjusting her scores to try to come closer in line with the other readers in the public context of training. Further, holistic scoring requires agreement between readers to be generated from trial scoring of sample papers, and thus depends on the readers involved on a particular day reaching an accommodation among them for the standards they will apply on that occasion. The weaknesses of this approach, both for equitable student evaluation and for program evaluation, are immediately obvious. Adaptations have arisen, most notably the development of essay scales and/or rating guides to accompany holistic scoring sessions, resulting in what is known as "modified holistic scoring" or "focused holistic scoring", and testing agencies, especially Educational Testing Service, have refined the technique into a very efficient and accessible tool. But holistic scoring still yields only one score to express the quality of the student's text.

Figure 1 is an example of an actual writing assessment question used in a statewide writing assessment at eighth grade level, and the scoring rubric, or guidelines, used to score student writing on the prompt.

Figure 1
Holistic Scoring

Task:

We are beginning to understand how important it is for everyone to help protect the environment. What can your school and your class be doing to help the environment?

Rubric:

none

Scoring Instrument:

6 High/Excellent
5 Good
4 High Average
3 Low Average
2 Weak
1 Low/Very Weak

Monitoring for reader reliability is facilitated by the use of two readers for each paper, and readers' scores are correlated. The kind of reporting on the performance of individual students that is possible is shown in Figure 2:

Figure 2
Score Reporting (1) Students

(Class X, Grade 8

COMPOSITION

Adams, J.J.	4
Brown, C.	3
Dong, K.K.	2
Gonzales, R.L.	1
Hunter, W.	5
Jackson, J.	1
Nguyen, M.	2
Rogers, B.	4
Smith, D.	4
Santiago, D.	3
Taylor, B.	3
Weissbaum, E.	5
(etc)	

There are a number of serious problems with holistic scoring in any context, but these problems are especially serious in ESL writing assessment contexts. Chief among these is that holistic scoring is not designed to offer correction, feedback, or diagnosis (Charney, 1984). The integration of evaluation and education is being increasingly recognized in all spheres, and the trend is certainly toward assessment instruments that can inform pedagogical decisions in quite specific ways: This is simply not possible with holistic scoring. We are increasingly coming to view this as a severely limiting feature of holistic scoring, and to demand a richer definition of a "valid" writing assessment. For LEP and other special educational needs students in particular, diagnostic feedback and correction have a central educational role to play. Many LEP students have had only limited exposure to instruction in English, and are only part way through their individual development of their potential mastery of English. Given appropriate instruction, interlingual development remains a real possibility for most of these learners. As Figure 2 suggests, a single score does not provide sufficient information for the student, the teacher or the administrator to decide on the best use of teaching provision in the form of course placement or curricular options, or to set up plans for special services such as tutoring, conferencing or workshops. These services can be especially helpful to LEP students.

Another weakness of holistic scoring is the limited potential it offers for meaningful program evaluation. Suppose two classes in

neighboring schools each use the same holistic writing assessment: the hypothetical data in Figure 3 might result:

Figure 3
Score Reporting (2) Program

	CLASS X (N=30)	CLASS Y (N=30)
SCORE		
6	0	2
5	2	5
4	8	13
3	13	8
2	5	2
1	2	0
	etc	etc

The two classes at the same level have very different results: that much is clear. However, the holistic score data provide no clues as to **why** that might be. Without a more fully-fleshed picture, any generalizations about the effectiveness of curriculum, materials, or teachers would be foolhardy.

Primary trait scoring

A second kind of holistic writing assessment is primary trait scoring, which is in fact, despite its name, more than a scoring method. Primary trait scoring is based on a view that one can only judge whether a writing sample is good or not by reference to its exact context, and that appropriate scoring criteria should be developed for each prompt (Lloyd-Jones, 1977). Primary trait scoring responds to what we have discovered about the influence of task and purpose on any learner's writing, by paying close attention to task specification and to establishing close congruence between writing goals, task demands and scoring. The theory is that every type of writing task draws on different elements of the writer's set of skills, and that tasks can be designed to elicit specific skills. One task might, for example, be designed to elicit the ability to write a formal letter of complaint, and another might elicit persuasion. Primary trait scoring also emphasizes appropriate content, and each task would be expected to elicit certain specific content depending on the exact topic and wording of the prompt. The primary trait scoring guide consists of: (1) the task, (2) the statement of the primary rhetorical trait to be elicited, (3) an interpretation of the task hypothesizing writing performance to be expected, (4) an explanation of how the task and primary trait are related, (5) a scoring guide, (6) sample papers and (7) an explanation of scores on sample papers. Clearly, development of the scoring guide and development of the prompt go hand in hand. I

am going to take as my example, here and in the next section on multiple trait scoring, the same example I used above, and sketch out for you how it might be developed into a better instrument using the primary trait approach or the multiple trait approach. I will not be able to offer you a full instrument because the development of a good writing assessment instrument is a skilled, careful, and time-consuming process, and one that depends absolutely on extreme responsiveness to context. These examples were constructed not for a real assessment but purely for the illustrative purposes of this paper. The examples I give should not, therefore, be taken as examples of excellence but as examples of the shape and direction that excellence might take. Consider Figure 4:

Figure 4
Primary Trait Scoring

Task:

We are beginning to understand how important it is for everyone to help protect the environment. Write a letter to your school principal making some suggestions about what the school and your class could be doing to help the environment.

Rubric:

When you are writing your letter remember that it doesn't help just to complain. You need to have some practical and well-described suggestions for how the school, and your class in particular, can take action to make a difference.

Trait Specifications:

PRIMARY TRAIT= suggesting a solution to a problem

TRAIT DESCRIPTION: The trait requires the identification of actual areas of present environmental concern that relate to the activities of a school (e.g., waste paper disposal). It requires specific language in identifying a problem area and in suggesting a solution (e.g. composting; paper recycle boxes in each classroom, and a class rota of recyclers). It requires use of clear structure to signal a suggestion, e.g., "I think we should..." "What we could do is...." It requires a clearly-made connection between the problem (e.g. a lot of paper gets wasted in schools) and the suggestion for a solution (e.g. recycle boxes), such as, "If we xxxxxx then yyyyyy would no longer happen" or "Using yyyyyy would mean that xxxxxx is not as bad as it is now."

Figure 4 (Continued)

Scoring Instrument:

6 High Writer identifies a real problem in school buildings and names it appropriately. She identifies a reasonable way of dealing with this problem. She shows how it would be possible for the class or the school to put the proposal into action with the resources already available, or she shows how it could be done with only minor additional resources.

5 Good (would be added)

4 HiAv (would be added)

3 LoAv (would be added)

2 Weak Real weaknesses are evident in identifying a problem and suggesting a solution. There is no attempt to show the proposal could be put into action.

1 Low (would be added)

Figure 4 shows, first, a revision of the task in Figure 1: the revision was necessary to fit the more specific tasks implied by the primary trait approach. Then, the trait is named and characterized. The scoring instrument has the same six levels as in the holistic scoring example, but this time a fairly detailed statement of the expectations on the trait to be assessed is provided (I have completed only two of the levels, for the purpose of illustration: note again that is not an operational instrument). When scores are reported for students and groups of students, still only a single number is reported, as shown in Figures 5 and 6, but the numbers are more meaningful than scores from a holistic scoring because they apply **only** to the skill or trait that was assessed. The opportunity to use the language of the scoring instrument to report individual student performance is an important benefit of primary trait scoring, especially in the LEP context. Parents of LEP children are usually LEP themselves, and anxious about their children's ability to succeed in school. Descriptive reporting permits them to see not only a number, interpretable only by reference to some "norm," which in mainstream classrooms is a native speaker "norm," but also some real explanation, which they can read or have a more fluent English speaker read for them, which reports their child's performance against a criterion, against expectations for real language use.

Figure 5
Score Reporting (1) Students

Either
Same as Holistic Scoring

Or
by text description, e.g.:

Farizah's score was 3: she has shown that she can identify a problem and name it but not describe it in full detail with clarity or suggest a reasonable solution to it.

For program evaluation primary trait scoring also offers the possibility of a more explanatory model, as Figure 6 suggests:

Figure 6
Score Reporting (2) Program

Either
same as Holistic Scoring

Or
by text description, e.g.:

In Class X most children identified a real environmental problem and suggested a solution. Five children suggested solutions that were not realistic. No child was able to show convincingly how the solution could be put into effect within the school's existing resources by providing full detail of the operation of their solution. The papers in the middle (levels 3 and 4) were characterized by vagueness of content, etcetera.

In Class Y, two children achieved the highest score by demonstrating a convincing and realistic implementation of the solution to the problem; several other children made a fair attempt at doing this but omitted some important aspect of a workable solution, etcetera.

I believe you can see that the primary trait approach permits a much richer picture of what children have done and how well than does a holistic scoring. The limit is that this information is available only for a single trait, but when students are given several primary trait tasks, the several scores that result can provide a rich diagnos-

tic picture of where that student's strengths and weaknesses lie, and this diagnostic information can be very useful to teachers and administrators as well as to the students themselves. Because of the careful development and detailed specification of the trait and the involvement of teachers and essay readers in test development, when readers use primary trait scoring, they make judgments with the support of an instrument that gives very clear and strong guidance, and the social pressure of the holistic scoring session can be avoided. But the advantages of this ecologically rich assessment are bought at the cost of an expensive development procedure. Whereas when most schools and colleges use a holistic scoring procedure, they transfer and adapt one from a large testing agency with expert personnel and a development budget, the principles of primary trait scoring make this impossible. The competencies specified and tested must be those found to be salient for the context in which the writing assessment takes place, which means very careful needs assessment must precede the test development. In the primary trait method, every writing task requires its own primary trait scoring guide. Not only must each school and college develop its own prompts and primary trait scoring guide, it must do so with almost the same expenditure of time and expertise for every new prompt.

As I developed writing assessment instruments, first for large scale second language writing contexts, then for a first language plus advanced ESL population, I looked for a compromise approach between the rich detail and uncompromising specificity of primary trait, which was beyond the financial possibilities, and the cheap but unacceptably uninformative holistic scoring approach. Building on the principles of primary trait scoring and rather outdated work in analytic scoring, and stimulated in particular by the work of Jacobs et al (1980), I developed what I have called a "multiple trait" approach.

Multiple Trait Scoring

The basic concepts of context-appropriate and task-appropriate criteria that underlie primary trait scoring underlie multiple trait scoring also, and I owe the concept of multiple trait scoring directly to Lloyd-Jones' primary trait approach. The development of multiple trait scoring procedures has been motivated by the desire, first, to find ways of assessing writing which in addition to being highly reliable would also provide some degree of diagnostic information, to students and to their teachers and/or advisers; and second, to find ways of assessing writing with the level of validity that primary trait scoring has, but with enough simplicity for teachers and small testing programs in schools and colleges to apply in the development of their own writing tests. While I have developed multiple trait instruments for English L1 contexts as well as for LEP contexts, and believe in their great value in both, I am convinced that limited English profi-

cient students stand to benefit particularly from a multiple trait form of writing assessment.

"Multiple trait scoring" implies giving separate scores for more than one facet or trait on any single essay. When proponents of holistic scoring object to methods that do this, they are usually reacting against the "analytic" scoring used in the 1960s and 1970s, which focussed on relatively trivial features of text (grammar, spelling, handwriting) and which did indeed reduce writing to an activity apparently composed of countable units strung together, hence the label "analytic," which came to have a derogatory connotation in writing assessment.

But what I am calling multiple trait scoring procedures are very different from the old analytic scoring. Like primary trait scoring, the multiple trait procedure is an approach to the whole writing assessment and not only the scoring. Reader training is the norm in all writing assessments these days, but a multiple trait procedure goes beyond this to include reader involvement in instrument development as a vital components. Like primary trait instruments, multiple trait instruments are grounded in the context for which they are used, and are therefore developed on-site for a specific purpose with a specific group of writers, and with the involvement of the readers who will make judgments in the context. Each is also developed as a response to actual writing on a single, carefully specified, topic type. However, because multiple trait instruments, at least as I have designed them, unlike primary trait instruments do not contain any content specifications, multiple trait scoring instruments can be applied to a range of prompts, as long as those prompts fulfil the initial design criteria for prompts for which the multiple trait instrument was developed, and as long as the context remains essentially unchanged. This makes them more viable for small but committed groups of teachers to develop, pilot, and monitor in their own context, thereafter adding new prompts and paying close attention that new prompts pursue the same writing goals as the original prompts. Of course, multiple trait instruments can be developed that **do** include content specifications, but the amount of work in both development and in training for scoring would be very great. Increasingly, the trend is to develop multiple trait scoring instruments to fit a particular view or construct of what writing is in this context, and to reflect what it is important that writers should be able to do with the written language. "Ideas" are found to be a salient trait in most contexts, but this trait is generally judged in the general rather than the specific (that is, of the nature of "pertinent and convincing ideas," "plenty of relevant ideas," "adequate quality of ideas," etc., rather than "contains ideas a, b, c and d" or "contains ideas a and b but not c or d").

Each of the characteristics of multiple trait scoring I have made brief reference to above is, I think, a significant difference between holistic scoring and multiple trait assessment. The on-site, contextual development of prompts and trait descriptors cannot be illustrated in a paper, but Figure 7, which shows our task again, this time in a multiple trait context, does suggest some of the outcomes to be expected of that development process. Note the explanatory rubric that students receive accompanying the task. Note also the task specifications which guide not only the readers' movement toward shared expectations on this task, but also the processes of communal development of new prompts of the same task-type to be scored on the same scoring instrument.

Figure 7
Multiple Trait Scoring

Task:

We are beginning to understand how important it is for everyone to help protect the environment. What can your school and your class be doing to help the environment?

Rubric:

There are a lot of different ways schools can help the environment, but you will do well on this task if you think of one of them, explain it clearly and show clearly what action the school could take. Be specific and realistic in explaining how your proposal would work.

Task Specifications:

Problem—>Solution. These tasks require the writer to make a clear specification of a/the problem, putting it into the appropriate context. They also call for a textual connection between the problem and a proposed solution. The solution should be explained in enough detail to give it credibility, and it should be convincingly argued. Opposition to or minor flaws in the solution need not be addressed.

Figure 8 shows the beginnings of a multiple trait scoring instrument for scoring this prompt and task-type. Note that, as I have stressed above, development of a multiple trait instrument should be a communal process; certainly it is a time-consuming one. In pursuing my purpose of illustrating the differences among writing assess-

ment methods I have taken a prompt from a holistic scoring and adapted it within each of the methods. Therefore I have only begun to sketch out how trait descriptions might look in the multiple trait approach. To do more would not only be too time-consuming for merely illustration purposes: it might also mislead readers to see this as an actual instrument that might be taken and used in a real assessment context. For a completed, piloted, and validated multiple trait instrument, I refer you to Appendix A and B.

Figure 8
Multiple Trait Scoring Instrument

Score	Trait 1 Problem/Solution text structure	Trait 2 Reasonable content	Trait 3 Development of specifics	Trait 4 Control of the language
6	Problem stated before solution; suggestion made before explanation. Text elements are logically related throughout.	Both problem and solution are reasonable and significant.	Neither problem nor solution is vague. Each is clearly explained. The proposal for how the solution would work is clear, detailed and rational.	Any language problems are too minor for the reader to notice.
5				
4				
3				
2				
1				

There are many positive differences between multiple trait scoring and holistic scoring, but the most obvious difference, and probably the most important, especially in the LEP context, is that in multiple trait scoring more than a single score is generated and reported. In the Michigan Writing Assessment, for example, the instrument I developed generates four scores, all of which are used in decision making, and the descriptive correlates of three of these are reported to the student herself or himself as diagnostic feedback and as a textual explanation of placement in the writing program. (Appendix 1 and 2) Like primary trait scoring, multiple trait instruments focus only on the most salient criteria or traits for the context, and do not claim to assess every facet of writing competence that may appear in the student's writing. This means that careful test development is essential to establish what features are salient, and this development must focus on careful data collection in and about the writing situation where the test is located. At the eighth grade,

for example, participant observation might reveal that teachers considered the ability to see problems outside the self as a salient feature, and one trait in a multiple trait instrument might attend to how far the writer builds comments about how individual choices lead to problems for larger groups into her text. Related to this is the important trait of problem solving, and another trait might focus on the ability to propose and describe solutions to problems. Another salient feature at this level is likely to be evidence of the student's developing control over sentence structure, the ability to use compound and complex sentences in appropriate rhetorical contexts. Discoveries about what features are salient may be made through discussions with teachers, practice scoring, and discussion of a range of essays, study of the marginal notations on in-class writing from the same context, discussion with teachers in other subjects in the school about the strengths and weaknesses **they** note in students' writing at that level, and so on. But the outcome of this data collection stage is always a statement of the salient features to be assessed in this context and on this occasion. The principles and the basic procedures do not change from the college context through the school grades because of its context-dependent nature, this approach is suitable for all levels and situations where writing is assessed.

Figure 9 attempts to illustrate the richness of information about individual performance that can be obtained from a multiple trait assessment (refer back to Figure 7 for the trait explanations):

Figure 9
Multiple Trait Score Reporting

(1) STUDENTS:
EITHER Numerical, e.g.:

Class X, Grade 8

	Problem/Solution		Content	Development	
Language TOTAL					
Adams, J.J.	4	3	5	4	4
Brown, C.	2	3	3	3	3
Dong, K.K.	2	5	2	1	2
Gonzales, R.L.	1	1	2	1	1
Hunter, W.	5	5	3	6	5
Jackson, J.	1	1	1	1	1
Nguyen, M.	2	2	2	2	2
Rogers, B.	6	5	4	3	4
Smith, D.	4	4	4	5	4
Santiago, D.	3	5	3	2	3
Taylor, B.	3	3	4	2	3
Weissbaum, E.	5	6	5	4	5
(etc)					

Figure 9 (Continued)

OR by text description, e.g.:

> Bajni's writing showed excellent control of problem/solution structure, with clear textual relationships. Bajni offered a reasonable problem and solution, but one or both of them might have been more significant. Bajni developed the material fairly well, although there is room for more detail in the writing. Bajni's language control is still developing, and readers are aware of a number of problems of use of language in the writing.

To recap: A multiple trait instrument is an attempt to build up a scoring guide that permits readers to respond to the salient features of the writing whether these are all at the same quality level or are at several different quality levels. The essential characteristics of the multiple trait instrument are its grounding in actual reading data from the context where decisions are to be made; the selection of facets of writing quality in that context shown to be most salient by readers in the context, which in turn permit the reader to attend to what is salient on future reading occasions; and the provision of scores on each of these facets for use in decision making such as acceptance into a program or placement within a program, or in diagnosis of specific problems to be addressed within the instructional context.

Multiple Trait Scoring and LEP Writers

Writing assessment measures very like multiple trait assessment have been used for over a decade now in assessing the writing of second language English writers. Jacobs, Zinkgraf, Wormuth, Hartfiel and Hughey (1981) developed the "ESL Composition Profile," a scoring procedure containing several clearly articulated scales for the scoring of different facets of writing and introducing the term "profile" which I have found so useful. The ESL Composition Profile became deservedly very widely-known and emulated, and has been transferred into and is still used by many college-level ESL programs today. Jacobs et al., worked as a team, they conducted a detailed literature survey, and piloted their instrument carefully; they did not, however, collect observational data from which to build their instrument: rather, they began with criteria previously established for the test and expanded and refined them. Weir (1983) developed a writing test for postgraduates in Britain based on extensive questionnaire data from many British universities coupled with observational studies of faculty at the University of Reading. The collecting of empirical data and building of scales in response to it takes Weir's work

closer to the development process I imply by the use of the term "multiple trait," but Weir did not work with readers as he developed his scoring procedure. Purves (1984) and a team of International Education Association researchers developed a large and complex set of scales for measuring the writing of high school writers in many countries against a common set of values. Although a number of useful insights have come from this work, the size and complexity of the instrument have meant that they are not used outside the IEA-funded studies. I have already referred to some of the insights which came from my work as a consultant to the British Council developing multiple trait instruments for two task types used in assessing the writing of ESL postgraduate entrants to British universities (Hamp-Lyons, 1984, revised 1986).

Each of the studies I have referred to has shown that reliable scores can be obtained using well-designed methods of holistic assessment that are more detailed than holistic scoring -- by which is meant a multiple trait scoring procedure with carefully developed and monitored prompts, a multiple reader system, reader involvement in the development process, and thorough initial and refresher reader training. Each of the studies I have referred to has focused on the assessment of the writing of nonnative writers of English.

Every writer would benefit from sensitive and detailed feedback on their writing, but LEP writers have a special need for scoring procedures that go beyond the mere provision of a single number score. First, for reasons that at present are unclear, LEP writers often acquire different components of written control at different rates. Every instructor of second language writers has encountered those students who have fluency without accuracy and those with accuracy but little fluency. We also sometimes see writers who have mastered a wide vocabulary but markedly less syntactic control; or who have syntactic control not matched by rhetorical control; and so on. With second language writers who already have some mastery of a specialized discipline, it is quite common to encounter texts that show very strong content while grammatical and textual competence lag far behind. De Jong & Henning (1990) have suggested, based on preliminary analysis of a very large data set, a pattern of language acquisition in which absolute non-users of the language have a single dimension to their performance -- zero on everything, and at the highest levels their performance on different tasks and skills once again converges so that they again show a single level of competence, this time a high one: But in between, they advance in different areas more quickly than in others (depending on language background, exposure to English, school and social context, and many other factors), so that their test scores appear divergent and multidimensional. We need writing assessment measures that provide the level of detail that allows such disparities to emerge.

Another argument for the use of multiple trait assessment is that the chances of significant improvement in writing, and the speed with which this can occur, are both greater for LEP writers than for most L1 writers. On one hand, growth in writing proceeds slowly for most first language writers of English after about eighth grade. Second language writers, on the other hand, are in the process of developing their language skills, of acquiring new areas of control and expanding their confidence in areas where they already have some control. LEP writing teachers have the joy of seeing their students make real progress, often in rather short periods of instruction, at any age. The potential for using writing assessment instruments to measure the real language gain of second language learners over a course of instruction (that is, achievement testing) is very real, but once again this means that a detailed scoring procedure is needed.

Another reason for a special kind of scoring of LEP writing is to help ensure that scores reflect the salient facets of writing in a balanced way. LEP writing typically contains significantly more language errors than L1 writing (McKenna and Carlisle, 1991), and the danger is that readers might respond negatively to the large number of grammatical errors found in many second language texts, and not reward the strength of ideas and experiences the writer discusses. This is especially likely to happen where LEP writers are part of a larger test candidate pool containing mainly L1 writers, and readers don't have special training in teaching LEP writing. The opposite can happen too: If the assessment emphasizes ideas and formal argument structures, readers may not attend sufficiently to language errors that would be seriously damaging in most school and college courses. Holistic scoring would obscure a pattern of consistent overemphasis or underemphasis on basic language control. These problems can be minimized by the use of a multiple trait instrument in which this facet is a trait to be judged, together with other facets found to be salient in the context, and where readers are freed to attend to the multidimensionality of ESL writing.

Advantages of Multiple Trait Assessment

While multiple trait instruments are less costly than primary trait instruments because they can be used with multiple prompts that fit the design parameters for the instrument, they are considerably more costly than holistic scoring because of the extensive development efforts involved. What, then, are their advantages?

Reliability When the scores on the multiple traits are combined to create a single composite score in use in making an administrative decision, that single score is highly reliable. In a study of an adapted version of the New Profile Scale developed for the British Council as applied to ESL essays from entirely different contexts, Grant Henning and I found that composite scores were consistently above

.90. (Hamp-Lyons & Henning, 1991). The use of composite scores increases reliability as follows: Assume a multiple trait scoring method with four traits: thus four scores are collected from each reader. Assume also that each essay is scored by two readers, as is the most common practice in writing assessment programs. The result is eight scores, four matched pairs. We may then obtain correlation coefficients for each pair of scores: each of these uncorrected correlation coefficients is an estimate of the reliability of the score on that trait if a single reader were to read each essay and give a score. Because two judges are used, scores will in fact be more reliable than that estimate, and we may use Spearman Brown's prophecy formula, also known as correction for attenuation, to estimate the increase in reliability[1]. Most programs also use a third reader in cases where the first two readers are far apart in their judgments; the way these third scores are used varies, but their result is an adjudicated score that is theoretically closer to a "true" score than the first two scores alone. Generalizability theory (Bachman, 1990) would fulfil the same function, but correction for attenuation can be done quickly by hand by the least statistically literate among us. Thus the multiple trait procedure possesses psychometric properties that enhance the reliability of single number scores built from its components, which can be used for making yes/no decisions such as whether or not to accept a candidate into a program of study where writing competence is required, and for setting cut points such as the level below which a student should be placed into a remedial writing program. While single scores are often used for these purposes, the reporting of the trait scores seems to me to be a vital part of the multiple trait assessment; I will discuss this in detail in the section on Increased Information below.

Validity No test can be valid without first being reliable: only when we have stable score data to look at can we usefully go on to ask questions about validity. But reliability does not imply validity: to judge validity, we need to look at other kinds of data. Following Anastasi, 1982, I take construct validity to be the overarching validity, and it is this type of validity which is central in writing assessment. When a test accurately measures the behavior which defines the construct, it has construct validity. Subsumed within this is content validity, for the traits in the multiple trait instrument derive from fairly concrete expectations in the college or workplace setting. Construct validity and content validity come from careful observation of a context and the shaping of the instrument to fit with those observations. If, when test design is complete, others can look at a test exemplar and see in it the appropriate behavior and values for the context, the test has achieved ecological validity. To ensure content and construct validity, test developers must pay careful attention to the evidence for what is valued in writing in the context to which the writing test applies, design prompts to elicit that kind of writing and scoring procedures to judge those values and ensure that

readers keep those values in mind. These judgments of prompts and scoring procedures are in large part content validity judgments (note that content validity can really only be measured by expert judgments). Cronbach (1949:48) called this "logical validity." This must be coupled with a clear sense of what is involved in the construction of written discourse, of the limitations imposed by the assessment medium -- keeping in mind what it means to write in these circumstances. The text construction in a one-hour impromptu is, after all, a very different matter from the text construction that is possible in a take-home assignment from a course. To then show empirical validity involves statistical validation to discover whether scores are closely related to other measures which are already known to measure the same, part of, or closely related, skills or behavior. This statistical validation is rarely done outside large testing agencies which employ full-time statisticians and researchers, and I would refer you to the Research Reports of ETS for examples of empirical validation.

Increased information A key statistical question that must be resolved when using a multiple trait scoring procedure is whether scores should be combined and if so, how. If diagnostic information is part of the purpose of assessment, clearly, each of the trait scores should be reported separately. If reliability is key, trait scores when combined result in highly reliable scores. In combining scores, we do not know enough (and may never know enough) about how facets of writing weave together and in what proportions, so that decisions about combining and weighing scores are always based on presuppositions and prejudices. If score combining is essential, in my view the safest way to combine scores is to weight each facet equally. If a development team feels a strong urge to weight one facet more heavily than others, that may be an indication that for this context a focussed holistic scoring would be sufficient. Score weighting for purposes of obtaining a single score should always take place with the advice of a statistical expert.

But it is when multiple trait scoring is combined with profile reporting that its chief advantage becomes clear. Profile reporting is the reporting of all the separate trait scores rather than, or in some contexts in addition to, a composite score. Scores exist not simply to **assign** decisions but also to **communicate** decisions. Scores are information which can be shared with the writers, their academic advisors, and other concerned parties and used by them to take various kinds of action in the context of the new information. Although at the University of Michigan we found the information helpful in relation to all students, it has proved especially useful for second language writers.

I have identified two types of profile which profile reporting can convey: the flat profile and the marked profile. In contrast to holistic scoring, where the reader who notices an unevenness of quality in

the writing has no way to report this observation, and must somehow reconcile it as a single score, multiple trait scoring permits performance on different components or facets of writing to be assessed and reported. When the writing in any one sample looks rather similar from any perspective, with no visible peaks or troughs of skill, I call the set of scores on multiple traits which result a flat profile. When the writer shows no extreme variations in performance, as in the example in Figure 10 below, her writing performance may reasonably be expressed as a single score of "6" on a nine point scale without significant loss of information. This is what I mean by a "flat profile": the profile and the averaged score say basically the same thing. But sometimes, and more often with LEP writers for the reasons I discussed above, the writing quality looks rather different from some perspectives than from others. I call the set of scores which result from this unevenness a marked profile (Hamp-Lyons, 1987; Hamp-Lyons & Prochnow, 1989a). In the example in Figure 11, below, the resulting averaged score of "6" does not well describe what the reader sees in the writing, nor does it signal to the teacher what she should expect to encounter when working with this writer in class.

Figure 10
Flat Profile

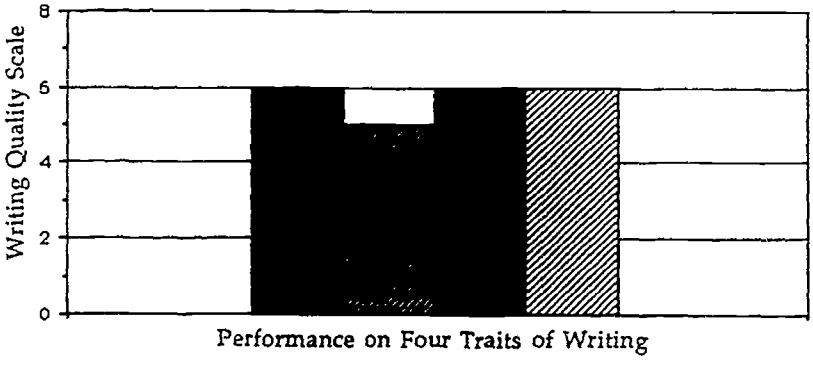
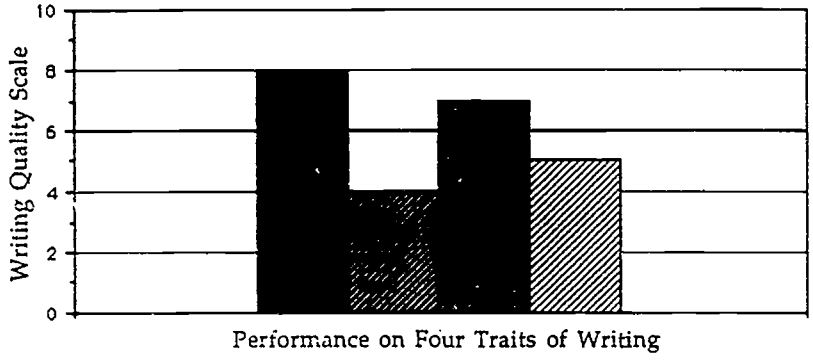


Figure 11
Marked Profile



Knowing the information in the profile is particularly important in two types of cases. If a writer's overall performance puts her into the category of those who will receive special courses or other special services, by looking inside the information provided by the multiple trait instrument, that is by looking at the score profile, the writer, the class teacher, and the program administrator can make good decisions about which course offering or other kind of service would most help this individual writer make progress. Clearly, the provision of special services is particularly likely in cases of special needs students, LEP writers among them. Second, when a writer has generally sound writing skills but a particular weakness in just one area, a single number score would almost certainly fail to reflect the extremely marked aspect of writing performance but separate trait scores would reveal it. While the overall score may not indicate that the writer needs any special help, program administrators, college counselors, the teacher and the writer himself can see the unusual pattern and decide whether to take action about it. Here too second language users of English are likely to be in this category.

These applications to diagnosis and specialized services are the greatest benefits of multiple trait scoring. As the federal government continues to reduce the amount of funding for LEP and other students with special educational needs, yet hypes up the rhetoric about failing schools and this country's resulting decline in world markets at each opportunity, we need to find forms of assessment that will provide more information about LEP students' needs so that the limited resources available for services can be well spent. A multiple trait form of holistic writing assessment does this.

Figure 12 is an attempt to illustrate the ways that the information-rich data generated by a multiple trait type of holistic writing assessment, which uses profile reporting, may explain differences across classes. This type of detailed reporting across classes could answer some of the questions about unsatisfactory results from LEP-funded programs that have been caused by the inability of non-experts to understand the complexities of the problems LEP learners and their teachers face.

Figure 12
Multiple Trait Score Reporting (2) Program

EITHER numerical, e.g.:

	Class X				Class Y			
SCORE	T1	T2	T3	T4	T1	T2	T3	T4
6	5	6	2	2	3	4	7	8
5	6	5	6	4	5	4	7	9
4	5	8	9	8	8	6	6	6
3	8	5	9	8	7	9	5	4
2	5	6	3	5	5	6	3	2
1	1	0	1	5	2	1	2	1

OR by text description, e.g.:

Students in Class X were generally fairly competent in discovering and stating a problem, solution, and the connection between them, and their suggested problems tended to be reasonable and realistic. Students in the class tended to do less well in developing their ideas, with 13 of 30 scoring in the lower half of the range. It was noted that a number of the students in Class X have serious language problems, scoring low on the Language Control category: In particular, five students scored only 1 for Language Control, and five more scored only 2.
Students in Class Y (etc.)

In hypothetical Class X there are a number of LEP students, and their unfamiliarity with writing in English and with the full spectrum of the grammar of the language (I use the word in its linguistic rather than its lay sense here) shows up on the Language Control trait, where their performance contrasts strongly with that of the total group in Class Y, in which (also hypothetically) there are only three LEP students. Not only does the multiple trait report allow the identification of Language Control as the problem area, it also allows

us to see that students in Class X as a whole are doing a good job on higher order cognitive skills such as problem solving, areas where they do not start from a disadvantage. If these data were combined and reported as though they came from a holistic scoring, all this information would be lost.

Salience and Wash back By "salience" I mean that the writing qualities evaluated, and the kinds of writing samples collected are those that have been found appropriate in the context where the assessment takes place. In the British Council writing test referred to above, for example, one writing task (known as the "convergent" task) called for students to read a text and prepare what was in effect a summary, selecting the correct factual content and putting it into a short text of their own, perhaps with graphical material, and using the appropriate vocabulary from the discipline. The multiple trait instrument I designed as a result of work with readers of this test contained the traits of content coverage, presentation format, linguistic features (especially register and lexis), and task fulfillment (see Appendix 3). This task is very unlike the writing task I have used as my example in this paper, where no special knowledge is assumed, no selection skills are called on, answers are expected to be all text, and a general vocabulary will suffice. Because the multiple trait procedure, like primary trait scoring, involves prompt specification and development as well as scoring and reader training, it is a prerequisite of a multiple trait instrument that there is a close match between the writing to be done and the skills and text facets to be evaluated. I argued earlier that all holistic writing assessment has positive wash back -- a positive effect on the teaching that goes on in the context leading up to the test. I believe that this positive wash back is greater for multiple trait forms of holistic writing assessment than any other. This comes from two primary sources: the careful, contextual test development which ensures congruence between teaching aims and testing values, and the provision of score consumers with descriptively informative and accurate test score information appropriate to their potential uses of it.

Improving on Multiple Trait Assessment

In developing writing assessment measures, I have always found myself in the situation of coming in after a good deal of water has flowed under the bridge, and trying to shore up the banks and reroute the waters through fertile lands. This means that certain desirable elements of excellence in a writing assessment are often not within practical reach. What are these? Some of them are commonly-accepted test characteristics that enhance accuracy of information by increasing the amount of information obtained. First, a basic principle of educational measurement is that the more items in a test the more reliable the information obtained will be: a writing

test where the writers write several texts will provide more information about the range of the writer's skills in the contexts and traits that are salient. Second, all modern teachers of writing regret the limited amount of time available for writers to respond to prompts, since these speeded tests run counter to what we know about how successful writers write and to the philosophies of the "process" school of teaching writing. We would like more tasks, and more time: In the trade-off between time and task, there is some evidence (Livingstone, 1987; Hamp-Lyons & Henning, 1991) that LEP writers do not perform significantly differently when they have one hour to respond to a prompt than when they have only 30 minutes to respond to a prompt. And, when Michigan's State Writing Committee experimented with giving several days (a day and an hour for students in third, sixth and eighth grades) to respond to a writing prompt, there was no clear pattern of advantage for any of these below the eighth grade, where the longer led to higher scores. There is, however, considerable evidence (Reid, 1989; Hamp-Lyons & Prochnow,1990) that writers' performances vary considerably across task types. With a school-age population and an hour for a writing test, my preference would be, then, to shorten the time available for writing each task and have two tasks. A better option, of course, would be to increase the total amount of time and have two or more tasks with varying time limits. Another desirable element would be to have writing test data collected in small "bites" on several occasions rather than in the context of a stressful formal test situation. This is, of course, especially important with LEP students who may not be confident in their writing to begin with. Collecting a 30-minute sample once a week for three weeks gives the opportunity for different task types and different contexts, and also for the teachers to build the assessment into the curriculum, making it less intrusive and more educationally meaningful.

The two other elements on my "wish list" may not contribute to making writing assessment more accurate, although each is so poorly understood I don't think we can say that yet, but they would certainly contribute to making it more humanistic. First, it never fails to amaze me how little we know about what the test takers think about the tests, what they do when faced with a test, and I would like to see test design pay more attention to test takers' views and responses. As an example, we often hear it said that LEP students need longer to write on tests because their writing is not yet well-internalized. But we also often hear that LEP writers do less revising, and less global revising than advanced writers, and therefore are unlikely to take good advantage of additional test time, and that has been my own experience (Hamp-Lyons, 1990). But these two statements provide conflicting suggestions for test design. I don't think we can resolve these issues until we spend time in close observation of and conversation with LEP writers as they engage in the writing test event. And second, I think we should put some serious

research effort into self assessment of writing. In my own classes, which typically contain both native and nonnative writers of English, I am becoming more and more courageous in introducing student self assessments into the assignment of end-of-course grades. I am finding that students who have taken a course with clear goals and pathways to achieving those goals finish the course with a very accurate internal sense of how good their writing is and where they need to improve, even though I never assign grades during the course. I find I rarely need to adjust the grade the student suggests for himself or herself by more than a half-grade: The exception seems to be in cases of long-term LEP residents who have made little progress in their English skills, typically because they have become absorbed into a local community of users of their first language and because they have avoided all situations where they might need to use English beyond the level they know they have already mastered. These students often greatly overestimate their writing competence. We have a great deal to learn about self-assessment, about what its benefits and problems are, but involving students in the assessment of their own competencies gives them a responsibility that may be repaid with greater understanding of their own strengths, weaknesses, and needs. It is when learners understand what they need, and take responsibility for filling their own needs, that they exercise the democratic citizenship rights we all believe in, that they move out from under the shadow of paternalism and condescension. We all, teachers and testers, must do all we can to help them make that move toward self determination.

Portfolio Assessment

A full consideration of portfolio assessment goes beyond the limits of this paper, but I must at least mention the rapid growth of interest in and practice of portfolio-based assessment of writing. I think the evidence is now strong that portfolio assessment will eventually become the preferred method for judging writing in many school and college contexts.

A portfolio is a collection of texts the writer has produced over a defined period of time to the specifications of a particular context. Portfolios, usually called "writing folders," have been used in formal assessment in England since the introduction of alternative school-leaving examinations in the early 1970s. Portfolios are used in many disciplines and at all school levels, but they seem to be especially appropriate both for the assessment of writing and for the assessment of the writing of LEP students. Individual high, junior high, and even elementary schools and school districts are using portfolios to monitor learning through the school year. Pittsburgh Public Schools have been developing portfolios in a range of subjects for some years, with a joint Rockefeller grant with ETS and Harvard Project Zero.

Having introduced an ambitious direct writing assessment in the late 1980s, California is now experimenting with portfolio assessment in consortia of schools. States such as Rhode Island are beginning to use portfolio assessment to obtain a picture of achievement in writing across the school system, and even a state with a very large school population such as Michigan has evaluated the need for and practicality of portfolio assessment at certain grade levels in order to obtain a "report card" of writing competencies statewide. Portfolio assessment is rapidly gaining ground at the college level too: at the University of Michigan, for example, they are used to assess exit competence from our pre-composition course (Condon & Hamp-Lyons, 1991; Hamp-Lyons & Condon, 1990), while schools such as Miami University of Ohio are beginning to use optional portfolios as part of entry assessment.

The portfolio usually does not contain writing produced under test conditions, although in some contexts such writing is also judged and considered in decisions such as whether exit competence standards have been reached. Some portfolios are simply a collection of responses to several essay test prompts, usually in different modes, while others incorporate drafts and other process data in addition to final products. The best portfolio assessments collect writing from different points over the course or year and take into account both growth and excellence. Such portfolios require students to include in their portfolio papers which have been revised over a period of time and to provide the original draft and all subsequent drafts. I know of no projects that explore portfolio assessment specifically as this applies to and affects nonnative writers at college level but, in the Michigan writing program exit assessment referred to above, we found that nonnative writers were more likely to be promoted to the next level than when promotion was based on impromptu writing alone. It seemed to us that the opportunities for multiple drafting, self-reflection, and receiving and responding to feedback implied by the portfolio mirror the reality of writing as it is taught these days and the ways students approach writing when it is required in their courses outside English class. Portfolios, because they contain several samples, and because they can be constructed so that texts written under different conditions are included, allow a more complex look at a complex activity, and are therefore generally considered to be more valid. Many problems, not only of reliability but also of the validity of readers' responses, training for portfolio reading, and others (Hamp-Lyons & Condon, 1990) remain to be solved, but the application of portfolio assessment in the ESL writing assessment context is an area that will repay attention in the next decade or less. I hope we will see many studies of portfolio assessment in LEP contexts before much longer.

Conclusion

My purpose in this paper has been to argue for direct, that is, holistic assessment of writing. Unlike some of my education colleagues, I **believe** in assessment, and I applaud President Bush's identification of assessment as a strategy for moving the country toward educational excellence. However, I agree with my colleagues Scott Enright and Mary Lou McCloskey, executive board members of TESOL, when they deplore the President's exclusion of teachers, the expert educators of the nation's youth, from primary input and participation in any of the national strategies including test design. I agree with them when they declare that "Our schools are already burdened by numerous standardized tests which put low-income and language minority students at a disadvantage" and that "we need new ways to recognize and utilize our students' genius, not new ways to label and sort students." (Enright & McCloskey, 1991, p.8). Most tests are based on a deficit model: they point out what the student cannot do, and special needs students are most in danger of suffering from the application of a deficit model to their educational needs. Multiple trait assessment in its most fully-developed form allows a description of both strengths and weaknesses, neither obliterating the other, an approach which holds great promise for LEP students.

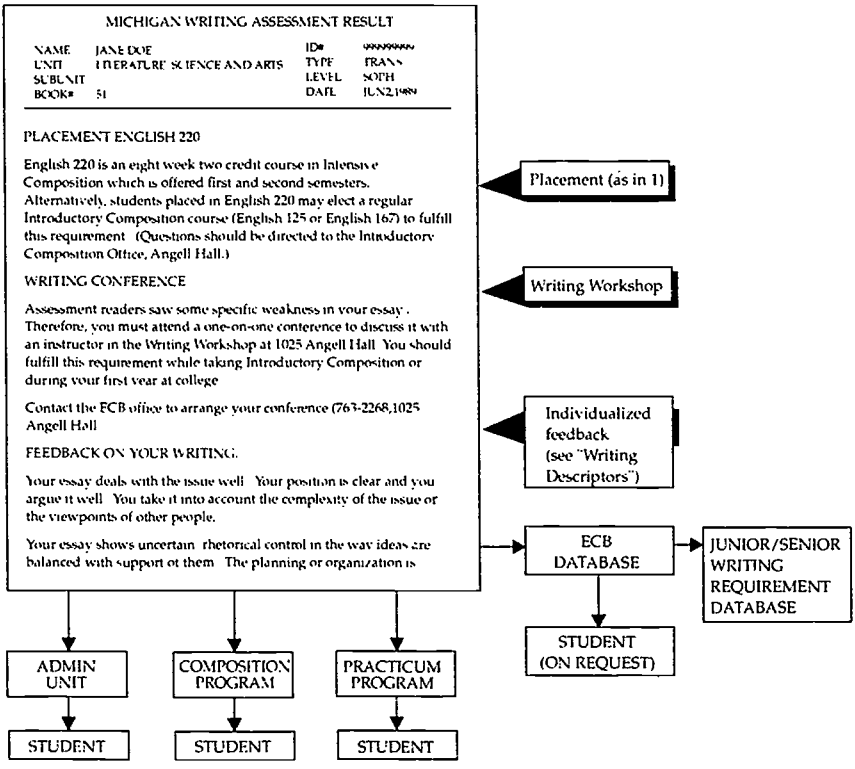
Enright and McCloskey have noted that students with special needs are mentioned only once in AMERICA 2000, and in that reference they are referred to as "at risk". They note too that nowhere in the report is there any mention of the language minority population which makes up about 10 percent of the school-age population nationally. These are discouraging signs for those of us committed to the education of this group and to their integration as fully functioning citizens. Still more discouraging is the lack of reference to the underlying problems in this country, to poverty, malnourishment, lack of affordable child care and health care, to racism and alienation, to the abandonment of millions of women and children by their men and by the welfare system. Assessment is not a quick fix or a cheap fix: good assessment costs money. I think that holistic writing assessment, especially multiple trait assessment, offers a great value for money. But if our LEP children are sick, or homeless, or afraid; if our LEP adult students are unemployed, drug or alcohol addicted, or alienated by and from society, even the best assessments cannot help them.

Appendix A
Michigan Writing Assessment Scoring Guide

English Composition Board: Criteria for Reading the Assessment

	Ideas and Arguments	Rhetorical Features	Language Control
6	The essay deals with the issues centrally and fully. The position is clear, and strongly and substantially argued. The complexity of the issues is treated seriously and the viewpoints of other people are taken into account very well.	The essay has rhetorical control at the highest level, showing unity and subtle management. Ideas are balanced with support and the whole essay shows strong control of organization appropriate to the content. Textual elements are well connected through logical or linguistic transitions and there is no repetition or redundancy.	The essay has excellent language control with elegance of diction and style. Grammatical structures and vocabulary are well-chosen to express the ideas and to carry out the intentions.
5	The essay deals with the issues well. The position is clear and substantial arguments are presented. The complexity of the issues or other viewpoints on them have been taken into account.	The essay shows strong rhetorical control and is well managed. Ideas are generally balanced with support and the whole essay shows good control of organization appropriate to the content. Textual elements are generally well connected although there may be occasional lack of rhetorical fluency: redundancy, repetition, or a missing transition.	The essay has strong language control and reads smoothly. Grammatical structures and vocabulary are generally well-chosen to express the ideas and to carry out the intentions.
4	The essay talks about the issues but could be better focussed or developed. The position is thoughtful but could be clearer or the arguments could have more substance. Repetition or inconsistency may occur occasionally. The writer has clearly tried to take the complexity of the issues or viewpoints on them into account.	The essay shows acceptable rhetorical control and is generally managed fairly well. Much of the time ideas are balanced with support, and the organization is appropriate to the content. There is evidence of planning and the parts of the essay are usually adequately connected, although there are some instances of lack of rhetorical fluency.	The essay has good language control although it lacks fluidity. The grammatical structures used and the vocabulary chosen are able to express the ideas and carry the meaning quite well, although readers notice occasional language errors.
3	The essay considers the issues but tends to rely on opinions or claims without the substance of evidence. The essay may be repetitive or inconsistent; the position needs to be clearer or the arguments need to be more convincing. If there is an attempt to account for the complexity of the issues or other viewpoints this is not fully controlled and only partly successful.	The essay has uncertain rhetorical control and is generally not very well managed. The organization may be adequate to the content, but ideas are not always balanced with support. Failures of rhetorical fluency are noticeable although there seems to have been an attempt at planning and some transitions are successful.	The essay has language control which is acceptable but limited. Although the grammatical structures used and the vocabulary chosen express the ideas and carry the meaning adequately, readers are aware of language errors or limited choice of language forms.
2	The essay talks generally about the topic but does not come to grips with ideas about it, raising superficial arguments or moving from one point to another without developing any fully. Other viewpoints are not given any serious attention.	The essay lacks rhetorical control most of the time, and the overall shape of the essay is hard to recognize. Ideas are generally not balanced with evidence, and the lack of an organizing principle is a problem. Transitions across and within sentences are attempted with only occasional success.	The essay has rather weak language control. Although the grammatical structures used and vocabulary chosen express the ideas and carry the meaning most of the time, readers are troubled by language errors or limited choice of language forms.
1	The essay does not develop or support an argument about the topic, although it may "talk about" the topic.	The essay demonstrates little rhetorical control. There is little evidence of planning or organization, and the parts of the essay are poorly connected.	The essay demonstrates little language control. Language errors and restricted choice of language forms are so noticeable that readers are seriously distracted by them.

Appendix B
Michigan Writing Assessment Scoring Report



Appendix C
British Council ELTS M2 Writing Sub-test: Convergent Task Scoring

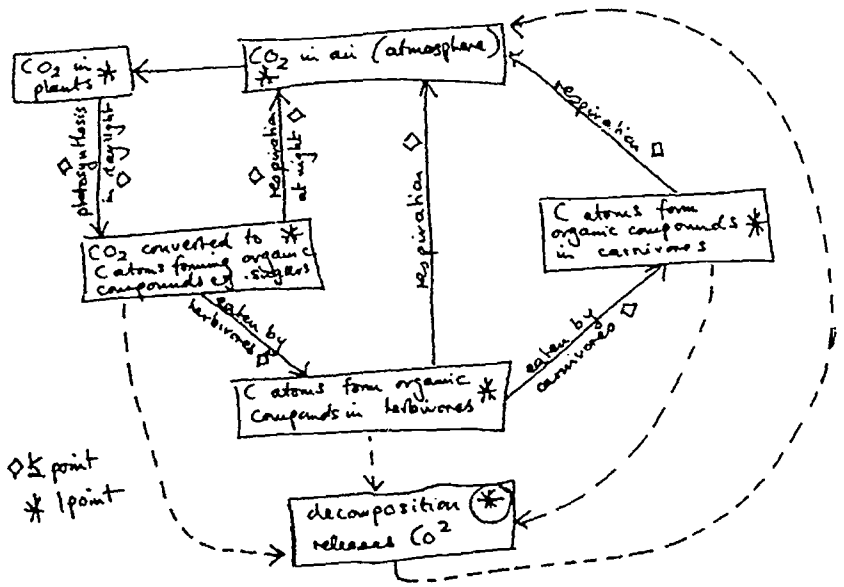
MARKING SUB-SCALES FOR QUESTION 1

CONTENT COVERAGE

THERE IS NO SUB SCALE FOR CONTENT COVERAGE
MARKERS SHOULD REFER DIRECTLY TO THE PROTOCOLS

SAMPLE:
LIFE SCIENCES PROTOCOL 1

(Questions 1 of Versions 4, 5, 6)



Points

----→ • this point is not overtly stated in the text. Candidates may receive 1 point if they include it and have not got all the other Content points.

Presentation Format
Flow diagram.
Outline: all stages must be clearly sequenced.

Appendix C (Continued)

SUB-SCALE for PRESENTATION FORMAT

BAND	DESCRIPTOR
9	The most suitable presentation format is used. It is applied in a way that shows full mastery of it in presenting main points and details.
7	A suitable presentation format is used. The format is applied effectively in general, although one or two inaccuracies in the application of the format to the details may be observed.
5	EITHER A suitable presentation format is used, but it is not applied effectively in the presentation of the information. OR An unsuitable presentation format is used, but it is applied effectively in the presentation of the information.
3	An unsuitable presentation format is used. There are many inaccuracies in the application of the format to the main points and details.
1	No evidence of control over a comprehensible presentation format can be observed.

Appendix C (Continued)

SUB-SCALE for TASK FULFILLMENT

BAND	DESCRIPTOR
9	The overall impression is of a set of notes which fulfills the task fully, clearly and with complete subject command and language control. No irrelevant or inaccurate information is included.
8	The overall impression is of a set of notes which fulfills the task fully, clearly, and with good subject command and linguistic control. No, or very little, irrelevant or inaccurate information is included.
7	The overall impression is of a satisfactory answer which fulfills the task with only occasional, minor, flaws in the subject or language control. Some irrelevant or inaccurate information may have been included, but the clarity of the answer makes it possible to ignore this.
6	The overall impression is of a mainly satisfactory answer although there are some minor flaws of subject or language which detract from the fulfillment of the task. Some irrelevant or inaccurate information may have been included, but this does not seriously impinge on the presentation of the essential material.
5	The overall impression is of an adequate answer, but failure to include some essential information, uncertainly in presenting the notes, language hesitancies, or the inclusion of irrelevant or inaccurate information detract from the satisfactory fulfillment of the task.
4	The overall impression is of an answer which, although it makes a valid attempt to fulfill the task, is too flawed by problems such as lack of information, an inappropriate or unclear approach to note-making, inappropriate transfer from the input text or task, irrelevance, inaccuracy or language weakness to be considered adequate.
3	The overall impression is of an answer which attempts the task but is so seriously flawed in several areas (as listed in band 4) that it does not approach a fulfillment of the task.
2	The seriousness of the flaws in this answer make it impossible to judge it in relation to the task set.
1	A true non-writer who has produced no assessable notes, either because of evident lack of command or because the answer has been lifted wholly or almost wholly from the input text or task (please note which category on the front of the candidate's answer paper).

Appendix C (Continued)

SUB-SCALE for LINGUISTIC FEATURES

BAND	DESCRIPTOR
9	There are no errors or omission in the candidate's application of conventions of register. Key lexis, if appropriate, is present and used correctly. No errors of accuracy or appropriacy in the candidate's linguistic control.
8	There are no errors in the candidate's application of conventions of register but the marker may be aware of certain features of register which would have been appropriate but which are not present. Key lexis, if appropriate , is present and used correctly. There is no inappropriate transfer of key lexis from the input text or task. There are no significant errors of accuracy or appropriacy in the candidate's linguistic control.
7	There may be one or two errors in the candidate's application of conventions of register, and/or the marker may be aware of certain features of register which would have been appropriate but which are not present. The candidate may fail to transfer key lexis when appropriate, but there is no inappropriate transfer of key lexis from the input text. There are occasional minor errors of accuracy or/and appropriacy in the candidate's linguistic control.
6	Several errors are noted in the candidate's application of conventions of register. The marker may be aware of restricted range of register features, or of a failure to transfer appropriate key lexis from the input text, but key lexis is not transferred inappropriately. There are a number of errors or linguistic accuracy and a limited ability to manipulate the linguistic system appropriately.
5	Several errors are noted in the candidate's application of register of conventions. The marker is aware of a restricted range of register features and of a failure to transfer key lexis when appropriate. One or two key lexical items may be transferred inappropriately. Linguistic errors of accuracy or appropriacy intrude frequently.
4	The marker notes a lack of overall command of appropriate register, although one or two appropriate features may be present. The candidate does not transfer key lexis when appropriate. One or two key lexical items may be transferred inappropriately. The control of the linguistic system is generally inadequate. The effect of these failures and omissions is to make retrieval of the information difficult.

Note

[1] If 10 percent of scores received a third score, for example, in the formula K would hypothetically be 2.10 and attenuated reliability would be enhanced: however, a third reader would only be needed in 10 percent of cases if the first two readings were quite unreliable or the standard for a discrepant score very stringent. Standards for recognizing a score as discrepant vary considerably: the TOEFL Program's TWE requires third readings on the basis of a two-point discrepancy on a six scale (33 percent discrepancy criterion), the MELAB uses a two-point discrepancy criterion on a nine-point scale (22 percent discrepancy criterion), and the Michigan Writing Assessment uses a six-point discrepancy on a thirty-six point scale (16.5 percent discrepancy criterion).

References

U.S. Department of Education (1991). AMERICA 2000: An Education Strategy. Washington, DC, Author.

Anastasi, A. (1982). Psychological Testing (5th ed.). London: Collier Macmillan.

Bachman, L. (1990). Fundamental Considerations in Language Testing. Cambridge: Cambridge University Press.

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

Condon, W. C., & Hamp-Lyons, L. (1991). Introducing a portfolio-based writing assessment: progress through problems. In P. Belanoff, & M. Dickson (eds.), Portfolio grading: Process and product. Portsmouth, NH: Boynton/Cook.

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: NCTE.

Cronbach, L. (1949). Essentials of psychological testing. New York: Harper & Brothers.

De Jong, J. H. A., & Henning, G. (1990, March). Test dimensionality in relation to student proficiency. Paper presented at the Twelfth Annual Language Testing Research Colloquium, San Francisco.

Enright, S. & McCloskey, M.L. (1991, Aug/Sep). America 2000: Two TESOL members respond. TESOL Matters, 1, (4), 1-8.

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). The measurement of writing ability, ETS Research Monograph, 6. Princeton, NJ: Educational Testing Service.

Hamp-Lyons, L. (1984; revised 1986). Assessment Guide for M2 Writing. London: The British Council.

Hamp-Lyons, L. (1987). Performance profiles for academic writing. In K. Bailey, R. Clifford, & T. Dale (eds.), Language Testing Research: Papers from the Ninth Annual Language Testing Research Colloquium. Monterey, CA. : Defense Language Institute.

Hamp-Lyons, L. (1990). Essay test strategies and cultural diversity. Pragmatic failure, pragmatic accommodation, and the definition of excellence. [ERIC ED 07-90-019.]

Hamp-Lyons, L., & Condon, W. C. (1990, Nov). Readers' responses to portfolios. Paper presented at the Eight Annual National Testing Network in Writing, New York.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. Language Learning, 41, (3), 337-373.

Hamp-Lyons, L., & Prochnow, S. (1989, March). Person dimensionality, person ability and item difficulty in writing. Paper presented at the Eleventh Annual Language Testing Research Colloquium, San Antonio, March.

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

Livingstone, S. (1987, April). The effects of time limits on the quality of student-written essays. Paper presented at the annual meeting of the American Educational Research Association, New York.

McKenna, E. & Carlisle, R. (1991). Placement of ESL/EFL undergraduate writers in college-level mainstream writing programs. In L. Hamp-Lyons (Ed.), Assessing Second Language Writing in Academic Contexts. Norwood, N.J.: Ablex.

Purves, A. C. (1984). In search of an internationally-valid scheme for scoring compositions. College Composition and Communication, 35(4), 426-439.

Reid, J. (1989). Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. In B. Kroll (Ed.), Second Language Writing: Research insights for the classroom. (pp.191-211). New York: Cambridge University Press.

Weir, C. (1983). Identifying the language problems of overseas students in tertiary education in the United Kingdom. Unpub. doctoral dissertation, University of London.

White, E. (1985). Teaching and Assessing Writing. San Francisco: Jossey-Bass.

Wiseman, S. (1949). The marking of English compositions in grammar school selection. British Journal of Educational Psychology, 19(3), 200-209.

Response to Liz Hamp-Lyons' Presentation

Denise McKeon
National Clearinghouse for Bilingual
Education, Washington, DC

It's a pleasure for me to be here today, and it's also a pleasure to respond to Liz Hamp-Lyons' paper. As a former bilingual and ESL classroom teacher, I have never been a very big fan of assessment and, in particular, standardized assessment. It's not that I believe that we don't need to know how students are doing. It's not that I believe that assessment is inherently bad. It's just that the assessments that have traditionally been used with limited English proficient students, such as standardized multiple-choice tests, had a way of neglecting to show all the things that my kids could do and all the things that they had learned.

In addition, testing always seemed to take away valuable time and resources from instruction and never seemed to give much back. What I always wanted from assessment was some kind of measure that would point me in the right direction, instructionally, with my students; something that would provide me and the students with some guidance as to how to move closer toward that illusive goal of becoming proficient in English. Liz Hamp-Lyons has shown me that there may be hope -- that assessment has really come a long way. She documents the move toward holistic assessment quite eloquently and echoes the concerns that most teachers and responsible test developers have expressed about the testing processes used in assessing writing skills.

I should point out that this movement away from multiple-choice tests of discrete writing skills is linked to a national movement, which parallels the call for development of national standards and school reform, as you've heard enumerable times today, and I'm sure, will continue to hear throughout the course of this program. While many from the school reform movement are calling for a national test, some of those charged with the responsibility of designing and implementing assessment, such as the New Standards project, are, thankfully, exploring ways of making testing more representative of what students really need to know and learn. They are looking for ways to put the instructional cart back behind the testing horse, having curriculum drive instruction, rather than the other way around. Holistic writing assessment is one component of this responsible testing movement.

As Liz pointed out in her paper, there have been some concerns expressed about the reliability of scoring such holistic assessments.

While the amount of information that we have with regard to scoring reliability in relation to LEP students, is quite small, recent evidence of scoring reliability with mainstream populations suggests that this might not be the problem that it has previously been thought to be.

The New Standards project, for example, convened a meeting of more than one hundred elementary, middle school, and high school teachers and educators from five states in July of this year. These teachers and educators, who were conducting direct writing assessments in their own states, met to score each other's sample student papers using his or her own state's rubrics for scoring. The purpose of this activity was to examine whether it was possible to calibrate or compare the results from prompts developed by different states and scoring rubrics developed by different states. The results were astounding. Cross state inter-scorer reliabilities in the range of .81 to .87 were obtained, leading Dan Resnick, one of those involved with the project to remark, "it appears that there are conditions under which human judgment can be trusted."

What seems to be needed in holistic writing assessment of ESL students is both the development of standards and scoring rubrics that are sensitive to students acquiring English as a second language, as well as some accompanying professional development of teachers and educators, which will nurture the type of trained professional judgment that has been shown to be so powerful. There also seems to be some need for ESL teachers and other educators to discuss what it means for an LEP student to be a 3 as opposed to a 4 on a writing test. There appears to be a need for teachers and other educators to discuss what it means to be a proficient writer of English as a second language. Is it the same as being a proficient writer of English? Would holistic writing assessment of LEP students in K to 12, for example, be tied to some measure of exit from ESL instructional programs? Teachers who have had little experience with LEP learners might find the unevenness of their writing surprising. Scoring rubrics that help to alert teachers to the types of unevenness that are to be expected across certain levels of language development could help to guide those teachers in making more accurate assessments of students' abilities.

Let me just say a word here. I have been talking with some people that have been working in writing projects in the Northern Virginia area. They've noticed some very interesting things about the way certain students respond to these types of test taking circumstances. When certain groups of LEP students, for example, who really do quite well in class on a regular basis, get into these test taking situations, they are very afraid of being wrong. They tend to hold back and, in holding back, they tend not to perform as well as they could have if they would have gone ahead and taken the risk, be-

cause risk taking happens to be one particular point of a given scoring rubric. It's a catch-22 which is very interesting.

The development of scoring rubrics becomes even more critical when students who are in the beginning stages of writing are encouraged to use their native languages in school programs. There has been very little work done in determining what certain levels of writing look like in languages other than English at the K to 12 level.

Teacher input into the development of such scoring rubrics is a source of professional development in and of itself. The more experience teachers and other potential scorers have, not only with varieties of ESL or L-1 writing, but also with how those varieties fit against some scale of second language writing, the more they will be able to rate those types of writing discerningly.

There is an additional issues that needs to be raised here, that of the amount of experience that teachers, themselves, have with writing. One question remains to be answered: What is the relationship of scoring patterns of teachers to their own writing experience and competence? In other words, do teachers, who write on a regular basis, score students differently than teachers who do not? Do teachers, who write well, score students differently than teachers who do not? To date, those who have been involved with assessment of holistic writing are those who have had great interest in writing. They believe writing is a valuable skill. They believe in practicing that skill though process approaches and conferencing, and they believe in the use of holistic measures as a viable assessment system. This is a very important feature of what has occurred in holistic language assessment to date.

I am not arguing with any of this. What I am suggesting, however, is that, as more and more teachers and educators become involved in such testing, many of those who become involved will be as crazy about whole language, process writing, and holistic assessment. What will happen as those less enchanted teachers are asked to administer and score holistic language assessments? It would seem important to compare scoring results between those who are "experts with writing" and those who are, for want of a better term, "novices." It would also seem equally important to compare scoring results of those who are fans of holistic language approaches and holistic assessment, and those who are not.

Portfolio assessment, which Liz talked about a good deal in her paper, is one area of writing assessment which has received a great deal of attention and shows great promise for judging writing in many schools and college contexts. The portfolio provides an opportunity for teachers to view multiple samples of student work includ-

ing work that has undergone revision. One important benefit of portfolio assessment is that both teachers and students begin to see the evaluation process as one which involves growth, rather than as one which is an immutable static measure of competence at some point in time. Portfolio assessment allows teachers and students to engage in collaborative examination, examination that provides students with some measure of control in the examination process. What is necessary to determine is how certain pieces, which contribute to the portfolio, are selected. Are the pieces selected by the teacher alone? By the student? What types of writing are determined to be necessary for inclusion? If portfolio assessment is to be used as a representative measure of student work, care must be taken to be as inclusive as possible of all the types of writing that a student is being asked to learn and practice as part of instruction.

Given the paradigm shift that has occurred in K-12 ESL instruction in recent years, this means attending to the emergence and presence of content-based ESL. As more and more programs begin to introduce content-based ESL or sheltered English, the presence of such subject matter must also begin to be addressed in portfolio assessment. Just as writing across the curriculum becomes an important part of content-based ESL classes, it must also be examined through portfolio assessment. The examination of student writing by both trained ESL and content teachers could help to build instructional bridges that result in more meaningful instruction for LEP students.

Another benefit of portfolio assessment deals with the notion of eliciting student work in naturalistic settings. These naturalistic settings allow three things to occur. Student work can be produced under "normal" classroom circumstances, in other words, on a non-timed basis. Student work can be seen as evolving, and data can be collected which reflects students' thinking about the nature of writing. The inclusion of multiple drafts of a particular piece of work allows both teacher and student to reflect on the effect of the instructional and learning process over time.

Since one of the ultimate goals of writing is to produce writers who can self-edit and self-evaluate, the representation of this process in the portfolio is critical. The naturalistic setting in which work for inclusion in portfolios is developed is further enhanced by the underlying assumption that conferencing is an important part of holistic writing approaches. Through conferencing, portfolios and the work which they contain become a reason for talking and thinking about the ways in which language and content interact.

One of the most important benefits derived from portfolio assessment by way of conferencing is the ability to explore metacognitive aspects of student writing. Students can and should be asked ques-

tions such as "how do you know a piece is getting better?, how can you tell that someone is a good writer?, what kinds of things do you usually do to make a story more interesting?" These expressions of student intent and understanding provide important clues about what students know and understand about writing. Additionally, they offer the teacher insight into students' conceptions of what writing is, further providing opportunities for teachable moments.

One additional use of portfolios may be to use them to train student judges of writing. One of the biggest drawbacks in writing has traditionally been that students rarely get to read the work of other students or, for that matter, the teacher. Portfolios provide an opportunity for students to interact with the work of others and to serve as editors to others, by offering suggestions that may ultimately serve as self-instruction. Perhaps the biggest benefit to be derived from portfolio assessment and other types of holistic writing assessment is that they may actually affect a change in how classes designed for limited English proficient students are taught.

While many ESL and bilingual classes have moved to whole language approaches, there are still many places where whole language is not readily accepted. This raises the question of whether holistic language assessment is a viable approach to use with those more traditionally taught ESL classes. While I generally deplore the notion that tests may drive instruction, a move toward holistic language assessment may actually have the effect of changing the way in which instruction gets delivered. You can't perform well on a writing test if you haven't had any experience with writing in class. This fact alone may induce certain districts and teachers who are reluctant about holistic writing approaches to try them.

Thus, performance based assessments may eventually nudge schools away from the reductionist "kill and drill" form of instruction to instruction which enables students to perform well, not only for the tests, but in real life. If nothing else, holistic language assessment will have assisted the processes of teaching and learning greatly if only this is accomplished.

Response to Liz Hamp-Lyons' Presentation

Joy Kreeft Peyton
Center for Applied Linguistics, Washington, DC

There is a great deal to celebrate about this paper, so my response begins with celebration of many of its points. I follow that with some comments about areas in which I think we need to push further, and I close with some questions that still remain for me and probably for most of us.

Celebration

It was extremely heartening, reading this paper and listening to Liz talk, to realize the progress we have made in our thinking about what writing is and how we can best assess its quality. A paper about writing assessment written 10 years ago might have begun with extensive discussion of what Liz calls objective tests (which could also be called indirect tests, since they don't assess writing itself but related sub-skills) and then as a wish, suggestion, or afterthought move to a brief discussion of assessing actual writing samples. This paper begins with the recognition that holistic scoring of actual pieces of writing is the only way writing can be assessed, offers a well-developed and much-needed critique of this approach, and moves us along further with a description of multiple trait scoring. For me, this reflects a great and long-in-coming leap forward in our thinking about writing and its assessment, even though, as Liz acknowledges, direct writing assessment is still a young field, and there is still a lot more work to do.

It is also heartening to realize how far we have come in understanding the importance of content, task, and context in the quality of writing products, and the need to take those into consideration when designing an assessment. We now know that a valid writing assessment must begin long before testing actually takes place, with a needs assessment to determine what the writing context and teaching aims are and what qualities of writing are desired. For far too long, we have designed, scored, and accepted the results of decontextualized writing tests, and we have had very little idea of what actually went on in the programs and classes involved or even what the participants were actually trying to accomplish.

In her "wish list" at the end of the paper, Liz mentions a number of ways that writing assessment might be improved even more:

- Involving teachers in test development and scoring.

- Providing for multiple, revised drafts.

- Collecting writing regularly during the year from different contexts and types of tasks, as part of instruction and not separate from it.

- Including portfolio assessment as an option for even large-scale assessments. (Liz says this will eventually become the preferred method for assessing writing, and I hope it does.)

- Observing students as they compose, to better understand how they approach the tests we design. (The methods for this are already well-developed, through the writing protocol research, and computer programs allow us to do it unobtrusively without imposing on students and without asking them to talk while they compose. For example, Recording WordStar, developed at the University of Minnesota (cf. Bridwell, Sirc, & Brooke, 1985), plays back a student's composing session, and the student and researcher can talk about what the student did and why.

Carmen mentioned this morning that this conference would be helping to set a research agenda, and I think these items on the wish list should be part of that agenda.

That these approaches are already being tried on a small scale in a number of places is another indication of the progress we are making, and I hope that as we continue to think about writing assessment, they will move to the beginning of our papers and the forefront of our thinking and research.

Finally, I celebrate something that Liz laments--the genuine and truly educational activities mentioned early in the paper: taking a field trip to the pond, carrying out an experiment on specific gravity, writing a poem about an important experience, and so on. Liz mentions, for example, that in the school district where her first grade son attends, he didn't take even one field trip during the year, and that if this is a trend in education, it's a lamentable one, and I agree. Although Liz bemoans the absence of these kinds of activities in our schools, they are precisely the kinds of activities now advocated by leading teachers and researchers across the country. They may not yet be hailed in discussions of educational goals at the national level and they may not have reached all school districts (they evidently haven't reached the district in which Liz's first grade son goes to school), but they are slowly gaining recognition and respect, and I believe they will eventually prevail over skill and drill exercises to help students pass some standardized test.

Comments

There are a couple of areas where I would like to see us push further:

First, in the discussion of whether students need more time to complete a writing task, presumably to allow them to draft and revise; Liz mentions research finding that limited English proficient writers do very little revising and don't make good use of additional test time anyway. Therefore, they don't perform differently when given 30 minutes, an hour, or even several days to write. I believe the reason for this is that students have not been taught how to revise. They are so accustomed to submitting first drafts as final products to be evaluated that they don't know what to do with time for revision when they have it. If we want students to benefit from time for producing multiple, revised drafts, we need to teach them how to draft and revise. Until that process becomes a regular part of instruction, we can't expect to see it in assessments.

Second, we may be asking too much of large-scale writing assessments, designed primarily to determine how schools across the nation are doing, to evaluate individual programs, or to make decisions about student acceptance or placement when we ask that they not only yield numbers that can be compared but that they also give correction and feedback to writers. I wholeheartedly agree that writers need "sensitive and detailed feedback on their writing," but no amount of score detail can provide that. Multiple scores on well-defined traits can certainly give a rough indication of where a student is strong or weak and needs to work more, but they cannot replace thoughtful qualitative response to writing. Decisions about how many and what traits to score, whether or not to weight the scores, and whether or not to report the full score profile or only the composite score are all important at the administrative or policy level, but they provide little help to a student working on his or her writing. In the quest for the most descriptive test scores, we need to assure that those scores don't replace actual responses that are relevant and meaningful to individual learners. Someone still needs to react to students' text with text.

Questions

Finally, I have some questions that I don't think any of us have answers to at this point.

First, I don't know how national or even district-wide writing assessments can be very context-specific. Student characteristics, teacher goals, and program exit criteria can be as diverse and numerous as the teachers, programs, and classrooms themselves, and I

don't see how a district or nationally developed assessment can possibly be sensitive to that diversity. The description and scoring of particular traits of writing seem extremely useful within a program or classroom, but can we expect agreement on which traits are important on any broader scale than that?

Second, I think we need to continually question what is the match between what we do and assess in school and the challenges that actually face students when they leave our programs. Whether we use "objective" tests, holistic scoring, multiple trait scoring, or writing portfolios, we still run the risk of focusing solely on school-based writing, which may have little relation to the literacy tasks demanded in the work place (see Harste & Mikulecky, 1984; Mikulecky, 1990). In deciding what students need to be able to do and, therefore what we will assess, we need to be sensitive and responsive to the continually changing situations those students will enter when they leave our programs.

For example, our discussions of writing assessment, whatever the format, revolve almost exclusively around the production of extended, usually expository text, by one author working alone. With the increasing emphasis on collaborative work both in school and in the workplace, is solitary text production really what students will do or need to be able to do? Or is this simply a vestige of our academic tradition, which no longer reflects the way we or our students actually work -- in collaboration with others? In future papers on writing assessment, I would like to see serious attention paid to the implications of collaborative writing practices.

Third, what do the students themselves want and feel they need to learn? Hunter and Harman (1979; cited in Wiley, 1991) note that assessment measures are not negotiated with those tested, but imposed largely by middle-class educators. Involving teachers in the assessment process or studying what students do with the tasks we design may be only first steps. In some portfolio assessments students not only select which writing pieces to include but also critique their own writing and prepare the portfolio for assessment. Is it possible to involve them even more, even possibly in deciding the kinds of writing they will do and helping to establish the evaluation criteria? Especially in programs for adults, it seems that our writing contexts and tasks need to encompass the contexts and tasks in which the students also find value.

Conclusion

We have come a long way in our thinking about writing and its assessment; but there is still more to do, and there always will be more to do, if we are going to be truly responsive to students' learn-

ing needs and desires and to society's changing needs for a literate population. Maybe I'm overly optimistic, but I believe that, as we continue to grapple together with that challenge of the linguistic and cultural diversity now prevalent in our schools and as we test and research new approaches to teaching and assessment now available to us, we will return to an understanding of "education" not as mastery of a set of specific skills, but rather, as Liz suggests, as preparation for life and citizenship and for social and moral responsibility.

References

Bridwell, L., Sirc, G., & Brooke, R. (1985). Revising and computing: Case studies of student writers. In S. W. Freedman (Ed.), The acquisition of written language: Response and revision (pp. 172-194). Norwood, NJ: Ablex.

Harste, J.C., & Mikulecky, L.J. (1984). The context of literacy in our society. In A.C. Purves & O. Niles (Eds.), Becoming readers in a complex society (pp. 47-78). Chicago, IL: The University of Chicago Press.

Hunter, C., & Harman, D. (1979). Adult illiteracy in the United States. New York: McGraw-Hill.

Mikulecky, L.J. (1990). Literacy for what purpose? In R.L. Venezky, D.A. Wagner, & B.S. Ciliberi (Eds.), Toward defining literacy (pp. 24-34). Newark, DE: International Reading Association.

Wiley T. (1991). Measuring the nation's literacy: Important considerations. Washington, DC: National Clearinghouse on Literacy Education.