Catharine Keech Lucas

Toward Ecological Evaluation

(This is Part One of a two-part article. It applauds the recent evolution of writing assessment toward “tests worth teaching to” but urges further movement, away from accountability models of evaluation toward an ecological model that supports rather than inhibits teaching and learning beyond the limits of what we can scientifically test.)

Test tyranny may well be the central fact of the environmental impact of evaluation in our time, at least in American public schools K-12. Although college teachers are less directly affected by a need to shape their efforts to an external test, they too reap the results of the K-12 dilemma, as they work with students who have spent twelve years trying to get right answers.

It is an old complaint—parents, students and teachers protest the arrangement. But many dedicated teachers I have encountered, especially those with uncertain job security, hesitate to adopt new practices that do not directly prepare students for district or state testing, so great is their fear of being found wanting when the scores come in. This counter-productive defensiveness results from an invasive, non-ecological approach to evaluation that is both non-responsive and non-responsible. If we can get good scores, valid and reliable, and if we can do it cost-effectively, it doesn’t matter what that effort does to the teacher, the learner or the curriculum. And if we insist that only what can be well-tested is worth teaching, we can move toward 100% certainty that we know what is being accomplished in the classroom. And that, in the accountability model of education, is all that matters.

Is it possible to bring ourselves as evaluators into a more positive relation with the learning environment without engaging in the kind of wishful thinking that requires the world to change before we can achieve anything? There are signs of hope. The general direction of movement for educational evaluators, especially in the field of writing assessment, has been from a period of all-consuming concern

continued on next page
Toward Ecological Evaluation

continued from page 1

with the test itself, tempered by an occasional but largely unexamined concern for its backwash effect on curriculum, toward a primary concern for what I will call "the ecological validity" of evaluation measures, as we acknowledge and begin to use creatively the impact of the test on teaching.

I see this movement as the beginning of a radical shift, from a worldview in which learning is done in the service of evaluation to one in which evaluation is done in the service of learning; from teaching to the test, to testing for teaching. The shift occurs I believe in two stages: the first stage involves us in trying to create tests worth teaching to—tests that reflect the not the least that ought to be happening in the classroom. This is a critical stage, one that holds promise for staff development and curriculum reform: a good test can have as positive an effect on mediocre teaching as a bad test has negative effects on good teaching.

But it is not enough. Somewhere along the way, for good teachers and enlightened program evaluators and administrators, a further shift must occur, when the teacher gains independence from the test, sees it as a source of information in her constant reinvention of curriculum and methods rather than as the sole authority in establishing what she and her students do in the classroom.

Are There Tests Worth Teaching To?

In one of the most dramatic reversals in education history, English educators have recently succeeded in reinstating "essay" testing on a wide scale, insisting that evaluators create tests teachers ought to teach to rather than ignoring the effects of measurement practices on curriculum. Throughout fifty years of psychometric suasion, teachers have remained unshaken in their conviction that the only valid method of testing students' writing is, in fact, to examine students' writing. In 1974 they were helped in their struggle by the discovery in the popular press of a "writing crisis," which some observers attributed in part to the general lack of demand for writing in school—especially on tests.

But in American education, the preferred cure for any perceived educational lack is, often, to throw a test at it—sometimes with, sometimes without additional support for educational reform. In this case, if the tests were partially to blame, then they could simply be changed. As agencies in charge of literacy education have come to understand that overuse of standardized multiple-choice tests may actually contribute to the decline of writing in the school, they have eagerly turned to direct assessment—"real essay tests"—hoping that the new tests will model more appropriate instructional aims than the old, and that continued, even increased testing, can reverse the damage done by past neglect of writing performance.

What has been the effect of all this testing on the writing curriculum? So far, the tendency to teach to the test has worked in favor of bringing back composing in the public schools from the edge of extinction. Where a writing sample is included as part of proficiency tests or State skills tests, teachers clamor for training in composition as they scramble to include writing in the curriculum, many for the first time in their careers.

Unfortunately, given the nature of externally administered tests and the problems inherent in all direct measures of writing performance, we are about to watch a re-run of an old drama complete with unhappy ending. Current writing assessment practices threaten to kill off kinds of writing, ways of writing, reasons for writing, as teachers borrow large-scale writing-test formats for guides to curriculum. Borrowing from the test as the best guide to teaching persists, like any serviceable habit, and many educational evaluators see a wisdom in using it rather than trying to change it. This course would present no problem had not the same writing crisis that spawned this generation of testing also given birth to new composition goals and practices that are at base incompatible with current testing traditions in our society. As a result, classroom practices arising out of the new understandings of writing are as much endangered by most current models of direct assessment as the old ways of teaching writing were once endangered by machine-scored testing. This is not an easy argument to offer in these days of fresh enthusiasm over more and better essay tests, but I feel a certain urgency in attempting it.

Teachers Versus Testers

Rexford Brown, Director of Communication for the Education Commission of the States, offers an analysis of why our current assessment dilemmas remain identical with and as serious as those of prior eras, even though for the time being the essay test appears to have made a comeback. He identifies the tensions that exist "between the values of composition and the values of assessment, the values of education and the values of policy," and reflects on the continuing gap between the evaluation demands of classroom practice, which arise from the needs of individual learners, and the evaluation demands of larger institutions, which arise from the needs of "policy-makers and other managers of the social enterprise" (1987, p. 44). It is the latter group, he argues, that "view the physical sciences as paragons of truth, (preferring) purportedly scientific and purportedly objective data to all other kinds of information whenever they need to rationalize a decision" (p. 49). This is a natural partnership, since "the hard scientist and the policy maker share . . . [a] belief that only purposive-instrumental rationality (Max Weber's term), with its
capacity for proof and disproof, deserves the name of reason.” Against this he sets the values of “people trained in the humanities,” who:

tend to prefer evaluation models based on experience and intuition . . . evaluation that is qualitative rather than quantitative, context-rich rather than context-free, naturalistic rather than laboratory oriented or artificial, aimed at understanding specific cases rather than general “truths,” involving multiple points of view rather than a single point of view presumed to be objective, and aspiring to persuasiveness and credibility rather than “certainty.” (p. 46)

The opposing rationality belongs to managers who have little use for the value-laden discourse indulged in by people like English teachers:

So thoroughly has this rationality driven normative discourse to the margins that we have no way of discussing excellence—a fine moral concept with a noble history—except in numbers! Excellence can only be understood as having something to do with higher test scores. We of the testing world are thus increasingly charged with the responsibility to define an essentially moral quality in essentially amoral terms . . . (p. 50, emphasis his)

Many of us in educational evaluation have devoted the past two decades to wrestling with these tensions, attempting to honor our allegiance both to the managers or policy makers and to teachers and their students. We remain torn between two desires: on the one hand, we want to control every variable in order to increase efficiency and insure interpretability of scores. On the other hand, we know that the more one controls writers to make their samples comparable, and the more one controls readers to make their scores reliable, and the more one reduces costs to make the test appealing to administrators, the less one knows that is worth knowing.

It is time to examine the optimism of new assessment programs against the realities of continuing discontinuities between teaching and testing. As long as they insist on direct measures, composition people can never be "test proud" in the way researchers are who study more controllable and easily measured phenomena. In language arts performance, "test proud" is "teaching poor." How much pertinent information will we continue to sacrifice in the interests of cost effectiveness and score accuracy, as we try to make essay tests yield the clear-cut results of standardized multiple-choice tests? How long will English educators keep trying to make direct measures produce yes-no, right-wrong certainties about what children can do,

continued on page 12

Editor's Note

Both teaching and researching writing demand that we periodically check our sights, that we focus past the trees in which we often entangle ourselves in order to re-see—and remember—the forest. The forest presents a formidable picture for those of us concerned with the way writing gets taught and learned and with the status of written literacy among our nation’s language learners: it is a picture that ties written discourse to individual and social needs, to cultural values and expectations, to a complex cognitive and social web. It is fitting, I believe, to begin 1988 with a Quarterly that steps back from the trees to consider the bigger picture.

We start this issue with Catharine Keech Lucas, who puts an historical perspective on writing evaluation and assessment and looks toward a farther-reaching assessment model that addresses why and how written language is acquired and developed in the process of classroom learning. We continue with Don Rothman, who reminds us that writing is always a socially situated act and that, for the oppressed and the forgotten, the personal message and the political message are uncomfortably and urgently intertwined. We come next to Wallace Chafe, whose look at punctuation, I suggest, only appears to take us closer to the trees again: Chafe takes us away from the grammarian’s usual concerns with the details of externally imposed rules to suggest that our own inner voices play a crucial role in determining how we want readers to “listen to” our writing. Roger Cherry gives us a review essay of a new book on composition research by Stephen North, and through his review Cherry negotiates with North the big picture of writing research. Miles Myers’s addendum review of the same book speaks directly to classroom teachers. Finally, Barbara Grant puts her own broad perspective on Nancy Atwell’s new book on teaching writing in the middle school.

From another part of the forest, I invite you now to a fruitful reading of The Quarterly.

—M.S.
Toward Ecological Evaluation

continued from page 3

before we admit that questions which yield only yes-no answers are often the ones that matter least?

Managers, testers and researchers can't help dreaming of the perfectibility of essay tests, wishing they were another kind of measure. As a member of that group I too have struggled to construct perfect test questions, create parallel topic demands in pre- and post-tests, make fool-proof scoring guides; I have wanted to train readers so well that inter-rater reliabilities would rival those of machine-scored tests; I have sought larger scoring ranges, huger sample sizes, to insure statistical significance of outcomes. Simultaneously, as a teacher, I have begun to acknowledge that these psychometric aims may be self-limiting: that it is not enough to replace machine-scored skills tests with teacher-scored essay tests, if the latter can only measure abilities so "scientifically" reduced they bear little resemblance to the goals of instruction.

The reduction of the writing task is not, of course, necessary, nor has it been common in most professionally developed essay tests. But many state and local assessments have, in their push for more reliable scores at lower cost, opted for simplified, even mechanistic tasks with such clearly defined criteria for success that all elements of uncertainty are removed. When "writing" as a testable construct is defined as creating a "type A paragraph" or its counterpart, the "five-paragraph essay," the test merely reinforces a kind of teaching that is just as restrictive as earlier training for short answers or multiple choice questions.

In the same spirit, professional large-scale assessments, with the best intentions, evaluate student writing ability by collecting single-sample, timed, impromptu responses, kept as consistent in task-definition as possible across test sittings and scored by consistent standards. This traditional test design evolved as an economic way to increase the predictive validity of machine-scored tests while satisfying public demands for face validity. The construct validity of such tests, however, has now been challenged to the extent that enlightened assessment programs are adopting more comprehensive formats. But only a rare few have shown adequate concern for what I mean by ecological validity: the extent to which a test reflects (and hence reports results from) the whole writing environment of the learner, and the extent to which it impacts that environment in positive rather than negative ways.

continued on next page
In arguing that a new theory of assessment is essential to the reform of teaching in the English language arts, one must be prepared to propose alternatives, to offer a vision of future directions that both includes and transcends the best of what we've learned so far. I believe that such a vision has begun to emerge—in what I have dubbed "ecological evaluation." In the second half of this two-part article I want to sketch its basic premises and the ways in which these differ even from some of our most advanced assessment practices. First, however, I will review briefly the recent evolution of those practices and ask, not only, "What's wrong with where we are?" but also, "What have we gained that is worth retaining as we move forward?"

As a framework for the following discussion, I chart the recent history of writing assessment, suggesting a still somewhat hypothetical movement toward the vision of assessment I would like to urge us towards. (The demarcation between stages is fairly arbitrary, intended to be suggestive in the way all divisions of history must be, granting the complex overlapping of emerging perspectives in each era.) Optimistically viewed, the first two phases can be seen as belonging to the recent past; the third and fourth to the present and future.

Phase One: "Reclaiming Direct Measures"

This phase represents the virtually unanimous position of leaders in English education from the '50s onward, after over twenty years of gauging the effects of standardized multiple-choice testing on school writing curricula: that direct measures must be included in all writing assessment for their intuitively obvious validity and the message they send teachers about what to teach.

In their introduction to ETS's 1987 study, Assessing Writing Skill, Breland and Camp, et al., quote several Phase One champions of direct assessment, and concede:

The logic of (their) arguments is so cogent that despite more than a half century of criticism by educational measurement specialists, the essay remains a principal means of evaluation in courses of instruction of all types. In recent years, the essay has gained more and more advocates as evidence of a decline in writing skills among high school and college students accrues... (Consider) the message that is implicitly sent to students and

continued on next page
Toward Ecological Evaluation

continued from previous page

teachers by direct assessment used on a wide
scale: If large numbers of students are required
to produce compositions for assessments ... students will be encouraged to learn com-
position skills and teachers to teach them. (p. 2)

But the authors shake their heads over the "bleak" history of essay testing, documenting the notorious idiosyncrasies of scorers and concluding that "writing skill is inherently difficult to assess: ... assessments (are) often unreliable and always labor-intensive, expensive, and cumbersome" (p. 3). One hears them sigh as they are forced to add: "Non-essay approaches, on the other hand, lack credibility among many who teach English composition." In the end, credibility with teachers becomes an inescapable criterion for policy makers and psychometricians—something teachers might well remember as they press for further reforms in testing.

The power of Phase One thinking is that it calls for essay tests regardless of their admitted statistical limitations, solely on the grounds of their "backwash" effect on curriculum. But Phase One thinking doesn't solve the problems of essay testing that cause managers to shun it.

Phase Two: "Improving Direct Measures"

Phase Two comprises the profession's repeated attempts to reconcile the demands of English educators with legitimate criticisms of earlier incarnations of essay testing. Phase Two announced itself in 1966, when research on "multiple marking of compositions using rapid impression readings" (the British Schools' Council Examination Bulletin, #12) was published simultaneously with an ETS report showing that "a 20-minute essay scored holistically...by three readers added...to the validity of the total score" (Godshalk, 1966, p. 5).

Further improvements in essay score reliabilities were reported for ETS by Brelan and Gaynor (1979). These researchers guessed at the ripple-effect of more "standardized" question types: that is, after several years of a more consistent and imitable essay topic, teachers became familiar with the genre and, through scoring, with the responses most honored by readers; consequently they were better able to prepare students to produce what the readers were looking for. Implied is another factor—rater experience—that continues to improve score reliability in stable assessment communities over time. Inevitably, as essay tests grow in popularity, the pool of totally inexperienced readers decreases while the presence of "old hands" allows newcomers quickly to assimilate group standards and ways of talking about writing. The results are twofold: improved score reliabilities and a wide and growing network of local assessments through which new research on topic effects and writer differences can be disseminated.

The charge that essay scoring is "cumbersome and costly" compared to multiple choice testing was initially rebutted for local test administrators when the large costs of development for machine-scored tests were added into the equation. And the cost of scoring essay tests was further reduced as experienced readers began to produce sufficiently reliable scores with only two rather than three readings. Finally, the complaints over costs are mitigated when local holistic assessments are seen as reasonably priced staff development programs that can be funded from other sources than an administrator's meager testing allocations.

By 1981, in a comparison of direct and indirect assessment models, Spandel and Stiggins could argue that, when possible, institutions should use both kinds of test in combination. This compromise position has become commonplace since it was first given scientific credibility by the Godshalk report for ETS in '66.

Phase Two achievements are impressive: in trying to perfect holistic assessment, testers have made it more cost effective and so more attractive to schools, while improving its scientific foundations; leaders of holistic readings are more conscious of short-cuts they dare not take; researchers in testing seek ways to test consistency of scoring criteria across assessment sites or years.

The work done in Phase Two has been essential in efforts to increase the use of direct measures in a society governed by a managerial rationality. Since all that managers need are summary statistics, and since those are provided more cheaply and apparently more "scientifically" by expertly-normed machine-scored tests, policy makers and managers are highly unlikely to opt for direct measures until and unless they become competitive in terms of cost, efficiency and accuracy—no matter how compellingly teachers argue that standardized tests don't help learners.

In part the spreading popularity of essay tests is driven by that unanswerable Phase One logic—there is no test of writing like a test that demands writing—in part, by the psychometric improvements of Phase Two; but in part, the movement can be attributed to the powerful practical spin-offs arising from teacher involvement in scoring the samples, involvement that carries the seeds of Phase Three activism.

Involving large numbers of teachers in the direct assessment process has proved to be a double-edged sword for professional testers. The obvious gains include greater teacher confidence in and understanding of the test measures, and a greater likelihood the measure will reflect the aims and
values of the profession. Teachers often come away from well-conducted scoring sessions feeling stretched, re-vitalized, with new respect for colleagues and student writers and a new sense of perspective on their own work. But this intimate acquaintance with the measure also gives teachers a way to discover how it does not reflect their aims and goals, how the topic or the timing work against certain kinds of students, how the "consensus" criteria imposed by a strong leader during a scoring session might over-ride their deepest concerns. They suspect some days that no one really knows where appropriate criteria come from; that sometimes scoring disagreements between readers shouldn't be resolved because they represent two distinct and equally legitimate responses to a piece of writing, just as two critics might disagree about the quality of a piece of published prose; that the writing before them is a put-on or a put-down of the test and reveals nothing about what this funky kid might really be able to do. (The London Board of Examiners, who have been rating essays for half a century or more, have an elaborate system of appeals, since even "multiple markings" don't guarantee that a student's score on an external exam bears an infallible relation to his proved ability in composition classes.) As they grow in sophistication and scoring expertise, becoming more efficient and reliable at producing the "right" score, teachers will occasionally feel the crazy and terrible inadequacy of that number, its irrelevance to the student whose voice breaks through the tedium with a cry that demands a human response they aren't allowed to give.

Inter-penetration between ETS and schools in designing and scoring essay tests may be one of several factors (including more vocal and articulate arguments from inside as well as outside the testing establishment) helping to break down any illusion that we can create "context-free" tests which are intended to have no effect on curriculum, and to repudiate the old argument that any unintended effects are the responsibility of those who misuse the test. Breland and Camp (1987) raise, for the first time in an ETS publication on this issue, questions about ways to "strengthen the relationship between the assessment process and the educational process" (p. 4). This represents a new kind of thinking among large testing establishments. For years we have been concerned with creating measures "uncontaminated" by vagaries in classroom process. Now, having seen that we are unable to keep the measure, in turn, from contaminating the classroom, we are beginning to redefine the professional evaluator's responsibility to include a concern for the testing-teaching relationship.

Phase Three: Creating Tests Worth Teaching To

Rexford Brown complains that as soon as an assessment program accepts some responsibility for "backwash" effects on curriculum and begins to collect writing samples, it is challenged to do something more or something differently:

"We (of the testing world) are constantly criticized: we are not collecting real writing; we aren't collecting finished writing . . . holistic scores don't tell us anything about writing; primary trait scores only tell us about one aspect of one kind of writing. . . ." (Brown, in Greenberg, p. 50)

Such criticisms are inevitable when any field repudiates old practices—in this case over-reliance on machine-scored selected-response tests—while still uncertain what better practices might replace them. But they also reflect an increasing consciousness of the opposing needs of practitioners and managers: teachers and institutions. More voices speaking from better documented observation and experience have been raised to argue that not only indirect but also poorly designed direct measures do not help the learner but, as used in our current system of testing, may contribute to depriving the learner of important experiences.

The current era of fresh efforts to create "tests worth teaching to" comprises Phase Three of the evolution of evaluation practice in our time, and addresses part of the problem of "ecological validity": the extent to which a test impacts the learning environment in positive rather than negative ways. Tierney and McGinley (1987), among others who hold this truth to be self-evident—that "changes in what is tested, for better or worse, lead to changes in what is taught"—go to the heart of the ecological issue. They chart a series of contrasts between what we know about reading and writing and what our tests actually measure, contending that current evaluation practices "artificially neaten" language arts processes and so misrepresent children's ability to read and write, as well as send misleading messages to teachers who look to the test for guidance.

The unecological practices common to traditional holistically-scored essay tests can be summarized under three broad headings (see Keech, 1981):

1. They fail to treat writing as a multiple construct;
2. They fail to treat writing as a process-in-context, different from the process of test-taking;
3. They fail to extract information from the writing samples that might be useful to teacher or learners.

All of these weaknesses are being addressed by Phase Three evaluators in a variety of ways, even as they continue to respect Phase Two psychometric demands and the needs of "managers" for scientifically and legally legitimate measures.

One of the most innovative large-scale attacks on problem number one—treating writing as a global construct in which

continued on next page
Toward Ecological Evaluation
continued from previous page

one kind of task predicts performance on other kinds of tasks—has been mounted recently by the California Assessment Program (CAP). The new CAP writing assessment was undertaken with full awareness of its potential ecological impact: the CAP staff knew that whatever kind of writing task appeared on the test would instantly become the curriculum norm for schools whose state performance ranking was determined by their students' scores on the test. Leaders in assessment and discourse theory were hired to hammer out a list of discourse domains that should be sampled by the assessment program, in order to assure school children an opportunity to write more than a single kind of essay. Prompts within each domain are administered unannounced; teachers cannot anticipate which single kind of writing will be tested and so are encouraged to prepare students for all types.

How does this improve on traditional single-topic, one-domain testing, in which one sample of writing is thought to be equivalent to any other? It quite simply promises no such equivalence. Through massive involvement of teachers in development of topics and scoring guides, through wide dissemination of assessment materials in in-service workshops, CAP demonstrates that there is more than one kind of writing, that different writing tasks demand different skills and processes, that these can be taught, but only if they are recognized and addressed. In time, if teachers get the message, students may no longer learn one year that the "only way to write" is "concisely," with minimum detail and elaboration, while the following year they are bluntly informed that "all good writing" uses many precise adjectives or elaborates with multiple details. In the best scenario, teachers who "teach to" the CAP writing tests will begin to look more closely at which text features are appropriate for which kind of writing task and help students make distinctions formerly blurred by the single-construct view of written discourse. (The worst scenario is that teachers will simply teach a dozen "set" formulas for students to execute on command, in place of the single all-purpose formula that once sufficed. This latter can be avoided, I believe, only by progressing to Phase Four, as I describe it below.)

The second problem area for most large-scale direct assessments is the traditional "shoot-from-the-hip" theming students must do on a timed, one-shot test, in direct contradiction of efforts by enlightened composition instructors to provide opportunities in the classroom for thinking, drafting, revising. The Maryland Writing Assessment has recently attempted to correct this weakness by providing for multi-draft writing in supervised settings. They have encountered interesting problems (one teacher over-supervised the drafting, holding students to planning stages until he approved their going on), which may reflect some confusion on teachers' parts about how this kind of test occasion differs from classroom writing where the teacher is free to become involved at various points. The Maryland program directors will no doubt solve this by calling on the old ethic for test supervision which forbids all teacher intervention; in doing so, of course, they will choose good measurement principles over good teaching principles. An appropriate choice, since there is a difference between practicing a skill under tutelage and performing it on one's own before judges. The moment of testing is not the same as the moment of teaching, any more than umpiring is the same as coaching. The distinction between learning in the classroom and performing on a test, or between teaching (coaching) and supervising (refereeing) is a good one—as long as teachers make it themselves. The danger to date has been that some teachers think classroom activities should be a series of big game simulations, all of which require them to take the role of umpire and ignore the role of coach altogether.

A few assessments are addressing the second and third problems simultaneously by experimenting with portfolio assessment based on collecting writing samples from students' work in the ordinary progress of a class. These early efforts appear as harbingers of the movement into Phase Four, which aims for ecologically, pedagogically and psychometrically sound evaluation: evaluation in the service of learning. Phase Four attacks a further (fourth) widely held view that weakens assessment design: that one of the primary aims of assessment is to produce a single, statistically reliable score for each student, and that production of this score under rigorous conditions is more important than any direct improvement of instruction as a result of assessment. Phase Four evaluators refuse this priority, arguing that such a summary score can be obtained just as easily from assessment procedures that are designed first and foremost to lead instructors and students into deeper insights and more refined perceptions about written texts and developing writers.

Phase Four: Recontextualizing Assessment

The move to Phase Four does not mean—as at first it may appear—a return to the old naive idiosyncrasy in teacher responses to student writing, responses encouraged by the former isolation of teachers from other teachers and from their profession. Recent developments in direct assessment have contributed to a breakdown of that isolation and the building of a new sense of profession for many teachers; movement into Phase Four requires a continuing commitment to team-assessment, to continual sharing and reshaping of assessment purposes, processes, tasks, holistic scoring criteria and methods of analyzing writing samples.

We must also go on working with professional testing firms
and others: the research must continue—on topics, on factors affecting writer performance and reader responses, on the source of criteria for evaluating texts, on statistical properties of change-scores, and the like. But Phase Four demands that, while keeping and extending what we now know about the test, we begin to look with new eyes at what happens before the test and around the test, because of and in spite of the test—what happens besides the test in a good writing classroom. Most importantly, Phase Four asks assessors and administrators to consider an alternative to the accountability model of assessment. It offers a model which requires, not scores, but the assessment process itself to provide the information teachers need in order to improve instruction. It offers new ways of observing and recording that improvement. And it challenges the current system of extrinsic rewards that can only encourage instruction to fit itself ever more precisely to the limited demands of imperfect measures.

What information do teachers need; where can they get it; how can they learn to use it? These are questions addressed by Phase Four evaluators. For decades, ignoring their own curiosity, wisdom, compassion—and frequently their lack of knowledge—teachers have focused narrowly on the pursuit of greater test-wiseness. We can alter the effects of test addiction in the teaching of language arts K-university by re-incorporating evaluation into the teaching and learning processes: by making both students and teachers into newly aware self-evaluators who reclaim the entire evaluation process, from goal setting to measuring outcomes, as part of the learning experience.

The changes in attitude and practice this implies and how such changes might come about is the subject of Part II of this article, to follow.

Catheina Keech Lucas is Assistant Professor of English at San Francisco State University. She is a Teacher Consultant with the Bay Area Writing Project.

References


