Joseph Williams

RE-EVALUATING EVALUATING

So many questions we ask about writing, about teaching it, about research in how to teach it turn on the problem of evaluation that we ought not be surprised at the energy we expend on devising reliable ways to measure our competence at putting one word after another. Who goes to which college, once there who remains, whether the quality of our national prose is sinking toward illiteracy—all such judgments depend on whether we can (1) identify what in a text is most salient to determining good and bad writing and (2) measure it consistently enough to make the measure more than a reflection of its inventor's good taste.

Nor ought we be surprised at those who wonder why the profession hasn't settled the question long before this. Not many other fields have devoted more effort at establishing clear-cut standards of evaluation with fewer results. The NCTE has published a whole collection of measures, none of which are unassailably reliable. The National Assessment regularly assures us that our intuitions about a decline in the writing ability of our students is not the product of irritable old age, but it continues to search for better criteria to evaluate student writing. The Educational Testing Service invites only those it is reasonably sure can grade essays consistently to read the College Placement exams, but still devotes large amounts of time to regulating the grading for consistency and reliability.

The search for reliable criteria has gone in two general directions. One is toward objectively quantifiable features of a text that might correlate with different levels of maturation. These include clause/T-unit and word/clause ratios, counts of errors in grammar usage, number of words written in time, and so on. The other is toward systems that would make more accurate, valid, and consistent the wholistic judgments of paper graders. This has taken the form of training graders to be consistent in

---

Joseph Williams is Professor of English and Linguistics at The University of Chicago.

looking for and evaluating whole essays or particular features of essays, of finding ways to sum differential responses into valid wholistic scores, and more recently, of weighting by various mathematical operations the responses of different graders so that the score of those who tend to grade too stringently and too leniently will be respectively raised and lowered to fit a median response.

Our inability to find simple and reliable measures stems partly from the complex nature of written texts and from our equally complex responses to them. Different components of a written text elicit our judgments and responses from a variety of conscious and non-conscious levels. At any moment, any one of those components might touch most saliently on any one of our responses and thereby dominate the final wholistic judgment. More personally, we may not be able to agree on explicit criteria partly because so few of us are qualified to make reliable judgments in the first place. I suspect that most teachers of composition themselves write fewer words per week than their students, and the vast majority among us never have to write for keeps, never have to produce anything as consequential as a production report or a planning memo.

What follows is not especially a critique of any of the specific methods we now use to evaluate student writing, much less a new one. It is intended rather to raise some questions that I don't think we have attended to as carefully as we might have. I wish I could say that I think the questions will help simplify this matter of evaluation, but in fact their answers, such as they are, seem to complicate it.

Let us suppose that we finally devise a system of training an English teacher to respond consistently within his own grading and with the grading of others, and that we can reliably count objective data such as T-units, errors of grammar and usage, and so on. When we have done this, we would have a means to rationalize and defend admissions procedures, grading, the adoption of better teaching methods, and judgments about any national decline in the writing ability of our college population.

But it is not at all clear that such a system would be more than a self-justifying instrument that had taken its values and hence its measures from those who have not demonstrated any special competence in distinguishing competent writing in any world except their—our—own. That is a harsh charge to make against a whole profession and by no means includes every member in it. But I think it is essentially true.

I want to begin indirectly. Consider for a moment, the American Heritage Dictionary panel on usage and its findings. The criticism heaped on them and their judgments by those familiar with the realities of

modern usage is deserved. It is not merely that they did not represent educated, literate writers. (They averaged a year past retirement age and were by and large, Eastern educated or Eastern employed or both, and for the most part no more technically qualified to pass judgment on good and bad usage than those who edit them.) More seriously, that their judgments were solicited and quantified virtually assured the most Neandralithic sort of majority opinion. No one who has spent a life-time tangling with editors, themselves steeped in 19th c. rules of usage, will easily contradict a body of knowledge it took them years of abuse to acquire. Asked point blank whether the verb *contact* meaning "to get in touch with" is appropriate in formal usage, what could a 66-year-old writer educated at an Eastern university and writing for an Eastern seaboard publication edited largely by others of the same sort be expected to answer, particularly when he knew his opinion would be recorded and printed? The very fact that a writer had achieved an editorial eminence sufficient to call his name to the attention of the AHD staff suggests that he had accepted the values his position implies. And the very acceptance of the solicitation to join the panel constituted the final step in guaranteeing that the panel would be a bastion of linguistic conservatism.

But even if the members of the panel did fairly represent those in the world of letters, their judgments, no matter how close to a consensus they might come, ignore two questions which all such overtly compiled evaluations fail to address. First, even if the proscribed items do not appear in edited, publically printed prose (and it is not the case that they do not), we do not know how often they may appear in that considerably more voluminous quantity of unedited and unpublished prose generated by educated writers in government, industry, commerce, and the professions for their purely internal and private institutional consumption.

Now on the one hand, our professional response is to assert that the standards of usage in studiously re-written, edited published prose should constitute the standard of usage for *all* prose. It is, after all, the sort of prose that is written and presented with the greatest care. But the concept of "care" here is a misleading one. There is no analogy to being careful in, say, medical practice or engineering, where carelessness can have immediately self-evident, objective consequences. Patients die and bridges fall. In writing for publications, the concept of "careful" in regard to a rule of usage has good or bad consequences only to the degree that a reader responds to a violation of that rule.

But if in private prose any rule that holds for public prose is broken

and not responded to as a violation, then to justify the rule we would have to assert that such readers "should" respond negatively, that if they don't then their education failed them. Now this is a curious argument. It requires us to accept the idea that we must arbitrarily generate consequences where none before existed. (The argument that by observing some set of rules we prevent the language from degenerating is, of course, empty.) The only non-arbitrary non-socially based argument for honoring a rule would be if the rule contributed to ultimate clarity. But we know that the overwhelming majority of the usual rules of usage we find in the manuals have nothing to do with clarity or economy, but represent only a set of items whose capriciousness guarantees their imperviousness to mere logic.

In truth, we have publicized a variety of linguistic items as distinguishing literate from illiterate speech, but we have accepted these rules without determining whether educated writing that is *not* edited by people especially trained to identify violations of rules displays those items. We do not know the degree to which these items of usage have been circularly perpetuated as a standard for educated writing because of our assumption that public, printed writing, self-consciously edited by those paid to perpetuate those items of usage, should constitute the standard for *all* educated writing.

Unfortunately, we cannot answer any of these questions by asking. We are all thoroughly familiar with the way almost any educated but linguistically naive person who is put on the spot about correct grammar begins to speak quickly and nervously about grammar being his worst subject, and so on. To directly ask educated but linguistically naive informants would invite only those answers that they could dredge up out of their most insecure memories of junior high school, particularly when it appeared that they were being interrogated by the types that trained those who terrified them in the first place. Nor can we ask them to correct papers in which we have inserted a variety of usage problems, for that would induce even greater uncertainty since such readers would not only have to worry about the correct answers but the correct questions, as well. And even if we examined the writing of this group and found few or none of the items of usage we were looking for, we could conclude nothing, because their absence says nothing about the possible responses of readers if those items were present.

Theoretically, the best way to determine what counts as an error in the minds of non-academic, non-print-oriented writers would be to have them read reports, memos, and so on that each reader had to approve

and send on to his own superior, memos and reports into which had been inserted one or two items of debatable usage, and to repeat this process with many such readers and several items until we found those items for which they would not risk their own prestige. Any more direct method is certain to call up the most regressive sort of response.

Three cases to illustrate what I mean:

(1) I am in the process of drawing up a program to evaluate the quality of writing in the investigative office of a Department of the Federal Government. This particular division has been having increasing problems with the reports prepared in the offices around the country for the rest of the divisions in the Department. In the last two years, according to the director of the office, some of the reports have been delayed for up to six months while their prose was being revised and re-revised into a modest degree of intelligibility. During those two years, the division set up tutorial writing programs staffed by English teachers from the areas around the regional offices. In discussing with the officials the sort of program this division might find useful, I asked to see the comments those teachers had made on the reports they reviewed. They were about what we would expect to find on a carefully marked freshman essay. I asked whether one of the corrections, faulty parallelism, was a serious problem among the report writers. First response: Hesitation; second response: "If he says so."

Now this is an interesting response. A problem exists if the English teacher says it does, even though it may not be felt on the nerves of those who read the reports. None of the administrators would need an English teacher to tell them which reports were disorganized or illogical or pointless or lacking in supporting evidence. Nor would they need English teachers to tell them which sentences might be manifestly nonstandard: *Don't nobody know what goin' on in them offices.* The English teachers were called in to address a perceived problem that seemed to fall between areas which are not the peculiar domain of English teachers. The problem is for us to understand what that domain peculiar to our profession properly includes.

It certainly includes style, particularly in those sentences so confused and prolix that they fail to express what the writer meant. And it ought to include all the rules of usage, both those that are observed by the best publications and those that are observed in literate non-published, non-edited private writing. The crucial problem is not to define literate by the rules germane to print.

But the response, "If he says so," suggests that some believe that there

are other problems which either impinge on our understanding of a text but are beyond the conscious articulation of naive readers or that there are problems which in some metaphysical way violate platonically defined rules of usage. Parallelism, at least in some of its manifestations, may be such a rule. Are the readers of those reports faulted for violating strict grammatical parallelism conscious of that violation? Always? Never? Only certain kinds?

(2) Professor Rosemary Hake of Chicago State University and I have been conducting some research into the ways English teachers respond to different kinds of styles, particularly what we have been calling nominal and verbal. Here are two contrasting examples.

Nominal: There is a need on the part of this office for a determination in regard to the resolution of these matters.

Verbal: This office needs to determine how it is going to resolve these matters.

Given these two sentences point blank, no English teacher reading this would recommend to his students the first as a prose model. And yet when pairs of essays differing only in these two styles were at different times given to English teachers from a variety of institutional backgrounds, most tended to grade the essays written in a nominal style higher than the essays written in a verbal style. What many of us claim we reject we seem tacitly to prefer. The connection between (1) and (2) seems fairly clear: Not only do we not know how readers outside our profession regard different features of language; we cannot even say that we are entirely confident that we know how we respond to them ourselves.

(3) We have replicated this research under a number of different conditions. We have given the papers to graders to take home and mark at their own convenience. We have brought them to the University of Chicago on two Saturday mornings to provide responses they knew we would examine. We slipped papers into a state-wide examination required of all graduates of public colleges. These three contexts set increasingly stringent demands on the graders. The first was entirely non-threatening. The graders had done exactly the task given to them with other papers on other occasions for other purposes. No one was watching them and so far as they knew, they had no reason to be insecure in their responses. The second situation was the campus of a prestigious university where a different set of graders were providing data for research they knew would reflect on them (though they did not know the

specific nature of the research). In the third situation, the graders knew they were constantly being reviewed by grading proctors supervising the entire state-wide examination, proctors who would recommend which of them would be invited back to grade again, a decision that would have both financial and professional consequences, for an invitation to read the essays is regarded as a significant mark of professional recognition.

As the pressure of explicit review increased, the overall average of paper grades declined. And it declined most markedly among those on the lower end of the totem pole: among high school and junior college teachers.

The conclusion that suggests itself would certainly seem to be that the more explicit and personally consequential the task, the more conservative and disapproving become the responses. In light of points (1) and (2), we must become even less certain of what we know. Most of our evaluation is done under self-conscious circumstances. Our own performance is subject to review, if not by our peers, at least by our students, who would like justification for whatever grades we give them. We are only too happy to find criteria to defend strict judgments, judgments which testify to our strict standards. But when we read as unself-conscious readers, we seem to respond rather differently from what we might predict. In what you have read so far, for example, there are a number of errors in usage.

One of the tasks in the preliminary evaluation of the government writing project is to answer as many of these questions as we can. We will circulate among a variety of officials reports into which we have inserted particular errors. We will ask them to read the reports for their content, and only incidentally, to suggest any changes in the texts they think appropriate. The primary task will be to read for overall quality. Of course, even if no one identifies any of the items we insert as errors, we cannot conclude that those items are entirely irrelevant to how readers actually respond, for it may be that they respond to them at some non-conscious level. For this reason, we will recirculate essentially the same documents with the "errors" corrected to determine whether the "corrections" raise their evaluation.

When we turn to the less objectively quantifiable and more subjective questions of style, the problems of evaluation become no less tangled. I have already mentioned the results of Professor Hake's and my research on responses to nominal and verbal styles. Despite the fact that we might all claim that we prefer a clear, concise, direct style with lots of strong

verbs and few abstract nouns, a very large number among us, if our findings are accurate, grade an essay in a nominal style higher than exactly the same content in a verbal style.

But we are faced with essentially the same problem here that we faced with problems of usage; We do no know what counts as good style in places not familiar to those of us in English departments. This is one of the problems our preliminary evaluation of the government writing project will also have to speak to. We are familiar with the turgid bureaucratese that all of us hoot at. Indeed, this is one of the problems of the division we are investigating: Its administrators refer to it as a lack of clarity, as confused sentences and so on. But much of the problem seems to derive from the most common feature of bureaucratese, indirect nominalizations. The deeper problem is why report writers may value this heavy, indirect style more highly than a simple direct style. (At least we tentatively assume that it is valued more highly, since that is the style they use.)

It may be that two systems of values are competing here. On the one hand, the administrators want something that they can read quickly and easily, but the report writers are unwilling—perhaps unable—to be simple and direct. It may be a consequence of bad writing habits, but it may also be the consequence of the first rule of a bureaucracy, not to make oneself responsible for anything. Findings and recommendations couched in governmentalese at least partly cover the writer's ass from recrimination.

Under these circumstances, there is no simple answer to what counts as a good style. In our scholarly innocence, we might value the simple and direct as transcendentally good, much as Thomas Spratt did in the 17th century when writing about the ideal style for scientific prose. But in the real world of government bureaucracies, GS 10's and 12's are—or may be—looking over their shoulders to see who might be watching. And considering the state of a good deal of academic bureaucratic prose, we might have a hard time deciding who among us should cast the first stone.

Questions such as these, of course, also touch on attempts to quantify syntactic maturity. If we can define bureaucratic prose as that hyper-mature writing with more than one nominalization every five or six words, then most recent pedagogical efforts seem to be directed more toward increasing the syntactic maturity of a writer in the direction of bureaucratic abstraction than toward the pellucid prose of an E.B. White. Despite Hunt's disclaimers that increased syntactial maturity is

not to be equated with increased quality, the sense of accomplishment in most recent research papers reporting such gains suggests that raising the syntactic maturity of a ninth grader to the level of a twelfth grader is an unqualified good. The fact that graders also reported an overall improvement in the papers only underscores the value they attribute to syntactic growth. And in the absence of any evidence or arguments to the contrary, there is no reason to disagree.

What follows is not that evidence nor even the argument as much as some questions about syntactic maturity and its unqualified use as a means of evaluation.

As a writer matures, syntax is not the only feature of prose that becomes more complex: organization, a sense of audience, clear intentions, close logic, and so on also mature. One important question is the order in which these mature. We know that projecting ourselves into the role of audience is something most of us never completely master. Nor are logical arguments as natural a level of achievement as, say, puberty or 11.5 words per clause. Thus we ought not accept quantitative measures of syntactic development as good indications of—what shall we call it—rhetorical maturity, regardless of the attractive objectivity that the quantitative measure seems to provide. In fact, syntactic maturity may be a misleading measure, at that.

The figures most often cited are these:

grade:	7	8	12	superior adults
words/T-unit	9.99	11.34	14.4	20.3
clause/T-unit	1.30	1.42	1.68	1.74
words/clause	7.7	8.1	8.6	11.5

Hunt and O'Donnell have suggested that of the two, words-per-clause most sensitively indicates growth. The main problem with this measure is that we have no idea what affective consequences these figures entail. Do we affectively discriminate between texts whose word/T-unit ratios differ by one word? two words? three words? Physiological maturity is ordinarily accompanied by a change in the ratio of cartilege to bone, but under most circumstances, the results of those changes have no appreciable consequences on how clothed adults relate to one another. Growth is a fact of maturation, but it makes no social difference. Word/clause ratios increase as a writer matures, but where is the threshold for perceived differences? There must be some difference at some point, but we have no idea where, and if we have no idea where, then we have no

valid way of making the evaluation relevant to our rhetorical concerns.

Furthermore, though Kellogg has figures that reflect the prose of what he calls superior adults, those whose self-consciously written, revised, and edited prose appears in the Atlantic Monthly and Harpers, we have no extensive figures for workaday world prose, the private prose I described earlier. A case in point: Professor Hake obtained a number of memos and reports from a large manufacturing concern in the Chicago area. We asked those administrators who had to act on the documents to rate the perceived quality of the prose on a scale of 1 to 10, according to whatever criteria seemed appropriate. We selected several from the extreme ends of the scale and analyzed their clause/T-unit ratio. Those rated low on the evaluation had a clause/T-unit ratio of about 1.5, roughly equivalent to the prose of a ninth grader. The documents rated high, on the other hand, had a clause/T-unit ratio of 1.3, about equivalent to the prose of a seventh grader.

Now at first glance, this would seem to contradict the figures that Hunt and O'Donnell gathered, but in fact, it tends to confirm them, unfortunately. A lower clause/T-unit ratio means a higher word/clause ratio, the figure they identified as most salient to maturity. When we recall how our evaluators responded to nominal and verbal styles, the pieces fit together. The memos with fewer clauses had more nominalizations, a construction which reduces the clause/T-unit ratio and increases the word/clause ratio. And a text written in a style with more rather than fewer nominalizations tends to be evaluated more highly than one written in a verbal style.

But doesn't this present us with a pretty problem? We English teachers —and virtually anyone else we might ask point-blank—would almost certainly prefer a verbal style for reasons none of us would find difficult to articulate: clarity, economy, directness, honesty, and so on. And yet when our preferences are probed indirectly, quite another set of values and responses seems to emerge, at least for a large number of us. If this is the case among writers of private prose as well, as the evidence slightly suggests, ought we English teachers adopt such criteria not merely as a measure of syntactic maturity but as explicit objectives in the teaching of style? Just as we have sentence-combining exercises we might have nominalization exercises that would by increasing the frequency of nominalizations lower the clause/T-unit ratio and raise the word/clause ratio.

An argument could conceivably be made that such an objective would not be entirely dishonest. As we have mentioned before, a heavily

nominal style sounds authoritative and judicious, but simultaneously allows a writer to avoid directly stating unpalatable or expensive truths. Every bureaucrat has learned to avoid taking responsibility not only for decisions but for the facts on which such decisions are based. Uncertainty leads to caution, abstraction and indirectness. If these two conditions are facts of bureaucratic life, of whatever industrial, commercial or governmental origin, then to persuade writers to write in clear, concise, and direct language is to ask them not just to change their habits of writing but, at least in their minds, perhaps, to risk their professional position.

Furthermore, we could find ourselves in exactly the situation I urged earlier: Just as we are perhaps wrong to insist that faulty parallelism and so on are mistakes if they do not elicit unfavorable responses in casual readers, so would it be a mistake to argue that a bureaucratic style is wrong, simply because it offends our sensibilities. But there is a difference: One of the problems with a bureaucratic style is that it resists easy reading. Often, it even resists strenuous reading. In virtually all matters of usage, the principle of clarity is rarely if ever invoked. *Data* as a singular, *irregardless* as a connector, *less* modifying count nouns—not one of them is obscure or ambiguous. But a sentence like this is virtually impenetrable:

> There is now no effective mechanism for introducing into the initiation and development stages of reporting requirements information on existing reporting and guidance on how to minimize burden associations with new requirements.

But one more inversion: From the bureaucrat's point of view, an opaque style is good, difficulty in understanding is good, confused meaning is good. Or is it? Is it really a bad habit that once corrected will give way to the concise style of an E.B. White?

We hope that at the end of our project with the government agency, we will know. What we know now is that we know very little; what we do know raises more problems than it resolves. One of the problems these considerations raise is that our understanding of good and bad, right and wrong, effective and ineffective may not be as straightforward as most rhetoric texts make them out to be.