DOCUMENT RESUME

ED 318 009 CS 212 295

AUTHOR Schriver, Karen A.
TITLE Evaluating Text Quality: The Continuum from Text-Focused to Reader-Focused Methods. Technical Report No. 41.
INSTITUTION Center for the Study of Writing, Berkeley, CA.; Center for the Study of Writing, Pittsburgh, PA.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Mar 90
NOTE 42p.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Information Analyses (070)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Audience Awareness; *Critical Reading; Discourse Analysis; Reader Text Relationship; *Revision (Written Composition); *Technical Writing; *Writing Evaluation; Writing Research; Writing Skills
IDENTIFIERS Text Factors; Textual Analysis

ABSTRACT

Arguing that writers must be able to evaluate the quality and effectiveness of the texts they produce, this paper begins by isolating some of the persistent questions raised by people in education, business, and government who want to judge how well their texts are working. The paper then compares the cognitive processes involved in "reading to comprehend text" with those involved in "reading to evaluate and revise text," stressing that even experienced writers often need help in detecting and diagnosing text problems. The paper then characterizes three general classes of tests for evaluating text quality: (1) text-focused; (2) expert-judgment-focused; and (3) reader-focused approaches. The paper reviews typical methods within each class and discusses the relative advantages of reader-focused methods over other approaches. (Four figures are included; 150 references are attached.) (RS)

Center for the Study of Writing

Technical Report No. 41

EVALUATING TEXT QUALITY: THE CONTINUUM FROM TEXT-FOCUSED TO READER-FOCUSED METHODS

Karen A. Schriver

March, 1990

University of California, Berkeley
Carnegie Mellon University

CENTER FOR THE STUDY OF WRITING

Technical Report No. 41

EVALUATING TEXT QUALITY: THE CONTINUUM FROM TEXT-FOCUSED TO READER-FOCUSED METHODS

Karen A. Schriver

March, 1990

Also appears in *IEEE Transactions in Professional Communication*, Vol. 4, December, 1989.

University of California
Berkeley, CA 94720

Carnegie Mellon University
Pittsburgh, PA 15213

The project presented, or reported herein, was performed pursuant to a grant from the Office of Educational Research and Improvement/Department of Education (OERI/ED) for the Center for the Study of Writing. However, the opinions expressed herein do not necessarily reflect the position or policy of the OERI/ED and no official endorsement by the OERI/ED should be inferred.

CENTER FOR THE STUDY OF WRITING

Abstract

To create texts that meet the needs of audiences, writers must be able to evaluate the quality and effectiveness of the texts they produce. Over the last sixty years, a variety of text-evaluation methods have been developed and writers can now choose among many alternative methods. This paper begins by isolating some of the persistent questions raised by people in education, business, and government who want to judge how well their texts are working. It then compares the cognitive processes involved in "reading to comprehend text" with those involved in "reading to evaluate and revise text," stressing that even experienced writers often need help in detecting and diagnosing text problems. The paper then characterizes three general classes of tests for evaluating text quality: (1) text-focused, (2) expert-judgment-focused, and (3) reader-focused approaches. It reviews typical methods within each class—examining the strengths and limitations of particular tests—and discusses the relative advantages of reader-focused methods over other approaches.

EVALUATING TEXT QUALITY: THE CONTINUUM FROM TEXT-FOCUSED TO READER-FOCUSED METHODS

by

Karen A. Schriver
Carnegie Mellon University

We frequently read texts by writers who fail to consider our needs as readers. Writers may forget to provide a necessary context, fail to include examples, obscure the purpose, leave out critical information, or write too abstractly. Writers of all ages from almost every profession share two questions: How can we anticipate and meet the reader's needs? How can we know if we were successful? Writers have been found to have genuine difficulty both in considering the reader's needs while planning and generating text as well as in judging their success during revision. Thus, it is not surprising that people in education, business, the health professions, and government have been looking for reliable ways to evaluate the quality of texts they create.

Since the 1930s, a variety of document-evaluation methods have been developed and writers are now in the position to choose among alternative evaluation methods. In this paper, I categorize typical methods for evaluating text quality into three general classes: *text-focused, expert-judgment-focused,* and *reader-focused* approaches. My aim is to give an overview of popular methods and to identify their strengths and weaknesses within the context of what is known about text evaluation.

Initially, I discuss research in reading and writing that has investigated the thinking processes of people as they engage in evaluating text with the goal to revise. In particular, I compare the cognitive processes involved in "reading to comprehend text" and "reading to evaluate and revise text." This research raises the issue that an adequate theory of text evaluation must account for what people do as they read with the intention of judging text quality. This work also points out that adequate testing methods must provide writers with what they need most for planning or revising: an image of the intended audience interacting with the text. I then discuss these issues in the context of the most frequently used methods within each of the three classes—text-focused, expert-judgment-focused, and reader-focused approaches—and show why reader-focused methods have relative advantages over other approaches.

QUESTIONS RAISED BY TEXT-EVALUATION RESEARCH

Text evaluation is a difficult and tangled issue. If you asked a room of researchers or practitioners in the area "What are the key questions in text evaluation" you would hear a wide range of issues:

- What are the characteristics of an effective text?
- Can we agree on a working definition of text quality?
- What are the key skills and abilities involved in text evaluation? What do experienced evaluators do that inexperienced evaluators do not?
- What do writers learn from repeated experience in judging text quality?
- How can we improve evaluators' abilities to judge text quality?

- What are the tradeoffs associated with particular methods for judging text quality? What methods produce reliable and valid judgments?
- What aspects of text evaluation can we automate using the computer?
- How can the computer help reduce the burden of text evaluation?

Underlying these questions are several themes: Can we identify benchmarks for characterizing quality text? Can we teach evaluators to judge the quality of text consistently and reliably? Can we identify ways to help evaluators improve their skills in judging text? How can technology help us in our efforts to assess text quality? Much of the work that is directed toward answering these questions has been conducted by theorists and researchers in reading, rhetoric, composition, and document design.

Reading researchers have been trying understand differences between what they term "considerate" and "inconsiderate" text [1-5]. They have been exploring the kinds of text structures that promote or inhibit comprehension and want to know more about what happens to the comprehension process when we encounter poorly written text. Such work sheds light on what readers do in constructing a representation of a text—whether the text is well formed or ill formed. They emphasize that we need more empirical work identifying the global and local textual relations which help readers to construct a coherent model of the text's information.

Studying literacy in the workplace is also helping us to understand the demands of reading, showing how dramatically work-type reading differs from school-type reading [6-10]. Such research makes it clear that to meet the unique needs of readers in nonacademic contexts, writers need detailed information about the kinds of reading that gets done, especially information about the diverse purposes, goals, and strategies for reading at work.

Research in rhetoric, writing, and document design has been trying to identify the key variables which underlie skilled performance in creating rhetorically effective text. There are now a number of studies which aim to characterize the processes involved in planning, writing, and revising text for readers [11-16]. Such studies are exploring the cognitive, social and cultural processes of writers as they engage in creating and evaluating text. The results show large differences in writers' abilities to judge text from the perspective of the audience. Both experienced and inexperienced writers have been found to have more difficulty evaluating texts they write themselves than those written by other writers. In other words, it is easier to identify the strengths and weaknesses of someone else's text than one's own. For such reasons, researchers have been particularly concerned with identifying text-evaluation methods that help writers judge text from the reader's point of view [17-24].

Taken together, work in these areas is changing our thinking about the problem of assessing text quality and is laying the foundation for a theory of the process of evaluation (see reference [25] for a review of the literature). Such efforts are helping us make more informed decisions about what makes a text-evaluation approach useful. Moreover, we are beginning to identify methods that have the advantage of enhancing both a *writer's* process of evaluating text as well as the *reader's* process of comprehending and using text.

READING TO COMPREHEND VERSUS READING TO EVALUATE TEXT QUALITY

To understand what an optimal text-evaluation method might look like, writing researchers have been examining the process of evaluation itself—that is, the writer's

cognitive processes of evaluating text with the goal of revising it for comprehensibility and/or usability. What is it that expert writers do when making revision decisions that improve the text from the reader's perspective? Do people "read differently" when engaged in revision? In a recent study of revisio.., Hayes, Flower, Schriver, Stratman, and Carey [14] asked the question: How is "reading for comprehension" different from "reading to evaluate?" Figures 1 and 2 present hypotheses about what some of the differences may look like. Figure 1 shows the cognitive processes in reading to comprehend text; it is a slightly revised version of the Hayes et al. model which was adapted from the Thibadeau, Just, and Carpenter "reader model" [26].

The intention of this model was not to enter the debate about whether reading is a bottom up or top down process, but rather to show that when one reads to comprehend, one's primary aim is to construct an integrated representation of the text. Put differently, during reading for understanding, most of our effort is devoted to "putting the text together" to construct an understanding of how ideas work as a whole.

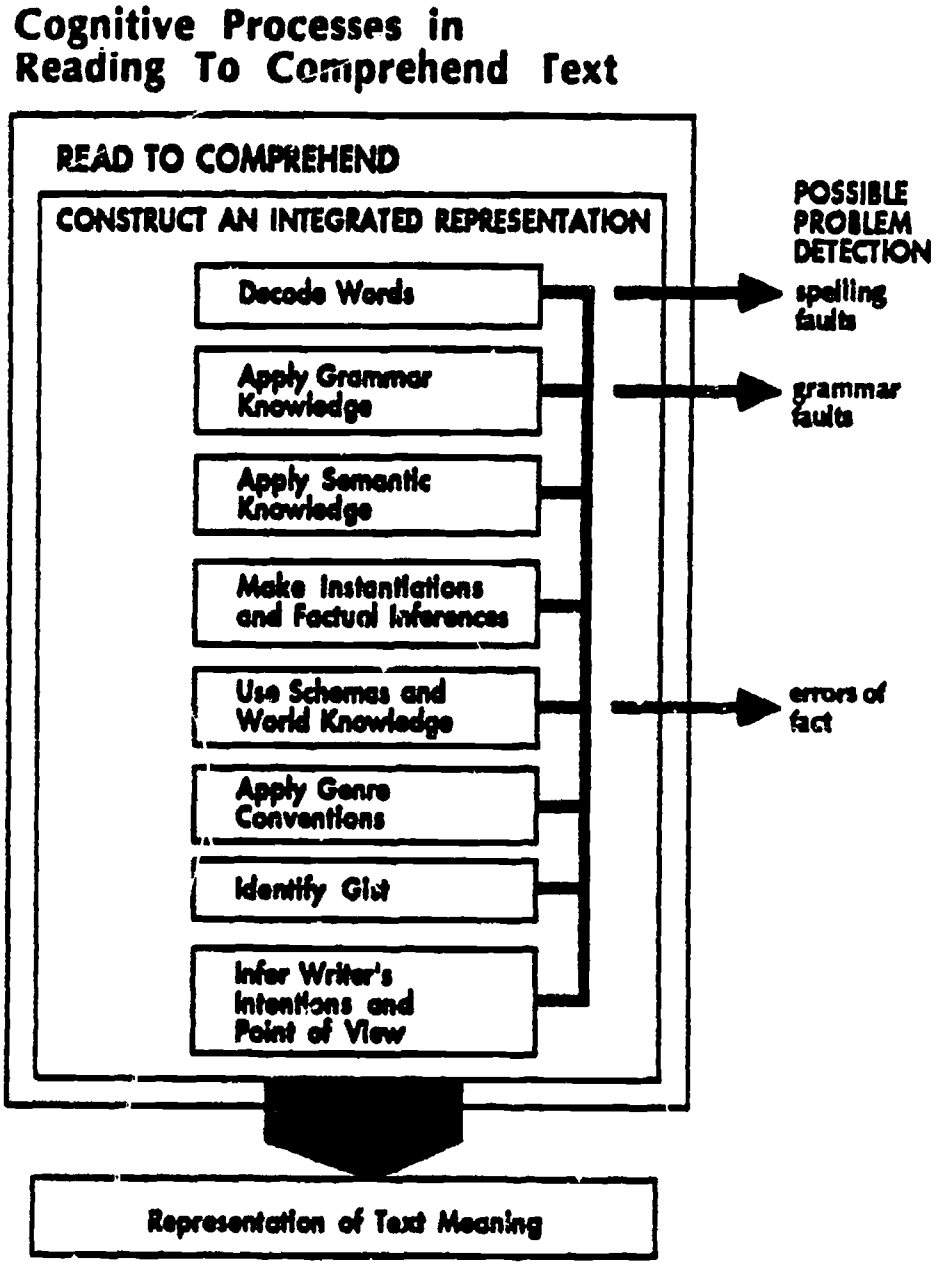

Figure 1. The Process of Reading for Comprehension (adapted from the Thibadeau, Carpenter, and Just model of reading [26] by Hayes, Flower, Schriver, Stratman, and Carey [14]).

Notice that during the process of comprehending, the reader also sometimes detects text problems without much thinking or conscious attention devoted to them. For example, it is common to notice spelling or grammar faults in what we read. When we encounter such faults during reading to understand, we typically ignore them. We pay more attention to them, of course, if the faults are bad enough to slow our reading or to make us reread. During reading to comprehend, we might also note errors or ambiguities in the text's information. For example, if we are familiar with the topic, we often have a good deal to say about the author's claims, logic, examples, anecdotes, and even choice of language.

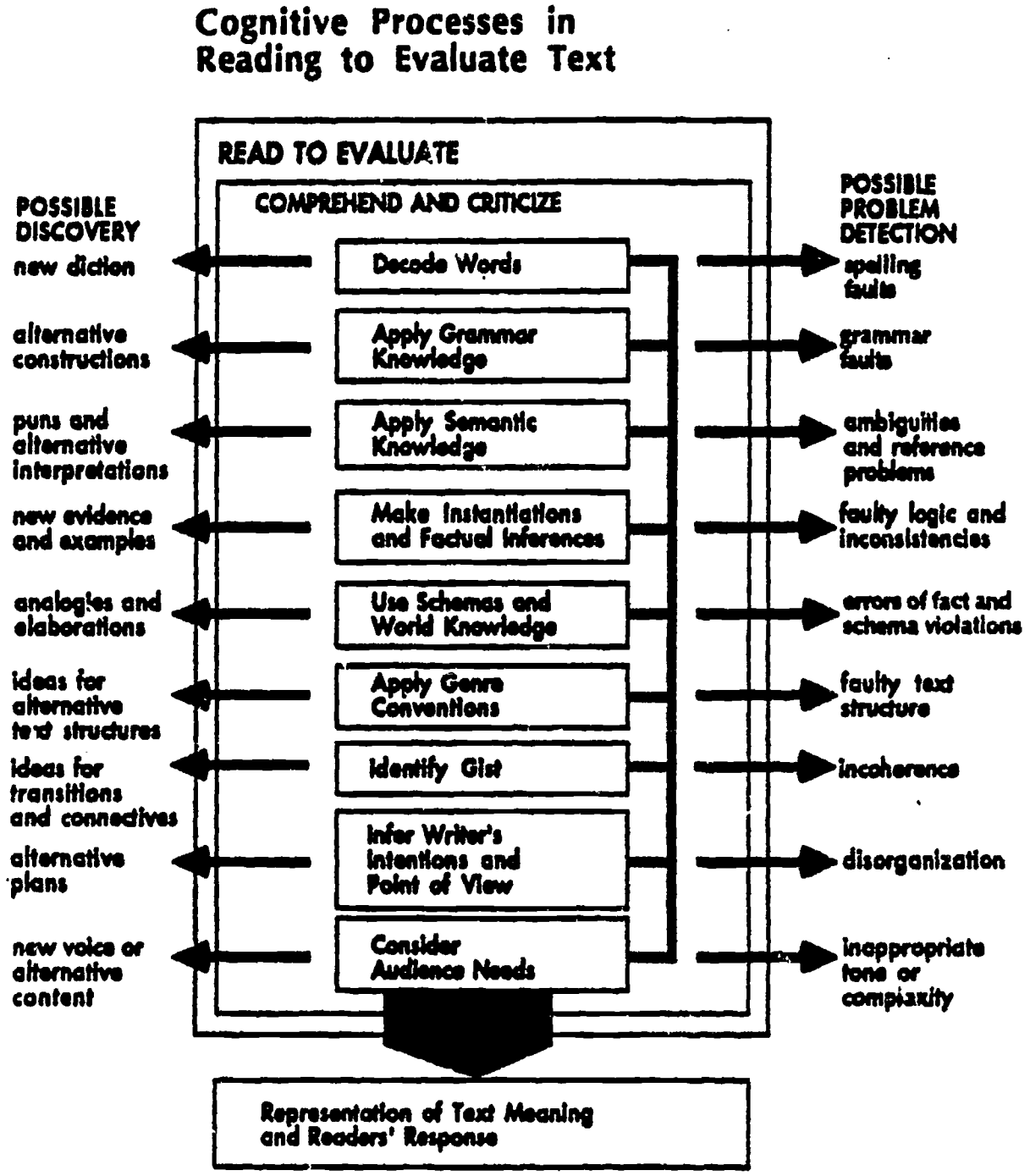

Figure 2. The Process of Reading to Evaluate Text Quality by Hayes, Flower, Schriver, Stratman, and Carey [14].

We can think of our active engagement with the author as conversation, sometimes playful while other times aggressive. On the other hand, when we have little or no background information on the topic, we are more likely to spend our attention trying to understand and connect what we have read with our prior knowledge rather than scrutinizing the author's claims.

Although the activity of reading to comprehend is a very complex process indeed, writers faced with the task of revising a poorly constructed text must go well beyond comprehending the author's ideas. Instead, when "reading to evaluate text" (Figure 2), our goal is to identify weaknesses in the text as well as to find solutions for them. Reading to evaluate text can be viewed as a cognitive process which is "built on top" of the comprehension process, but with the added top-level goals of *comprehending and criticizing the text from the point of view of its effectiveness for the intended audience.* Thus, when engaged in "reading to evaluate," the writer consciously looks for problematic text features and attempts to discover alternative solutions. Furthermore, instead of simply trying to understand the text as best one can, the revisor must ask, "Is this the most rhetorically effective way to present these ideas to the intended audience?"

One of the key differences between the models shown in Figures 1 and 2 is that in reading to evaluate, the writer's problem detections (some examples are shown on the right side of the model) become a source for possible discoveries (some examples are shown on the left side of the model)—that is, alternatives for improving the text. For example, when writers recognize that the audience may not have the appropriate background knowledge to follow the text's major claims, they often create new examples and add supporting evidence to make the text more understandable. Choosing among revision strategies once a problem has been noted is often difficult because changing one aspect of the text changes others. It is usually hard to decide if one should keep the text basically as it is written but simply to change the surface structure (that is, make changes to the phrasing) or delete sections of the text as written and make wholesale meaning changes.

COGNITIVE PROCESSES IN REVISING

Figure 3 presents a modified version of the revising process developed by Hayes, Flower, Schriver, Stratman, and Carey a few years ago [14]. The model, derived from observing experienced and inexperienced writers at work, is intended to capture the thinking processes of writers engaged in text revision.

As shown, text revision calls on a range of hierarchically organized subprocesses:

- *Representing the task*—characterizing the text's goals, the goals for the intended audience, the writer's goals, the goals of others with influence over the text (editors, bosses, clients), the purpose for writing, the context (social, organizational, historical, cultural) in which the text is being revised, the constraints under which the revision is taking place, and the criteria being invoked for judging success.

- *Detecting*—seeing or noticing problems.

- *Diagnosing*—characterizing or describing the text's problems.

- *Selecting strategies*—choosing among optional methods for solving identified problems (rewriting or editing).

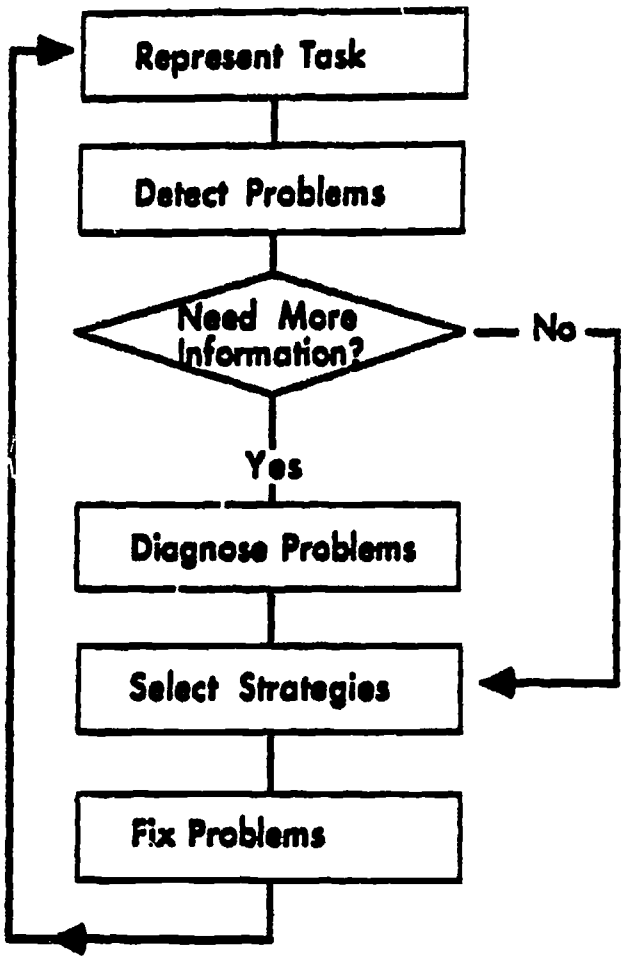

Figure 3. The Process of Revision (adapted from the Hayes, Flower, Schriver, Stratman, and Carey model of revising [14]).

- *Fixing problems*—taking action to solve the problems.

The research from which this model was developed revealed dramatic differences in the abilities of experienced and inexperienced writers to engage in and carry out these processes. Within each of these subprocesses, writers have a variety of options. The ability to recognize available options and to make changes that actually improve text was found to distinguish experienced from inexperienced writers.

Research on revision has been remarkably consistent in isolating two major differences between experienced and inexperienced revisors:

- *Experienced writers are skilled in evaluating global aspects of text quality* such as rhetorical stance, organization, logic, cohesion, persona, and tone. Inexperienced writers are not. Inexperienced writers tend to focus on local- level errors such as word choice, grammar, and syntax.

- *Experienced writers are skilled in taking action to meet the needs of the audience,* that is, making revision moves that improve the text from the reader's perspective.

> Inexperienced writers often identify the same problems as experienced writers but they are frequently unsuccessful in taking action to solve them. In fact, in some cases inexperienced writers' revisions introduce new problems and make the text worse instead of better [27].

From the research in writing, we can conclude that in choosing among methods to evaluate text, we need to draw on those that can help us act more like experienced writers. An optimal text-evaluation method should provide writers with two sorts of information: (1) information about whole-text or global aspects of text quality, and (2) information about how the audience may respond to the text.

THE CONTINUUM OF TEXT-EVALUATION METHODS

When one examines the kinds of document-evaluation methods currently in practice, we find a great deal of diversity both in the level of text problems they help writers to see and in the amount of actual reader feedback they provide. Figure 4 presents a continuum of text-evaluation methods. It classifies some of the most popular evaluation methods used in education, business, the health professions, publishing houses, and government—organizations which produce everything from textbooks to computer manuals to pamphlets on life-threatening diseases to mystery stories to tax forms.

The continuum is divided into three sections—*text-focused, expert-judgment-focused, and reader-focused methods*—which are separated by how explicit the feedback from the intended audience is. My assumption here is that text-focused methods, while sometimes created from information about readers, never use direct reader response; that experts—through their experience—provide surrogate-reader feedback; and that reader-focused methods make explicit use of audience response. I have listed a variety of kinds of tests and/or the people who have developed or elaborated them (the list is not exhaustive). Under each test (or group of tests) are the typical concerns of evaluators using the method. If a group of tests tend to address similar issues, I list the concerns only once. Some of the concerns are ideas that evaluators keep in mind, as they judge text quality, for instance, principles of style for visual or verbal text; in other cases, the concerns are variables for evaluation, perhaps the number and kind of errors a text leads a reader to make. Notice also that the tests within each class vary in the scope of text problems they help writers to identify, ranging from word-level to whole-text level problems.

Text-Focused Evaluation

On the left side are *text-focused methods* or those which operate by asking a person (or sometimes a computer) to examine a text, attend to a set of text features, and assess text quality by applying principles or guidelines that have been developed from ideas (and sometimes from research) about how readers at a certain level and background will probably respond. Thus, the reader's input, when used to develop such tests, is indirect at best. Text-focused methods include readability formulas, computer-based stylistic analysis programs, guidelines and maxims, and checklists.

Readability Formulas

Readability measures, such as the Flesch [28], Fog [29], SMOG [30], Dale and Chall [31], Fry [32], or Kincaid [33] formulas operate by analyzing word frequency and sentence length. Such procedures have been discussed and severely critiqued at length by many researchers [34-38] and it is not my purpose to belabor their obvious deficiencies

Evaluating Text Quality: The Continuum from Text-Focused to Reader-Focused Methods

TEXT-FOCUSED ⟵ **EXPERT-JUDGMENT-FOCUSED** ⟶ **READER-FOCUSED**

Readability Formulas
√ Flesch Reading Ease Score
√ Fog and Gunning Index
√ SMOG Formula
√ Dale and Chall Formula
√ Fry Formula
√ Kincaid Formula
- word frequency/length
- length of sentences

Computer-based Stylistic Analysis Programs
√ Writer's Workbench (UNIX™)
√ Epistle - 1 (IBM)
√ Critique, formerly Epistle - 2 (IBM)
√ Star (GM)
√ Grammatik III (Reference Software)
√ MacProof 3.2 (Lexpertise Linguistic Software)
- proofreading
- readability/grade level
- grammar & style (sentence-level)
- sentence complexity
- personalized dictionary

Guidelines & Maxims
√ Felker et al.
√ Harcourt, Brace, & Jovanovich
√ Hartley
√ Strunk & White
√ Williams
- principles of style for visual or verbal text
- visual or verbal text features to avoid or use sparingly
- audience analysis

Checklists
- recommended or required visual or verbal text features
- visual or verbal text features to avoid or use sparingly
- audience analysis
- consistency & completeness
- probable translatability
- adherence to conventions

Peer Review
- style (local & global issues)
- audience analysis
- graphics & typography
- organization & access features
- adherence to conventions
- consistency & completeness

Technical and/or Subject-Matter Expert Review
√ Content Evaluation
- accuracy (visual & verbal text)
- completeness and depth of text
- match with functionality of a machine, product, etc.
- match to stated goals for text

√ Presentation and Delivery Critique
- market/audience analysis
- competitive analysis

Editorial Review (in-house)
- style & copyediting
- adherence to conventions, specifications, boilerplate or guidelines (visual & verbal text)
- consistency & completeness

External Review
√ Text Features Evaluation
√ Holistic Rating (primary trait scoring or general impression marking)
- style (local & global issues)
- audience analysis
- graphics & typography
- organization & access features
- persuasiveness & believability
- competitive analysis
- persona & corporate identity

√ Consumer Advocate Review
- competitive analysis
- truth & ethics
- legal, health & safety implications

√ Gatekeeper Review
- appropriateness of content
- context(s) for use/dissemination

√ Document Design Process Critique
- document development cycle
- audience & task analyses
- style guides/standards/tools
- project management
- communication channels
- education and training methods

Concurrent Testing
√ Cloze Testing
- lexical predictability

√ Keystroke Protocols
- number of keystrokes
- time on task
- number & type of errors/assists
- error recovery behaviors
- number of failures to recover

√ Eye Movement Protocols
- number & location of fixations, saccades, and regressions
- total gaze duration

√ User Edits & Performance Testing
- reading time
- time on task
- number & type of errors/assists
- error recovery behaviors
- number of failures to recover
- cognitive load
- access & retrieval behaviors
- memorability/recall/retention

√ Protocol-Aided Revision (think-aloud verbal reports during reading and/or using text)
- problem solving strategies
- comprehension
- miscues and error recovery
- access & retrieval behaviors
- inference & predictions
- satisfaction/preference

Retrospective Testing
√ Comprehension (true/false, etc.)
- paraphrase
- recall/summary/gist
- recognition
- inference

√ Surveys, Interviews & Focus Groups
- rank/rate visual & verbal text
- comprehension
- persuasiveness & believability
- satisfaction/preference
- attitudes & beliefs
- inference

√ Critical Incidents/Storytelling
- key events & incidents
- relevance/severity judgments

√ Reader Feedback Cards
- comprehension
- satisfaction/preference
- attitudes & beliefs

Key
√ = an example of a particular method or individuals who developed or elaborated a method
• = a typical focus or dependent measure during evaluation

Figure 4. Evaluating Text Quality: The Continuum from Text-Focused to Reader-Focused Methods.

again. Research about how people use readability formulas has shown that they are often misused and misunderstood. Rather than using them as a *gross index* of the readability of a final draft, evaluators tend to use the formulas for specifying how writers must plan, write, and revise. Thus, "meeting the readability level" becomes the primary criteria for judging text quality. Unfortunately, there is no evidence to support this practice; in fact, just the opposite is true. To understand how loose the relation between comprehension and readability formulas is, one need only notice that a passage will get the same readability score whether its words are arranged in normal or backward order.

Indeed, research shows that writing to a readability level is an extremely questionable means for improving the comprehensibility of text. In discussing the use of readability formulas in the assessment of textbook difficulty, Singer and Donlan assert that sentence complexity and word frequency are only partial indicators of text difficulty because

> . . . a text may be relatively difficult because it has a high density of ideas and a high degree of interrelatedness or coherence among ideas. But, whether these characteristics of a text are difficult or not also depends upon the reader's prior knowledge, vocabulary ability, reasoning processes, purposes, and goals in reading the text. For example, if a text is densely packed with ideas but the reader's purpose is only to get the general idea of the text, the reader is likely to find the text easier than if his or her purpose was to comprehend the text fully. Hence . . . the difficulty level of a text as computed by the Fry and Flesch formulas . . . is only the average or general level of difficulty of a text.
>
> To determine the difficulty of a text for a particular reader, for example, a student who was having difficulty in reading and learning from a text, we would examine factors not only within that text but also within the reader. In short, *reading difficulty for a particular individual depends upon an interaction between the text and the individual* [39, 330].

But because they are relatively easy to automate and cheap to employ, many organizations use readability formulas exclusively, despite the lack of empirical support for their validity in assuring text quality. In discussing methods that are likely to be important in the future of prose processing research, Voss, Tyler, and Bisanz [40] dismiss the future impact of readability research, devoting less than a paragraph to the topic.

Computer-based Stylistic Analysis Programs

Computer-based style programs (for example [41-43]), such as UNIX's Writer's Workbench [44, 45] or the GM Star program [46] typically operate by assessing readability using one or more of the standard formulas and by counting passive constructions, misspellings, numbers of simple, compound, or complex sentences and then by providing the evaluator with a statistical summary of the text problems by assigning particular features an average score by comparing the use of the text feature, e.g., number of passive sentences, against the proportion used in a "good text" template. As Figure 4 shows, the focus of critiquers has been proofreading at the word or sentence level.

For some time, companies have been trying to improve on the range of problems computer-based style programs check. Lance Miller, in describing the "space of possible critiques," describes a number of key distinctions that are important in evaluating the goodness of a style program:

(1) the examination text-unit, (2) the report text-unit, (3) the critique type, [and] (4) the strength of the critique report. . . . The examination text-unit refers to the unit of text which is examined for the presence of some target. If the critique is that of spelling-checking, then the examination text-unit is a word. . . .

The report text-unit is the unit at which the critique is made, and this unit is either the same as the examination unit or else larger. An example of the latter instance is when a text is critiqued for low frequency words (examination-unit = word) and the results are summarized on a paragraph basis (report-unit = paragraph), e.g., "This paragraph contains the following low frequency words." . . .

The third distinction, critique type, refers to the manner in which the critique is made, and the two options are isolated vs. relative. In an isolated critique, a particular examination-unit is compared against a standard, and the judgment can be rendered without taking into account the characteristics of that unit relative to other units. Thus checking for spelling errors, incorrect capitalization, overly-long sentences . . . involve an isolated critique. In contrast, a relative critique checks the characteristics of one text-unit (having certain features) against the characteristics of another text-unit (having different features); the logic of the comparison is along the lines of "if the first unit has an aspect of X, then the second unit must have an aspect of Y." Most ungrammaticalities, such as disagreement in number between subject and verb, involve a relative type of critique.

> The fourth distinction concerns critique strength for which there are also two possibilities: right-wrong vs. threshold. A right-wrong judgment is one in which one can say "Right!" or "Wrong!" without fear of contradiction (from experts), as is the case of the majority of grammatical errors. . . . On the other hand, questions of style are not only matters of taste but . . . need to be reported with some deference and sensitivity to the fact that the author and critiquer may not share the same standards. One means of systematically handling the problem of varying stylistic standards is to arrange to have each stylistic evaluation result in the computation of a single number whose value grows with the severity of that particular gaffe; this value can then be compared against the threshold for a particular enterprise, and, if it exceeds that threshold, a suitable commentary is provided [47, 195-196].

It is not surprising that most early style programs looked at the word and sentence level, summarized at the sentence and paragraph level, focused mainly on on isolated critiques, and on right-wrong judgments. Miller argues that the primary challenge for developers of computer-based style programs is to go beyond the basics and to increase the space of critiques provided. Similarly, Richardson, Creed, and Chandler point out that most stylistic programs cannot address the kinds of grammatical problems that poor writers often create; the fundamental drawback of most programs is that "they rely too much on lookup tables instead of a parser to determine the roles words play in a sentence" [48, 57].

One program that aims to go well beyond the basics is IBM's Epistle system, now called Critique. It is developed by linguists and artificial intelligence experts at IBM's Watson Research Laboratory [49-51]. Recently (June 1989) IBM released Critique. Reporters from the machine translation magazine from the Netherlands, *Language*

Technology Electric Word, who put the prototype through its paces in July 1988, described its features in this way:

> Identification of unrecognized words or awkward phrases, checking for spelling errors, grammar and style errors and the generation of statistical information. It appeared to be fast and reliable.
>
> The program is written with Penelope, Heidorn's Programming Language for Natural Language Processing, and is based on colleague Karen Jensen's PEG (PLNLP English Grammar.) It parses a sentence, provides a syntactical representation, then employs hundreds of grammar rules to check the sentence's grammatical structure, before it highlight [sic] problems on the screen. Users will be able to establish individual profiles so that Critique will also reflect personally selected criteria [52, 7].

Currently, Critique runs as a new feature of IBM's mainframe editing software Process Master 1.3 (running on a VM/CMS operating system). Reporters speculate that there may be a PC version under development. For information on how Critique is being used in writing classes, see Richardson, Creed, and Chandler's summary of a pilot program at the University of Hawaii at Manoa. They point out three virtues of the program:

- Writers can use it interactively.
- It has three levels of help screens that provide information about principles of grammar and usage.
- It provides parse trees for each sentence it processes, thus allowing writers to see the structure of their sentences [48, 58].

Two other style checkers are worth note (they won the 1989 State-of-the-Art *Electric Word* Awards for Technical Excellence): Grammatik III for the PC and MacProof for the Macintosh:

> Grammatik III made by Reference Software Inc. proofreads documents for errors in grammar, style usage, punctuation, and spelling. Grammatical errors identified include improper use of homonyms (its/it's, they're/there/their) and possessives (you/you're, who's/whose) transpositions (form/from), disagreement between subject and verb (the government think) redundant comparatives (more better), incomplete sentences, double negatives, and split infinitives . . . also checks jargon, sexist terms, redundant phrases, neologisms, and overused phrases . . . also flexible enough to allow you to turn off rules and even add new ones of your own . . . and the documentation is so well written that even the layperson can make such modifications.
>
> MacProof checks on what its makers, Lexpertise Linguistic Software, call mechanics, usage, style, and structure . . . "mechanics" refers to spelling, punctuation, capitalization, and double words; . . . dictionary contains 120,000 entries. The "usage" dictionary contains 10,000 terms to be flagged for such barbarisms as offensiveness, imprecision and verbosity. "Style" means little more than flagging the verb "to be" . . . and "structure" is essentially about counting words in sentences and lines in paragraphs . . . it checks for logical transitions between paragraphs . . . [53, 35].

Guidelines and Maxims

Guidelines and maxims are perhaps the most popular text-focused method used. They are usually aimed at giving writers advice on the linguistic, stylistic, or graphic features of text (for example [54-57]). From a writer's perspective, most guidelines are frustrating to use either because they are vague and generic, e.g., "omit needless words" (Strunk and White [58]) or because they force us to assume that all writing tasks are alike and require the same simplistic prescriptions (e.g., "use short sentences"). Put differently, guidelines often fail to help writers adapt their texts to the unique features of the given rhetorical situation.

Furthermore, evidence suggests that writers have difficulty recognizing when and how to apply guidelines [23, 59-61]. When guidelines are invoked too rigidly, they function as rules and can have the effect of stifling creative solutions to rhetorical problems. Although there are genuine difficulties associated with the guideline approach to judging text quality, there have been some very good examples of the effective use of guidelines, such as Williams' well-known text on style [57].

Checklists

Checklists, another text-focused method, typically work in one of two ways. On the one hand, the evaluator is asked to use the checklist as a reminder of issues to consider. For good examples of checklists, see Price's "giant checklist" for writing computer documentation [62] or Spencer's "usability considerations checklist" for testing computing systems [63]. Many checklists focus on recommending visual or verbal text features to employ or those to avoid or use sparingly. Other checklists are essentially additive weighting procedures which ask the evaluator to rate the text's features along a "goodness" scale and then to assign a quality score to the text. (See Hayes [64] for a discussion how to design an additive weighting scale.)

A drawback of checklists lies in the difficulty of deciding what text features are most important and in assigning weights or numerical values to text features. Writers usually disagree about the values assigned to any given feature. And checklists, like guidelines, usually fail to ask evaluators to judge the use of text features in relation to the given rhetorical context. For example, there are many rhetorical situations in which the passive voice is the most sensitive linguistic choice, yet most checklists remind writers to avoid using passives. Such situations leave the writer with the questions: How "bad" is a text feature that is rated average or below average? If two texts receive the same low score but are intended to serve different rhetorical purposes, are they equally poor? How should text feature ratings be used in revision? Should all poorly evaluated text features be revised extensively?

It should also be pointed out that most checklists are not based on data from readers or users of the text under evaluation. Rather they are often created by consolidating an organization's conventions and accumulated folklore about the features of good and bad texts. Thus, checklists may simply codify an organization's misunderstanding of the audience.

Summary

Advantages of **text-focused methods** are that they are inexpensive to use, some can be automated, and they can be helpful in detecting certain obvious classes of error. The inherent weakness of these methods lies in their predominant focus on word and sentence-level features of the text. Typically, their output provides little, if any, information about

how the document is working at the paragraph and whole-text level. Perhaps the biggest weakness is that their output provides no information about the reader's needs. When text-focused methods are used as the only guide for revision, research by Swaney, Janik, Bond, and Hayes [22] shows that revisors may actually make the text worse instead of better.

Expert-Judgment-Focused Evaluation

Expert-judgment-focused methods constitute another widely used set of evaluation procedures. By *expert* judgment, I mean individuals who possess high knowledge about the text, its audience, or writing itself. Expert-judgment-focused methods include peer reviews, technical and/or subject-matter expert reviews, editorial reviews, and external reviews.

Peer Review

Peer review is one of the more standard expert judgment methods employed by education, industry, and government [65-68]. With peer review, people who share a common background are called upon to evaluate texts for issues of style, consistency, tone, and the like. Peer reviews can be very informative in pointing out text problems, allowing the writer to draw on the multiple perspectives of other writers. Peer reviewers tend to be quite good at recognizing stylistic issues at both the local- and global-level, and writers find that peers are helpful in making suggestions to solve organization problems.

However, some writers report that peer review can also be a frustrating experience. When the writer receives divergent opinions about the problems the text will create for readers (or when personalities enter into decisions about what is problematic) it is often difficult to determine which problems to solve and which suggestions for revision to use. This difficulty is magnified when the revisor is operating under severe time constraints.

Peer reviews can also suffer from evaluators who work too frequently with texts of similar genres and subject matter. Writers who always evaluate the same sort of text—for instance, proposals—may not improve in their skills over time, but may actually erode their skills by doing too much of the same kind of text evaluation all the time. When evaluators always work with the same kinds of texts, they can become insensitive to the audience's likely response to texts of that sort. Researchers who studied experienced U.S. government writers at the IRS, for example, found that evaluators were particularly insensitive to language and stylistic issues that bothered readers outside that institution [69]. Indeed, peer review is a way of socially constructing and institutionalizing certain styles.

Peer review has also come under question by authors who submit articles to professional journals that use peer review for judging manuscripts for publication [70, 71]. Authors whose work is evaluated by peer reviewers sometimes question the criteria used for making decisions about what gets published and what does not. They suspect that it is almost impossible to conduct a truly "blind" review since often the peer can guess the author's identity by carefully examining the reference list [72, 73]. Because peer reviewers for journals serve such a critical gatekeeping function, authors are concerned that peer reviewers invoke consistent standards for all manuscripts received.

Technical and/or Subject-Matter Expert Review

Technical and/or subject-matter expert (SME) reviews usually conduct *content evaluations* of text, aiming to find deficiencies in coverage, accuracy, authenticity, or

completeness. In many industrial contexts, for example, technical reviews are conducted by engineers or computer scientists who assess a text's content in terms of its match with the functionality of a product or a machine. Technical reviews are intended to provide writers with detailed information about the ways in which text content is inaccurate or misleading. While a technical review can be conducted by a technically-oriented person, like a computer programmer who is verifying the procedures presented in a user's manual, this is not always the case. The phrase *technical review* is also used to refer to evaluations by subject-matter experts who verify text adequacy, like a museum historian who is verifying the accuracy of facts presented in a brochure. Those who participate in subject-matter expert reviews are typically extremely knowledgeable about the content, the information medium, the audience, or the rhetorical situation in which the text will be read or used.

Subject-matter expert reviews conducted by marketing experts, for example, may conduct a *presentation and delivery critique,* checking for features such as the tone and mood created by the integration of the visual and verbal text. Thus, they may evaluate the presentation and the delivery of the content in terms of its match to a set of articulated goals (for example, the text must be short; it should present a theme; it should use vibrant color and visuals) or against a set of esthetic criteria (for instance, the text should convey seriousness and warmth).

Although both technical and/or subject-matter expert reviews do give valuable feedback about difficulties with a text, it may be unwise to use such reviews in isolation. Research is beginning to show that topic knowledge is sometimes a detriment instead of a help and that experts are not always the best people to ask about text quality. Hayes, Schriver, Blaustein, and Spilka [74] found what they term "the knowledge effect in writing": readers with high topic knowledge were very poor in judging how lay readers would understand the topic.

Similarly, in another study, I found that writers with 2 to 3 years of experience with word processing were extremely insensitive to judging the kinds of problems new users would have with poorly written procedural instructions for a word processor [15]. To help writers recognize and overcome their insensitivity, I asked them to study the transcripts of think-aloud protocols from a group of new users which demonstrated numerous comprehension and usability problems. After reading users' comments illustrating their unsuccessful attempts to invoke simple commands, some writers reported that the users' errors seemed stupid and that it was hard to remember what it was like to be a newcomer to computers. Such research reminds us that writers, technical experts, or subject-matter experts with high topic knowledge may find it especially difficult to anticipate the needs of readers with low topic knowledge.

Editorial Review

Editorial in-house reviews, another expert judgement evaluation procedure, are typically carried out by senior writers or copy editors who check for such issues as style, consistency, specifications, or use of conventions. Traditionally, editorial reviews focused on grammar and mechanics. Bourns and Grove point out that in many settings, editorial reviews used to be quite mechanical and tended to be extremely rule-oriented [75]. More recently, the province of editorial reviews has been expanded to issues of organization, presentation, readability, coherence, retrievability, and accuracy. Put differently, editors have moved away from a one-dimensional view of what they do and now see their work as a complex hierarchy of skills and perceptual abilities [76-79].

Another way that editorial reviews are changing lies in the kinds of advice they provide. In the past, most editorial reviews were viewed as activities designed *to find errors* in text. Today, most editors consider their role much broader than the wordsmith who looks for problems. Instead, they view their role as *discovering ways to improve text* (see Henke [80] for a brief discussion of the usefulness of tabulating editorial contributions rather than number of errors found). In effect, the definition of an editorial review is slowly changing from *editing* to *revising*.

A similar evolution in thinking has occurred in the research on composing. Although early research in composing focused on studying editing and mechanical correctness, today's work looks at the process of whole-text revision. Studies show that expert writers are much more than standard good editors; they are able to "resee" text in ways that standard good editors cannot [14, 81-84]. Put differently, expert writers are revisors, not editors.

Although we have seen dramatic practical improvements in the editorial review process, we have seen almost no research in the area. Longitudinal studies need to be done which track the editorial review process over many writing tasks and which focus on particular writers working alone and collaboratively. Such work might find that some skills get much better with time while others get worse. As mentioned above, research investigating the "knowledge effect in writing" [74] provides us with reason to suspect that some editors may have an "in-house effect": they have been editing within the same context on similar text types too long. Alternatively, we may find what we already believe: Experienced editors, unlike many writers, are much more skilled in recognizing the audience's needs and in making effective linguistic and rhetorical choices that meet those needs.

External Review

In many contexts, it is impractical and even undesirable to judge text quality using people who are insiders to the context, like peers or technical and/or subject-matter experts. In such cases, external reviews are used for judging text quality. Organizations often turn to external reviews when they recognize that something is wrong with the texts they produce but are uncertain how to pinpoint the problems and need to gain a fresh perspective on the quality of their document design. Thus, many document design and graphic design consulting agencies are retained by organizations who want critical feedback about how their texts are functioning from a competitive standpoint. External reviews vary in the methods employed to conduct them and the people who carry them out.

One type of external review, a *text features evaluation*, criticizes the relative goodness of a text by assessing the design of visual or verbal features. Text features evaluations typically involve selecting a representative set of an organization's texts and then analyzing them in terms of key features, such as style, tone, content, format, grid systems, logos, and so on. In this way, text features evaluations aim to characterize how the integration of the visual and verbal text shapes the organization's public image. From such a diagnosis, a new plan can be derived that better matches the organization's goals.

Another kind of external review uses *holistic rating* methods to judge text quality [85-89]. According to Charney, "holistic rating is a quick, impressionistic qualitative procedure for sorting or ranking samples of writing. It is not designed to correct or edit a piece, or to diagnose its weaknesses. Instead, it is a set of procedures for assigning a value to a writing sample according to previously established criteria" [85, 67]. Holistic rating refers to the set of methodologies used to arrive at a total impression of a text. Testing agencies such as the Educational Testing Service (ETS) use holistic scoring to judge

student essays for the Scholastic Aptitude Tests (SAT) and the high school Advanced Placement Examinations. There are many variations on how to derive a holistic rating; two of the more typical methods are *general impression marking* and *primary trait scoring*.

General Impression Marking is a method in which "rater fits a writing sample into an ordered ranking on the basis of the total impression created by the paper" [85, 71]. The "defining characteristic of this approach is that it weighs sample papers against each other, rather than against a predetermined set of criteria" [85, 72]. The criteria are arrived at inductively by either test organizers or by the evaluators themselves. Often test organizers using general impression marking will select a set of "anchor texts" which represent "the range of good to poor texts" the judges can expect to see. Evaluators are then trained to judge a set of texts against the anchor papers.

Primary Trait Scoring, developed by Lloyd-Jones [90], is different in that it gives raters a scoring guide carefully adapted for the judging task; thus, it uses a set of explicit criteria to judge text quality. Raters are then trained to evaluate texts using the agreed-upon set of text features, e.g., style, organization and coherence. Although the procedure sounds quite straightforward, studies show that it is extremely difficult and sometimes impossible for a group of evaluators to agree on a set of criteria and to invoke such criteria consistently and reliably [91-93]. Charney cites a number of studies which show that "in spite of training, readers' judgments are strongly influenced by salient, though superficial, characteristics of writing" (spelling, length, unusual words, and the quality of handwriting) [85, 75]. Although raters' say that they agree on the predetermined criteria, they tend to fall back on other criteria while they are engaged in evaluation. For such reasons, Charney and others have raised serious questions about the reliability and validity of holistic scoring procedures.

Another type of external review is the *consumer advocate review* conducted by people who are concerned with judging text quality from the perspective of the consumer. For example, the U.S. Office of Consumer Affairs has evaluators who judge the clarity of instructions, warranties, and contracts (see the *Consumer Resource Handbook* [94]). They are concerned with legal, health, and safety implications of poorly designed text. Government administrators such as the late Malcolm Baldridge, former U.S. Secretary of Commerce, and Lee L. Gray, former U.S. Director of Consumer Affairs, went to great lengths to stress that "talking or writing in plain English is a challenge to both the private and public sectors" [95, preface]. Their important work, some of the fruits of which are described in *How Plain English Works for Business: Twelve Case Studies* [95], provides concrete evidence of the enormous practical and financial benefits associated with producing easy-to-read warranties, credit contracts, insurance policies, and product information booklets.

Consumer advocate reviews usually use weighted scoring methods or scaled surveys so common to publications such as *Consumer Reports*. More and more publications are providing consumer reviews about text quality than ever before. For example, early in 1989, *MACazine* introduced a feature called "Reader Reports" in which readers evaluate computer products along various dimensions, and one of the key features rated is the quality of documentation [96]. Surprisingly, in their first survey, over 1300 readers responded, highlighting that consumers of high technology want to know more than the manufacturers' facts about a product's key features, they want to know how other users rate those features.

A *gatekeeper review* is one in which a text is evaluated by a group of individuals who are responsible for disseminating a text. According to the U.S. Department of Health and Human Services:

> Often, public and patient information education materials are distributed to their intended target audiences through health professionals or other intermediary organizations. These intermediaries act as gatekeepers, controlling the distribution channels for reaching target audiences. Their approval or disapproval of materials is a critical factor in a program's success. If they do not like a poster or a booklet, it may never reach the intended audience. . . . Questions may include such areas as overall reactions to the materials and assessments of the appropriateness, completeness, and utility of the information [97, 25].

Along with gathering information about whether a given final draft "will fly" in the particular context in which it is intended, gatekeeper reviews are sometimes used to help writers plan their texts. Floreak presents an interesting case study describing how extensive interviews with gatekeepers in a small town's community services organization provided valuable insight into the target audience for a poster campaign designed to help low literate parents care for their youngsters [98]. Gatekeeper reviews then can be helpful in both planning and revising text.

Another type of external review is the *document design process critique*—an evaluation procedure that focuses on identifying predictors of poor writing quality [99]. It is designed to help identify weaknesses in the ways in which a writer, a group of writers, or an organization, engages in the process of creating text. The idea is to try to predict (and prevent) poor writing before it occurs. Process critique evaluators examine the approach to planning, generating, revising and evaluating text. They look at the way people collaborate, the guidelines writers follow, the kinds of feedback that goes into the shaping of a text—in effect, evaluators pay particular attention to the way typical writing tasks get done, assessing project management, observing the nature of communication channels (for example, between writers and technical experts) throughout a writing project. The goal is to identify the strengths and weaknesses in the process along with recommending education or research that will help remedy the weaknesses.

Summary

Although **expert-judgment-focused** evaluations are useful and can provide a wealth of information for the writer, they often suffer from the evaluators being *too close* to the text or product the text describes. In many contexts, the only readers who participate in evaluating a text are the readers within an organization who *know most* about the text and/or the product it describe—peers, technical experts, and subject-matter experts. The result is that the text may work well for people such as engineers, computer scientists, and marketing specialists—people who developed or influenced the creation of the text—but may fail miserably for the average reader. Certainly external reviews are quite helpful in supplementing standard inhouse evaluation procedures. But expert-judgment focused evaluation methods should not be used in isolation; they need to be supplemented with other document evaluation procedures, particularly those which are reader-focused.

Reader-Focused Evaluation

Reader-focused text-evaluation methods—on the right end of the continuum—are procedures which rely on feedback from the intended audience. There are two general classes of reader feedback methods: *concurrent tests* (which evaluate the real-time problem-solving behaviors of readers as they are actively engaged in comprehending and using the text for its intended purpose) and *retrospective tests* (which elicit feedback after the reader has finished with reading and using the text). Concurrent reader feedback

methods include cloze testing, behavior protocols (sometimes called motor protocols), performance testing, and thinking-aloud verbal protocols. Retrospective tests include comprehension methods, surveys, interviews, focus groups, critical incidents, and reader feedback cards.

Concurrent Testing

The *cloze test* [100-102] presents readers with text which has had words systematically deleted, asking readers to try to fill in the missing words. The idea is that quality text should have a high degree of lexical predictability. Thus, if a text is "good," readers should be able to fill in the blanks. To use the cloze technique, evaluators:

> . . . simply delete or omit every fifth word from a passage of approximately 250 words, but the sentence before and after the passage is left intact. A total of 50 words will be deleted from the passage. The reader's task is to infer from the remaining content what the missing words are, retrieve the exact words from vocabulary stored in his or her memory, and insert them into the passage. In scoring, only the exact, original word is counted as correct. The cloze technique places a premium upon the reader's ability to infer the missing words from the semantics and syntax of the remaining words in the passage and upon the reader's vocabulary repertoire and ability to retrieve words from storage in memory [39, 311].

The cloze test is interesting because it does take real readers into account and surprisingly, the activity of filling-in the blanks does appear to draw on many levels of the reading process—word recognition, knowledge of syntax and semantics. However, it seems to be limited in the genres to which it can be applied. It seems best suited for narrative and expository text and seems most unsuited for procedural or reference texts. For example, the cloze test would be a very bad test to evaluate the quality of a telephone book. It also fails to provide any feedback about how the text is working from a visual perspective.

Another kind of concurrent testing involves collecting *behavior protocols*, that is, recordings of readers' actions and behaviors. The primary feature of behavior protocols is that participants do not talk aloud while performing a task—they simply do the task while either a human evaluator and/or a computer program records what they do. Evaluators collecting behavior protocols are often interested in such issues as the following:

- How people comprehend information and solve problems with text that is presented in prose and/or with diagrams, illustrations, or pictures.
- How quickly and accurately people can perform a task using only printed instructions as their guide (for instance, using a manual to assemble a bicycle or to operate a VCR).
- Where readers look for information in lengthy texts such as reference guides (in indexes, in tables of contents, in glossaries).
- How frequently readers refer to printed instructions (whether in hardcopy or online) to perform computing tasks, along with how users recover from errors as they try to operate machinery (for example, the steps taken to undo a mistaken deletion of a computer file).

- How computer interface design features such as color, windowing, or display rate influence people's ability to use computers (evaluating the differences between a small CRT screen and a large bit-mapped display).

Behavior protocols include keystroke logs, eye movement studies, and user-edits. *Keystroke logs,* which can be collected automatically during interaction with a computer, provide detailed information about users' error and error-recovery patterns and can be used to develop models of users' behavior [103, 104].

Eye movement protocols have been used to determine the effect of colors, display rate, and cursor movement in online documentation and interface design [105]. They have also been used to study how people read scientific texts involving prose and diagrams [106]. At this point, most of the work in this area is concerned with studying the behavior of the eyes during reading from a computer screen rather than using the method for text evaluation. Voss, Tyler, and Bisanz point out that:

> Although there are some problems with interpretation of what eye movements reflect (see McConkie, Hogaboam, Wolverton, and Lucas [107]), most research has validated the assumption that the position of the eye at any given time corresponds to what is currently being processed (Just and Carpenter [108]). The measures obtained from eye movement data can include the number of fixations within a given text portion, the number of saccades, the number of regressive eye movements, or simply the total gaze duration, independent of the number of fixations. Rayner [109] provides a good summary of these various approaches [40, 380].

Another type of behavior protocol, the *user-edit,* first described by Atlas [17], involves observing readers directly while they work and interact with a machine, using only its operations manual as a guide. The observer (who sits either near the user or in another room while observing through a two-way mirror) pays close attention to how readers use text, when they use text, and how the text helps or hurts understanding. User-edits are now widely used in industry to evaluate usability of text.

Performance testing characterizes the class of tests in which evaluators monitor factors such as readers' task performance, retrieval and access behaviors, error recovery strategies, cognitive load, and general ability to use a text [24, 63, 110, 111]. Thus, user-edits are a type of performance test. Evaluators using performance testing are often concerned with obtaining benchmark information about speed and accuracy [112, 113]; thus, talking aloud is an undesirable activity because it adds to the time on task. However, since it is often hazardous to infer problem solving strategies without more explicit indicators of thinking such as those gained through verbal reports, many evaluators use performance testing to look at large numbers of participants and supplement their evaluation with case studies using think-aloud protocols. As Evans points out:

> Used as part of a wider research project, case studies can provide material to illustrate or test a theory, and they may . . . help to humanize, what, without such additions, might be an arid statement of observations or facts. Research which has been reduced to mere statistics can seem very remote from the flesh and blood world we know, and case studies, judiciously used, can reclothe the bare bones . . . [114, 11].

Clearly, performance testing has been and will continue to play a major role in text evaluation in the future. See Schumacher and Waller [115] for an excellent review of frequently-used methods in document design.

Thinking-aloud protocols ask participants to perform a task while thinking aloud as they interact with a document and/or with a machine [22, 116-123]. When people experience difficulty in comprehending or in using the document, their comments typically reveal the location and nature of the difficulty [20]. Unlike participants in behavior protocols, think-aloud participants are asked to verbalize anything that comes to their mind as they are engaged in the task. Because thinking-aloud protocols are collected while the person is reading and is engaged in the process of comprehension, they provide much more explicit and complete information than do readers' comments collected after reading is finished. The advantage of think-alouds is that participants often say how and why they are having a difficulty with the text. Therefore, the writer has both *locative* and *diagnostic* information that will help guide revision decisions. In addition, think-alouds often highlight both visual and verbal text problems caused by either *what has been written* or by *what has been omitted*—an important advantage over other document-evaluation procedures. Thus, think-alouds are typically used when the goal is to assess how people understand, solve problems with, draw inferences about, use, or read text [21, 119, 124-127].

In the early 1980s, Hayes and his colleagues at Carnegie Mellon University's Communications Design Center pioneered a technique using thinking-aloud protocols called *protocol-aided revision* to revise texts such as insurance forms, apartment leases, computer manuals, and medical consent forms [22, 116, 118, 128]. Protocol-aided revision is a process in which evaluators videotape or audiotape readers as they think aloud while comprehending a text and/or while interacting with machines, toys, devices, equipment, and the like. The transcripts are then analyzed for evidence of readers' problem-solving strategies, comprehension, miscues and error recovery, access and retrieval behaviors, inferences and predictions, along with comments indicating satisfaction or preference. Such information is then used to guide revision activity. Protocol-aided revision is an iterative process involving testing a text with members of the intended audience, revising based on the problems readers experience, followed by more testing and revising until the text satisfies the reader's needs and the writer's goals.

In 1986, Diehli compared think-aloud protocols with some other methods (guidelines, a computer-based style program called "Murky," and checklists called revision filters) to determine the kind of information provided by each [59]. Results showed that no single method was best but that guidelines were worst, reiterating that writers need to consider the costs and benefits associated with alternative evaluation methods. And Holland and her colleagues [119], who studied writers revising procedural instructions after watching videotapes of readers using their texts, found that writers who observed readers-in-action were much more able to solve text problems that were specific to the rhetorical situation—problems for which guidelines were too general to be helpful.

Although think-aloud protocols have obvious advantages over other methods, it is important to recognize their limitations as well. Glass, Holyoak, and Santa raise the following issues:

- Often a protocol will seem to have "gaps" in which the participant forgets to speak.
- Sometimes participants will take a "mental leap" reaching some conclusion without mentioning any intermediate steps.
- Sometimes the protocol will be ambiguous and difficult to interpret.
- They are time-consuming.

- They are verbal and are difficult if not impossible to conduct with children.
- If participants are using visual imagery or some other nonverbal representation, they may be unable to talk about what they are doing.
- Participants may use a more systematic method for solving problems than they would normally because they know they are being watched [129, 416-417].

Despite these limitations, protocol analysis remains one of the most informative methods for studying problem-solving behavior.

A few years ago, I observed that writers working at Carnegie Mellon's Communications Design Center who had extensive experience using protocol-aided revision seemed better able to anticipate a reader's interaction with their texts than were other professional writers with years of on-the-job professional writing and editing experience. When I questioned these writers about why they were so good, they claimed that protocols changed not only the way they revised text, but the way they planned. Indeed, these writers had collected and evaluated the transcripts of dozens of think-aloud protocols. Their claim both intrigued and puzzled me. I found that writers were unable to articulate in what way(s) protocols had changed their writing.

I wondered if their superior skill in evaluating and revising text resulted from their frequent and direct experience with reader feedback. I thought that if this were true, a sequence of lessons that took writers through a similar experience might help them increase their sensitivity to readers' needs. To this end, I refined the protocol-aided revision methodology, characterized the cognitive processes involved in using the method [20, 21], and developed and evaluated a protocol-aided revision pedagogy. The aim of the teaching method (described elsewhere in detail) was to give writers the benefits of protocols *without* the need to collect protocols on every text [15].

After training in the protocol-aided revision pedagogy, writers were tested on their ability to accurately predict readers' problems with texts in which protocols were unavailable. Five classes of writers taught with protocols were compared with five classes of writers taught using guidelines, audience analysis heuristics, and peer review procedures—that is, with more standard text-focused and expert-judgment-focused approaches. In particular, writers were compared for their ability to detect and diagnose readers' problems along three dimensions:

- *Commission* versus *omission*, that is, problems caused by what the text says versus what it leaves out.
- Problems characterized from the perspective of the *reader*, the *self* (i.e., the writer), or the *text*.
- Problems at the *global* or *local* level of the text.

Results show that writers taught to anticipate readers' problems with poorly written instructional text, using the protocol-aided revision pedagogy, improve significantly ($p \leq .005$) in their ability to judge readers' problems accurately. More specifically, *writers taught with the protocol-aided revision method improve in their ability to predict problems of omission, problems from the readers' point of view, and global problems*. For each of the three types of diagnostic categories, experimental writers improved more than did control writers ($p \leq .005$). Writers in the experimental group made dramatic gains in their

ability to detect and diagnose problems caused by difficulties such as poor organization, ambiguous purpose statements, missing illustrations and diagrams, faulty analogies, and unclear procedures.

In addition, writers who were taught to anticipate readers' problems by studying the protocol transcripts of lay readers comprehending instructional texts (in this case, computer manuals) were able to transfer their knowledge to anticipating lay readers' problems with elementary science texts. Thus, learning about how readers responded to one genre helped writers anticipate readers' problems with another. Such results also underscore the benefits of using protocol-aided revision not only for improving texts under evaluation, but for enhancing writers' skills generally.

Retrospective Testing

Retrospective methods are the more frequently used of the reader-focused methods. They include a wide range of comprehension tests, along with methods such as surveys, interviews, focus groups, critical incidents, and reader feedback cards. The problems associated with retrospective reports have been well documented by Ericsson and Simon [124]. Aside from the drawback of asking readers to reflect on their remembrance of comprehending the text, the primary disadvantage of retrospective tests is that they frequently fail to pinpoint specific text features that need revision, and instead, often give the revisor vague and often uninterpretable feedback, e.g., respondents on a reader-feedback card may write, "it was pretty easy to read except for some of the procedures."

Comprehension testing has been a widely-used retrospective measure in evaluating text quality. Basically it involves asking readers to *paraphrase, recall, summarize, recognize,* or *draw inferences* about particular text items or textual features through having them engage in activities such as true/false, fill-in-the-blank, essay, or multiple choice tests. Typically, text evaluators using comprehension testing look for readers' abilities to make judgments and inferences about the text's content. As with other evaluation methods, the success and value of comprehension measures is directly related to the quality of the test itself. Results obtained by the use of poorly-constructed questions are likely to produce trivial results.

Besides the very familiar types of recall and recognition testing used in school settings and standardized test situations, other ways that comprehension is often assessed focus on summary, paraphrase, or inference measures. With these tests, participants are asked to read a text (or portions of it) and then to summarize or paraphrase the main ideas. Researchers are often interested in the number and importance of idea units recalled, the number and type of elaborations and integrations made, the number and kind of inferences drawn, and the number and type of errors made. Such tests are often very useful in pinpointing people's reactions to subtle cues in the text.

For example, in evaluating how people understand texts such as unemployment compensation brochures and policy statements, writers have found it useful to study what people infer as they read. Such testing shows that people tend to draw elaborate (and often incorrect) inferences from statements about benefits that are made in such policies. Inference testing is likely to become a frequently-used method in the 1990s, especially with so many companies worried over lawsuits related to the misunderstanding of written information [130, 131]. For instance, tampon companies have been trying to determine what they must do in creating warning labels and package inserts to limit their liability in cases of toxic shock syndrome.

In assessing participants' performance on comprehension tests, evaluators typically use either *criterion-referenced* or *norm-referenced* approaches. Dick and Carey explain that the difference between these approaches lies in how tests results are interpreted [132]. In criterion-referenced tests (sometimes called mastery tests), the performance of all participants is compared to a preestablished criteria for success. For example, in testing the effectiveness of a procedures manual for operating a computer, one might set a criterion that users must be able perform the procedures with 85 percent accuracy. Thus, testing and revising would take place until all participants were able to meet the criterion using the text.

On the other hand, norm-referenced testing compares the performance of participants with each other (either within a group or between groups). The participants' rank or position in the group becomes a reference point for determining the quality of performance rather than a meeting a specified mastery level. Since many contexts for assessing text quality are ones in which it is impractical (and irrelevant) to set rigid criterion levels, norm-reference testing is a useful alternative. For example, evaluators may want to know which of two texts is better for conveying detailed visual information, e.g., a full color photograph or a black and white line drawing? Similarly, evaluators may want to know which of several groups of readers respond most favorably to particular text features—for example, do experts retain more information from line drawings than do novices? The idea is to judge the relative quality of the text by looking at readers' performance in comparison to each other.

Surveys and *interviews*, perhaps the most commonly used methods for evaluating text quality, range from face-to-face procedures to pen and paper questionnaires to telephone and online surveys [133-137]. With surveys and interviews, participants typically respond to a mix of open-ended and close-ended questions designed to elicit opinions about the use of visual and verbal text features along dimensions such as comprehensibility and persuasiveness. The advantages of surveys and questionnaires are that they are relatively inexpensive, they can be self-administered, they do not require much time, and respondents can remain anonymous. (For a brief discussion of some of the types of survey scales, see Davis and Mentecki [138].) A major disadvantage is that quite often the participants are self-selected, thus biasing the results. From a revisor's perspective, surveys also have the drawback that if readers rate the text poorly, evaluators must conduct other tests to determine the particular text features or portions of text that caused problems for readers [139]. In addition, survey response rate may be low and participants often ignore some questions (especially open-ended questions that require time to respond). For a discussion of how surveys have been used in learning about writing in the workplace, see Anderson [140] and for some of the problems associated with survey research done in the field of technical communication, see Isakson and Spyridakis [141].

Interviews, on the other hand, do provide participants with the opportunity to discuss a text at length and allow the evaluator to probe individual responses in detail. See the U.S. Department of Health and Human Services' approach to conducting individual in-depth interviews or central location intercept interviews—interviews conducted in locations frequented by representative members of a text's target audience [97]. For example, they describe a pilot maternal and child health care program in which interviewers went to several clinics in large metropolitan areas to talk with the intended audience of pregnant women and pretest a bilingual (Spanish/English) booklet on breast feeding [97, 17-18]. They point out that interviews are an extremely rich data source about how a text is working because people often feel more comfortable answering interview questions than objective test items. Disadvantages of interviews include that they are time-consuming to conduct and the data are often very difficult to analyze, thus making it hard to generalize from them.

Focus groups, a method using group interview procedures for evaluation, has been a very popular means of pretesting the marketability of consumer products [142-145]. Focus groups use open-ended interviews to solicit people's attitudes, perceptions, and opinions about a single text or sometimes a group of texts, such as a new science textbook for a particular grade level or a new science textbook series for several grade levels of an entire school district. Focus groups in such a case could be helpful in discovering the kinds of text features teachers pay attention to when using a textbook and the range of factors that influence their choice of one text over another. (Unfortunately, up to this point, most focus groups aiming to evaluate text quality are actually subject-matter expert interviews—in this case, interviews with "expert" teachers or school system administrators.) Although in this example, the teachers are an important audience for judging text quality, it would be
better to conduct the focus groups with the students who will be reading the science texts. See Markle [146] or Pepper [147] for a discussion of the value of using student feedback to improve instructional materials.

Nonetheless, writers can use the kind of information generated by focus group discussions in planning and revising their texts. Under ideal circumstances, "the focus group presents a natural environment where participants are influencing and are influenced by others—just as they do in real life" [142, 30]. According to Krueger, focus groups have several distinct advantages and disadvantages:

- It is a socially-oriented research method capturing real-life data in a social environment.
- It has flexibility.
- It has high face validity.
- It has speedy results.
- It is low in cost [142, 47].

But focus groups have limitations that affect the quality of the results:

- Focus groups afford the researcher less control than individual interviews.
- Data are difficult to analyze.
- Moderators require special skills.
- Differences between groups can be troublesome.
- Groups are often difficult to assemble.
- The discussion must be conducted in a conducive environment [142, 48].

Critical incidents, a method which asks participants to remember salient aspects of their interaction with a text, is designed to elicit readers' memories of positive or negative experiences associated with reading or using text [148, 149]. For example, Williges [150] has used it as a method for software design and its accompanying documentation. He asks participants to describe a positive or negative incident using the computer, to discuss how many times the incident occurred and then to rate the relevance and severity of the incident in terms of "How much does this factor matter to you?" A similar technique is

called "storytelling"; participants are asked to tell the evaluator a narrative that reveals their attitudes and experiences related to text type or genre. Sometimes participants are provided with a scenario and are asked to complete the story discussing how and when they might use the text under evaluation. A key drawback of these methods is that they place an enormous burden on memory and may predispose participants to exaggerate, thus not providing very accurate or reliable data.

Another common retrospective test is the *reader feedback card* which is usually found at the end of a book or an instructional guide. The idea is to gather perceptions about text quality through having readers fill in a series of close-ended and/or open-ended survey questions. Again, reader feedback cards have the inherent bias of self-selected participants who are lavish with praise or condemnation for a text.

Summary

Overall, **retrospective testing** can provide extremely useful data for revising text. It is clear, however, that most researchers agree that concurrent measures provide the most reliable data. For this reason, retrospective methods should be used in conjunction with concurrent methods for greater reliability.

CONCLUSION

Earlier I argued that an optimal text-evaluation method should provide writers with two sorts of information: (1) information about whole-text or global aspects of text quality, and (2) information about how the audience may respond to the text. Clearly, research and experience show us that reader-focused testing methods have the advantage on both counts. When practical considerations such as time and expense allow, reader-focused methods are preferable to text-focused and expert-judgment- focused methods because they shift the primary job of representing the text's problems from the writer or expert to the reader. Thus, reader-focused methods help minimize the chances of failing to detect problems. In addition, reader-focused methods expand the scope of text problems that get noticed, shifting the evaluator's attention to global problems, especially problems of visual and verbal omissions. Most writers and readers would agree that perhaps the biggest problem with poorly written text lies not in what it says but in what it *fails to say*. Overall, reader-focused methods such as protocol-aided revision can help writers achieve a better model of readers actively engaged in meaning construction. Such a model of readers is helpful not only in revising the text under evaluation, but in planning and revising future text.

References

1. Grice, H. P., "Logic and Conversation," in P. Cole and J. J. Morgan (eds.), *Syntax and Semantics*, Vol. 3: Speech Acts, New York: Academic Press, 1975.

2. Kantor, R. N., "Anomaly, Inconsiderateness, and Linguistic Competence," in D. M. Lance and D. E. Gulstad (eds.), *Proceedings of the 1977 Mid-America Linguistics Conference*, Columbia, MO: University of Missouri, 1977.

3. Armbruster, B. A., "The Problem of 'Inconsiderate Text'," in *Comprehension Instruction: Perspectives and Suggestions*, G. Duffy, L. Roehler, and J. Mason (eds.), New York: Longman, 1984, pp. 202-217.

4. Anderson, T. H., and Armbruster, B., "Readable Textbooks, or, Selecting a Textbook is Not Like Buying a Pair of Shoes," in J. Orasanu (ed.), *Reading Comprehension: From Research to Practice*, Hillsdale, NJ: Erlbaum, 1986.

5. Meyer, B. J. F., "What is Remembered from Prose: A Function of Passage Structure," in R. O. Freedle (ed.), *Discourse Production and Comprehension*, Norwood, NJ: Ablex, 1977, pp. 307-336.

6. Sticht, T., "Understanding Readers and their Uses of Texts," in T. Duffy and R. Waller (eds.), *Designing Usable Texts*, New York: Academic Press, 1985, pp. 315-340.

7. Mikulecky, L., and Strange, R. L., "Effective Literacy Training Programs for Adults in Business and Municipal Employment," in J. Orasanu (ed.), *Reading Comprehension: From Research to Practice*, Hillsdale, NJ: Lawrence Erlbaum, 1986, pp. 319-334.

8. Smillie, R. J., "Design Strategies for Job Performance Aids," in T. Duffy and R. Waller (eds.), *Designing Usable Texts*, New York: Academic Press, pp. 213-243.

9. Redish, J., Battison, R., and Gold, E., "Making Written Information Accessible to Readers," in L. Odell and D. Goswami (eds.), *Writing in Nonacademic Settings*, New York: Guilford Press, 1985, pp. 129-153.

10. Wright, P., "Functional Literacy: Reading and Writing at Work," *Ergonomics 11*, 1988, pp. 265-290.

11. Schumacher, G. M., Scott, B. T., Klare, G. R., Cronin, F. C., and Lambert, D. A., "Cognitive Processes in Journalistic Genres: Extending Writing Models," *Written Communication 6* (3), July 1989, pp. 390-407.

12. Burtis, P. J., Bereiter, C., Scardamalia, M., and Tetroe, J., "The Development of Planning in Writing," in G. Wells and B. M. Kroll (eds.), *Explorations in the Development of Writing*, Chichester, England: John Wiley, 1983, pp. 153-174.

13. Hayes, J. R., and Flower, L. S., "Identifying the Organization of Writing Processes," in L. W. Gregg and E. R. Steinberg (eds.), *Cognitive Processes in Writing*, Hillsdale, NJ: Erlbaum, 1980, pp. 3-30.

14. Hayes, J. R., Flower, L., Schriver, K. A., Stratman, J., and Carey, L., "Cognitive Processes in Revision," in S. Rosenberg (ed.), *Advances in Applied Psycholinguistics, Volume II: Reading, Writing, and Language Processing*, Cambridge, England: Cambridge University Press, 1987, pp. 176-240.

15. Schriver, K. A., "Teaching Writers to Anticipate the Reader's Needs: Empirically Based Instruction," Doctoral Diss. in Rhetoric, Pittsburgh, PA: Carnegie Mellon University, Department of English, 1987. For a summary of this work, see Schriver, K. A., "Teaching Writers to Anticipate the Reader's Needs: A Classroom Evaluated Pedagogy," *Technical Report*, Berkeley, CA: University of California and Carnegie Mellon, Center for the Study of Writing, in press. A detailed version of this work is in preparation for the *NCTE Research Monograph Series*.

16. Spilka, R., "Studying Writer-Reader Interactions in the Workplace," *The Technical Writing Teacher 15(3)*, Fall 1988, pp. 208-221.

17. Atlas, M. A., "The User Edit: Making Manuals Easier to Use," *IEEE Transactions on Professional Communication*, PC-24, 1981, pp. 28-29.

18. Duffy, T. M., and Kabance, P., "Testing a Readable Writing Approach to Text Revision," *Journal of Educational Psychology 74 (5)*, 1982, pp. 733-748.

19. Hartley, J., *Designing Instructional Text*, 2nd ed., London: Kogan Page Ltd., 1985.

20. Schriver, K. A., "Revising Computer Documentation for Comprehension: Ten Exercises in Protocol-Aided Revision," *CDC Technical Report No. 14*, Pittsburgh, PA: Carnegie Mellon University, Communications Design Center, 1984. Also, see Schriver, K. A., and Mehlenbacher, B., "Predicting Readers' Problems with Text," a hypercard version of *CDC TR 14* created for the Apple Macintosh computer and designed to provide online instruction for document designers in predicting and diagnosing usability problems created by poorly written documentation, Pittsburgh, PA: Carnegie Mellon, Communications Design Center, alpha version, 1989.

21. Schriver, K. A., "Plain Language Through Protocol-Aided Revision," in E. R. Steinberg (ed.), *Plain Language: Principles and Practice*, Detroit, MI: Wayne State University Press, in press. Also, see "Writing for Expert or Lay Audiences: Designing Text Using Protocol Aided Revision," *CDC Technical Report No. 43*, Pittsburgh, PA: Carnegie Mellon University, Communications Design Center, 1989. A shorter version of this text is under consideration with *Written Communication*.

22. Swaney, J. H., Janik, C., Bond, S. J., and Hayes, J. R., "Editing for Comprehension: Improving the Process through Reading Protocols," *Document Design Project Report No. 14*, Pittsburgh, PA: Carnegie Mellon University, Communications Design Center, 1981.

23. Wright, P., "Quality Control Aspects of Document Design," *Information Design Journal 1*, 1979, pp. 33-42. Also, see Wright, P., "Usability: The Criterion for Designing Written Information," in P.A. Koler, M.E. Wrolstad, and H. Bouma (eds.), *Processing Visible Language, Vol. 2*, New York: Plenum Press, 1980, pp. 183-206.

24. Wright, P., "Is Evaluation a Myth? Assessing Text Assessment Procedures," in D. H. Jonassesn (ed.), *The Technology of Text, Vol. 2*, Englewood Cliffs, NJ: Educational Technology Publications, 1985, pp. 418-435.

25. Schriver, K. A., "Document Design from 1980 to 1990: Challenges that Remain," *Technical Communication*, 4th Quarter, 1989, in press.

26. Thibadeau, R., Just, M., and Carpenter, P. A., "A Model of the Time Course and Content of Reading," *Cognitive Science 6*, 1982, pp. 157-203.

27. Bracewell, R., Scardamalia, M., and Bereiter, C., "The Development of Audience Awareness in Writing, in *Resources in Education*, 1978, ERIC Doc. Service No. ED 154 433.

28. Flesch, R., *The Art of Readable Writing*, New York: Harper and Row, 1949 [revised, 1974]. Also, see R. Flesch, "A New Readability Yardstick," *Journal of Applied Psychology 32*, 1948, pp. 221-233.

29. Gunning, R., *The Technique of Clear Writing*, 2nd ed., New York: McGraw-Hill, 1968 [originally published in 1952]. Also, see R. Gunning, "The Fog Index After Twenty Years," *The Journal of Business Communication 6*, Winter 1968, pp. 3-13.

30. McLaughlin, H. G., "SMOG Grading—A New Readability Formula," *Journal of Reading 12*, May 1969, pp. 639-646.

31. Dale, E., and Chall, J. S., "A Formula for Predicting Readability," *Educational Research Bulletin 27*, January 21 and February 17 1948, pp. 11-20 and pp. 37-54.

32. Fry, E. B., "A Readability Formula That Saves Time," *Journal of Reading 11*, 1968, pp. 513-516, 575-578.

33. Kincaid, J. P., Fishburne, R. P., Rogers, R. L., and Chissom, B. S., "Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy Enlisted Personnel," *Research Branch Report 8-75*, Millington, TN: Naval Air Station Memphis, February 1975.

34. Duffy, T. M., "Readability Formulas: What's the Use?" in T. Duffy and R. Waller (eds.), *Designing Usable Texts*, New York: Academic Press, 1985, pp. 113-143.

35. Huckin, T., "A Cognitive Approach to Readability," in P. V. Anderson, R. J. Brockmann, and C. R. Miller (eds.), *New Essays in Technical and Scientific Communication: Research, Theory, Practice*, New York: Baywood, 1983, pp. 90-108.

36. Kintsch, W., and Vipond, D., "Reading Comprehension and Readability in Educational Practice and Psychological Theory," in L. Nilsson (ed.), *Proceedings of the Conference on Memory*, Hillsdale, NJ: Erlbaum, 1977.

37. Klare, G. R., "Assessing Readability," *Reading Research Quarterly 10*(1), 1974-1974, pp. 61-102.

38. Seltzer, J., "What Constitutes a Readable Technical Style?" in P. V. Anderson, R. J. Brockmann, and C. R. Miller, (eds.), *New Essays in Technical and Scientific Communication: Research, Theory, Practice*, New York: Baywood, 1983, pp. 71-89.

39. Singer, H., and Donlan, D., "Readability: A Text Selection Strategy," in *Reading and Learning from Text*, 2nd ed., Hillsdale, NJ: Erlbaum, 1989, pp. 308-343.

40. Voss, J., Tyler, S. W., and Bisanz, G. L., "Prose Comprehension and Memory," in R. C. Puff (ed.), *Handbook of Research Methods in Human Memory and Cognition*, New York: Academic Press, 1982, pp. 349-395.

41. Coke, E. U., "Computer Aids for Writing Text," in D. H. Jonassen (ed.), *The Technology of Text,Vol. 2*, Englewood Cliffs, NJ: Educational Technology Publications, 1982, pp. 383- 399.

42. Cherry, L. L., Fox, M. L., Frase, L. T., Gingrich, P. S., Keenan, S. A., and Macdonald, N. H., "Computer Aids for Text Analysis," *Bell Laboratories Record, 10-16*, May/June 1983, pp. 269-284.

43. "Computer Aids for Authors and Editors: A Natural Extension of Word Processing and Typesetting?" *The Seybold Report on Publishing Systems, 13*(10), 1984.

44. Frase, L. T., "The UNIX Writer's Workbench Software: Philosophy," *Bell Systems Journal 62*, 1983, pp. 1909-1921.

45. Macdonald, N. H., Frase, L. T., Gingrich, P., and Keenan, S., "The Writer's Workbench: Computer Aids for Text Analysis," *IEEE Transactions on Communication, Com-30*(1), 1982, pp. 105-110.

46. Taft, M.W., "Online Readability Testing of Software Manuals," *IEEE Professional Communication Society Conference Record: The Many Facets of Computer Communications*, New York, NY: The Institute of Electrical and Electronic Engineers, Inc., 1983, pp. 44-45.

47. Miller, L. A., "Computers for Composition: A Stage Model Approach to Helping," *Visible Language, XX* (2), Spring 1986, pp. 188-218.

48. Richardson, S., Creed, W., and Chandler, R., "Critique as a Teaching Tool for Writing Classes," in *The Dynamic Text Guide, 9th International Conference on Computers and the Humanities (ICCH) and 16th International Association for Literary and Linguistic Computing (ALLC) Conference*, Toronto, Canada: University of Toronto, Centre for Computing in the Humanities, June 5-10, 1989, pp. 57-58.

49. Heidorn, G. E., Jensen, K., Miller, L.A., Byrd, J. R., and Chodorow, M.S., "The EPISTLE Text-Critiquing System," *The IBM Systems Journal 21*(3), 1982, pp. 305-326.

50. McCord, M. C., "Semantic Interpretation for the EPISTLE System," in *Proceedings of the Second International Logic Programming Conference*, Uppsala, Sweden, 1984, pp. 65-76.

51. Miller, L. A., "Project EPISTLE: A System for the Automatic Analysis of Business Correspondence," in *Proceedings of the First Annual Conference on Artificial Intelligence*, Stanford, CA: Stanford University, August 1980, pp. 280- 282.

52. Language Technology Electric Word, "IBM Critique Out," *Language Technology Electric Word*, July/August 1989, pp. 3, 7.

53. Language Technology Electric Word, "1989 Annual Award for Technical Excellence: Style/Grammar Checker for the PC and MAC," *Language Technology Electric Word*, July/August 1989, pp. 34-35.

54. Felker, D. B., Pickering, F., Charrow, V. R., Holland, M., and Redish, J. G., *Guidelines for Document Designers*, Washington, DC.: American Institutes for Research, 1981.

55. *Harbrace College Handbook*, 10th ed., J. C. Hodges, and M. Whitten (eds.), New York: Harcourt Brace and Jovanovich, 1986.

56. Hartley, J., "Eighty Ways to Improve Instructional Text," *IEEE Transactions on Professional Communication 24*, 1981, pp. 17-27.

57. Williams, J. M., *Style: Ten Lessons in Clarity and Grace*, 2nd ed., Glenview, IL: Scott, Foresman, 1985.

58. Strunk, W. Jr., and White, E. B., *The Elements of Style*, New York: Macmillan, 1968, p. 23.

59. Dieli, M., "Designing Successful Documents: An Investigation of Document Evaluation Methods," Doctoral Diss. in Rhetoric, Pittsburgh, PA: Carnegie Mellon University, Department of English, 1986.

60. Wright, P., "Five Skills Technical Writers Need," *IEEE Transactions in Professional Communication PC 24*, 1981, pp. 10-16.

61. Wright, P., "Can Research Assist Technical Communication?" in *Proceedings of the 36th International Technical Communications Conference*, Washington, DC: Society of Technical Communication, 1989, pp. RT-3 — RT-7.

62. Price, J., *How to Write a Computer Manual: A Handbook of Software Documentation*, Menlo Park, CA: The Benjamin & Cummings Publishing Co, 1984.

63. Spencer, R. H., *Computer Usability Testing and Evaluation*, Englewood Cliffs, NJ: Prentice-Hall, Inc., 1985, pp. 150-174.

64. Hayes, J. R., *The Complete Problem Solver*, 2nd ed., Hillsdale, NJ: Erlbaum, 1989, pp. 218-223.

65. Cole, S., Rubin, L., and Cole, J. R., *Peer Review in the National Science Foundation: Phase One of a Study*, Washington, DC: National Academy of Sciences, 1978.

66. Keddie, J., "Peer Review: Guaranteeing Document Quality," in *Proceedings of the 36th International Technical Communication Conference*, Washington, DC: Society for Technical Communication, 1989, pp. WE-51 — WE-54.

67. Maggiore, J. G., "A Three Step Approach for Conducting Formal Peer Reviews," in *Proceedings of the 35th International Technical Communication Conference*, Washington, DC: Society for Technical Communications, 1988, pp. MPD-84 — MPD-87.

68. Sherwood, A. C., and Cowan, J., "A System of Formal Peer Review for Documentation," in *Proceedings of the 34th International Technical Communications Conference*, Washington, DC: Society of Technical Communication, 1987, pp. MPD-70 — MPD-72.

69. Bond, S. J., Hayes, J. R., and Flower, L. S., "Translating the Law into Common Language: A Protocol Study," Document Design Project *Technical Report No. 8*, Pittsburgh, PA: Carnegie Mellon University, Communications Design Center, April 1980.

70. Crandall, R. "Peer Review: Improving Editorial Procedures," *Bioscience 36*, 1986, pp. 407-409.

71. Kupfersmid, J., "Improving What is Published: A Model in Search of an Editor," *American Psychologist 43*, 1988, pp. 635-642.

72. Ceci, S. J., and Peters, D., "How Blind is a Blind Review?" *American Psychologist 39*, 1984, pp. 1491-1494.

73. Surwillo, W., "Anonymous Reviewing and the Peer Review Process," *American Psychologist 41*, 1986, p. 218.

74. Hayes, J. R., Schriver, K. A., Blaustein, A., and Spilka, R., "If Its Clear to Me, It Must Be Clear to Them: How Knowledge Makes it Difficult to Judge," paper presented at the American Educational Research Association (AERA) Conference, San Francisco, CA, April, 1986. Also discussed in Hayes, J. R., "Writing Research: The Analysis of a Very Complex Task," in D. Klahr and K. Kotovsky (eds.), *Complex Information Processing: The Impact of Herbert A. Simon*, Hillsdale, NJ: Erlbaum, 1989, pp. 209-234.

75. Brouns, V. L., and Grove, L. K., "Comprehensive Editing: A Solution to Some Typical Editing Problems," in *Proceedings of the 35th International Technical Communication Conference*, Washington, DC: Society of Technical Communication, 1988, pp. WE-119— WE-121.

76. Buehler, M. F., "Controlled Flexibility in Technical Editing: The Levels-of-Edit Concept at JPL," *Technical Communication 24(*1), 1977, pp. 1-4.

77. Buehler, M. F., "Defining Terms in Technical Editing: The Levels of Edit as a Model," *Technical Communication 28*(4), 1981, pp. 10-14.

78. Dressel, S., and Prasad, S., "Error Classes and Editorial Accountability," in *Proceedings of the 36th International Technical Communication Conference*, Washington, DC: Society of Technical Communication, 1989, pp. WE-59 — WE-62.

79. Van Buren, R., and Buehler, M. F., *The Levels of Edit*, 2nd ed., JPL Publication 80-1, Pasadena, CA: California Institute of Technology, Jet Propulsion Laboratory, 1980.

80. Henke, K. A., "Measuring Editing Quality Contributions," in *Proceedings of the 35th International Technical Communication Conference*, Washington, DC: Society of Technical Communication, 1988, pp. WE-34 — WE-36.

81. Berkenkotter, C., "Understanding a Writer's Awareness of Audience," *College Composition and Communication 32*, 1981, pp. 388-399.

82. Flower, L., Hayes, J. R., Carey, L., Schriver, K. A., and Stratman, J., "Detection, Diagnosis, and the Strategies of Revision," *College Composition and Communication 37*, 1986, pp. 16-55.

83. Schriver, K. A., "Moving from Sentence-Level to Whole-Text Revision: Helping Writers Focus on the Reader's Needs," in K. A. McCormick (ed.), *Reading to Write: Practical Approaches for the Teaching of Writing*, Reading-to-Write Report No. 11, Technical Report No. 30, Berkeley, CA: University of California and Carnegie Mellon, Center for the Study of Writing, May 1989, pp. 46-57.

84. Sommers, N., "Revision Strategies of Student Writers and Experienced Writers," *College Composition and Communication 31*, 1980, pp. 378-387.

85. Charney, D., "The Validity of Using Holistic Scoring to Evaluate Writing: A Critical Overview," *Research in the Teaching of English 18(*1), 1984, pp. 65-81.

86. Cooper, C., "Holistic Evaluation of Writing," in C. Cooper and L. Odell (eds.), *Evaluating Writing*, Urbana, IL: National Council of Teachers of English, 1977.

87. Diederich, P., *Measuring Growth in English*, Urbana, IL: National Council of Teachers of English, 1974.

88. Odell, L., "Defining and Assessing Competence in Writing," in C. Cooper (ed.), *The Nature and Measurement of Competency in English*, Urbana, IL: National Council of Teachers of English, 1981.

89. Odell, L., and Cooper, C., "Procedures for Evaluating Writing: Assumptions and Needed Research," *College English 42*, 1980, pp. 35-43.

90. Lloyd-Jones, R., "Primary Trait Scoring," in C. Cooper and L. Odell (eds.), Evaluating Writing, Urbana, IL: National Council of Teachers of English, 1977.

91. Freedman, S., "How Characteristics of Student Essays Influence Teachers' Evaluations," *Journal of Educational Psychology 71*, 1979, pp. 328-338.

92. Freedman, S., "Influences on the Evaluation of Expository Essays: Beyond the Text," *Research in the Teaching of English 15*, 1981, pp. 245-255.

93. Grobe, C., "Syntactic Maturity, Mechanics and Vocabulary as Predictors of Quality Ratings," *Research in the Teaching of English 15*, 1981, pp. 75-86.

94. U.S. Department of Consumer Affairs, *Consumer Resource Handbook*, Pueblo, CO: Consumer Information Center, Office of Consumer Affairs, 1988.

95. U.S. Department of Commerce, Office of Consumer Affairs, *How Plain English Works for Business: Twelve Case Studies*, Washington, DC: U.S. Government Printing Office, March 1984.

96. LeVitus, B., "Reader Report #1: Word Processors," *MACazine*, January 1989, pp. 64-65.

97. U.S. Department of Health, *Pretesting in Health Communications: Methods, Examples, and Resources for Improving Health Messages and Materials*, NIH

Publication No. 84-1493, Bethesda, MD: U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute, 1984.

98. Floreak, M. J., "Designing for the Real World: Using Research to Turn a 'Target Audience' into Real People," *Technical Communication*, 4th Quarter, 1989, in press.

99. Schriver, K. A., Steinberg, E., Ogura, A., and Bloch, J., "Anticipating Poor Writing Before It Occurs: The Process Critique," *CDC Technical Report*, Pittsburgh, PA: Carnegie Mellon University, Communications Design Center, in preparation.

100. Bormuth, J. R., "Readability: A New Approach," *Reading Research Quarterly 1* 1966, pp. 79-132. Also, see Bormuth, J. R., "Cloze Test Readability: Criterion Scores," *Journal of Educational Measurement 5*, 1968, pp. 189-196. Also, see Bormuth, J. R., "Development of Readability Analyses," *Univ. of Chicago Final Report, Project No. 7-0052*, Contract No. OEC-3-7-070052-0326, Washington, DC: U.S. Office of Education, 1969.

101. Taylor, W. L., "Cloze Procedure: A New Tool for Measuring Readability," *Journalism Quarterly 30*, Fall 1953, pp. 415-432. Also, see Taylor, W. L., "'Cloze' Readability Scores as Indices of Individual Differences in Comprehension and Aptitude," *Journal of Applied Psychology 41*, 1957, pp. 19-26.

102. Panackal, A. A., and Heft, C. S., "Cloze Techniques and Multiple Choice Technique: Reliability and Validity," *Educational and Psychological Measurement 38*, 1978, pp. 917- 932.

103. Card, S. K., Moran, T. P., and Newell, A., *The Psychology of Human-Computer Interaction*, Hillsdale, NJ: Erlbaum, 1983.

104. John, B. E., and Newell, A., "Cumulating the Science of HCI: From S-R Compatibility to Transcription Typing," in K. Bice and C. Lewis (eds.), *CHI '89 Human Factors in Computing Systems Conference Proceedings*, New York: Association of Computing Machinery, 1989, pp. 109-114.

105. Hendrickson, J., "Performance, Preference, and Visual Scan Patterns on a Menu-Based System: Implications for Interface Design," in *CHI '89 Proceedings*, New York: Association of Computing Machinery, 1989, pp. 217-222.

106. Hegarty, M., Carpenter, P. A., and Just, M., "Diagrams in the Comprehension of Scientific Text," in R. Barr, M. Kamil, P. Mosenthal, and P. D. Pearson (eds.), *Handbook of Reading Research, Vol. II*, New York: Longman, in press.

107. McConkie, G. W., Hogaboam, T. W., Wolverton, G. S., and Lucas, P. A., "Toward the Use of Eye Movements in the Study of Language Processing," *Discourse Processes* 2, 1979, pp. 157-177.

108. Just, M., and Carpenter, P. A., "Eye Fixations and Cognitive Processes," *Cognitive Psychology 8*, 1976, pp. 441-480. Also, see Just, M., and Carpenter, P. A., "Inference Processes During Reading: Reflections from Eye Fixations," in J. W. Senders, D. F. Fisher, and R. A. Monty (eds.), *Eye Movements and the Higher Psychological Functions*, Hillsdale, NJ: Erlbaum, 1978. Also, see Just, M., and Carpenter, P. A., "Theory of Reading: From Eye Fixations to Comprehension," *Psychological Review 87*, 1980, pp. 329-354.

109. Rayner, E., "Eye Movements in Reading and Information Processing," *Psychological Bulletin 85*, 1978, pp. 618-660.

110. Bennion, B., "Performance Testing of a Book and its Index as an Information Retrieval System," *Journal of American Information Science 31*(4), 1980, pp. 264-270.

111. Mills, C. B., and Dye, K., "Usability Testing: User Reviews," *Technical Communication 32*(4), 1985, pp. 40-44.

112. Lewis, B., and Crews, A., "The Evolution of Benchmarking as a Computer-Performance Evaluation Technique," *MIS Quarterly*, March 1985, pp. 7-16.

113. Roberts, T. L., and Moran, T. P., "The Evaluation of Text Editors: Methodology and Empirical Results," *Communications of the Association for Computing Machinery 26*, 1983, pp. 265-283.

114. Evans, K. M., *Planning Small Scale Research*, 3rd ed., Berkshire, England: NFER-Nelson Publishing Company Ltd., 1984, p. 11.

115. Schumacher, G., and Waller, R., "Testing Design Alternatives: A Comparison of Procedures," in T. Duffy and R. Waller (eds.), *Designing Usable Texts*, New York: Academic Press, 1985, pp. 377-403.

116. Bond, S. J., "Protocol-Aided Revision: A Tool for Making Documents Usable," paper presented at the IBM Academic Information Systems University AEP Conference, Alexandria, VA, 1985. (Available from the author at: Software Engineering Institute, Fifth Avenue, Pittsburgh PA 15213.)

117. Knox, S. T., Bailey, W. A., and Lynch, E. F, "Directed Dialogue Protocols: Verbal Data for User Interface Design," in *CHI '89 Proceedings*, New York, NY: Association of Computing Machinery, 1989, pp. 283-288.

118. Hayes, J. R., "The Use of Reading Protocols in the Editing of Text," paper presented at the International Reading Association (IRA), Chicago, IL, April, 1982.

119. Holland, V. M., Rose, A. M., Dean, A. R., and Dory, S. L., "Processes Involved in Writing Effective Procedural Instructions: Final Report," *Technical Report No. 1*, Washington, DC: American Institutes for Research, 1985. Also, see Holland, V. M., "How to Write Procedural Instructions: Assessing Problems and Solutions," paper presented at the American Educational Research Association Convention, New Orleans, LA, April 5-9, 1988. (Available from the author at the U.S. Army Research Institute for the Behavioral and Social Sciences, 5001 Eisenhower Ave., Alexandria, VA 22333.)

120. Lewis, C., "The 'Thinking Aloud' Method in Interface Evaluation," paper presented at the Computer and Human Interaction (CHI) 1983 Conference on Human Factors in Computing Systems, 1983.

121. Lund, M. A., "Evaluating the User Interface: A Candid Camera Approach," *Computer and Human Interaction (CHI) Proceedings*, New York, NY: Association for Computing Machinery, 1985, pp. 107-113.

122. Mulcahy, P. I., "Testing Documentation Using Protocol Analysis," in *Proceedings of the Society of Technical Communication*, Washington, DC: Society of Technical Communication, 1988, pp. WE-27 — WE-29.

123. Schriver, K. A., Hayes, J. R., Danley, C., Wulff, W., Davies, L., Cerroni, K., Graham, D., Flood, E., and Bond, E., *Designing Computer Documentation: A Review of the Literature—Hardcopy, Online, General Applications, CDC Technical Report No. 31*, Pittsburgh, PA: Carnegie Mellon University, Communications Design Center, 1986.

124. Ericsson, K. A., and Simon, H. A., *Protocol Analysis: Verbal Reports as Data*, Cambridge, MA: The MIT Press, 1984.

125. Hayes, J. R., and Simon, H. A., "Understanding Written Problem Instructions," in L. W. Gregg (ed.), *Knowledge and Cognition*, Potomac, MD: Erlbaum, 1974.

126. Hayes, J. R., Waterman, D. A., and Robinson, C. S., "Identifying the Relevant Aspects of a Problem Text," *Cognitive Science 1*(3), 1977, pp. 297-313.

127. Soderston, C., "The Usability Edit: A New Level," *Technical Communication*, 1st Quarter, 1985, pp. 16-18.

128. Hayes, J. R., and Flower, L. S., "Uncovering Cognitive Processes in Writing: An Introduction to Protocol Analysis," in P. Mosenthal, L. Tamor, and S. Walmsey (eds.), *Research in Writing: Principles and Methods*, New York: Longman, 1983, pp. 207- 220.

129. Glass, A. L., Holyoak, K. J., and Santa, J. L., *Cognition*, Reading, MA: Addison-Wesley Publishing Co., 1979, pp. 416-417.

130. Specter, H. A., *Products Liability: The First Twenty Five Years*, Washington, DC: The Association of Trial Lawyers of America Education Fund, 1983.

131. Wilson, C. M., "Unsubstantiated Claims," in *Proceedings of the Society of Technical Communication*, Washington, DC: Society of Technical Communication, 1988, pp. WE-103 — WE-106.

132. Dick, W., and Carey, L., *The Systematic Design of Instruction*, 2nd ed., Glenview, IL: Scott, Foresman, 1985, pp. 118-119.

133. Babbie, E. R., *Survey Research Methods*, Belmont, CA: Wadsworth Publishing, 1973. Also, see Babbie, E. R., *The Practice of Social Research*, Belmont, CA: Wadsworth Publishing, 1975, pp. 268-274.

134. Fink, A., and Kosecoff, J., *How to Conduct Surveys: A Step-by-Step Guide*, Newbury Park, CA: Sage Publications, 1985.

135. Frey, J. B., *Survey Research by Telephone*, Newbury Park, CA: Sage Publications, 1989.

136. Erdos, P. L., *Professional Mail Surveys*, New York: McGraw Hill, 1970.

137. Institute for Social Research, *Interviewers Manual* (rev ed.), Ann Arbor, MI: The University of Michigan, Institute for Social Research, 1978.

NATIONAL ADVISORY PANEL
The Center for the Study of Writing

Chair
Fred Hechinger
The New York Times Foundation

Alonzo Crim
Professor of Urban Educational Leadership
Georgia State University, Atlanta, GA

Sister Regina Noel Dunn
Teacher
Villa Maria Academy, Malvern, PA

Marcia Farr
Associate Professor of English
University of Illinois, Chicago, IL

Abraham Glassman
Chairman
Connecticut State Board of Education

Bill Honig
California Superintendent
of Public Instruction

The Honorable Gary K. Hart
California State Senator

Sibyl Jacobson
Executive Director
Metropolitan Life Foundation

John Maxwell
Exeutive Director
National Council of Teachers of English

Roy Peña
Principal
Andrews High School, El Paso, TX

Carol Tateishi
Teacher
Ross Elementary School, Kentfield, CA

Richard C. Wallace, Jr.
Pittsburgh Superintendent of Schools
and Secretary, Board of Education