ED 387 500		TM 023 627
AUTHOR	Witt, Elizabeth A.	
TITLE	Issues in Constructing an Analytic Scoring Scale for a Writing Assessment.	
PUB DATE	Apr 95	
NOTE	53p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).	
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)	
EDRS PRICE	MF01/PC03 Plus Postage.	
DESCRIPTORS	Educational Assessment; Elementary Secondary Education; *Holistic Evaluation; *Scores; Student Evaluation; Test Bias; *Test Construction; Writing (Composition); *Writing Evaluation	
IDENTIFIERS	*Analytic Scoring; *Iowa Writing Assessment; Topic Selection; Writing Test Prompts	

ABSTRACT

Analytic scoring was added to the 1994 Iowa Writing Assessment as a complement to focused holistic scoring. Four trait scores are provided: ideas/content, organization, voice, and conventions. Scoring protocols were developed on the basis of students' actual writing samples, with particular attention to factors affecting the fairness and objectivity of the scoring process. Challenges encountered in the construction of the protocols raised a number of issues regarding the definition of analytic traits, prompt-related problems, and topic-trait interactions. For example, the relative importance of each trait was found to vary with the specific topic, so that a student's choice of topic has a greater impact on analytic scoring than on focused holistic scoring. Furthermore, the particular manifestation of each trait that leads to a successful performance was sometimes found to vary with the topic, prompt, or mode. This paper describes these and other issues and explains how they were dealt with in constructing the protocols; suggestions are made for modifying prompts to enhance fairness in future writing assessments. (Contains 11 references, 1 table, and 2 appendixes giving prompt descriptions and scoring protocols.) (Author)

Issues in Constructing an Analytic Scoring Scale for a Writing Assessment

by

Elizabeth A. Witt
The University of Kansas

April, 1995

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California, April 18-22, 1995.

ABSTRACT

Analytic scoring was added to the 1994 *Iowa Writing Assessment* as a complement to focused holistic scoring. Four trait scores are provided: ideas/content, organization, voice, and conventions. Scoring protocols were developed on the basis of students' actual writing samples, with particular attention to factors affecting the fairness and objectivity of the scoring process. Challenges encountered in the construction of the protocols raised a number of issues regarding the definition of analytic traits, prompt-related problems, and topic-trait interactions. For example, the relative importance of each trait was found to vary with the specific topic, so that a student's choice of topic has a greater impact on analytic scoring than on focused holistic scoring. Furthermore, the particular manifestation of each trait that leads to a successful performance was sometimes found to vary with the topic, prompt, or mode. This paper describes these and other issues and explains how they were dealt with in constructing the protocols; suggestions are made for modifying prompts to enhance fairness in future writing assessments.

The increased popularity of "authentic" performance assessments in all areas of academic achievement is perhaps most evident in the widespread use of large-scale direct assessments of writing proficiency among American schools. In recent years analytic scoring, which yields individual scores for several components of writing skill, has become increasingly desired because of its perceived diagnostic capabilities. In response to the demand for this type of scoring, the *Iowa Writing Assessment (IWA)* (Hoover, Hieronymus, Frisbie, & Dunbar, 1994; Feldt, Forsyth, Ansley, & Alnot, 1994) has developed an analytic scoring system to complement the focused holistic scoring for which the *IWA* is already well known.

Among the questions that must be addressed in the development and evaluation of any performance assessment are those pertaining to *fairness* and *objectivity* (Baker, 1994). The scoring protocols that constitute the analytic scales for the 1994 *Iowa Writing Assessment* were therefore developed with particular attention to factors affecting the fairness and objectivity of the scoring process. Challenges encountered in the construction of the protocols raised a number of issues regarding the definition of analytic traits, prompt-related problems, and topic-trait interactions. This paper describes those issues and explains how they were dealt with, when possible, in construction of the protocols; suggestions are made for modifying prompts to enhance fairness in future writing assessments.

Development of the Scales

Analytic scoring systems have been produced and refined over the last twenty years, building on the work of Diederich (1974) and others (e.g., Spandel & Stiggins, 1990). The fruits of their labor were incorporated into the construction of four trait scales for the *Iowa Writing Assessment*: **ideas/content, organization, voice, and conventions**. Traits were defined in such a way as to include all skills considered important to good writing under the umbrella of one or more of these traits. (The four traits are described in the sample protocols in Appendix B. See Witt [1993] or the *IWA* manual [in press, no reference] for actual definitions.) In keeping with the focused holistic scale used in the earlier forms of the assessment, scales were designed with a four-point range, with 1 representing minimal proficiency and 4 an outstanding performance. Scoring protocols for raters were created with the goal of measuring performance on each trait apart from contextual considerations (unique demands of the mode of discourse, the topic, or the prompt-style). It was not possible to achieve this goal completely, since different modes, and sometimes different topics, elicited very different types of responses. Furthermore, concerns for fostering interrater agreement necessitated a certain amount of mode specificity and occasionally prompt specificity in the protocols. Therefore, a balance was sought between the specificity required to ensure both sufficient interrater reliability (objectivity) and the generality required to foster greater generalizability of analytic scores to other writing contexts (fairness).

Although scales were developed for higher grade levels, this paper explicitly addresses only the construction of the three lowest levels, appropriate for grades 3/4, 5/6, and 7/8. Similar issues were also manifest in the construction of the high school level scales.

Table 1 lists the prompts used in the standardization of 1994 *Iowa Writing Assessment* by mode of discourse and level. Prompts were administered by grade; the level designations indicate the approximate age range of students responding to each prompt. At each level, two prompts were administered to measure each of four traditional modes of discourse: narrative, descriptive, persuasive, and expository. These are the familiar modes of discourse addressed in the theory and practice of writing instruction (Ruth and Murphy, 1988) and are described in detail in the manual accompanying the *IWA* (in press, no reference available). In brief, a narrative tells a story, whether fictional or true; it features character and plot development. A description attempts to create so rich an image that the reader can vicariously experience the event (person, place, or thing) the writer describes. A persuasive essay presents an opinion and supports it with material designed to convince the reader to adopt the author's point of view. An exposition presents information or ideas; it seeks to explain, inform, or instruct. (Cantor, 1986; Hieronymus, Hoover, Cantor, & Oberley, 1987a, 1987b)

Insert Table 1 about here

The fall standardization of Forms K/L of the *Iowa Tests of Basic Skills (ITBS)* took place in the autumn of 1992. Approximately one-third of this nationally representative sample also took part in the standardization of the *Iowa Writing Assessment.* The construction of the analytic scoring protocols was based on a subsample of essays from this group.

Two prompts were used in the standardization for each level in each of the four modes of discourse. Four separate scoring protocols were developed for each same-mode pair of prompts: one protocol for each analytic trait. Eighty essays were sampled from the complete standardization sample for each prompt at each level (40 essays from each grade responding to that prompt). Half of each group of 40 were written by girls, half by boys; two girls and two boys were sampled from each decile of academic ability (as estimated by concurrent *ITBS* vocabulary score). Furthermore, insofar as possible, the scale construction sample was selected in such a way that geographical regions and public/parochial schools were represented at each ability decile in proportions similar to the proportions in which they occurred in the full standardization sample. Each protocol was thus developed on the basis of 160 essays written by a balanced diversity of students in response to two different prompts in the same mode of discourse and designed for the same two-year grade level.

Issues Raised in Scale Construction

The construction of the analytic trait scales was a challenging task, presenting both anticipated and unexpected difficulties. The scales, represented as scoring protocols, were built on the basis of actual essays composed by students in the scale construction sample, and those responses raised a number of issues for consideration. Although it was expected that mode-to-mode differences would provide some challenges for scaling, most issues seemed to be related instead to the prompt--or more precisely, to the pair of prompts or the prompt style. Even differences in student responses across age/grade levels sometimes appeared to be due more to differences in prompt style than to increased maturity and writing ability. (This is not to say that there was no improvement in trait skills with increased level; however, improvement was predictably modest, and other kinds of differences were prominent.)

The remainder of this paper will present some of the general problems related to the trait scales and to prompts, then discuss issues raised in connection with specific pairs of prompts. The manner in which certain difficulties were addressed in the protocols and other training materials is also described; some issues, however, were not dealt with in training, but were later taken into account in selecting prompts for the final form of the assessment. Descriptions of the 24 prompts--indispensable for understanding the discourse below--are presented in Appendix A. Appendix B contains samples of typical scoring protocols. The full texts of both prompts and protocols are available from the Iowa Testing Programs and Riverside Publishing Company.

Trait Scales

General Issues

Scoring protocols for all 48 analytic scales (four trait scales for each same-mode, same-level pair of prompts) display similarity in the descriptions of each trait across both modes and levels. This similarity occurs by design. There are general statements that can be made about each trait without taking mode into account. The more general the scoring guides, the easier it is to handle the scoring of responses that seem to be partially or completely off-mode. On the other hand, for each trait (although to a lesser extent for conventions), there are mode-to-mode differences in the importance of certain aspects of writing. For example, a sound conclusion is likely to carry greater weight in distinguishing among responses to a persuasive prompt than it would among descriptive essays. Furthermore, the more specific the scoring guide is to the prompt as well as to the mode, the easier it is to build a scale that is likely to elicit high rater reliability. Of course, the potential generalizability of trait scores may suffer as protocols become more specific. It is a familiar trade-off. The finished protocols represent an effort to find the right degree of mode-specificity, one that balances concerns for interrater reliability and generalizability. This effort appears to have been successful, since interrater reliability coefficients, correlations between scores on pairs of essays written in the same mode of discourse, and correlations between scores on pairs of essays written in different modes all reached reasonably high levels (Witt, 1993).

Two prompts were used to build the scales for each mode-level combination in order to minimize prompt-specificity and to support the ability to make inferences about students' skills that would generalize to similar types of writing assignments. However, it should be noted that most prompt pairs were somewhat ?arallel in format, style, and degree of explicitness. In the one case where the two prompts were quite distinct (Level 13-14, Expository), component skills were manifested differently in the sample essays depending on the prompt, and it was necessary to address each prompt specifically in the original protocols. This will be discussed in greater detail below. Ultimately, only prompt from each pair was selected for inclusion in the final form of the assessment. Nevertheless, constructing each scoring protocol on the basis of two prompts not only supported the generalizability of the scores but also helped to identify issues related to the production and evaluation of writing samples that might have otherwise gone unnoticed.

Among the sample essays there appeared to be a strong relationship between performance on ideas/content and the overall quality of the essay. In addition, writers who scored high on ideas/content often scored high on either organization or voice, or both. In part, this is a simple matter of correlated skills: many writers appear to develop skills simultaneously on different components. In part, however, the traits are intertwined, dependent on one another. For example, the writer who generates few ideas has little to work with on organization. The writer who generates many irrelevant or loosely related ideas will have a hard time creating a logical progression of thought and smooth transitions. Smooth transitions, a characteristic of a well-organized essay, are often created by carefully chosen words and skillfully constructed sentences--elements that may also contribute to the expression of a personal voice (and, to an extent, conventions). These interrelationships of component skills were kept in mind in composing the protocols; where confusion of traits might arise, the potential confounding is explicitly pointed out and raters are encouraged to concentrate only on the contribution of various elements to the trait of interest. For example (from organization scales for descriptive essays):

> ". . . be careful to concentrate only on the manner in which ideas are organized; do not judge the quality of the ideas themselves. Pay attention only to the contribution of organization to development."

General instructions to raters also contain a reminder to concentrate on a single trait and avoid allowing scores to be influenced by other traits or by an overall impression of writing proficiency.

Inasmuch as possible, organization, voice, and conventions were scaled without regard for the purpose of the essay as stated in the prompt. That is, protocols attempt to describe varying degrees of proficiency on each of these traits without reference to the specifics of the pair of prompts on which the scales are based. When an essay is on topic and on task, the resulting trait scales do essentially measure the effectiveness of component skills for the given purpose. When an essay is off task, the student's component skills in context (in an appropriate response to the mode and prompt) may be overestimated. (For example, students often respond to a non-narrative prompt with a narrative essay. In general, students display greater skill in organization and voice when writing in the narrative mode.) However, the student's skill in achieving the purpose of the prompt is already measured (to a degree) by the focused holistic score, and essays that are clearly off task would be designated "unscoreable" on ideas/content. By not directly considering the prompt-imposed purpose, raters can provide scores on the other three traits to yield some additional information about the student's writing ability. In addition, if the specific purpose is ignored, the entire essay may be rated, rather than just the prompt-relevant part. Some analytic assessments direct raters to score only the mode-appropriate portions of an essay (e.g., Gardner, Rudman, Karlsen, & Merwin, 1983). This leaves some gray areas open to subjectivity when essays are partially off mode, and it also results in a rating that is based on a smaller sample of writing. Scoring the entire text of the essay provides a more accurate estimate of the writer's abilities in organization, voice, and conventions in general--even if it does not give an accurate picture of the writer's skills as applied to a particular purpose. Perhaps more importantly, since it is common to test students' writing with only one or two prompts (modes), and since human beings appear to have a natural tendency to stretch the limits of generalization (at least when it's convenient or profitable), trait ratings that are not bound to the purpose of the prompt may be more suited to the kinds of generalizations that people are likely to make about them.

Criteria for designating an essay "unscoreable" (U) differ somewhat for the four trait scales. Any essay that is completely illegible, blank, or written in a foreign language is considered unscoreable on all four traits. In general, raters are directed to score an essay if at all possible. Thus a partially legible essay might be scoreable on everything except conventions, and very brief papers might receive a U on voice or organization. In addition, some prompts elicited responses among the scale construction sample that could not be scored on organization because the sequence of ideas was unjudgeable. For example, in describing a room, the writer might list many objects in the room. In some cases it is impossible to tell whether the writer is describing the objects in order of location, in order of importance, or simply in no particular order at all. If there is no evidence of a planned sequence yet no evidence of *no* plan, it is not possible to make a fair judgment regarding that student's organization skills.

Ideas/Content

The mode-dependent nature of the ideas/content scales resulted in stricter rules regarding which papers can be scored on that trait. Essays that are off-mode are also off-purpose; the writer would have to receive a low score because of the irrelevance of the essay content to the prompt purpose. Yet if the writer is writing to a different purpose, this low score may not accurately reflect her or his ability to produce relevant ideas. Off-mode essays are therefore considered unscoreable. In addition, it was decided to consider a paper unscoreable if it was clearly off-topic (not a response to the prompt). It would be possible for a student to rehearse and memorize a story, an exposition, or a description with which to respond to any prompt in that category. This would not be a sample of that student's ability to write "on demand" with a time limit. By considering off-topic essays unscoreable, we avoid scoring most such prepackaged responses. Protocols for ideas/content require consensus with the chief rater in determining whether an essay is sufficiently off-topic that it should not be scored on ideas/content.

Guidelines requiring the designation of an essay as unscoreable on ideas/content are somewhat more lenient for narratives than for other modes of discourse. A story may be inspired by the prompt yet have little to do with the plot recipe suggested there. (Television mystery fans may recall how Ray Bradbury picks an object from his cluttered office and builds a story around it. The object inspires the story but is often far from the center of it.) In fact, the prompts themselves do not require the student to stick closely to the prompt topic. (E.g., "Use these pictures to help you write a story.") This gives creative writers the opportunity to produce their own plot without being bound to a suggested story that may fail to capture their interest and bring out their abilities.

Unfortunately, students occasionally respond to narrative prompts with stories that are not their own. Familiar campfire tales, fairy tales, books, and contemporary films (especially teen-appeal movies that adults are not likely to watch) are some sources of these stolen plots. In initial drafts of the ideas/content protocols, rather than marking such essays unscoreable, raters were instructed to penalize students on the ideas/content scale when plagiarism was recognized. However, this was found to introduce too much complexity; raters had difficulty determining whether or not a story with a familiar ring was in fact a borrowed plot. Because only a small percentage of sample essays (less than 1%) exhibited this problem and because the ability to recognize a stolen plot depends largely on the rater's personal experience, protocols were revised so that raters are encouraged to go ahead and score all essays if possible, without penalty. In the case of any uncertainty, they are instructed to consult with the chief rater.

Incorrect "facts," inconsistencies, and logical flaws occasionally appeared in the sample essays. Theoretically, such flaws could be said to mar the quality of content. However, they actually appear to reflect a lack of knowledge rather than a lack of skill in generating relevant ideas. Given the timed nature of the assessment, it is not possible for writers to check their facts. In the context of a portfolio assessment, where samples of writing have been collected over time and opportunity for revision has been provided, an ideas/content scale might include attention to accuracy and logical consistency of ideas. Such attention was not deemed appropriate here.

Because of the different purposes of each mode of discourse, aspects of writing that are important to the ideas/content scales were found to differ from mode to mode among the sample papers. Relevance of ideas and adequate supporting detail contribute to this trait for all modes. For narratives, however, the object is to tell a story; originality and strong development of ideas are emphasized. A description seeks to create an image or atmosphere; therefore vividness of content is an important quality. Persuasive essays require ideas that are forceful, plausible, and potentially appealing to the intended audience, while clarity is emphasized in judging the content of expository papers. These differences were taken into account in describing characteristics of essays that exemplify each score point on the scales for ideas/content.

Organization

Similarly, students' essays were distinguished by different aspects of organization in different modes. For narratives, the mode in which children tend to develop skills first, both the sequence and flow (manner in which the writer relates ideas to one another) were found to be important to the development of a successful story. Sequence and general structure were more important than flow among descriptive essays. For persuasive essays, sequence and flow were largely superseded in emphasis by the grouping of ideas and the presence of a conclusion. Among the responses to some expository prompts, an effective sequence was vital to good organization, and simplicity of transitions was more likely to raise an expository organization score than lower it. (Usually complex transitional phrases enhance the flow of the text. But when the goal is a clear explanation, the deliberate use of simple transitions may signal strong skills in organizing ideas effectively.) In general, the organization scales therefore emphasize the sequence and/or grouping of ideas, with the smoothness and sophistication of flow (transitions) often making the distinction between papers earning a 3 or a 4.

Many sample essays that could not be scored on ideas/content were scoreable on organization. However, additional papers considered unscoreable for organization include those composed entirely of unjudgeable sequences, mentioned above, and those that are simply too brief to judge the writer's organization skills. Organization protocols suggest that a minimum of three to four lines of text, or two to three sentences, is required to make a judgment.

Efforts were made to address the problem of unjudgeable sequences in the protocols for organization, but initial efforts to score the entire standardization sample indicated that raters sometimes had difficulty with the increased complexity of scoring that resulted. Therefore, raters are instructed to assign a low score if the organization of an essay is not effective and to assign a higher score if the order, sequence, and flow can be easily followed. In the case of uncertainty, training materials direct raters to consult with the chief rater.

Voice

The voice scales were perhaps the most difficult to develop, and it was suspected during scale construction that they would be the least reliable of the trait scales. (This turned out to be true for persuasive and expository essays [Witt, 1993].) There was not much of a problem with mode-to-mode differences in the aspects of voice that distinguished between essays. However, the styles in which students chose to express themselves differed from mode to mode, as did the distribution of performance on the voice scale among the students in the scale construction sample. In writing narratives (whether in response to a narrative prompt or to a prompt in some other mode), students seemed to be more comfortable (and skillful) about revealing their involvement and individuality in their style of writing. Outside of a narrative setting, however, most students appeared to have little command over the expression of a personal voice. The majority followed one of two options: either self-expression without purposeful control, or enough control to employ the style they may have perceived as expected but without individual, personal expression.

Third and fourth graders were especially weak in voice skills. A certain level of skill in vocabulary, sentence structure, and the like is necessary before the writer can effectively exercise a personal voice. Children at this age are still struggling with the mechanics of production; it is questionable whether the measurement of voice at this level is even meaningful, especially outside the context of the narrative mode of discourse.

Elementary and junior high students are often quite adept at including humorous allusions in their writing. These are aspects of voice which, if possible, should not be overlooked. However, the egocentric nature of youth may cause the young writer to overlook the fact that the essay's audience (the raters) may not understand the allusion. It's a private joke to which the raters may not be privy. The classroom teacher may actually be a better judge of voice than the stranger-rater in this case. Simply by knowing the student (and the values of the student subculture), the teacher may catch important sparks of individual expression that scoring companies will miss. Fortunately, it is not likely that such allusions will occur frequently enough--or make great enough changes in voice ratings--to be a major concern. Yet, for the voice scales, this is a weakness of the scoring process that cannot be averted by carefully written protocols.

Nearly all legible sample essays were scoreable on the voice scales. A few, however, were too brief to measure the writer's skills on this trait. As with the organization scales, protocols suggest a minimum of three to four lines or two to three sentences before an essay can be scored on voice.

Conventions

In some respects, the scoring guides for conventions are more vague and open to subjectivity than the other trait scales. Raters are asked to consider the distractiveness of errors, based on the quantity and seriousness of departures from convention. However, language conventions have been heavily and explicitly addressed in the classroom. Although some may disagree regarding specific microskills that ought to be included in a conventions scale, the trait is relatively well understood and conventions scales have generally produced reliable ratings (Witt, 1993).

The process of scoring conventions was simplified somewhat by deliberately downplaying the role played by the complexity or difficulty of the language and mechanics employed. Although errors in basic skills and simple words are considered more serious than errors in more complex language, writers are not necessarily rewarded for attempting to employ higher-level language nor penalized for keeping it simple. Thus students are rated (and ranked) on the basis of the accuracy and correctness of whatever conventions they choose to employ. A three-sentence essay with a simple structure may receive a high conventions score if it is substantially correct. At the same time, a relatively low score might be earned by a writer who has mastered the mechanics of simple sentence structure but incorrectly uses some more complex skills. While this may seem to reward some students for not taking risks, it also avoids contaminating the conventions scales with more general language production abilities (which are likely to show up in ideas or voice). Raters are further exhorted to rely on a general impression of the overall distractiveness of mistakes, rather than making a count of errors.

Aspects of conventions important to measuring the trait differed little from mode to mode among the sample essays, although quotations were more likely to be employed (correctly or not) in narrative essays. Problems related to unscoreability were usually the result of poor legibility. For example, poor handwriting, a light pencil, or a weak photocopy might obscure spelling and punctuation while leaving some words decipherable enough to score ideas, organization, and voice.

Prompts

Prompts were designed to give directions as clearly and explicitly as possible. Nevertheless, many of the students in the sample did not follow directions. Not a few essays were off mode or off topic, at least to some degree. Especially among the third and fourth graders, some papers read as if the writer had only seen the first paragraph of the prompt. ("Look at the pictures. Who is in the pictures? What is happening?") The writer responded by answering the questions rather than thinking about the answers and then following the remaining instructions, which describe the actual writing assignment.

Occasionally the writer seemed not to understand what the prompt called for. When students were asked to choose their own subtopic, some listed a series of possible choices rather than making a selection and building a description, an explanation, or a persuasive essay around it. When several pictures suggested possible choices, some writers seemed to think they were required to choose one of those shown (and perhaps defend it against the other possibilities pictured). Some students seemed compelled to cover every detail shown in the pictures. Some tried to artificially wedge in every bit of information suggested by the prompt (e.g., using all five senses in a description when two or three would do, or taking pains to include the words, "first," "next," "finally," even though they did not fit the context). Many students seemed to have trouble interpreting directions such as "Use details . . ." or "Explain your reasons." And of course, some students responded as if they had never seen the prompt; they simply did their own thing.

Many of these aberrant responses simply reflect the ability (or willingness) of the students to follow directions or to read the prompts and comprehend what is expected. Others, however, could possibly be avoided by modifying the prompts. For example, prompts at the lowest level could be even more explicit: "First think about this . . . Now write a story about . . ." Pilot testing and subsequent modification of prompts was not feasible from a practical perspective in this case. However, it is highly recommended that a piloting step be built into the process of developing the next form of the *Iowa Writing Assessment* in order to minimize the occurrence of aberrant responses and bring out the students' best efforts. Although the prompts for this assessment were developed with attention to content, wording, and format, some seemed to elicit unintended response characteristics. One can never be sure how a prompt will function until it is field tested.

Prompts that do not require students to choose a subtopic might also be piloted and the responses compared with those from prompts that do require a choice. The latter, open prompts, may have the advantage of freeing some students from the burden of writing on a topic of no interest to them, thus allowing greater personal involvement in the writing and drawing out their best performance. On the other hand, the burden of choosing may limit the time and thought that some students put into the substance of the essay, thus inhibiting their best performance. Also, since some students eliminate the choice issue by refusing to choose, their essays do not reflect the intention of the prompt a.d are therefore harder to score. Speculation suggests that the most effective prompt may be one that requires no choice but focuses on a very carefully selected topic, one that is likely to be of interest to nearly all examinees. Another solution may be to develop prompts that *allow* the writer to choose a subtopic, but do not *require* a choice. For any new prompt style, however, pilot testing should be considered essential; there is no way of knowing what peculiarities of response a prompt may elicit without a tryout.

If analytic scoring is to be provided for the next assessment, prompts should also be developed with the analytic scales in mind. Some of the standardization prompts for the 1994 *Iowa Writing Assessment* are fairly explicit in suggesting an organization, reminding students to include details, etc. The explicitness of tie prompts must be considered in interpreting analytic scores. Some prompts tended to elicit essays that showed little variability on some traits (e.g., Level 9-10 on voice) or that were simply hard to score (e.g., unjudgeable sequences on organization). For some prompts, high performance on a given trait was not essential to the purpose and was therefore seldom elicited (at least among the sample responses) by the prompt. (See the discussion of the Level 13-14 expository prompts, below.) If analytic scoring is to be retained, prompts should be developed that will elicit responses showing a range of ability on each of the traits of interest; otherwise some trait scales may not be very meaningful.

Prompt-Specific Issues

Narrative

Level 9-10. Among the essays selected for the scale construction sample, the Level 9-10 narratives drew quite a few responses in which the writers appeared to have read only the first part of the prompt. They simply stated who, what, and where were shown in the pictures, then stopped. Some simple modifications to the prompts might minimize the occurrence of this type of response.

The Race prompt (See Appendix A.) generated some very dull and predictable stories, and very few of the sample essays on this topic received 4's on any of the trait scales. (Otherwise, the distribution of scores on each trait scale was similar for both prompts.)

Level 11-12. The Level 11-12 narratives of the scale construction sample elicited the same types of responses as Level 9-10, although performance was better on all traits. At this level, a fair number of humorous allusions appeared in the stories, often as references to rock stars. Logical inconsistencies and incorrect ideas also appeared, but never in such a way as to interfere with the development of the stories. A few papers, for whatever reason, were lacking a beginning or an end. These papers were scoreable on all traits on the basis of the text present. (Such papers would probably be designated unscoreable on the focused holistic scale.) There was more plot and character development than at the lower level. Since both ideas and organization are vital to plot and character development, many of the students in the scale construction sample received similar scores on the ideas/content and organization scales, suggesting that a high correlation between these two traits might be expected for narrative essays at this level and higher levels when the full standardization sample is scored. (This expectation proved true; disattenuated correlations were all perfect [Witt, 1993].) Organization skills appeared to play a stronger role in plot development among responses to the Suitcase prompt than to the Carnival prompt. (The kinds of stories related tended to be more dependent on sequence.)

The Carnival prompt tended to generate a lot of responses that might be titled, "What We (They) Did at the Fair"--no real plot, just a description of the activities of a family or some other group visiting a carnival. Some students entirely missed the salient feature of the prompt picture: one child was too short for some rides. Others seemed compelled to include every detail shown in the picture, even if they had to bend and stretch the story to make everything fit. (Similar problems occurred with the Suitcase prompt, but far less frequently.) These kinds of responses again raise the question of whether relatively open prompts harm the performance of as many students as they help. Some students did, in fact, produce very good stories by taking advantage of the leeway allowed by the open prompt. Others, however, appear to have been uncertain about what was required of them.

Level 13-14. The style of narrative prompts at Level 13-14 was somewhat unusual in that newspaper headlines were used in place of pictorial events. In response, many students wrote news articles rather than stories. This was not the intention of the prompt, and it made scale construction a bit more difficult, particularly in describing characteristics of papers typical of each score point. Narratives in the news article style tended to include details that would otherwise be considered irrelevant; they also tended to be concise, without much complexity in plot and character development. Nevertheless, such papers were scoreable, and many students used the style quite effectively.

These prompts left little opportunity for students to respond with plotless narratives. In comparison with the lower level narratives, plot development was less important and originality became more important in distinguishing among essays on the ideas/content scale. At the same time, the prompts were more open-ended than the prompts at the lower levels; a greater variety of stories was generated, and the prompts appeared to invite better stories with greater writer involvement. Undoubtedly the greater maturity of the students is involved here as well. (Nearly all students in the scale construction sample at least seemed to understand the assignment.) However, the sample responses suggest that good writer-topic relationships were at least partly a function of prompt style. Because prompt styles differ from level to level (and mode to mode) and because different prompt styles may elicit different types of responses, many of the differences in characteristics that distinguish between essays within level and mode--and that inspire differences in scoring protocols across levels and modes--may be due as much to *prompt style differences* as they are to actual differences in age and mode of discourse.

It was at this level that most of the narratives with stolen plots appeared. In fact, nearly all were among the responses to the Discovery prompt, the more open-ended of the pair. This may suggest that some students had difficulty dealing with the open-ended item; unable to produce an original plot on demand in the time allotted, they chose to write up a plot they already knew.

Performance of the scale construction sample on all four trait scales was notably higher for Level 13-14 narratives than for any other pair of prompts. Most students seem to have developed some ability in all four components and are able to apply their skills in producing narrative essays by grades 7 and 8.

Descriptive

Level 9-10. For Level 9-10 descriptive prompts, differences among sample essays on the ideas/content and organization scales were difficult to describe. Distinctions on voice could be made only by closely attending to the writer's attempt to exercise some control of expression in order to appeal to the audience. Individuality and involvement were otherwise hard to judge. The prompts seemed to invite a flat listing of objects or activities, making it extremely difficult to build scales for organization and voice. Again, this may have been a problem of students not discerning what was expected. Whether modifying the prompts would help is uncertain. If students do not have experience with descriptive writing, they can be expected to have trouble understanding the assignment.

Again, some students appeared to have read only the first paragraph. For the Special Room prompt, they simply listed favorite things and/or stated what they did in the room. For Summer Day they simply answered the first set of questions. Other students made a point of forcing the inclusion of all five senses--some rather successfully, but others with a ring of artificiality that actually detracted from the effectiveness of the description. Quite a few papers were partially off mode, nearly all taking a narrative form. Many of these nevertheless succeeded in presenting a vivid description (leading to a high score on ideas/content), and all of them may well have received higher scores on voice and organization as a result of choosing a more familiar format. In general, compared with narrative essays at the same level, scores on all traits but conventions were dismally low (although Summer Day elicited reasonably high scores on voice).

Summer Day seemed to inspire better descriptive essays than Special Room. Special Room elicited too many lists, often in unjudgeable sequences and expressed in a flat voice. Summer Day was not without its own problems, though. The prompt is very open-ended about what should be described--a place, a schedule of activities, the weather. This may have helped some students and hindered others. Many students apparently responded by falling back on a familiar assignment: write an essay on "What I Did on My Summer Vacation." Although scoreable on all four traits, many such responses were scored low on ideas/content because of some degree of mode-departure.

Level 11-12. The ideas/content scale for Level 11-12 descriptions was very difficult to develop because students tended to respond in unusual ways, especially with the Special Possession prompt. Both prompts seem to demand a physical description, and Special Possession explicitly asks for feelings as well. The types of ideas that lead to an effective description (one that creates an image or atmosphere) differ depending on the topic chosen. Does anyone care how a baseball card collection looks, feels, and smells? More important are the writer's feelings about the collection and relationship to the collection; these things may be conveyed by ideas other than a physical description and an explicit statement of feelings. Some students apparently knew this and produced good descriptions by *not* precisely adhering to the requirements of the prompt. Others attempted a physical description where it was not particularly appropriate. Still others chose a topic that seemed to require both physical description and feelings, but fell short on providing one of the two. For example, one child gave an excellent description of her teddy bear's physical appearance, but included not a word about her feelings, not even where the teddy bear came from or how it acquired its rips and tears. (Some of the best papers conveyed feelings implicitly by describing settings or events related to the object.)

Some ultimately very good descriptions had focus problems. The writer began to tell about some object, then ended up concentrating on a description of something related instead. For example, one student began to describe a gift from his grandfather, then gave equal space and effort to describing his grandfather. Another, intending to describe his pet snake, devoted most of his text to describing snakes in general (and he did it very well). Still another tried to describe books, apparently found the topic limited what she could say, and ended up with a well-written essay on why she loves to read. One writer responding to Show Visitor started with a single object and ended up describing the whole town. The sheer variety of unusual student responses and the unique demands of the topic chosen made scale construction difficult for these prompts on ideas/content. (In fact, ideas and content did not play a big role in scoring descriptive prompts on the focused holistic scale in the previous assessment. Perhaps the above examples illustrate the reasons why.) Overall, the Level 11-12 sample descriptive essays were interesting to read, but difficult to score analytically, especially on ideas/content. Other than directing raters to consider whether supporting details and elaboration are sufficient and appropriate for the topic chosen, little can be done in composing protocols to account for the effects of topic choice. In responding to these prompts, some students undoubtedly will penalize themselves by selecting a topic that is inherently difficult to describe according to the demands of the prompt.

The organization scale was somewhat easier to construct. Unlike the Level 9-10 descriptions, sample essays at Level 11-12 often revealed sequence problems. There were fewer unjudgeable sequences, and the scaling of organization skills was less dependent on flow.

Voice was not nearly as difficult to score as at the lower level (responses were more variable), but it was difficult to describe verbally. For both prompts, many 1's and 3's sound very similar on the first reading. It is difficult to describe what makes a writer sound "interested" and "involved" when the typical style is a straightforward presentation of facts. The voice protocol for this pair of prompts suggests that the difference may be found in the presence of a personal point of view and/or more creativity and appeal among the better papers; sample essays provided with the training materials help to clarify the difference.

Curiously, Special Possession did not seem to elicit strong expressions of personal style. More students in the scale construction sample received higher scores on both organization and voice in response to Show Visitor than to Special Possession.

Level 13-14. Descriptive essays at this level created problems for the voice scale similar to the problems at the lower levels. Alone Place did not generate any particular peculiarities, but Costume tended to elicit lists, often in unjudgeable sequences, creating problems for the organization scale. For Costume, attempts to group or sequence ideas were generally mild, half-hearted--and understandably so, since they were often unnecessary. A costume need not be described in order (say from head to toe) in order to create a vivid image. A low score on organization may only mean that the student did not choose to exercise a skill that was not particularly necessary. Of course, if the student was responding to Alone Place, rather than Costume, a low organization score might be more directly meaningful. (Yet even with Alone Place, some essays with unjudgeable sequences were found.)

Overall, Costume essays tended to score lower than Alone Place papers, especially on organization and voice. Costume also provided less opportunity to describe feelings (although the prompt explicitly requests this) and to create an atmosphere. In addition, it elicited more unscoreable responses, most of which were narratives. Of the two prompts, Costume is clearly less suitable for analytic scoring.

Persuasive

Level 9-10. Young students seem to find persuasive writing at least as difficult as descriptive. In general, the · are not bad at generating ideas, but they are weak in providing supporting arguments. Among the sample essays, emotional involvement with the issue sometimes appeared to guide the content, at least in responses to the Cafeteria prompt. Once they began to respond to this personally relevant topic, some writers became immersed in their feelings about cafeteria rules or the limited menu, and their essays became models of self-expression. The persuasive task and the intended audience (those with the power to change the rules or the menu) were forgotten as these writers expressed their frustration with the current policy and their fervent desire to (for example) sit with their best friends and eat pizza every day. It appears that for persuasive writing by young students, a high-interest prompt may actually interfere with task.

In addition to the usual Level 9-10 problems (responding to only the first paragraph, addressing every detail of the prompt, etc.) most students in the scale construction sample did not seem to understand the phrase, "Explain your reasons." (This was true for Levels 11-12 and 13-14 as well.) This phrase was apparently employed in an effort to explicitly ask for elaboration in terms the young students could understand, but it missed the mark. Writers tended to restate the same reasons in different words or give additional examples that were so similar to previous examples that they added nothing substantial to the argument. They seemed to understand that something extra was expected, but they did not know what. Some appeared to deliberately ignore the prompt's list of points to remember, as if they didn't know what to do with it. Quite a few responded with off-mode or otherwise unscoreable papers. Similar problems also occurred at Levels 11-12 and 13-14, but less frequently and to a lesser degree.

Many writers at this level (9-10) provided a list of choices (rules to change, places to go, etc.) rather than choosing and defending a single idea. This may have hindered the performance of some on the ideas/content scale, since they spent their time listing choices rather than elaborating the reasons for their choices. Raters are instructed in training not to penalize students directly for not making a single choice, but to consider the quality of support for the choices given. (Again, any questionable situation is resolved in consultation with the chief rater, who is familiar with the kinds of issues that can be expected to arise.) The significance or force of ideas (given the intended audience) receives greater emphasis on the ideas/content scale for persuasive essays than for other modes.

The construction of the voice scale for Level 9-10 persuasives was a painful task. There was almost no purposeful expression of personal style in many of the sample essays. Following some mental perspiration, distinctions were found (mainly in the authors' maturity of expression and attempts to appeal to the perceived audience), and these are described in the protocol. However, it was quite clear that the third and fourth graders in the scale construction sample showed little skill in the expression of a personal voice as well as little skill in composing persuasive essays. The combination of the two is something that probably should not be measured at this age level. (This sentiment extends to the measurement of voice in descriptive and expository modes as well.)

Level 11-12. Unscoreable essays occurred even more frequently in response to persuasive prompts at Level 11-12 than at Level 9-10. Many of the same problems of the lower level also appeared here; the effects of emotional involvement on content and voice were perhaps greater than at Level 9-10.

The particular audience also had a major effect on the writing. In some respects, this is good; the sample essays reflected a strong sense of audience awareness at Level 11-12. However, an awareness of the audience does not necessarily imply an understanding of what is likely to be effective with that audience. Weak and selfish reasons and whining, pleading voices filled the pages in response to the Privilege prompt. (It is hard to say whether these were inappropriate since the audience was a parent. Selfish reasons may appeal to parents because they care about the well-being of their children. Pleading and whining may be more effective with some parents than reasoning.) For the School Program prompt, voices leaned more toward a "rational appeal" style, but often the reasons given were based on the narcissistic assumption that what appeals to me appeals to everyone. School Program also occasionally elicited "parrot" voices: the writers sounded as if they were just repeating propaganda they had heard. This was particularly common with topics such as drug education and ecology. Such differences in responses to the two prompts, apparently resulting from the specification of two different audiences, made it relatively difficult to describe papers typical of each score point for the ideas/content and voice scales.

As with Level 9-10, students at Level 11-12 showed relatively little skill in composing persuasive essays on all trait scales except conventions.

Level 13-14. Sample persuasive essays at the junior high level also differed in the styles of voice elicited by the two prompts. Spend Money tended to draw out the "rational appeal" voices, while Change Rule generated a lot of complaining tones. In addition, Change Rule is very open-ended regarding the identification of the audience, and there was some variability in the audiences students selected. Again, differences between the prompts created some difficulties in composing the voice protocol. It was expected, however, that a protocol based on two such different manifestations of a trait would result in scores with greater generalizability.

Many responses to the 13-14 persuasive prompts appeared to dance around the persuasion, almost as if the students would rather write anything but a persuasive essay as outlined in the prompt. Some writers went into great detail about how the new rule would function or how the money would be appropriated. Both prompts elicited quite a few responses that described, and perhaps defended, several choices rather than just one; some students seemed to concentrate their efforts on generating as many ideas as possible rather than trying to persuade. Others, especially in response to Change Rule, expended themselves in complaining; their only effort to persuade was in describing how bad things were under the current system. Here, even more than at the lower levels, emotional involvement often seemed to guide the writing, especially in terms of content and voice. Responses unscoreable on ideas/content also appeared, usually as narratives. (Actually, a fair number of students at each level responded to persuasive essays with narratives that were in fact scoreable because some persuasive content was included.)

A student's performance on ideas/content in a persuasive essay may have been affected by the choice of topic. Some choices are simply harder to defend than others. This may especially be a problem with the Change Rule prompt at Level 13-14. Some of the school rules addressed by junior high students, if accurately described, appear to have been invented by neurotics of modest intelligence. What difference does it make, for example, in terms of educational outcomes, if a boy's hair extends over the top of his collar by a quarter-inch--an infraction for which he might be suspended? Or why are girls required to purchase a uniform dress in a specific style from a specific distributor, while boys may wear a white shirt and blue slacks with a great deal of leeway in style and cost? (Obviously, the latter example efers to a private school setting; such a rule would be illegal on several counts for a public school.) It is difficult to refute the reasons behind a policy when there do not appear to be any reasons behind it. Once the ludicrousness of the rule has been pointed out, there is little left to be said. Students who choose a rule that can be confronted with a point by point refutation of the opposing viewpoint will have an easier time of supporting their arguments than those who choose a rule that is simply unreasonable.

Expository

Level 9-10. Topic choice was also an issue for the ideas/content scale among expository essays at Level 9-10. The amount of supporting information required, as well as the amount available, depends very much on the specific game or chore the writer chooses to describe. The protocols for all expository prompts therefore encourage the rater to consider whether the support provided is sufficient to enable the reader to duplicate the procedure, join in the activity, or comprehend the concept that the writer is attempting to explain. There was no way to make allowances for the fact that some topics require more skill and effort than others to earn a high score on the ideas/content scale.

The usual kinds of aberrant responses occurred among the sample essays: some writers appeared to have read only part of the prompt, some tried to force the essay to fit the prompt's literal instructions, some responded with narratives, and some listed several choices instead of choosing one game or chore. Actually, for the Chore prompt, the listing of a number of chores was a legitimate response to the prompt, which states, "Explain . . . how to do your job." To some children, "*my job* is to take out the trash, feed the dog, keep my room clean, and stay out of trouble." Nevertheless, the listing of several choices presents special problems for the ideas/content scale regardless of the mode and level. Support tends to be limited in such papers, and focus may be ambiguous.

Choice of topic complicated the construction of the organization scale for 9-10 expositions in very much the same way it affected ideas/content. Some chores and some games have a natural sequence that imposes itself on the exposition; others do not. The former require less skill and effort to earn a high score on organization. They are also easier to rate, since it is very apparent when ideas are out of sequence. Chores and games without a natural sequence sometimes resulted in unjudgeable sequences among the sample essays. In responding to the Chore prompt, some writers imposed a sequence, instructing the reader to do things in a certain order or at a certain time. It is sometimes difficult to tell whether a writer is imposing a reasonable order on a task that has no inherent sequence (typically a 2 or 3) or merely spewing out ideas in a haphazard manner and attaching words like "first," "next," and "finally" because the prompt suggested them (typically a 1). In general, if time words or similar phrases are used and there is no evidence that they were inappropriately placed or pasted in, the writer is considered to have made a genuine attempt to follow a plan.

No major modifications were made to the prompts at this stage of the standardization process, but because some children apparently interpreted the instructions of some prompts as *requiring* the inclusion of specific words like "first" and "finally" whether appropriate or not, the suggestion to employ these words was removed from the expository prompts.

The role of transitions in the organization scale had to be adjusted for the expository mode at all three levels. Organization protocols place more emphasis on simple transitions for expository essays than for other modes. In general, complexity of transitional phrases contributes to a high organization score by enhancing the smoothness of flow from one idea to the next. This was true of some essays composed in response to expository prompts. Other writers, however, earned high scores by displaying a purposeful simplicity of transitions in order to enhance the clarity of the explanation. The difficulty for raters is in determining when simple transitions constitute purposeful organizational techniques and when they merely reflect a lack of sophistication. Sample papers used in training sessions help raters to develop skill in making this distinction.

The voice scale was particularly troublesome for Level 9-10 expository essays. Problems were similar to the problems with voice at this level for descriptive and persuasive essays, but were compounded by the fact that the expository prompts invited a common or conventional tone. It did not help much to concentrate on the purposeful control of voice in order to generate audience appeal; the audience was most often another child. Therefore the style of voice that aimed for audience appeal (typically a 3) mirrored the style generally considered immature, conventional, and lacking in individuality (typically a 2) in the other modes of discourse. Distinctions were found and are described in the protocol, but it was suspected that rater reliability would be difficult to achieve with this scale. (Interrater reliability was expected to be relatively low for all the voice scales, but especially for the 9-10 expository scale. Indeed, reliabilities turned out to be lower for voice than for the other trait scales for expository essays at all levels [Witt, 1993].) Furthermore, the common voice may be the kind of voice that best serves the purpose specified by the prompt. A low voice score would not necessarily suggest limited success in producing an expository essay.

Level 11-12. The How To expository prompt inspired essays that were more problematical for scale building than the School Procedure prompt did. How To resulted in more unscoreable essays and produced many papers describing chores and games and therefore exhibiting the same problems seen at Level 9-10. In addition, the problems described below occurred with greater frequency and severity among the sample's responses to the How To prompt.

For both Level 11-12 expository prompts, the choice of topic again affected performance on both ideas/content and organization. Some students chose procedures that were easy to explain (e.g., how to make Kool-Aid); others chose procedures that were inherently more difficult to describe (how to make a bed or twirl a baton). Some students encountered difficulties simply because they chose a relatively complex procedure. In addition, unlike Level 9-10, the prompts left open the possibility of explaining a very general procedure (e.g., how to cook, or the daily routine at our school). Although they are legitimate responses to the prompt, such general papers would rarely merit more than a 2 on ideas/content because clarifying details are lacking. In most cases, details could not be appropriately inserted, except perhaps by citing specific examples (a skill students at this age have apparently not yet mastered.) The same writers might have easily earned higher scores if they had chosen topics that invited greater detail.

This again suggests that open-ended prompts may create more problems than they solve; it might be better not to require students to choose, or at least to limit their choices.

The voice scale for Level 11-12 presented problems similar to those occurring with expository essays at the other levels. Most writers in the scale construction sample tended to adopt the style of an instructor, especially that of one child instructing another, with little individualistic expression. This type of voice is usually adequate to the expository purpose even though it typically merits a score of 2. There was relatively little variation in voice among the sample essays. The How To prompt, in particular, seemed to invite a flat tone that could reflect either a purposeful, straightforward style of expression or simple uninvolvement, depending on the rater's subjective judgment. The protocol addresses this issue by instructing raters to discern whether the voice "seems authentic, individual, and consistent." Again, sample essays used in training help raters to understand what this means in practice.

Level 13-14. The expository prompts for Level 13-14 constitute the only pair of standardization prompts that are not even remotely parallel. One asks the writers to explain a procedure; the other calls for explanation of an abstract concept. One specifies an unusual audience (a robot), while the other leaves the audience completely undetermined. In addition, mode-explicit instructions are completely different for the two prompts. The difference in prompt styles made it difficult to write common protocols without inserting examples of typical responses separately for each prompt. Indeed, the initial drafts of protocols for this group of essays illustrate the influence of prompt style in constructing scoring guides. While general descriptions of trait manifestations tend to be very similar across levels and even across modes, a certain level of specificity is necessary in order to produce instructions that will be clear enough and precise enough to create a high rate of agreement among raters. The style of the prompt may have even greater influence on the content and wording of the protocols than does the level or mode of discourse. When prompt styles are highly similar, it is relatively easy to compose scoring guides that are general enough to be used with any prompt that is fairly parallel (similar in format, degree of explicitness, and topic category) to those used in generating the protocol. To use such protocols in scoring prompts that are significantly different in style would almost certainly result in lower interrater reliabilities because raters would have to individually and subjectively resolve uncertainties in interpreting the protocols in a slightly different context.

(Ultimately, protocol construction problems arising from the differences in these two prompt styles were resolved when a single prompt was selected for the final form; specific instructions pertinent to the selected prompt style were retained. At the same time, the construction of the original draft of the protocols on the basis of two such different prompts helped to form a more general base for these scales; theoretically, this should enhance the potential generalizability of the scores.)

The Friend prompt, in calling for an explanation of an abstract concept, was different from all of the other expository prompts in the standardization. Of the two prompts at Level 13-14, Friend presented more problems for the analytic scales.

Sample essays responding to this prompt tended to use a lot of trite phrases ("never let you down," "be there for you"). Because of the nature of the topic, it was sometimes hard to tell when the writers were elaborating and when they were merely free-associating. The prompt specifically instructs the writer to "use examples," but students seemed to have difficulty understanding how to apply that instruction in this context. Some listed friends and mentioned what they liked about each; others related activities or listed traits they considered important to friendship; still others simply rephrased what they had already written. In general, specific examples illustrating general principles were rare. Most examples were either very general themselves or they were not used to clarify and illustrate but simply to extend the discussion. The Friend prompt seemed to require different skills--or at least different applications of skills--than the other expository prompts, and students were not always able to respond appropriately.

If the ideas/content protocol were written on the basis of the Friend prompt alone, the quality, originality, and force of ideas would receive greater emphasis. These are the main aspects of the trait that contribute to distinctions between essays, along with clarity and relevance (aspects that are important to the other ideas/contents scales for expository essays). For the scale construction sample, the Friend essays that truly excelled on ideas/content tended to be those that employed original examples.

If originality and force of ideas were emphasized, however, the protocol would not be suitable for rating responses to the Robot Chore prompt. Robot Chore requires clarity expressed through simplicity and thoroughness; originality is not important for a successful response to the prompt, nor did it serve to distinguish among essays in the scale construction sample. An emphasis on originality of ideas among the Robot Chore prompts would amount to measuring creativity (in the production of original ideas) more than the writer's ability to generate appropriate ideas and support them. For these reasons, originality was downplayed in composing the ideas/content protocol, although it is employed to distinguish between a 2 and a 3 for responses to the Friend prompt.

Another problem with the Robot Chore prompt is the specified audience: a robot. How much detail does a robot need to be able to perform a task? Different writers made different assumptions regarding the robot's "cognitive" capabilities. In rating the sample essays, it was often difficult to determine when there was too little detail--or too much. Some of the papers judged best on ideas/content included numerous, minute details that might be considered distracting in another expository context, but in response to Robot Chore prompt, they revealed an awareness of the audience and an attempt to provide supporting details at the level required for that audience.

Ideas/content was also difficult to score because, as with the expository prompts at Level 9-10, the amount of elaboration required also depends on the chore chosen. In addition, many writers did not seem to understand that the prompt called for a detailed, step-by-step explanation. Many responded with lists of chores or general descriptions of tasks. Problems with following directions were also apparent in the frequent failure to address the robot, as the prompt requests. Many writers, for example, wrote as if describing to a friend what their robot might do; some simply listed advantages of having a chore-performing robot in the house.

The two prompts elicited very different distributions of performance from the standardization sample on the organization scale (Witt, 1993). In fact, Friend essays tended to receive lower scores on ideas/content, organization, and voice, but the larger proportion of lower scores on organization, in particular, is notable. Responses to Robot Chore exhibited organization problems similar to those with expositions at Level 9-10. The Friend prompt produced some unjudgeable sequences, but primarily it seemed to elicit unplanned sequences of free-association; many Friend essays received a score of 1 on organization. Apparently this reflects the fact that a strong organization is not necessary for an effective explanation of the concept. Some of the lowest scoring papers on the organization scale were, in fact, well-written expositions that clearly conveyed the writer's understanding of the concept.

This is perhaps the best illustration of the importance of developing prompts with analytic scoring in mind if analytic scoring is to be used--or at least subjecting prompts to a pilot test. It seems ludicrous to score Friend essays on organization. A score of 1 may not reflect an inability to organize one's writing effectively. It may simply mean that the writer failed to employ a skill that was neither elicited by the prompt nor necessary for a well-written, mode-successful response. (Nevertheless, because of considerations related to focused holistic scoring, Friend was the prompt selected to elicit Level 13-14 expository writing in the final form of the assessment. The organization protocol was then "softened" so that this trait, relatively unimportant in responding to this prompt, is less stringently scored here than it is at other modes and levels.)

The construction of the voice scale for Level 13-14 expository prompts was hampered by the same problems present at the lower levels. Few writers in the scale construction sample scored a 4 on voice in response to either prompt. In addition, the differences in prompt style made it difficult to compose a common protocol. For example, a similar level of control over voice skills often produced a "chatty" tone in response to Friend and a "bossy" style in response to Robot Chore. Especially for Robot Chore, it was sometimes difficult to judge whether the tone was appropriate. (What is proper etiquette in addressing a robot?) Students chose a variety of styles. The protocol attempts to focus on whether the writer tried to purposefully employ *some* appropriate voice.

The Friend prompt was somewhat limited in variability on the voice scale, with many sample essays scoring only a 2. Again, this seems to be a function of this unusual prompt, which apparently invited an "easygoing" approach: the free-association of thoughts on friendship presented in a "chatty" style. It should be stressed that this does not imply that the Friend essays were poorly written--only that they were not particularly amenable to analytic scoring and, as a group, they exhibited features that were quite dissimilar from those of essays written in response to other expository prompts.

Summary

Obviously writing is a complex skill, even among those who are just beginning to write. Judging the quality of a sample of writing is perhaps even more complex. The importance of analytic component skills to the success of a particular writing task varies not only with the demands of the prompt, but also with the specific topic chosen by the writer. The level of difficulty in displaying component skills also may depend on the specifics of the content and instructions. The whole is indeed more than the sum of the parts (Spandel & Stiggins, 1990).

Some of the problems encountered in scoring the sample essays are undoubtedly unavoidable in a large-scale assessment. Inappropriate responses, such as unreadable essays, stolen plots, perfunctory responses (flatly listing answers to questions in the prompt), and off-topic essays may occur for a variety of reasons--including low motivation, anxiety, failure to understand the assignment, or difficulty in following directions--that are beyond the control of the assessment. However, the occurrence of inappropriate responses can be minimized if the instructions of the prompt are clear and explicit, the content is of high interest to most examinees, and topic choice does not become a burden. (It should be pointed out that the prompts for the *Iowa Writing Assessment* were designed to be explicit and interesting, and they appear to be effective; although a sufficient number of inappropriate responses occurred among sample essays to raise questions, the majority of responses were in fact on task and scoreable, and many writers appeared to find the task engaging.)

Difficulties in measuring voice are also hard to avoid, especially for the very young, and especially outside of a narrative setting. This is probably because most third and fourth graders are not yet sufficiently advanced, developmentally and/or educationally, to exercise this relatively abstract writing skill.

Some of the issues raised (including those related to voice) present problems only for analytic scoring and do not hamper the scoring of overall writing quality on the focused holistic scales. For example, the "unjudgeable sequence," which occurred rather frequently among sample responses to some prompts, rendered some otherwise good papers unscoreable on the organization scale. Some prompts invited responses for which a strong organization was not essential to the purpose; still others allowed a choice of topic affecting the difficulty of the organizational task. These issues could not be effectively addressed in the protocols without sacrificing objectivity. Many of these writers may receive low organization scores, not because they lack the relevant skills, but because the prompt and the topic require only minimal exercise of those skills. Conversely, others may receive a high score on organization, not because they've mastered organizing skills, but because they happened to choose a topic that virtually organized itself. Because of problems like these, unless the assessment is to be used for diagnostic purposes, focused holistic scoring may provide a more accurate picture of a student's writing proficiency. In comparison with analytic scoring, focused holistic scoring has a built-in flexibility that better enables raters to adjust the emphasis on component skills in order to account for factors specific to the prompt and the choice of topic.

It is not, however, the analytic scales themselves that create such problems, but the relationship between the trait scales and the prompts. The prompts used in the standardization of the *Iowa Writing Assessment* were originally created with only holistic scoring in mind. The analytic scales were not yet defined, but were constructed afterwards on the basis of the essays written in response to these prompts. If analytic scoring is to be employed in future forms of the assessment, many of the problems addressed here can be avoided by composing prompts that not only require appropriate exercise of all four component skills, but also aim to elicit variable performances on all four traits. Meanwhile, the meaning of analytic trait scores (and the validity of inferences made from these scores) can only be understood in light of the particular prompts and the kinds of responses they invite.

Perhaps the most controversial issue raised by the sample essays--and one that affects both holistic and analytic scoring--is the question of topic choice. The problems that arise when different students elect to write on different topics are most apparent when attempting to evaluate performance on individual traits; however, to the extent that topic choice affects performance on skills that are emphasized in focused holistic scoring, these problems are also present in judging the overall quality of the writing sample. Particularly among less mature and less capable writers, who do not yet understand how their selection of topic may affect the difficulty of the writing task, topic choice may have an adverse effect on fairness. Topic choice can result in different students essentially being evaluated on different items of unequal difficulty. Ideally, if each student chose a topic that would showcase her or his strengths at a level of difficulty commensurate with his or her level of proficiency, topic choice would tend to elicit "best performance" and ultimately enhance fairness. In fact, topic choice aims to encourage all students to exhibit maximum proficiency by not forcing anyone to write on a topic that is not of interest. Unfortunately, many students at these age levels are not skilled at choosing topics that are both interesting and appropriate in difficulty. Furthermore, the very act of choosing seems to be burdensome for some, and they respond in ways that are not likely to earn high scores (listing many choices, addressing all options suggested by the prompt, etc.). Thus in aiming to enhance fairness, topic choice may actually backfire by reducing the probability that the response will reflect the student's best performance. A possible solution might be to create prompts addressing a pre-selected, high-interest topic, but add a statement allowing students who have no interest in that topic to choose another, within certain parameters. Students for whom choice is a problem could then respond to the topic given, while anyone whose performance might be adversely affected by a negative interaction with the topic would have the opportunity to select more personally relevant subject matter. Whether this would be effective is a question that can only be answered empirically.

Pilot testing of tasks may contribute important information in any performance assessment and certainly should be done if a new prompt format is created. In seeking new ways to measure human behavior, it is probably wise to expect the unexpected, use a pilot to elicit the unexpected, and then take steps to account for its effects. The use of the standardization sample essays served as a semi-pilot study with regard to the analytic trait scales of the *Iowa Writing Assessment*. Although prompts could not be significantly modified after the data were collected, some of the issues that arose in examining the writing samples could be, and were, addressed in the scoring protocols and rater training materials for both the analytic and focused holistic scales. In spite of the fact that essay prompts were created without attention to the specific features of analytic scoring, most of the challenges encountered in trait scale construction appear to have been met effectively, since the analytic scales do not appear to suffer from a lack of reliability or validity resulting from scoring difficulties (Witt, 1993). Certain prompts, however, exhibited anomalies that suggest they may be less than optimal for use with one or more analytic trait scales. These issues, along with others related to focused holistic scoring, were taken into account in selecting prompts for the final form of the writing assessment.[1]

[1] The reasons for selecting particular prompts cannot be addressed in this paper because the author was not involved in this step in the development of the *IWA*. Information on prompt selection can be obtained from the Iowa Testing Programs, 334 Lindquist Center, Iowa City, IA 52242.

REFERENCES

Baker, E. L. (1994). Making performance assessment work: The road ahead. *Educational Leadership, 51(6)*, 58-62.

Cantor, N. K. (1986). *The reliability and validity of the Iowa Tests of Basic Skills writing supplement*. Unpublished master's thesis, The University of Iowa, Iowa City, IA.

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

Feldt, L. S., Forsyth, R. A., Ansley, T. N., & Alnot, S. D. (1994). *Iowa writing assessment.* (Levels 15-18). Chicago: Riverside.

Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. (1983). *Stanford writing assessment program guide*. San Antonio: Psychological Corporation.

Hieronymus, A. N., Hoover, H. D., Cantor, N. K., & Oberley, K. R. (1987a). *Handbook for focused holistic scoring: Writing*. Chicago: Riverside.

Hieronymus, A. N., Hoover, H. D., Cantor, N. K., & Oberley, K. R. (1987b). *Writing: Teacher's guide*. Chicago: Riverside.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1994). *Iowa writing assessment.* (Levels 9-14). Chicago: Riverside.

Ruth, L., & Murphy, S. (1988). *Designing writing tasks for the assessment of writing*. Norwood, NJ: Ablex.

Spandel, V., & Stiggins, R. J. (1990). *Creating writers: Linking assessment and writing instruction*. New York: Longman.

Witt, E. A. (1993). *The construction of an analytic score scale for the direct assessment of writing and an investigation of its reliability and validity*. Unpublished doctoral dissertation, The University of Iowa, Iowa City, IA.

APPENDIX A

PROMPT DESCRIPTIONS

Narrative

Level 9-10

The "**Package**" prompt presents a series of drawings showing a package being delivered to two children. The children begin to open the package, which is addressed, "To Pat." The final drawing frame holds a large question mark. Students are asked to consider what is happening in the pictures and write a story based on what they see. Explicit comments remind the writer to pay attention to details and order of events.

The "**Race**" prompt is very similar in format. The first picture shows three children standing before a line on a track. A man with a whistle and clipboard looks on. The next drawing is a close-up of a shoe falling off a foot. The final frame again holds a large question mark. The directions to the student are presented with exactly the same wording used for "Package".

Level 11-12

The text for the two prompts designed for fifth and sixth graders is very similar to that used in the lower level prompts. Again the directions explicitly remind the writer to attend to organization and details, and the wording is identical for the two prompts. At this level, however, prompts contain a single picture rather than a series of drawings. In "**Suitcase**" a child is shown packing (or possibly unpacking) a suitcase in the bedroom. A checklist lies on the bed. In "**Carnival**" a girl and boy stand facing a sign that displays a horizontal line defining a height requirement for certain rides. The boy is tall enough, but the girl is not. The background is fairly detailed, showing a number of rides and carnival activities. Both prompts ask the student to write a story about the picture.

Level 13-14

Directions to the student at Level 13-14 are nearly identical to those at the lower levels, including reminders to provide details and an appropriate organization, but at this level the writer is asked to use a newspaper clipping as a starting point for a story. In "**Discovery**" oniy the headline is presented ("Teen Makes Surprising Discovery"), while "**Youth and Dog**" includes the first sentence of an article reporting thc finding of a lost 12-year-old with a dog. These prompts are more open-ended than those designed for younger children.

Descriptive

Level 9-10

"**Special Room**" asks students to describe a real or imaginary room where they can do whatever they want. The accompanying illustration depicts a boy imagining various objects and activities in his room. "**Summer Day**" asks writers to describe summer as they experience it. Illustrations suggest ideas, showing a beach, a city street, a canoe on a river, etc. Again the directions in both prompts are very explicit, reminding writers to include details, describe feelings, and use all five senses if appropriate.

Level 11-12

"**Special Possession**" asks students to describe a possession th,y care about. Illustrations show various toys and other objects to help inspire ideas. "**Show Visitor**" requires writers to select something in their homc town that they would like to show to a visitor. Accompanying pictures suggest a variety of possibilities, from historical landmarks to objects of personal significance. Explicit instructions in both prompts are nearly identical to those for younger students.

Level 13-14

The "**Costume**" prompt suggests a hypothetical dress-up day at school. Students are asked to pretend they are someone else and to describe their costume (including how they are affected by wearing it). Illustrations suggest a variety of props and accessories. In "**Alone Place**" writers are told to describe a place where they would like to go when they want to be by themselves. Illustrations suggest that activities (e.g., on a bike) as well as stationary locations may be suitable topics. Explicit directions closely resemble those given at the lower levels.

Persuasive

Level 9-10

"**Cafeteria**" requires choosing something about the school lunch program that the writer would like to change and persuading a teacher or administrator to make the change. Ideas are suggested in the text and illustrated with cafeteria scenes. Explicit instructions remind the writer to state what change is desired, list a number of reasons, and elaborate on ("explain") the reasons given.

"**Field Trip**" is identical in format and wording, except that students are asked to convince the teacher to go along with their idea for a class field trip. Illustrations suggest some possible locations.

Level 11-12

Persuasive prompts at this level are very similar to those at Level 9-10. "**Privilege**" asks students to persuade someone in power to give permission to try some new activity. To inspire ideas, a number of sports, skill lessons, and recreational activities are presented in pictures and as part of the verbal instructions. "**School Program**" asks writers to decide what they would like to do for a special school-wide program and persuade their fellow students to go along with the idea. Again, illustrations and verbal suggestions inspire ideas. Both prompts instruct students to state their choice, include a number of reasons, and explain their reasons persuasively.

Level 13-14

Persuasive prompts for seventh and eighth graders are similar in format and wording to those designed for younger students. In "**Spend Money**" students are presented with a situation in which their class has a sum of money to spend for the benefit of the school or community. They are told to decide how the money should be spent and to convince a committee to spend it thus. Illustrations depict a variety of possible activities and objects worth a potential expenditure.

"**Change Rule**" asks writers to consider a school rule they would like to change and persuade someone in power to make the change. Apparently the authors expected students would have no trouble generating ideas; the prompt is accompanied by a sole illustration: a rule book. (No specific rules are legible.) Again, explicit directions instruct students to state their choice, give many reasons, and explain their reasons convincingly.

Expository

Level 9-10

In the "**Game**" prompt students are asked to explain a favorite game to someone who has never played it before. Directions explicitly instruct students to pay attention to clarity, completeness, and organization. Drawings suggest a variety of possible games.

The "**Chore**" prompt directs writers to select a job or chore and write instructions for a friend who will take over that chore temporarily. Explicit directions are the same as those used for "Game." A number of illustrations suggest chores that could be chosen.

Level 11-12

Students are asked to explain some school procedure (e.g., what to do in a fire drill) to a new student in the "**School Procedure**" prompt. Ideas are suggested verbally and pictorially. Explicit instructions remind students to plan before writing, strive for precision in wording, pay attention to organization, and check their work for completion and clarity.

"**How To**" asks writers to choose a task or skill and explain it to a novice. Explicit instructions are the same as those for "School Procedure," and ideas are suggested in the text as well as by drawings.

Level 13-14

"**Friend**" is the only expository prompt that asks students to explain a concept rather than a procedure. Writers are instructed to consider various aspects of friendship and to write about what it means to be a friend. Drawings depict pleasant-looking people of various ages and races. Writers are explicitly reminded to organize their essays, use examples, and strive for clarity.

"**Robot Chore**" is similar in content to the "Chore" prompt at Level 9-10. However, the audience is a robot rather than a friend. Explicit reminders are identical to those given in the Level 11-12 prompts; they focus on planning, precision, organization, clarity and completion. A single illustration depicts a robot with cleaning attachments.

APPENDIX B

SAMPLE PROTOCOLS

ANALYTIC SCORING PROTOCOL

IDEAS/CONTENT
Narrative
Level 9-10

In scoring ideas and content, you should focus mainly on the *quality, relevance, and support* of the ideas presented. The quality of a narrative is primarily dependent on the originality, creativity, and complexity of the ideas. Ideas are relevant if they contribute to the development of a story. The story is supported well if details are sufficient to develop plot and/or characters.

In rating writing on this scale, be careful to concentrate only on the content (ideas, information) presented. The development of a good narrative will also be dependent on the manner in which ideas are organized, and the reader's interest will be affected by the writer's style or voice. The contributions of organization and voice will be rated on separate trait scales. Many writers show similar levels of performance on different traits, but many do not.

U Unscorable.

The ideas/content scale is somewhat more dependent on the mode of discourse than are the other trait scales; therefore, a response that does not attempt to produce a narrative (story) must be designated unscorable (U) on the ideas/content scale. However, a narrative that tells a story seemingly unrelated to the topic of the prompt may be scored. (See score point 1.)

Papers may also be rated unscorable if they a) are blank, illegible, or not written in English or b) cannot even broadly be construed as an attempt to respond to the topic of the narrative prompt. (The chief evaluator should confirm responses that seem to fall into the latter category.) If a story is incomplete, rate it on the basis of the content that exists.

1 No real story or an extremely brief or unclear story.

Narratives rating a score of 1 in the I/C scale suffer from some very major weakness in development. They may be very brief, doing little more than describing what's happening in the prompt drawings. Some"1's" offer a flat listing of events with no development of plot or character. Some present a plot outline with very few or very vague supporting details while some responses may go off on a tangent, losing the original plot entirely.

2 A dull, trite story or inadequately developed plot.

A story is present, but it is not supported well. Details are few, rambling, unoriginal, or vague. 2 writers sometimes get caught up in recitations of tedious, unnecessary details. ("Pat got up and brushed her teeth. She got dressed. Then she ate breakfast. She had oatmeal and orange juice..."). The 2 paper often relates a series of events, but attempts very little character development; the reader has little reason to care about them. Because there is so little supporting information, the text may present an ordinary, predictable story.

3 A reasonably good, well-developed story.

The 3 writer succeeds in telling a fairly good story. The story has a main point with some fairly good development of characters and/or action. Support is sufficient; enough details are provided to make the plot clear and the story interesting; some originality in the ideas adds to the plot or character development. There may be some rambling, but overall the details are relevant to the plot.

4 An engaging or exceptional story.

The 4 writer succeeds in telling an interesting story. The narrative has a clear plot with a main point. Characters and plot are well-developed. Details support the action, provide a setting or background, and add interest. Stories may have a lot of action or conversation. Some papers may contain very original, imaginative, or entertaining ideas--perhaps a twist of plot or an unexpected event or particularly satisfying conclusion. Overall, 4 papers stand out from all the rest.

ANALYTIC SCORING PROTOCOL

ORGANIZATION
Description
Levels 11-12

In scoring organization skills, you should focus on the *grouping* of ideas and the overall flow of the text. In a well-organized descriptions, material is ordered logically, elaboration is grouped with its central idea, and the reader has no trouble following the writing as it moves from one image to another. The writer's fluency--word choices, sentence structure, and reference to antecedent ideas--contributes to the smooth flow of the text. Together, grouping and flow contribute to the overall clarity of the text: the smooth, clear manner in which ideas are related to build an overall description.

Be careful to concentrate only on the manner in which ideas are organized; do not judge the quality of the ideas themselves. Pay attention only to the contribution of organization to development. While personal voice is often expressed, in part, by the same words and phrases that create effective transitions, focus on the organizational purpose of creating a smooth, logical progression of ideas. Paragraphing skills should be ignored since they can be taken into account in the conventions scale.

U Unscorable.

Responses that were designated unscorable (U) for ideas/content can usually be scored for organization. The rule of thumb is, if you *can* score the paper using this protocol, do so.

Responses cannot be scored if they are blank, illegible, or not in English. In addition, a response is unscorable for organization if it is too brief to judge organization skills. In general, at least 3-4 lines of text or 2-3 sentences are required to make a judgment.

1 An extremely weak plan.

There is little evidence of a plan. Ideas are not grouped logically and there is no connection between them. Details may seem irrelevant or misplaced, jumping from one idea to another and back again, or the writer may repeat ideas for no apparent reason. The writer may begin to describe something, then wander completely off the point. A score of 1 may be also be given to responses with very little content.

2 A weak or ineffective plan.

Typically, because details are generally grouped with the points they are intended to support and there may be some sort of opening and/or closing statement, there is evidence of a plan in a 2 response. However, the organization is not particularly effective. There may be disorganized or misplaced material or an overall choppiness to the text. Transitions may be lacking or simplistic. The text may wander away from the descriptive purpose. Or the text may reasonably well organized but relatively brief.

3 A reasonably effective plan.

The text follows a plan and flows reasonably well; details are elaborated and grouped logically. Usually there is a fairly effective opening and closing. There may be minor flaws such as missing information or gaps in logic or intrusive and/or immature transitions, but generally the 3 paper displays clear organizational skills.

4 A very effective and/or creative plan.

The 4 description is very well-constructed and/or clearly planned, and the text reads easily and smoothly. Together, the construction and flow of the text enhance the clarity of the description. The description may be effectively economic, focusing only on very telling details. Relatively sophisticated use of language creates smooth, clear connections between ideas. Usually there is a strong and effective opening or closing or both.

ANALYTIC SCORING PROTOCOL

VOICE
Expository Writing
Levels 13-14

The presence of strong voice may be described as writing that is sincere, deliberate, purposeful, and individual. At its best, voice is the combination of style and tone that makes a piece of writing lively and engaging. The absence of voice, however, results in a flat tone that lacks purpose or energy. In scoring voice, you should concentrate on an overall impression of the writer's individuality, conviction, and naturalness of expression. It may help to consider how the paper would sound if read aloud.

A writer's personal voice or style can be created by a variety of techniques. It results from choices of vocabulary and phrasing, sentence structure and sequence, and an inclusion of a personal point of view. Voice may not necessarily reflect the genuine personality of the student since the writer may adopt a persona. However, the voice for that persona should seem authentic to the reader. Seventh and eighth graders may interject bits of humor and sarcasm into their writing these should be considered legitimate expressions of voice.

Be careful to consider voice only, however, and not an overall impression of the entire response. An interesting or unusual response may be related in a conventional voice while a very poor response may be written in a unique and natural style. Avoid assigning a low rating on the basis of poor usage; usage skills are covered in the conventions scale, and it is possible for writers to display discernible voice despite poor usage skills.

At this level few students have had the opportunity to develop the expression of a personal voice outside of the narrative setting. Do not be surprised if you find that few papers merit a high score. The expository prompt tends to elicit responses that adopt an "instructor" voice. The rater needs to judge if this voice seems authentic, individual, and consistent.

U Unscorable.

Responses that were designated unscorable (U) for ideas/content can usually be scored for voice. In general, if you *can* score the paper using this protocol, do so. Unscorable responses include those that are blank, illegible, or not in English. Responses that are too brief to judge voice should also be rated "U." (These may or may not be the same responses that are unscorable on one or more other traits.)

1 Little evidence of voice.

The writing is functional, perhaps informative, but the style is flat. Many 1's are characterized by simple, unvaried sentence structures, repeated over and over, often with the same subject in every sentence. Vocabulary is often limited and, coupled with unvarying sentence structure, this gives the writing a detached, monotonous tone. In some 1 papers, the writer may seem to be on the verge of more interesting expression, but is held back by limited abilities in language production.

2 A conventional or immature voice.

In 2 responses, the voice is ordinary and uninteresting. The writer may use a generic style that makes the writing sound as if it could have been written by any student. Common phrases, trite expressions, popular slang, and simple sentence structures are hallmarks of the 2 response. An occasional indication of personal point of view or originality may be present, but for the most part, the writing is bland and mundane.

3 An adequate, emerging, or sincere voice.

The 3 writer begins to sound like an individual, perhaps more mature, perhaps more focused on the topic. The response is characterized by sparks of originality, more creative or sophisticated use of vocabulary and sentence structure, and/or the inclusion of a personal point of view. Somehow we get a sense of the writer's personality.

4 An interesting, appealing voice.

There is something unusually natural, unique, or appealing about the style of a 4 that engages the reader's interest. Perhaps the writer employs a mature and rational style; perhaps his/her style of expression seems exceptionally natural and sincere; perhaps the writers' voice seems particularly purposeful and consistent. The 4 writer may use words, phrases, and sentence structures that are varied and original, insert a personal perspective, include amusing anecdotes, and/or employ a unique turn of phrase to add interest to his/her writing.

ANALYTIC SCORING PROTOCOL

CONVENTIONS
Persuasive Writing
Levels 11-12

The conventions scale measures writers' control of mechanics and language conventions (capitalization, punctuation, spelling), usage, word choice, syntax, and sentence structure. The number of errors and the seriousness of the errors together should be used to determine the rating. Aim for an overall impression: given the amount of text and the seriousness of the errors, how *distracting* are the errors?

Keep in mind as well that these responses are drafts; errors that appear to be oversights (especially errors clustered toward the end of a paper) should be considered less serious than those that are consistent and seem evidence of a lack of skill. Be careful not to let your rating be affected by poor handwriting, which can be very distracting in itself. Base your rating on genuine errors in language conventions.

Serious errors include highly distracting errors and errors in low-level skills such as:

- Lack of subject-verb agreement (he don't)
- Glaring grammar errors (she seen)
- Problems with case and order of pronouns (him and I)
- Failure to structure simple sentences properly: run-on sentences (complete thoughts run together with no punctuation or capitalization) and sentence fragments (incomplete sentences) used without apparent purpose
- Misspelling of easy words and/or misspellings that are neither close nor phonetic
- Failure to capitalize the first word of a sentence or common proper nouns and pronouns
- Missing or improper terminal punctuation
- Quotations not set off by quotation marks and/or not paragraphed appropriately
- Incorrect habits of speech such as the use of "ain't" or "gonna" unless they seem to be a purposeful expression of voice.

Errors made in connection with more complex language may be expected at the fourth and fifth grade level. The following might be considered elements of complex language:

- Spelling of uncommon words
- Complex syntax: compound and complex sentences, imperatives, subordinate clauses; higher-level punctuation and usage skills
- Complex grammar: more than simple past, present, and future verb tenses
- Paragraphing, unless it seriously interferes with the clarity of the text

U Unscorable.

Unscorable texts include those that are blank, illegible, not in English, or too brief to judge conventions. Some responses that are scorable on other traits may be unscorable here because they are marginally legible; the message may be readable, but the punctuation, spelling, etc. are uncertain.

1 Extremely limited skills

Errors are serious, pervasive, and associated with basic language conventions. Sentences are structured improperly; glaring grammar errors are frequent; common words are misspelled; capitalization is inappropriate; and/or punctuation is missing or incorrect, making a smooth reading of text difficult.

2 Limited skills

Errors in 2 papers may be frequent, serious, and glaring enough to interfere with a smooth reading. However, the writer demonstrates some ability to manage basic language conventions. The writer is able, though perhaps inconsistently, to structure and punctuate some simple sentences, spell some common words, and control some basic conventions.

3 Fair to good skills

Errors are numerous and serious, but do not interfere with reading. Spelling errors are usually phonetic; generally simple sentences are structured and punctuated properly; grammar errors are neither frequent nor glaring. Many of the errors in 3 papers are connected with more complex endeavors.

4 Very good to excellent skills

Errors are infrequent. Sentences are consistently structured and punctuated correctly; spelling is generally correct. Errors that do occur are primarily associated with complex endeavors. Occasional errors that occur in spelling of common words or basic punctuation may be considered, because of strong skills demonstrated in the rest of the paper, to be draft errors or oversights.

TABLE 1

Summary of Prompt Titles Used in the Standardization of the 1994 *Iowa Writing Assessment*

NARRATIVE

Prompt	Level	Title	Prompt	Level	Title
#1	9-10	Package*	#5	9-10	Race
	11-12	Suitcase*		11-12	Carnival
	13-14	Discovery*		13-14	Youth & Dog

DESCRIPTIVE

Prompt	Level	Title	Prompt	Level	Title
#2	9-10	Special Room*	#6	9-10	Summer Day
	11-12	Possession*		11-12	Show Visitor
	13-14	Costume		13-14	Alone Place*

PERSUASIVE

Prompt	Level	Title	Prompt	Level	Title
#3	9-10	Cafeteria	#7	9-10	Field Trip*
	11-12	Privilege		11-12	School Program*
	13-14	Spend Money*		13-14	Change Rule

EXPOSITORY

Prompt	Level	Title	Prompt	Level	Title
#4	9-10	Game	#8	9-10	Chore*
	11-12	School Procedure		11-12	How To*
	13-14	Friend*		13-14	Robot Chore

*Prompts marked with an asterisk were selected for inclusion in the final form of the assessment: issues raised in creating the focused holisitic scale, as well as those discovered in constructing the analytic scales, influenced the selection of prompts.