DOCUMENT RESUME

ED 294 187 CS 211 170

AUTHOR Baker, Eva L.
TITLE Time To Write: Report of the US-IEA Study of Written Composition.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY Department of Education, Washington, DC.
PUB DATE Sep 87
GRANT NIE-G-85-0006
NOTE 30p.; Invited presentation at the IEA General Assembly (New York City, NY, September 1987). Project also partially supported by the MacArthur Foundation. For related documents, see ED 271 762 and ED 286 194-195.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Standards; Comparative Analysis; *Cross Cultural Studies; Grade 6; Grade 10; Grade 12; Instructional Effectiveness; Secondary Education; *Writing (Composition); *Writing Evaluation; Writing Research
IDENTIFIERS *IEA Study of Written Composition; *International Assn Evaluation Educ Achievement

ABSTRACT

Examining the status of the United States National Study of the International Evaluation of Educational Achievement Study of Written Composition (IEA-SWC), this report discusses the results of SWC, which attempted to assess, nationally and internationally, the quality of student writing in schools by collecting and evaluating student compositions. After a discussion on the nature of the writing process, the report presents a brief overview of SWC, noting its focus on sixth, tenth, and precollegiate twelfth grades, and describing the range of writing tasks and topics collected. The report also examines challenges which confront the study, including fundamental doubts about the quality of writing assessment, the validity of comparative educational effects assessed through writing compositions, and topic selection control. The report concludes that the US students' compositions met or exceeded the standard for minimally competent writing on major discourse tasks (narratives, persuasive essays, and reflective essays). Four tables are included in the report, and 40 references are appended. (MM)

Time to Write:

Report of the US-IEA Study of Written Composition

Eva L. Baker

UCLA Center for the Study of Evaluation

Invited Presentation

The IEA General Assembly
Teachers College Columbia University
New York, September 1987

This work was partially supported by UCLA, the MacArthur Foundation, and the US Department of Education (NIE-G-85-0006). The opinions expressed herein do not necessarily reflect the position or policy of the aboved named, and no official endorsement by them should be inferred.

Time to Write: Partial Report of the US-IEA Study of Written Composition

Eva. L. Baker
UCLA Center for the Study of Evaluation

For today's session, I have been asked to describe the US National Study of the IEA Study of Written Composition (SWC). My major purpose, however, is less to present our own fascinating results than to describe and argue for the approach that the SWC used for assessing student performance. This approach attempted to capture complex cognitive performance in a valid, economically feasible way. Specifically, our study required students to write: Not to answer multiple choice questions related to aspects of writing, but actually to write complex, lengthy, and time consuming essay answers. Even today such data are not thought to fit neatly into the the quantitative approaches often used in large scale comparative assessment. Eight years ago, when we started the IEA-SWC, the technical and experiential base for such assessments was much weaker. IEA took a risk then. Now, by identifying some of the barriers and benefits of tackling the measurement of the complex performance of writing, I hope to encourage the IEA and its financial angels to support other studies that use constructed responses to address deep understanding by students, perhaps new, direction-setting studies of subject matter learning. For surely the trend now is to try to match the sophistication of our measurements with our increased understanding of deeper student expertise. When Ralph Tyler (1966) articulated his ideas for the US National Assessment of Educational Progress (NAEP), he suggested that a central NAEP goal should be to provide new models for the measurement of student performance. The IEA is in the central position to take on this leadership role for the international educational community.

In this presentation, I will try to give you a sense of the following: what the writing process is about, the characteristics of

the SWC, some insight into the challenges to evaluation the SWC confronted, how we responded to these challenges, a special status report on the US efforts, and the implications of our work for future IEA studies.

What is the Status of SWC?

We are still puzzling through our data analyses, both for the US study and for the international comparisons. However, well before its conclusion, we can declare our study a success. That the study happened at all and children around the world in unimaginably diverse settings wrote on the same topics is success indeed. The SWC has provided ideas, approaches, and techniques that are new and reportedly useful to every participating country. Better assessment of writing will occur and increasing instructional attention should follow. And, not surprisingly, our study has generated a host of methodological problems and opportunities that will enrich the research community in measurement, teaching, linguistics and language. Our major failure has been the lack of a euphonious acronym for the study. Stable financial support is a close second.

What is Writing About?

Some insight into the complexities of the writing process may be in order. In any writing, the author juggles a variety of requirements and desires in order to accomplish specific goals. My goals included providing enough information about the U.S. study to be responsive to the agenda, but no more than I believed the audience cared about, or for that matter, no more than the level that our data analysis could support.

My major rhetorical problem was this: I needed to provide depth on selected aspects of the study that supported my thesis, but not to sink the presentation with too much detail. I oscillated

between a role in which I would tell about and illustrate the intellectual basis of the decisions for each aspect of our work, and a role in which I would focus the audience's attention on my main issue: that studies similar to the SWC are in the international community's best interest. I experienced a classic conflict between writing as a knowledge-telling process and as a knowledge-transforming process (Bereiter and Scardamalia 1987).

In contrast to much student writing in school, I was writing with specific purposes and audiences in mind, and I was well informed (after eight years prewriting experience) with the topic. In addition, I had an extensive writing history, reasonably good control over syntax and diction, and I, for one, had always regarded my own excellent spelling as a marker of deep intelligence.

Consequently, I wrote and revised, talked intermittently with my colleagues about what I was saying, and required my spouse to listen as I read with expression one or another version. The juggling involved explicit requirements, conflict between rhetorical plans and my intermediate judgments of effectiveness, and, of course, the deadline, too soon come and gone. My activities created a writing product; assessing products like it was the problem for the SWC.

Critical Elements of the Study of Written Composition

Let me turn to the SWC and provide a brief overview of what the study is about. SWC attempted to assess, nationally and internationally, the quality of student writing in school. Our focus was the type of writing that normally occurred within courses in the mother tongue, for instance, in US English classes. But all participating countries agreed to the same writing assignments and scoring criteria.

We designed the study so that comparisons could be made for students in the equivalent of US sixth, tenth and precollegiate twelfth

grades. Every decision in this study was painful, and the selection of these grade-age ranges was no exception. Sixth grade students (or 11-12 year olds) were chosen because they represent the point at which elementary (and sometimes compulsory) education ends. Developmental analysts also suggest that this age is when it is possible for written facility to catch up to oral language development (Farr, 1985). We chose 10th grade (15 or 16 year olds) in "average" classes in order to capture another breakpoint in schooling, a second juncture where compulsory education may end in various countries. In the US, this 15-16 year-old range coincides with the timing of high school minimum competency tests. The twelfth grade, college-bound students were selected for study because they represent the best and most academically well prepared students. We wanted to assure that our comparisons were based not only on the average student but on the top group each system prepared. Our populations for study, then, were chosen to provide both cross-national and within country policy-relevant information.

In another significant and relentlessly discussed decision, we adopted a domain-referenced approach to the design of our writing measures. (Baker & Herman, 1983; Baker, 1982; Millman, 1980). This approach demanded that we carefully specify task and scoring domains. Other evaluators and researchers could then generate comparable tasks according to our specifications. The study could thereby be extended for more intensive measurement of the same writing tasks or for use in the design of measures for future longitudinal or cross-national comparisons. Domain-referenced approaches resulted in a clear and exportable framework including performance standards that allowed student writing to be interpreted both descriptively (how well they met our scoring criteria) and comparatively (who did the best).

The descriptive analyses could also provide the foundation for instructional diagnosis and improvement, or formative evaluation (Bloom, 1981), since explicit features of writing were measured. Targeting this use for SWC data and instrument design was an

enormous departure for a large scale assessment study, particularly eight years ago.

SWC students wrote on a range of tasks and topics that varied along a number of dimensions: from informal functional tasks, like writing a note to explain a missed meeting, to tasks which required greater student control of the writing process, such as a longer narrative. We also chose tasks to reflect common writing experiences in school and to reflect a model of writing consonant with language and with cognitively-based analyses (Vahapassi, 1982; Purves & Takala, 1982). Because school assignments often focus on different types of writing, we assessed common writing genre, including descriptive, persuasive, and reflective writing. In an exciting departure from current practice, we also required each student to write on three different tasks. Now separate aspects of the individuals' writing skill could be estimated: 1) their general competence across genre; 2) their facility in the subtasks of writing (such as organization); and, 3) the details of their performance in a particular task or genre, such as description. Ten different compulsory tasks were used, distributed across the three populations of students. Like compulsory figure skating, we also used a well defined approach to rating.

Written compositions have no right answers, and the quality of each and every essay must be judged. SWC used an extensively researched rating scheme (Quellmalz, Capell & Chou, 1982; Quellmalz, in press). This scale was developed with years of NIE support. Our system was an integration of rating schemes that focused on general judgments of writing quality and those that provided separate scores on features of the essay. Our scheme resulted in a score of the overall impression of a student's essay quality and three component scores for the writing subtasks of content, organization, and style. These components emphasized the students' use of language to achieve the goal of the paper. No universal scoring of syntax, handwriting, or spelling occurred. The scale ranged from excellent

(5) to poor (1), with details spelled out for each score point on the scale.

Because we employed the same scale for all grade levels, and because certain essay tasks were linked across these grade levels, we could now determine whether our tasks were sensitive to gross maturational and instructional effects attributed to years in school.

In the IEA tradition, the study collected additional information about student characteristics, instructional experiences, and attitudes from questionnaires. Teachers also completed questionnaires on details of their training, interests, instructional emphases and characteristics of their classes. Administrators provided data on the school and community setting. We hoped to use these results to condition statements about student performance, to support validity claims, and to provide a more refined basis for comparisons among countries.

Because we have student essays as our data, we and other scholars will be able to study them in depth and to attend to discourse features from a variety of disciplinary perspectives. We can study the facility students display in rhetorical choices and the extent to which they control their writing. We can study the content they choose as a way to understand the scope and depth of their ideas. We can also use these essays to make inferences about the components of high quality writing for teaching and measurement purposes. We have accumulated a special resource for continued research, for the improvement of practice, and for policy formation.

Challenges to the SWC

In my eight year exposure to IEA I have learned that no study is easy to develop and sustain. But special challenges confronted the SWC and continue to do so. Like secret fears and nightmares, telling them might make them go away. These challenges grew from

legitimate intellectual concerns and from less informed suspicions. Let me run through the main entries in this dark catalog and share briefly the SWC response to each.

Was Writing Assessment Good Enough Science?

Eight years ago, the consensus answer was "probably not." First and foremost, writing assessment was not thought to be a sufficiently scientific form of testing to use to estimate student performance. I hasten to add that this was not a perspective unique to IEA, but was shared by federal and state agencies at that time. (For instance, in 1979 we were told no more federal support in writing assessment would be available for the UCLA center, since writing assessment wasn't really a form of testing. We heard the same story when we first proposed US federal support for the SWC.) Times have clearly changed.

Much of the early skepticism about writing assessment may have come from its lack of a secure intellectual home, with a credible core of proponents. Until very recently, writing assessment really wasn't measurement-ish enough for psychometricians. Students wrote on typically one test item (the essay topic) and statistical routines dependent upon multiple items seemed not to apply. (Coffman, 1986, remains a notable exception.)

Nor were the researchers in writing instruction much help in providing support for writing assessment. They tended to be concerned with the development of writing skills (Farr, 1985), or the writing process, including planning and revising (Freedman, 1985; Gray & Myers, 1978). Learning theorists focused on the cognitive demands of various writing tasks, (Bereiter & Scardamalia, 1987; Scardamalia and Bereiter, 1986), the utility of cognitive models of process for explaining writing behavior (Hayes & Flowers, 1980), and, along with linguists, attended in depth to the analysis of writing created under glass by relatively few students (Whiteman, 1981;

Frederiksen & Dominic 1981; Clark & Florio, 1982; Staton, 1982; Langer, 1986). All of these activities have importance, but assessment of writing products is not a priority for them.

Even those who regarded themselves as members of the roving band of writing assessors contributed mightily and unfortunately to the doubts about the level of technical quality possible for writing assessment. Partly, it was a two-worlds problem. In the US we had scholars with good disciplinary roots (Lloyd-Jones, 1977; Cooper & Odell, 1977) entering into the measurement world and trying to adapt their unique views of language to requirements for large-scale assessment. They were complemented by measurement folks, with well developed quantitative habits, who were demonstrably short on patience when the discussion turned to the niceties of rhetoric. A few technical people tried to integrate measurement, disciplinary and instructional concerns (Quellmalz, in press; Applebee, 1981), but were skeptically regarded by everyone else*. The result of this mix of assessment interests and expertise was not reasoned compromise, but instead, vituperative, sometimes public conflict among specialists about the advantages and shortcomings of various scoring schemes (the "holistic wars"). I will shield you from the pain of recounting the details of these discussions. Their effect, however, was to undercut the credibility of <u>any</u> writing assessment approach. As noted earlier, SWC responded to these concerns by adopting an integration of holistic and analytic ratings.

Another source of uncertainty about the rigor of writing assessment centered on statistical issues. Much attention tended to concentrate on some of the folklore about reliability of ratings, and how to achieve it. Because writing assessment had been largely conducted outside the mainstream of measurement and was used principally for selection (into college) or placement (into levels of

* Because of my own ecumenical spirit, in these meetings I took to dropping the fact that I had been an undergraduate and graduate student in the English department; it wore thin.

college courses), standards of reliability were assumed to be the same for both individual measurement and group assessment. We believe the SWC approach lays this issue to rest and supports the use of different technical criteria for inferences applying to individuals and to groups.

Yet, at the time, the reliability problem created unnecessary anxiety about many of the wrong issues: how many raters would be needed for each paper, should disagreements in scores be adjudicated by a third rater, should the rating scheme deal with easy-to-measure characteristics of writing such as syntax and spelling in an effort to improve the "numbers." Resolving these questions conservatively would have clearly destroyed the economic feasibility of the SWC study (and other writing assessments as well) simply because every rater costs resources. More importantly, these problems tended to pull concern from validity toward simple consistency, although both were obviously necessary conditions for successful measurement.

Solutions ultimately adopted by SWC were of fairly recent vintage, and although they had been tried in statewide, local, and university assessments, they were anything but common practice (Quellmalz et al, 1980). These options emphasized:

- careful training of raters to a standard of agreement;
- the use of benchmark papers to assure that a standard of validity had been set in advance of the scoring of each task;
- the use of frequent check papers to assure that the raters continued to use the scale validly and did not succumb to fatigue and boredom;
- the use of many rather than only a few raters (expanding the training requirement but avoiding idiosyncratic interpretation of the rating scale);

- the emphasis on speed and accuracy of judgment to keep costs down.

Could Any Study Based on Language Products Yield Fair Comparisons?

Another challenge to SWC derived from fundamental doubts that the comparative effects of education could be assessed in any case through the use of the written composition study. This challenge separated into several distinct worries.

One concern fastened upon the integral relationship between language and culture. What would then be the implications of comparisons among countries? More pointedly, were inferences about language facility separable from inferences about culture? (Heath, 1981; Scribner & Cole, 1981; Cazden, 1986.) How would such comparisons be qualified? Many answers to these questions await the final international analyses, but partial controls were instituted by assuring that comparable standards and training occurred for the essay rating process, by making system input and process data available to condition results, and by taking considerable care in the identification of topics for writing tasks. Specifically, the SWC approach to topic selection provides some safeguards.

Topic Selection as a Control

In topic selection we confronted the archetypal IEA challenge. First, in order to be valid indicators of educational services, our tasks needed to represent the regular writing experience of students in school. But because writing is not a well-structured discipline, our study was considerably more subject to differences in instruction, within large countries like the US or among countries.

Take the issue of writing purpose. The kinds of assignments students receive in school range widely. Nations differ in the extent to which writing assignments are contextualized and functional. Some countries emphasize writing to a real audience for a real reason. Others conceive of writing assignments as exercises to demonstrate command of the formal requirements of written communication. The SWC team needed to make topic choices that would reflect both purposes and audiences plausible to students if we were to elicit their serious effort and valid data.

Now for the matter of topic content. Unlike most IEA efforts, our study was focused on general skills and thought processes, unanchored to specific school subject matters. Nonetheless, in order to write, students must have something to say. Command of a body of content is a necessary condition for adequate writing. Research in writing suggests that topic knowledge strongly influences performance (Quellmalz, Baker & Enright, 1980). Just try to write coherently on a topic about which you know very little. Yet, topic selection presented manifold problems.

How could we balance the requirement for comparability among individual students and countries against the vastly different experiences that students have with topics? Our study would certainly fail if the core measurement was thought be unfair. In the past, one solution had been to chose a topic where ignorance was generally distributed. That way no one would have special advantage. We could be fair but not valid. Although a common choice, we discarded this option.

Another alternative was to provide specific information to learners to assure they shared a minimal set of knowledge to draw upon for writing (Baker & Quellmalz, 1981). Decisions splintered into choices about the amount of the information that could be provided, the form in which the information would be available, the time required for its transmission, the prewriting instruction required, and again, whether standardized administration could be maintained.

Some authors contend that such provision of information changed, either dramatically or subtly, the nature of the task as experienced by students. We were mindful of their arguments that the communication of meaning required internalization and elaboration processes (Langer, op cit).

Questions considered in our deliberations were numerous and included the following. What if information were provided in an abbreviated fashion? Would students have enough opportunity to digest it and then be able to elaborate its meaning? What if a large amount of information were presented in a discursive, complex form? Such a stimulus might encourage deeper thought as well as more refined content selection and elaboration. Yet, we worried that long passages would change the task from one that was principally a writing task to one with a strong component of reading comprehension, thus reducing validity. Employing a time-consuming task would also mean either less time for other tasks or less likelihood of securing cooperation of schools for the study. If information were presented in an highly organized way, we feared that students would simply mirror the structure and style of the prompt (a phenomenon repeatedly reported by many language teachers). What about avoiding verbal prompts? Could we use pictures to convey information efficiently? How common was this form of writing prompt to students? We certainly wished to avoid practices that would be unfamiliar to students and that would introduce new sources of error into our analyses.

The topic knowledge problem gave us an opportunity to explore options in writing domains (Vahapassi, op cit). In the most constrained tasks, students were given a core set of content, such as a story to retell or a picture that mandated careful description of detail. Students were also asked to describe themselves, a task that required some selection and organization of detail. Other tasks provided less information, or gave a story title or a cue to write a letter for a particular purpose. Finally, one of the tasks required relatively abstract processes. Students needed to choose a topic that

they had some feelings (and presumably) some knowledge of, and write a persuasive document. The measurement agenda set by these decisions is relatively clear. If we can develop multiple instances for each form of task (such as five different pictures to describe), then we will be able to assess the utility of that approach unconfounded by the particular subject of the picture. Moreover, if we move, as I believe we are, to the systematic linkage of written composition and subject matter understanding, certain of these tasks will be especially useful.

The US Results

Let's turn to a rapid consideration of the US results. The data for the study were collected in early 1985 for all three populations (6th, 10th and precollegiate 12th grade students) in a nationally representative sample of public and private schools. The percentage of targeted classrooms providing student essays averaged 94%. This high response rate was attributed to the close cooperation we received from the Council of Chief State School Officers and the importance US teachers place on student writing. Because the study required a minimum of two class periods, we are all the more pleased by the strong level of participation we received.

*How Did the US Students Perform?**

Our judgment is that, on the whole, they did quite well. If we focus attention on the major discourse tasks, a narrative, a persuasive essay, and a reflective essay, the performance of US students exceeds our expectations. To provide some meaning for the findings, we will report the general impression scores, rather than

*We are reporting only partial results for the major writing tasks. A full report is under preparation and, subject to funding, will be available in the Fall of 1988.

deal with each of the component scores. The score range was five points. In the US, we defined the score of "3" to be competent writing without major problems; "4" represented a higher level of skilled performance.

How Many US Students Met or Exceeded the Standard of Competent Writing?

Table 1
Task 5: Narrative Essay
Percent Scoring Competent

	Minimally Competent ≥Score 3	Skilled ≥Score 4
Grade 6 (Population A)	69.2	22.7
Grade 10 (Population B)	89.9	46.3

For narrative writing, sampled at the 6th and 10th grades, 69% and almost 90% of the 6th and 10th graders met or exceeded the standard for minimally competent writing.

Table 2
Task 6: Persuasive Essay
Percent Scoring Competent

	Minimally Competent ≥ Score 3	Skilled ≥ Score 4
Grade 6 (Population A)	24.2	2.8
Grade 10 (Population B)	72.4	26.3
Grade 12 (Population C)	85.6	48.4

Persuasive writing was sampled across all three populations. Notice the progressive improvement of scores with age group.

Table 3
Task 7: Reflective Essay
Percent Scoring Competent

	Minimally Competent ≥ Score 3	Skilled ≥ Score 4
Grade 10 (Population B)	78.8	26.2
Grade 12 (Population C)	89.1	40.7

In reflective writing, perhaps the most conceptually difficult task, almost 79% of the 10th graders and 89% of the precollegiate 12th grade students exceeded our competence cut score.

These results are somewhat surprising, and if accurate, show that US students' writing competencies are happily better than we had thought. Our immediate problem is to understand the meaning of our findings and to estimate the confidence we can place in them. Without international data analyses, we are unable to peg US students in the international scale. How else can we explore the validity of our findings, or, put in a more probabalistic vein, where are our findings vulnerable to attack? Let's consider alternative targets.

Maybe Our Sample Was Peculiar.

Our sample was randomly drawn (probability-proportional-to-size) with guidance from the Center for Educational Statistics, and an attempt was made to sample both public and private schools. The final distribution of responses overrepresented private schools by 12 schools across the entire sample. However, the entire sample was then weighted to assure appropriate representation by school stratum (public/private), size (total enrollment), and numbers at each grade level. At the student level our sample of major essays involved about 1500 students each at 6th and 10th grade and about 1100 at 12th grade. As a basis of comparison, NAEP in 1984 obtained about 1500 essays for each grade level (Applebee, Langer, & Mullis, 1986a, 1986b). However, it is true we sampled many fewer schools than NAEP (for instance, the SWC sample had 192 schools at 10th grade level compared with 539 for NAEP 13-year-olds). Because we collected multiple writing from each student, any bias would be replicated across tasks. On the surface, we feel our sampling is not a likely culprit. We will continue our detective work, however.

Maybe Our Raters Were Lenient and the Rating Process Was Biased.

Without the international calibrations, we do not know for sure. But we believe the rater leniency explanation to be very unlikely. The rating scales used in the study were based on ten years of work at UCLA. Versions of these scales had been used repeatedly in large scale state assessments and local assessments of writing. Systematic procedures were instituted for training raters and for assuring the validity of their scores. The ratings were completed immediately after training and qualification by raters. There was no opportunity for protracted "forgetting" or for the scale to be redefined over time. Twice a day for the three and one-half day rating period, the raters completed prescored sets of ten papers. Their ratings were immediately checked to determine adherence to the international scoring scale (see Quellmalz, 1980). Table 4 reports the agreement (within one point) of the raters with prescored papers.

Table 4
Rater Agreement Levels
Percentage of Agreement with Prescored Check Papers

Essay Type	Descriptive	Narrative	Persuasive
Scoring Element			
Overall Impression	100	96	97
Content	100	95	97
Organization	100	96	95
Style/Tone	100	98	98

Twenty percent of our papers were double scored. The percentage of agreement (within one score point) for the major tasks of narrative, persuasion and reflective essay was 95% among rater pairs. This number is all the more remarkable because we used 18 different raters in the study. Median alpha coefficients were .72, .74, and .74 respectively for the general impression scoring. These numbers were undoubtedly affected by range restriction. On the basis of our experience and our data, the quality of essay rating was one of the aspects of the study in which we have the most confidence. However, we will continue to investigate rating as a source of bias.

Maybe the Measures Provide Students with a Special Advantage.

How to defend the measures themselves is a complex task. To establish validity, one must first see whether certain basic regularities appear in the data. Clearly, the data indicate that the measures show sufficient sensitivity to increased proficiency of students with more schooling. This is a regular finding across all tasks. On average, younger students never do better than older students.

Another validity indicator is in the major essay responses of 6th graders. Forty-six percent of sampled teachers reported that narrative is a frequent assignment in 6th grade and almost 70% of the students met or exceeded the standard. But for persuasive writing, a task much less frequently reported as being taught to elementary school students, only 24% of the students met or exceeded the standard.

Only about ten percent of our sample was drawn from private schools, and although the numbers are absolutely small, these findings lend some validity to our measures. We found that private school students outperformed public school students at the 10th

grade, where, you will recall, we assessed "average" classes. Yet, when the 12th grade precollegiate performance is examined, the differences wash out. This pattern is consistent with our expectations.

Another obvious tack to explore the quality of our findings would be to try to establish construct validity by using NAEP data directly. Because we tested different age ranges with different tasks and different scales, we need to suffer major contortions to make any connections. At best, our 12th grade sample could reasonably be compared with NAEP 17-year-olds on the persuasive task. Table 5 presents the cumulative frequencies for this comparison.

In this comparison, the SWC sample demonstrates consistently superior performance. Notice, however, that the shape of both distributions is similar. The higher values for SWC are expected, because our sample was drawn from precollegiate 12th graders rather than from 17-year-olds in general and provides another basis of mild encouragement on the validity of our data. On the other hand, it appears that our standard, a score of "3," required more out of students than the NAEP score of "2." Clearly the further investigation of SWC-NAEP performance is desirable. Rescoring NAEP essays with the SWC scales seems to be the most sensible option.

What About the Judged Difficulty of the Tasks?

We also classified the writing tasks in terms of their cognitive demands as either hard, medium, or easy. Students also were asked, at the conclusion of each essay, to assess the difficulty of each task. Students, experts and data generally agreed on this classification. There were two interesting specific findings: First, that the letter of application was harder for high school students than any other task, including extensive essays, and despite minimum competency testing, they report little experience with this task. Second, 6th

graders had to struggle more when retelling a story than when making one up, a finding we attribute to the burden of reading a story and then writing about it in a short period of time.

Do Teachers' Responses Support our Findings?

Another source of validity may be derived from reports of teachers' practices. We ran correlations of teachers' reported instructional emphases and student performance. Significant correlations were found across grade levels for student writing performance and teachers' reported emphasis on writing in their teaching. For students at all grade levels, teachers' emphasis on reading related positively to writing performance. For instance, at the 6th grade an emphasis on prose fiction significantly correlated with students' narrative writing -- students who read stories could write them. Reading also correlated significantly 10th and 12th grade with the persuasive writing task. Reading obviously provided both content (what to say) and schema (how to represent such thoughts cognitively).

For 6th and 10th grade teachers, emphases on essay planning and group work also positively correlated with working performance. Other significant correlations involved the teaching of literary analysis, such as criticism, the study of varieties of language, and the study of esthetics. Although our analyses are not complete, at this point we believe these variables are proxies for high ability classrooms, and would explain our findings.

Certain teaching practices consistently negatively correlated with student performance, and we believe these lend our study credibility as well. For example, in eight of ten tasks, emphasis on parts of speech was significantly and negatively related to writing performance. Furthermore, for the longer persuasive and reflective tasks, emphasis on spelling was a negative correlate with

performance -- which leads me to reevaluate my appraisal of my own spelling prowess.

We believe these instruction-outcome relationships help substantiate our present claim that the US-SWC data are valid. Our ratings focused on full discourse, on essay- rather than on sentence-level skills.* Our instructional findings highlight a special set of classroom activities. These activities support the exposure to and development of ideas—either as something to write about or as models that help students internalize the communication process.

Our inferences about the importance of planning and reading are also corroborated by findings in the field of cognitive psychology. Given future confirmation by subsequent data analyses, these instructional practices should inform teachers' classroom behavior.

Looking Good

My initial reaction was that our findings were unseemly positive. We are too used to hearing about all of the flaws our students demonstrate, from not knowing simple historical facts to lacking competitive strength in mathematics. If true, can the US handle such bounty? What are its implications? Certainly, I cannot ascribe our findings directly to the effects of the Commission on Excellence reports, state level reforms, or beefy curricula. Unfortunately, it is too much too soon. More likely explanations involve the increased attention to writing process, a phenomenon at least ten years old. Our data may reflect, as well, the general upturn in achievement reported by Koretz (Congress of the United States Congressional Budget Office, 1987).

---

*We were not specifically interested in spelling, sentence structure, or usage, although we know, anecdotally, that these influenced the general impression scores.

Time to Think -- Time to Write

But there is another reason why the SWC study may have produced the kind of results it did: the nature of the measures we used. The tasks provided to students allowed them an adequate period of time to write. Consider that some large-scale assessment exercises are as short as 12 minutes and that the longest may only be 30 minutes. How much can a student say and how fast can he say it? Testing student proficiency with such tiny time periods for writing is like educational fast food—it provides only short term satisfaction. The SWC study permitted students about a full class period to create longer essays. Even this period is woefully short compared to the usual practices in some European countries and abbreviated when the full cycle of planning, writing, revision and editing is considered. But in explaining our findings, we believe that giving students a reasonable amount of time is critical: time to think about what to say, and then, time to write.

Next Steps

We clearly have many more analyses to conduct. First, we will search for funds to do so. Assuming we are successful, then we will be looking more intensively at predictors and instructional conditions associated with strong writing performance. We also will be looking deeply at the substance of what students have written to get a better idea of the different ways they approach the problem-solving tasks of writing. Finally, we will analyze the within-student data, their attempts across writing tasks. If possible, we would like to do explicit NAEP comparisons, even on a relatively small number of papers.

We also wish to assess the assumptions we have made on the quality of domain specifications for the writing tasks. This work will take the form of generating additional topics within each genre and assessing students' ability across topic, for instance, in narration. A

second related research interest is to assess within and between genres alternative approaches for providing students with common information upon which to base their writing. The SWC confounds information provision with genre and we need additional work to disentangle effects.

We are also pursuing the development of distributed technology-based rater training. The use of microcomputer networks will allow training and rating of essays to occur at a distance and with economy. Since we believe the more measurement of writing the better, such a system ought to increase student proficiency in the long run.

We are already at work, with the assistance of the US Department of Education, in exciting and new approaches to assess students' understanding of subject matter using systematic writing tasks as a major option. In the US, we are working in the area of history and are designing assessment tasks, using some of our SWC procedures. Our students are given primary historical documents to read and expository writing tasks requiring explication and comparison. Our goal is to develop a subject matter scale, analogous to the rating scheme used in the SWC. Then, we believe, writing assessment can properly marry the facility of expression with the substantive ideas students wish to convey. We will have created measures that stimulate the understanding and transformation of content and that also provide enough time to write.

Conclusion

Moving measurement beyond multiple choice constraints was part of what SWC was about. We expect that the results of our work will continue to stimulate discussion and some controversy. If IEA wishes to continue the exploration of measures that better match how ideas are shaped and how students learn them, we believe the experiences of the SWC have much to offer. We also advocate

continuing to study how instruction and outcomes are related. At minimum, we have learned that students need clearly specified tasks, something to write about (content), and time to write.

References

Applebee, A.N. (1981). *Writing in the secondary school: English and the content areas.* Urbana, IL: National Council of Teachers of English.

Applebee, A.N., Langer, J.A., & Mullis, I.V.S. (1986a). *Writing Report Card: Writing Achievement in American Schools* (NAEP Report No. 15-W-02). Princeton, NJ: Educational Testing Service.

Applebee, A.N., Langer, J.A., & Mullis, I.V.S. (1986b). *Writing: Trends Across the Decade, 1974-84* (NAEP Report No. 15-W-01). Princeton, NJ: Educational Testing Service

Baker, E.L. (1982). Specification of writing tasks. *Evaluation in Education: An International Review Series, 5*(3).

Baker, E.L., & Herman, J. (1983). Task structure design: Beyond linkage. *Journal of Educational Measurement, 20*(2).

Bereiter, C., & Scardamalia, M. (1987). *Psychology of Written Composition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Bereiter, C., & Scardamalia, M. (1982). From conversation to composition: The role of instruction in a developmental process. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 1-64). Hillsdale, NJ: Lawrence Erlbaum Assoc.

Bloom, B.S. (1981). *All our children learning: A primer for parents, teachers, and other educators.* New York: McGraw-Hill.

Britton, J. (1982). Spectator role and the beginnings of writing. In M. Nystrand (Ed.), *What writers know: The language, process and structure of written discourse* (pp. 149-169). New York: Academic Press.

Cazden, C.B. (1986). Classroom discourse. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 432-463). New York: MacMillan Publishing Co.

Coffman, W. (1986). *Recommendations on writing assessments for future NAEPs* (ERIC commissioned paper). Princeton, NJ: Educational Testing Service.

Congress of the United States Congressional Budget Office (1987). *Educational achievement: Explanations and implications of recent trends.* Washington, DC: Author.

Cooper, C.R., & Odell, L. (Eds.) (1977). *Evaluating writing: Describing, measuring, judging.* Urbana, IL: National Council of Teachers of English.

Faigley, L., Cherry, R.D., Jolliffe, D.A., & Skinner, A.M. (1985). *Assessing writers' knowledge and processes of composing.* Norwood, NJ: Ablex Publishing Corp.

Farr, M. (Ed.) (1985). *Advances in writing research: Vol I: Children's early writing development.* Norwood, NJ: Ablex Publishing Corp.

Florio, S., & Clark, C.M. (1982). The functions of writing in an elementary classroom. *Research on the Teaching of English, 16,* 115-130.

Flower, L., & Hayes, J.R. (1981). The pregnant pause: An inquiry into the nature of planning. *Research in the Teaching of English, 15,* 229-244.

Frederiksen, C H., & Dominic, J.F. (1981). Introduction: Perspectives on the activity of writing. In C.H. Frederiksen & J.F. Dominic (Eds.), *Writing: The nature, development, and teaching of written communication: Vol. II: Writing: Process, development and communication* (pp. 1-20). Hillsdale, NJ: Lawrence Erlbaum Assoc.

Freedman, S. (Ed.) (1985). *Acquisition of written language.* Norwood, NJ: Ablex Publishing Corp.

Gray, J., & Myers, M. (1978). The Bay Area Writing Project. *Phi Delta Kappan, 59,* 410-413.

Hayes, J.R., & Flowers, R.S. (1980). Identifying the organization of writing processes. In L.W. Gregg & E.R. Steinberg (Eds.), *Cognitive processess in writing* (pp. 3-30). Hillsdale, NJ: Lawrence Erlbaum Assoc.

Heath, S.B. (1981). Toward an ethnohistory of writing in American education. In M.F. Whiteman (Ed.), *Writing: The nature, development, and teaching of written communication: Vol. I: Variation in writing: Functional and linguistic-cultural differences* (pp. 25-46) Hillsdale, NJ: Lawrence Erlbaum Assoc.

Langer, J., (1986). *Children reading and writing: Structures and strategies.* Norwood, NJ: Ablex Publishing Corp.

Lloyd-Jones, R. ((1977). Primary trait scoring. In C.R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging.* Urbana, IL: National Council of Teachers of English.

Millman, J. (1980). Content domain specification/item generation: Computer-based item generation. In R.A. Berk (Ed.), *Criterion-referenced measurement: The state of the art* (pp. 32-43). Baltimore: Johns Hopkins University Press.

Purves, A., & Takala, S. (Eds.) (1982). *Evaluation in Education: An International Review Series, 5*(3).

Quellmalz, E.S. (in press). Writing skills. In R.A. Berk (Ed.), *Performance assessment: Methods and applications.* Baltimore: Johns Hopkins University Press.

Quellmalz, E.S. (1980). *Test design: Controlling rater drift* (CSE Report to NIE). Los Angeles: UCLA Graduate School of Education.

Quellmalz, E.S., Baker, E.L., & Enright, G. (1980). *Test design: A comparison of modalities of writing prompts* (CSE Report to NIE). Los Angeles: UCLA Graduate School of Education.

Quellmalz, E.S., Capell, F., & Chou, C. (1982). Effects of discourse and reponse mode on the measurement of writing competence. *Journal of Educational Measurement, 19*, 241-258.

Quellmalz, E.S., Spooner-Smith, L, Winters, L., & Baker, E.L. (1980). *Test design: The University of California writing study* (CSE Report to NIE). Los Angeles, UCLA Graduate School of Education.

Scardamalia, M., & Bereiter, C., (1986). Research on written composition. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd. ed., pp. 778-803). New York: MacMillan Publishing Co.

Scardamalia, M., Bereiter, C., & McDonald, J.D.S. (1978). Role-taking in written communication investigated by manipulating anticipatory knowledge. *Resources in Education* (ERIC Document Reproduction Service No. ED 151 792).

Scribner, S., & Cole, M., (1981). Unpacking Literacy. In M.F. Whiteman (Ed.), *Writing: The nature, development, and teaching of written communication: Vol. I: Variation in writing: Functional and linguistic-cultural differences* (pp. 71-88). Hillsdale, NJ: Lawrence Erlbaum Assoc.

Staton, J. (1982). Writing and counseling: Using a dialogue journal. *Language Arts, 57*(5), 514-518.

Stein, N.L., & Trabasso, T. (1982). What's in a story: An approach to comprehension and instruction. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 213-267). Hillsdale, NJ: Lawrence Erlbaum Assoc.

Tyler, R.W. (1966). Development of instruments for assessing educational progress. *Proceedings of the 1965 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service.

Vahapassi, A. (1982). On the specification of the domain of school writing. *Evaluation in Education: An International Review Series, 5*(3).

Whiteman, M F. (Ed.) (1981). *Writing: The nature, development and teaching of written communication: Vol I: Variation in writing: Functional and linguistic-cultural differences.* Hillsdale, NJ: Lawrence Erlbaum Assoc.

Wittrock, M.C. (Ed.) (1986). *Handbook of research on teaching* (3rd ed.). New York: . MacMillan Publishing Co.