The Evaluation of General Education: Lessons from the USA State of Washington Experience

Gerald M. Gillmore*

Office of Educational Assessment, University of Washington

Abstract — Assessment and accountability are presented as contrasting models for evaluating outcomes in higher education. While both models are concerned with quality and improvement, the accountability model stresses externally imposed evaluation goals, methods, and criteria and comparisons among institutions. The assessment model stresses faculty-determined, institutionally-specific goals and methods with a focus on improvement. The report next discusses general education and the outcomes that one should consider measuring. Two State of Washington (USA) studies, relevant to the evaluation of general education, are presented; one under an accountability model that used standardized tests, and one under an assessment model that used student writing from courses. The report concludes that in evaluating general education: 1) one should measure the skills that one would want all students graduating from a University to have mastered rather than particular content. 2) The diversity of educational institutions in a state or nation should be recognized and cherished. It is a mistake to think that tests or other measures can be developed and applied uniformly to all institutions and their results interpreted equivalently. 3) Measures should not be developed with the intention of applying them uniformly across students, irrespective of their majors. 4) One must give as much attention to how assessment results will be used as to the measures themselves. 5) Faculty must be intimately involved in the process of identifying goals and the measures to assess their attainment. Otherwise, the goals and measures are apt to lack validity and results are apt to be ignored by faculty. 6) The best way for an institution to be accountable is for it to develop a culture of assessment and reflection.

(Received on September 21, 2004)

1. Introduction

My discussion of the evaluation of general education will be oriented toward student learning outcomes. Certainly, one can evaluate the general education program of a university by looking at such things as course requirements and curricular integration. As important as these aspects of a general education program may be, I would argue that they are a means to a more important end. The bottom-line issues are those associated with student learning: what students know, what students can do, and how students have changed as a result of attending college.

This report begins by presenting definitions of the terms evaluation, assessment and accountability and how I have chosen to use them in addressing this topic. Next, I discuss general education and the outcomes that one should consider measuring. I then present two State of Washington studies that are relevant to the evaluation of general educa-

*) Correspondence: Office of Educational Assessment, University of Washington, Seattle, Washington, USA

tion. I close with what we can learn from these studies about evaluating general education.

2. Evaluation, Assessment, and Accountability

Before discussing general education, I would like to define and draw distinctions among three related terms: evaluation, accountability, and assessment. Of the three, I will use evaluation as the most general. I shall use the terms "assessment" and "accountability" as forms of evaluation that differ in their goals, strategies, and tactics. The definition of evaluation is simple: it is a determination of the value or worth of something. In many contexts, including higher education, evaluative judgments are highly subjective and apt to be controversial because consensus is often lacking on what constitutes value or worth.

The New Oxford Dictionary offers evaluation as a synonym for assessment and the two terms are indeed difficult to distinguish. For this report, I am going to use a specialized definition of assessment that I will attempt to describe here, but that will become clearer as the report unfolds. I define assessment as an evaluation model that focuses on the attainment of valued student learning outcomes and that is internally driven by an institution. When an institution or faculty do assessment they essentially go through four steps.

1. Determine what the desired outcomes are, possibly for a class or even one class session, for a major, for general education, or for an entire degree program.
2. Design measures, hopefully more than one, to determine the extent to which these outcomes are being met.
3. Make judgments about what the data indicate concerning successes and particularly what needs to be improved.
4. Make changes suggested by the data and start the process all over again.

Several aspects of this model should be emphasized. First, it is the faculty, individually or in groups, possibly with consultation from assessment experts, who determine the valued outcomes. Secondly, the process is focused on improvement. Finally, it is iterative in nature, much like the scientific method.

Accountability is an evaluation model that focuses more on the values of an external constituency. All institutions of higher education are rightly accountable to multiple constituents, including students and parents, of course, who are footing part of the cost, but also the public at large who depend upon higher education to deliver adults who will make positive contributions to the society. Public institutions are particularly accountable because they are more direct beneficiaries of public money. The accountability model tends to operate through the following steps:

1. A public agency, outside of but in consultation with the institutions themselves, determine the goals and values to be measured.
2. Measures are developed or chosen by the outside agency and applied in order to evaluate these outcomes.
3. Institutions are ranked or otherwise compared as a result of these measures.
4. Institutions with favorable outcomes are rewarded, often with budget enhancements, while institutions with unfavorable outcomes are punished, possibly with budget reductions.

Both models are concerned with quality and improvement. The accountability model assumes that if we make institutions accountable, they will become better. The flaw in this assumption, as we will see illustrated below, is that it assumes that the accountability measures are consequentially valid. In other words, it assumes that improved performance on the measures is indicative of increases in valued outcomes. Unfortunately, measures that can be equivalently applied across institutions in such a way that institutions can be readily compared often assess superficial aspects of education. Just as testing students only on recall of facts is apt to lead to superficial learning and unsatisfactory studying behavior, so to can superficial or poorly-considered accountability measures lead to unsatisfactory institutional behavior. For example, if an accountability measure were the number of student instructional hours per faculty, an institution could improve on that measure simply by increasing class size. The apparent result would be improved institutional quality. The real result might be lowered student learning.

The true goal of accountability is to have high quality institutions, and the assessment model assumes that the best way to assure this is for each institution to strive to improve itself. Furthermore, if there are continuous efforts toward assessing our programs for the improvement of their quality, the institution will be accountable. The flaw in the assessment model is that it depends on improving the right things and on good will efforts. How an institution prioritizes its assessment efforts may not match what various constituencies view as proper priorities. For example, an institution may try to increase the number of grants received and thus reward fundable research activity over teaching quality, which parents will see as wrong-headed and a demonstration of a lack of accountability. In addition, assessment efforts can be hard work and take one away from activities that are more interesting or more rewarding.

The two models need not be antithetical. Accountability measures can communicate the values of external constituents to institutions and provide motivation to improve

valued outcomes. The assessment orientation can help assure that improving an institution's standing on a given measure does indeed improve quality. Unfortunately, problems are inherent in combining the two models. One can easily see the conflict between accountability's need for simple, comparable measures and assessment's need for complex, heterogeneous measures. The results of assessment activities are typically tentative and incomplete, and therefore difficult to translate into useful accountability measures for government-level policy making. Thus, the assessment model requires the government's patience and tolerance of complexity, something for which it is not noted. Other inherent differences include the assessment model's favoring of cooperation across institutions and the accountability model's favoring of competition. From an institutional point of view, the assessment model tends to point one toward the bad in order to maximize the effect of improvement efforts. (If it is not broke, why fix it?) The accountability model points one toward the good and tries to promote and publicize it. Thus, the orientation of assessment naturally leads an institution "shine lights in dark corners," while the orientation of accountability naturally leads institutions to a defensive posture that tries to control where the light shines.

It might seem at this point that assessment and accountability should each go their own way. In the Oxford model of education, faculty who grade a given student are different from faculty who teach that student. It is felt that the specter of summative evaluation can detract from the proper teacher-student relationship. Clearly, accountability can get in the way of the genuine efforts toward improvement that are guided by assessment. Furthermore, those whose purpose is to evaluate institutional performance can be co-opted by program personnel when the former are also involved in improvement activities. Unfortunately, a separation of assessment and accountability can result in meaningless accountability measures and little motivation for improvement efforts. If accountability measures have no relationship with those aspects that the institution strives to improve, they are unlikely to lead to "good" institutional behavior. If doing "deep" assessment does not lead to rewards from external constituencies, genuine assessment activities may be hard to sustain.

3. Measurement of General Education Outcomes

General education is usually contrasted with education in a student's major. The former is assumed to be roughly uniform across all students while the latter is idiosyncratic to each discipline. A further assumption is that general education occupies the first half of a college education while the major occupies the last half.

General education outcomes can be placed into two overlapping categories: what students know and what they can do with what they know. The first, which we will label "knowledge," is basically factual in nature. Does the student know basic national and international history? Can a student knowledgably discuss the great literature of a country? Does the student know the difference between an atom and a molecule and how to define the scientific method? Between meiosis and mitosis? Between theories of Freud and Jung? The second, which we will label "skills," relates to how the students can make use of knowledge. Typical skill categories include writing, applying quantitative methods to solve problems, and thinking critically.

Thinking of general education as a knowledge set evokes an obvious problem: who decides what should be included in that set? Even in a culture as relatively homogeneous as Japan's, there would be certain and intense arguments among faculty and government officials about the relative importance of certain facts and concepts. Would not a biologist favor biological concepts over an artist who would favor artistic knowledge? Would not a feminist sociologist think all students should be able to demonstrate a firm understanding of gender inequalities, while a medieval historian would ascribe the same importance to the lineage of shoguns? Even in high school education, where the curricula are more centrally controlled and relatively homogeneous across and within schools, controversy over what should be taught and learned is present. A second problem is that it is unclear that an important purpose of a college education is to teach students facts. More likely, one would want to hold the high schools responsible for this level of learning; however, even high schools might balk at such a characterization. Thus, while not completely disregarding the importance of knowledge, it seems more promising to assess skills – or students' ability to apply the knowledge they have acquired.

The two State of Washington initiatives that I will describe below each attempted to measure what can be labeled as general education skills, but in very different ways. However before going to that discussion, I would like to first challenge assumptions about general education with which I began this section. General education cannot be divided from education in the major as cleanly as many people assume, either in its timing or in its content. The University of Washington is just ending a four year, longitudinal study of undergraduate student learning (UWSOUL). In this study, 191 students completed a survey each quarter over their first three years of college. (The data from the fourth year is yet to be completely collected). One set of questions asks students to rate on a 4-point scale how much they had learned about 25 skills or understandings for that given quarter (e. g., "Information, theories, and perspectives from your classes," "Understanding more about who I am and what I value," "Thinking critically about issues, and Working and/or learning independently").

Averages were computed for each year for each item, and on 23 of the 25 items, significant differences were found across areas of majors[1][2]. About as much variability was evident among the eventual majors in the first year as in the third. Or to put it another way: the students' major affected what they learned as much in the first year, when they were presumably taking only general education classes and had not yet even declared a major, as in the third when presumably their general education was over and they were within their major[3]. This result suggests that while students may not declare majors in their first year, they start on the path toward specialization this early. This specialization can be affected in two ways: the most obvious is through course selection, but subtler may be differences in what students with varying interests tend to learn within the same courses. It validates what we have seen in other contexts: that is, the idea that the first half of college is about general education and all students learn basically the same thing is a myth.

This result makes a powerful argument that general education and education in the major are inextricably linked, not only in timing but in substance as well. The skills one might label as a part of general education are heavily mediated by one's major in how they are learned and how they are appropriately carried out. For example, good writing in psychology is not the same as good writing in economics. The way one thinks about quantitative methods is not the same in chemistry as it is in business. Critical thinking in education is not the same as it is in engineering. This mediation of general education skills by the ways of thinking of the disciplines is a lesson we have learned over and over again in conducting the UWSOUL study.

Let me hasten to add that I do not wish to imply that learning the concepts of the major is more important than what I have labeled general education skills. Indeed, results of our surveys of alumni at the University of Washington suggest that the general education skills are more important than the concepts of the major, and that they increase in relative importance as time goes on.

4. The Washington State Test of Sophomores

The Higher Education Coordinating (HEC) Board of Washington State is an agency of the state's legislature whose purpose is to oversee and offer recommendations about public higher education. Its initial master plan (December, 1987), set in the early days of the assessment and accountability movement in the USA, "…recommended that both two-year and four-year institutions conduct a pilot study to evaluate the appropriateness of using standardized tests as a means for measuring the communication, computation, and critical thinking skills of sophomores."[4] The initial thinking of the HEC Board members was to administer the tests without the benefit of a pilot test, and their goals were to "…strengthen the curricula, improve teaching and learning, and to provide accountability data to the public."[5]

It is not clear how the three skills came to be chosen above all others. Operationally, communication was effectively the recognition of good writing, and computation was effectively math skills. Critical thinking is a concept that is talked of widely in all levels of education but is illusive in definition and operationalization. The only shared meaning is perhaps the vague idea of "high-level" thinking. One somewhat narrow definition is "The intellectual evaluation of a variety of assumptions and empirical data and the ability to arrive at logical conclusions on the basis of that evaluation."[6] In choosing to focus the study on sophomores at the end of their second year, so-called rising juniors, the HEC Board implicitly assumed that the learning of general education skills was complete at that time, and it also conveniently allowed students at two-year and four-year colleges to participate.

As recommended, the pilot study was conducted in 1988. To this day it remains the most thorough study of its kind and represents a watershed in the history of the assessment movement in the State of Washington and in the United States. Its impact was great in illuminating many problems with assessing general education outcomes with standardized tests. But before discussing these problems, we must present an overview of the study.

To conduct the study, two task forces were formed_one representing the public, baccalaureate (four-year) colleges and one representing the community (two-year) colleges. Each included members of the HEC Board staff. The task forces worked in parallel and ultimately conducted a joint study in an unprecedented level of cooperation.

Three nationally standardized tests were chosen as the most appropriate for the study's purposes:

- The Academic Profile (AP) (Educational Testing Service)
- The College Outcome Measures Program (COMP) (American College Testing Program)
- The Collegiate Assessment of Academic Proficiency (CAAP) (American College Testing Program)

The AP and CAAP were two-hour tests and the COMP was a four-hour test.

Test administration took place in the spring. Random samples of sophomores were invited to participate and over 1300 students from the five public four-year universities and from eight of the two-year colleges took two of the three tests. Thus, each test was taken by over 800 students. Students whose pair of tests included the four-hour COMP were paid $35.00 to participate, while others were paid $25.00.

In addition, more than 100 faculty from the same institutions took shortened versions of each test and then cri-

tiqued them for appropriateness of content. Subsequently half of these faculty met with test company representatives who described the meaning of the tests and provided the faculty with sample output. Faculty then evaluated the usefulness of the test results for their own teaching.

Pertinent results were as follows:

- The AP and CAAP tests were designed to measure reading comprehension, writing, mathematics usage, and critical thinking, and the COMP test was designed to measure knowledge and skills related to successful functioning in society. In fact, the three tests primarily measured verbal and quantitative aptitude or intelligence.
- The tests added little reliable information about students' academic performance. Results essentially reiterated what was already known from entrance scores and grades. Furthermore, there was a fairly high correlation between the test results and entrance tests scores. Across four-year schools, there was a perfect correlation between the average scores on these three tests and average scores on admission tests. To reward universities for their students' performance would be tantamount to rewarding them for their selectivity.
- Test scores were not related to specific aspects of the college experience, including number of general education credits, estimated time spent studying, or estimated time in the library.
- No test was rated by the majority of the faculty participants as a valid measure of general education. For the most part, faculty believed the range of skills and the methods for assessing the attainment of general education skills were too narrow. Especially, faculty saw the writing sections as falling short, because they measured editing skills rather than production skills.
- Perhaps of greatest importance, faculty did not find test results useful for their course planning or instruction. Therefore, they were unlikely to use the results in any way to improve their curricula or instruction.

I categorize this study as an accountability approach to evaluating general education because it was imposed upon state universities and colleges by an outside agency, and because it had as one of its major goals holding the institutions accountable. There was a clear intention to compare institutions to each other and to a national set of norms, which was a major motivation for using nationally standardized tests. And while the HEC Board did not choose the tests, the parameters were defined by the agency such that very few degrees of freedom were available for the choice.

To the HEC Board's great credit, as a result of the study, they changed their approach from the accountability model to the assessment model. They gave explicit recognition to the diversity of institutions in terms of student body and mission, to the importance of involving faculty in all assessment activities, and to the larger scope of higher education, from matriculation through alumni status. Perhaps of greatest importance, judgments of the quality of an assessment program became based on the number of changes made in the curriculum as a result of assessment studies, rather than by the results themselves.

5. The Senior Writing Study

In retrospect, it might be said that higher education in the State of Washington went through a decade-long, golden age (well, perhaps a bronze age) of assessment. The legislature allocated funds for its accomplishment. The HEC Board staff assumed a facilitative rather than an adversarial role. Projects were done on campuses, and there was a spirit of cooperation across campuses. A statewide assessment conference drew over 400 participants annually. Then, accountability again reared its head.

There are a few aspects of the accountability initiative of the late 1990's that need to be mentioned to give the proper context for the Senior Writing Study. The HEC Board and the institutions worked together to develop objective measures to evaluate institutional performance.(7) Four measures were in common and institutions could define four to seven idiosyncratic measures. Most but not all of the measures were related to instruction. One of the measures, the Graduation Efficiency Index, was developed by Phil Hoffman and me and is basically a ratio of the number of credits required for graduation to the number of credits actually taken.(8) We argued successfully that this measure was superior to time-to-degree measures in not punishing institutions for part-time students. For all measures, current performance was used as a benchmark and goals were arbitrarily set that required an equal increase for each subsequent year. Institutional budgets were cut for any goal unmet and NOT increased for goals that were met. Fortunately, this draconian system of punishment stayed in place only two years.

Under this accountability system, the measures had to share several characteristics. First, they had to be quantitative, unambiguous in computational formula, and reducible to single values. Secondly, the measures had to be such that improvement goals could be specified. These restrictions are formidable. While the committee formed by university and HEC Board personnel made a good faith effort and is to be lauded for their efforts, these restrictions proved too great and the resulting measures failed to measure the significant values held by the higher educational community. Or to put it more baldly, none of those measures dealt

directly with learning and some of them, it could be argued, were inimical to learning. After several years, in a very significant semantic shift, the measures became referred to as efficiency measures rather than accountability measures.

After the accountability measures were put into operation, I was asked by the Provosts of the four-year institutions to form a committee of assessment experts and people who had worked on the accountability initiative to see if we could come up with measures that addressed important aspects of student learning and that could also serve as accountability measures. This challenge was daunting, but one that spawned the senior writing study.

I was essentially asked to bring assessment principles to accountability concerns, and I began with the following three basic principles

- Assessment has at its heart evaluating programs such that information is fed back to constituents at all levels in order to improve teaching and learning. Thus, our measures should make a positive difference in how we teach and how students learn. This standard is the ultimate test of value.
- Faculty need to be heavily involved both because their expertise is needed to define and validly measure the student outcomes and because they are the ones who will use the results to make a positive difference in learning and teaching.
- If an accountability measure must emerge, it is better as an imperfect measure that is on the right track than as a psychometrically perfect measure that has little relevance for the improvement of learning or that may have potentially bad consequences. One implication of this principle is that it calls on us to simplify our measures of complex outcomes no more than we have to.

Early on at the first meeting of representatives from the baccalaureate institutions, we recognized that all our institutions have a great deal of self-report data from current students and alumni on an array of learning outcomes and satisfaction measures. However, these data, while important, are flawed in that they are based on subjective opinion and may be influenced in complex ways by student expectations and other uncontrollable variables. Thus, a more direct measure of student ability was sought. At this point we turned to the assessment of writing.

Why writing? Assessing writing, though very difficult, has two immediate advantages. First, there is universal agreement that writing is an important skill. Second, writing offers us the best window we have into student reasoning ability. The difficulty of separating thinking and writing skills can be an advantage because assessing student writing, important in and of itself, can also help us think deeply about students' critical thinking abilities.

The Senior Writing Study assumes that programs should be judged by the best writing that students within these programs can do when they leave college. (This affirmation should not be confused with the writing of our best students, which is not consistent with our purposes.) By best writing we mean the following:

- Writing that students are motivated to do well;
- Writing about a subject that students should know and care about; and
- Writing done in response to a challenging and well-formulated assignment.

"Best" writing is done in the context of fields of study. There is considerable agreement that the characteristics of good writing differ from discipline to discipline. For example, what constitutes a valid argument varies considerably from analyses of Japanese literature to arguments about issues in psychology, to lab reports in chemistry. Furthermore, by using graded papers students had written for the courses, we could assume some reasonable level of motivation. Thus, we chose papers for study that students did in courses in their majors, preferably in capstone courses in their senior year. Writing at this level also clearly differentiates college writing from high school writing.[9] The accountability question we faced was, can the quality of an important component of an institution's educational program be validly judged by reading a sample of senior-level papers? The six public four-year institutions joined to conduct pilot studies of this question over four summers, 1998 – 2001.

The studies we completed required three basic steps: developing a rubric that could be used to assess student writing, collecting student papers, and scoring these papers. In our first year, faculty in the disciplines from which papers had been collected, writing instructors, and assessment specialists met initially for a two-day session to develop scoring criteria, or rubrics. Rubrics were developed using an iterative process with a sample of actual student papers. One surprising result was that all disciplines were satisfied with the same criteria; however, many elements were interpreted differently across disciplines. Even so, the set of scoring criteria was a major benefit of the study in and of itself. The scoring criteria, or rubrics, are found in **Table 1**.

For the collection of papers step, we decided that the best way to proceed would be to identify courses that required students to write "good" papers in a limited but representative set of majors and take a random sample of the papers produced.

For scoring the papers, we decided that we needed faculty from the corresponding disciplines, writing specialists, and members of the community who were working in those fields in which papers were collected. The latter were in-

Table 1. Indicators of Effective Writing

Please note that the areas listed below and their numbered indicators must be viewed through the lenses of specific disciplinary and classroom contexts when used to evaluate specific texts. Evaluators from these contexts may define indicators and their general categories in particular ways that differ from those of evaluators in other contexts. In addition, disciplines may ascribe differential importance and assignments may dictate differential weighting to the various indicators. Finally, these indicators are intended for use in assessing papers and the effectiveness of writing programs in specific disciplines. However, they may be useful guides for students in writing papers and for faculty in grading papers and in discussing good writing practices with students.

A. Conventions/Presentation

1. The text reveals evidence of crafting.
2. The text reveals evidence of editing.
3. The text reveals evidence of proofreading.
4. The information is cited accurately and completely.
5. The documentation style is appropriate.
6. The format used, including visuals and diagrams, is effective.

B. Content

1. The topic is appropriate in terms of the assignment.
2. The purpose for writing is evident.
3. The evidence/information is relevant.
4. The evidence/information is accurate.
5. The evidence/information is necessary.
6. The evidence/information is complete.

C. Organization

1. The overall organization captures the designated purpose.
2. The ordering of information/evidence leads the reader through the text (e.g., transitions, signposts, headings, bullets).
3. The parts connect well with each other and with the governing idea.
4. The visual and verbal elements are well-integrated.

D. Reasoning

1. The claims, ideas, and purpose are significant.
2. The evidence is of quality.
3. A sufficient context is provided.
4. Assumptions are recognized and made explicit.
5. The interpretation and analysis of evidence/information/visuals shows depth of thinking.
6. The interpretation and analysis of evidence/information/visuals shows logical reasoning.
7. The interpretation and analysis of evidence/information/visuals shows complex reasoning.
8. The interpretation and analysis of evidence/information/visuals shows accurate conclusions.
9. The interpretation and analysis of evidence/information/visuals shows informed recommendations.

E. Rhetoric of the discipline

1. Sufficient knowledge of the subject is demonstrated.
2. Use of specialized concepts demonstrates understanding.
3. The genre is appropriate to the discipline.
4. The format is appropriate to the discipline.
5. The language is appropriate to the discipline.
6. The tone is appropriate to the discipline.
7. There is evidence of disciplinary ways of thinking and an appropriate sense of audience.

Table 2: Disciplines Represented in the Four Years of Study

1998 and 1999	2000	2001
Sociology	Psychology	Psychology
Biology	Biology	Chemistry
English	Education	Education
Engineering/Technology	History	History
Business	Business	Health/Nursing

vited because they provide the important perspective of employers and of the writing that will be expected on the job. Readers evaluated papers blindly: names, faculty comments, grades, and the institution of origin were removed from all papers. Thus, papers came from five disciplines, from six institutions, and from a number of classes within each of the institutions.(10) The disciplines represented across the four years are indicated in **Table 2**.

A second two day session of the 1998 study and the subsequent three day sessions in 1999, 2000, and 2001 were spent introducing the rubric, training raters, and scoring papers. In each disciplinary group, several papers were read and discussed and then each paper was read independently by two raters. When scoring papers, readers were asked to respond to each element on a four point scale (Strong, Acceptable, Weak, and Not Acceptable), as well as to give an overall rating to each category and to the entire paper, using the same four point scale. At rater insistence, ratings at mid-points between these scale points were allowed. These sheets were then use by pairs to reach consensus and another form was completed that indicated the consensus ratings. Usually, consensus was easily attained and the papers were rated with acceptable reliability. A third rater was used only when consensus could not be reached, which seldom happened. About 30 to 35 individuals read papers in each session. In the first, second, third, and fourth years, 83, 169, 225, and 196 papers were read and evaluated, respectively.

Even though this research was motivated by accountability concerns, I would classify it as being predominantly conducted under the assessment model for the following reasons:

- It was designed with faculty input at every step of the way.
- While one of its goals was holding the institutions accountable, improvement of curriculum and instruction were held as at least equally important.
- There was no clear intention to compare institutions to each other and no possible way to compare our results with any national norm group.
- We examined actual student efforts from on-going classes rather than an externally imposed task.
- We were given great latitude as to how the study should be conducted.

Among the average ratings of papers, two results have stood out. First, the lowest rated category was consistently "Reasoning," while the highest was consistently "Content." This result is hardly surprising since reasoning is a higher level skill than the mere presentation of content. However, it does point faculty in the direction demanding the display of clear reasoning skills in papers rather than mere recitation of facts. In addition, it suggests that while we are helping students "to know" things, we are having less success in helping them use what they know. A second finding of these rating sessions was that marked differences in quality were found in comparing disciplines.

The most clearly valuable aspects of these sessions might be termed faculty development. There was an essentially unanimity of opinion that reading the papers and discussing them with colleagues from around the state changed the way faculty participants taught in positive ways, and many carried their insights and enthusiasm to their colleagues. This result is clear from direct comments and also from the conversations that took place throughout the sessions. Two quotations from evaluation forms illustrate this faculty development value:

> Carmen Werder (Assoc. Director, University Writing Center Programs, Western Washington University): *This project has already proven itself in terms of faculty development. The conversations around those stacks of papers were some of the most valuable ones I have experienced anywhere. Everyone I talked to agreed. Any activity that gets faculty across disciplines and from many schools in the same room reading and discussing real student writing deserves support.*

Janet Ott (Professor of Biology, The Evergreen State College): *This has been the most useful four days that I have ever spent on education in general and on writing in particular.*

Everywhere, faculty discussed their teaching and student learning, and insights were frequently shared. There are several reasons for this positive response. First, faculty seldom get an opportunity to read the papers written by students in courses other than their own courses. One powerful lesson that came from reading papers from many different classes was the importance of the writing assignment for producing good writing. One cannot escape the conclusion that bad assignments yield bad writing. Second, faculty seldom get or take opportunities to talk to their colleagues, especially colleagues from other campuses, about student writing. Finally, discipline specialists and writing specialists were able to learn from each other. The former were able to learn a useful vocabulary to talk about and think about writing from the latter, while the latter were able to learn the writing expectations and conventions from the former.

Each campus has received the results of ratings of their papers as well as the overall ratings. These data provide information to be fed back to departments about the writing of their students. It is important to remember that the purpose of this project is to evaluate writing programs and not individual students. The fact of poor assignments leading to poor writing found an interested audience in departments, along with the fact that papers were rated lowest, on average, on the reasoning dimension, suggesting that students need more instruction and practice with reasoning in their disciplines. At one year's scoring session, papers from one discipline were rated particularly low. This result drew a great deal of attention from the corresponding departments across the state, and we saw a significant improvement in papers of students from this discipline the following year. The impetus for the Senior Writing Study was accountability considerations, and taking a narrow view of accountability, the project failed. For one reason, the number of papers rated from any particular department is small and the number of departments viewed in any one year is small, as well, so generalizations are hard to make. One must recall that often accountability is indexed by single numerical values whose function is to represent an entire institution's performance on a given dimension. To increase the number of papers and departments to a level that a reliable and representative single score could be gained for each campus would be very expensive. Even if the state decided to fund an expanded program, the thought of reducing all of the rich information that derived from the Senior Writing Project down to a single number or a few numbers by which the quality of education in writing is judged would be understandably worrisome.

However, one can conclude that our sense of accountability to the state is nowhere better demonstrated than by the very conduct of this project; the process was more important than the product. Participants recommended that the basic methodology of the Senior Writing Study should be extended to individual institutions: that is, for academic departments collect and read papers as a group on particular campuses, including writing specialists and community representatives in the process. Indeed, that has been the next step in the evolution of this evaluative procedure. It also should be stated that the Washington State Legislature has begun a another accountability process, which begins next year with a task force studying the ways that institutions can be compared using another set of accountability measures. This means that in the future, the scarce resources in the state may not be aimed at improvement in student learning.

6. Conclusions: What Have We Learned?

Evaluating general education from a student outcomes perspective is more elusive than it might at first appear because the concepts of a shared general education curriculum and shared goals are largely illusionary. Students begin to specialize early and what they learn and how they learn it is filtered through the lens of their major discipline. Thus, general education is a lot different for science majors than for humanities majors, and it is even a little different for physics majors than for chemistry majors. From my experience through fifteen years of assessment and accountability initiatives in the State of Washington, and in particular through being intimately involved in the two studies that were described in some detail above, I have reached the following conclusions about evaluating general education.

1. There is no well-established corpus of content that all students should know, nor should there be. To ask a set of individuals to make such a determination, given the enormous array of choices, would give that group too much power and too much burden. A far better approach is to think in terms of the skills that one would want all students graduating from a University to have mastered.
2. The diversity of educational institutions in a state or nation should be recognized and cherished. Institutions develop different goals, cultures, and student bodies. It is a mistake to think that tests or other measures can be developed and applied uniformly to all institutions and their results interpreted equivalently.
3. Measures should not be developed with the intention of applying them uniformly across students, irrespective of their majors. It is only at the lowest level of

intellectual functioning that such measures have a chance of being valid. As we move closer to the skills that university faculty really care about, such as critical thinking and effective writing, we find more differentiation by discipline.

4. One must give as much attention to how assessment results will be used as to the measures themselves. The ultimate proof of the efficacy of any assessment or accountability system is the long term effects it has on the quality of instruction and the extent to which other goals are being met. Systems that are imposed from external agencies are apt to be misdirected even with the best of intentions and met with hostility and suspicion and be undermined by the faculty. One must especially be vigilant with regard to unintended negative consequences.
5. Faculty must be intimately involved in the process of identifying goals and the measures to assess their attainment. Otherwise, the goals and measures are apt to lack validity and results are apt to be ignored by faculty.
6. The public has every right to expect and demand accountability from its institutions of higher education. The best way for an institution to be accountable is for it to develop a culture of assessment and reflection. Institutions must keep asking the questions about what students can and cannot do when they leave. At the classroom level, this shift can be viewed as a change in focus, from "What do I plan to teach?" to "What do I want students to know and be able to do, and how will I know they have accomplished those ends?" At the campus level, it is a shift from all attention on what courses should be offered and what requirements put in place to at least equal attention to actual competencies of graduates. In other words, the best evidence for accountability is the presence of a strong and vital assessment program.

Notes

1. Students were classified into one of the following seven major areas: Arts, Humanities, Social Science, Science, Business, Engineering, and Other Professional.

2. Because the study is still active, we have delayed publication of results.

3. Students at the University of Washington typically declare a major near the end of their second year.

4. Council of Presidents and State Board for Community College Education. (May, 1989). *The validity and usefulness of three national standardized tests for measuring the communication, computation, and critical thinking skills of Washington State college sophomores: General report.* Bellingham, WA: Western Washington University Office of Publications. (page iii)

5. ibid. page iii.

6. Unger, Harlow G. Encyclopedia of American Education (2nd Edition). New York: Facts on File, 2001. (Vol. 1, page 294).

7. Accountability Measures (System-wide measures are in bold letters. University of Washington measures are in italics)

1. Five Year Graduation Rate
2. Graduation Efficiency Index
   Freshmen
   Transfers
3. Retention (All undergrads; Fall to Fall)
4. Faculty Productivity
   (1) Efficiency (Student demand/course offerings)
   (2) Quality of Instruction (%Faculty >= 3.0 on student ratings)
   (3) Funding for research (External $/Faculty)
   (4) Quantity (Student credit hours/Faculty FTE)
5. Institution Specific
   (1) # of undergrads intensively involved in faculty research
   (2) % of upper division Credits Taken as Individualized Instruction
   (3) # of undergrads involved in public service Internships
   (4) % of undergrads having a research experience with faculty

8. Gillmore, G.M. and Hoffman, P.H. The graduation efficiency index: validity and use as an accountability and research measure. Research in Higher Education, 38, 1997, 667-698.

9. Writing ability is often tested by having groups of students write essays on a topic they know little about, without reference materials and within a narrow time period. The writing that these tests spawn is certainly not the best writing a student can do, nor is it even typical writing, or the kind of writing that we teach students to do in college. One can see why faculty would have little interest in the performance of students on this kind of meaningless task. Students aren't invested in the writing they produce for such "tests." This method—with no time to draft or revise—is in conflict both with research on what produces good writing and with the ways writing is taught.

10. Our goal was to have two papers, randomly chosen, from each of five classes within each discipline within each institution. This goal was seldom reached for a number of reasons including some disciplines not being represented at certain institutions and cases of too few senior level courses that assigned major papers.