A FIVE-YEAR STUDY OF STUDENT WRITING AT HAMILTON COLLEGE

(*Sharon F. Williams, Director of the Writing Center;*
Daniel F. Chambliss, Mellon Project Director)

1. Design and Implementation of the Writing Study

(Sharon F. Williams)

Overview of the Writing Assessment Study

The goal of the Mellon Foundation Assessment Project, a multi-faceted, five-year study of the Hamilton College class of 2005, was the assessment of the effectiveness of liberal arts education. The purpose of the Writing Assessment Study, one portion of the Assessment Project, was to assess whether the quality of Hamilton students' writing improves over time. The study's main focus was the evaluation of four years of written assignments submitted by a randomly selected group of 100 students in the class of 2005 (identified as "the Panel"). Starting in 2001-02 and continuing through 2004-05, Panel students were asked to provide an example of their best writing for each year; in addition, as part of a series of multi-topic interviews, Panel students were interviewed about their writing experience at Hamilton. In addition to papers written by Panel students, papers from other, non-Panel students also were used in the study. With student permission, faculty submitted entire class sets of non-Panel papers. Inclusion of the non-Panel papers increased both the sample size and the statistical power of the findings.

Over four years, an archive of student papers was developed. The final archive consisted of the following categories of student papers: original high school papers written by Panel students and provided by the Admission Office, self-selected papers submitted by the Panel students for some or all of the four college years, and class sets of non-Panel papers collected by faculty with student permission. All students submitting papers were asked to submit their best example for a given year of a standard essay, three to ten pages in length. The majority of papers submitted fit these criteria. Some papers included in the study did not match the requested form or length, particularly Panel papers.

Collection of papers

The process of collecting papers was a more complicated endeavor than originally envisioned, both in the amount of time and effort required to collect papers and the annual success of the collection effort (see "Limitations" discussion, below). After four years, the total number of students represented in the final data set was 541; the total number of papers collected was 1,100.

Preparation of papers for evaluation

To prepare papers for evaluation, all identifying information was removed, and each paper was given an eight digit code: six digits for the individual student ID, one digit for the category of paper (described above), and one digit for the year the paper was written and evaluated. An additional code for each of the six evaluators also was added. The coding system allowed for longitudinal assessments of individual students and cross-sectional comparisons of cohorts of students across years.

The scoring rubric

The evaluators used a nine-item evaluation scale, "the rubric," to score each of the papers evaluated over the four years of the study (final version attached). The language of the initial version of the rubric was preserved throughout the study; some additions and clarifications of rubric items were added each year. For example, when upper level papers began to be assessed, item #8 was added ("Author demonstrated complexity of intellectual reach"). With revision, the evaluators found the rubric to be flexible enough to apply to all types and levels of writing assessed. See below for further discussion of rubric changes.

Summary of papers evaluated

Approximately 1,100 papers were evaluated over four years. Almost all papers were written by Hamilton College students in the class of 2005; some senior papers from the class of 2002 were included in the first year of the study to make an initial comparison of freshmen to seniors. The 1,100 papers represented a wide range of undergraduate writing assignments, from high school essays to senior theses. Effort was made to include writing from a range of disciplines. For example, of the 186 senior papers evaluated in 2005, 70 were papers self-selected by Panel members, and 116 papers were collected from philosophy, history, economics, biology, and sociology classes.

Summary of papers evaluated each year

2002: 351 papers evaluated
73 first year Panel papers; 72 high school papers; 128 first year English 110/150, non-Panel papers; and 78 senior non-Panel papers, class of 2002

2003: 300 papers evaluated
60 sophomore Panel papers, 120 Sophomore Seminar non-Panel papers, and 120 sophomore non-Panel papers

2004: 228 papers evaluated
53 junior Panel papers; 96 junior non-Panel papers; 47 high school non-Panel papers (class of '05); and 32 non-Panel senior theses, class of 2004

2005: 189 papers evaluated
70 senior Panel papers; 3 Panel papers from previous years; and 116 non-Panel senior papers (entire senior theses were not included, but some thesis sections were included.)

The assessment process

For each of the four years of the study, six outside evaluators and the Director of the Writing Assessment Study met at Hamilton College for three days in June for an assessment workshop. The group first participated in a group norming session, consisting of reading, scoring, and discussing the scoring of one or more student papers and discussing the applicability of the scoring rubric. Over the next two days, the six evaluators read and scored student papers, each reader evaluating approximately 75 pages of student writing each day. The total number of papers read varied somewhat depending on the length of the papers to be read. Across the four years of the study, the papers became longer, and the number of papers evaluated decreased.

In this study, evaluators used a nine-item rating scale to assess student writing quality across five years. Despite limitations to the study's design, the findings provide an unusually complete picture of the quality of college student writing across time for students at a highly selective liberal arts college. Discussion of the specific strengths and limitations of the study follows.

Strengths of the Study Design

The evaluators

The outside evaluators were highly experienced writing program faculty and administrators with recognized expertise in the field; all were from institutions with student bodies and curricula similar to Hamilton's. All six of the evaluators participated for all four years, a factor greatly increasing the reliability of the scoring. The evaluators expressed strong loyalty to the study, in large part due to their recognition of the study's potential value for writing administrators and faculty at other liberal arts institutions. The evaluators formed a strong group bond. They recognized the value of the study; they shared a common interest in the teaching of undergraduate writing; and they enjoyed being together. The group's conversations extended far beyond the specifics of the study, and individuals continue to contact one other for advice on professional and other concerns.

Multiple years of study

Four years of student papers were needed to measure the learning of individual students across time. There were additional benefits to a four-year project. Data collection errors decreased over time. In addition, multiple years of working together allowed the evaluators to develop a sustained connection to each other and to the study. Finally, multiple years offered the evaluators the opportunity to modify the rubric to fit the changing nature of student writing across four years (see discussion "Rubric flexibility").

If the assessment had been scheduled for a single year, with multiple years' of papers collected in advance, perhaps the data collection errors could not have been addressed. The evaluators would not have connected as closely with each other and would not have felt the same commitment to the study, and evaluator burnout might have occurred if all 1,100 papers had been evaluated in a single year. In addition, the rubric would not have been tested and adjusted. For these reasons, it is likely that taking multiple years to evaluate multiple years of student writing produced more reliable findings.

Over time, the evaluators developed remarkable consistency in scoring. At the very first group scoring session, they produced a wide range of scores due to differing interpretations of the rubric items. Through four years of collaboration, they came to a shared understanding of what each item measured. At the time of the final group scoring session, scores were remarkably consistent.

Rubric flexibility

Because the study spanned four years and the scoring rubric had some flexibility in language and design, the evaluators were able to refine the rubric over time. As a result, they created a fair, usable tool for assessing undergraduate papers ranging from high school essays to senior papers in a number of disciplines.

The initial version of the rubric was a more mechanical, barebones rubric, although words such as "effectively" and "wise" allowed for some flexibility in scoring. During the initial meeting of the evaluators, the evaluators scored and discussed several student papers. As a result of this exercise, the initial version of the rubric was revised prior to beginning the assessment of the first year's paper set. Each year thereafter, evaluators made some small changes during the initial group scoring session. Care was taken to add only additional explanatory language; the original language of the first rubric was maintained across the four years. The rubric changes allowed the evaluators to assess complexities of composition that were not fully accounted for in the initial rubric. In other words, the evaluators molded the rubric over time to reflect writing professionals' understanding about how to evaluate writing, even when using a quantitative scoring rubric, which is not the typical way writing professionals evaluate writing.

For some rubric items, the evaluators agreed verbally what the item measured. For example, for item #7 ("Author developed an interesting theme or argument"), the final version of

item #7 has the identical wording as the initial version, but the evaluators agreed verbally on the application of the item. The group concurred that the emphasis should be on the author's *development* of an idea, not on the *interesting* aspect, which is subjective. The evaluators agreed that the mechanics of the punctuation of quotations fell under item #1, but the use of textual evidence in an argument fell under item #5 ("Author used evidence effectively"). One entirely new item was added two years into the study; item #8 ("Author demonstrated complexity of intellectual reach") was added as higher level papers were introduced. For item #8, the evaluators agreed that the emphasis was on the word *reach* to measure the writer's attempt to work with serious sources and to attempt significant analysis.

A-rhetorical assessment

The evaluators assessed each paper apart from the paper's rhetorical context. The evaluators had no knowledge of the assignment, the intended audience, the class, or the student (year, major, etc.). A significant advantage to this feature was that the bar for writing excellence was set quite high: papers had to succeed strictly on their own merits. The evaluators assessed only the writing; the texts had to be complete for the reader in order to meet standards. An additional advantage was that the evaluators were not responsible for weighing factors apart from the text, a responsibility that would have been a daunting challenge. Logistically, with the number of papers used in the study, it would have been very difficult to collect and manage all contextual information.

Range of disciplines

The focus of the study was assessment of student writing across four years, but the study also assessed writing across a range of disciplines. For example, as described above, the 186 senior papers evaluated in 2005 included 70 papers self-selected by Panel members and 116 papers collected from philosophy, history, economics, biology, and sociology classes. There was a comparable distribution of papers across a range of disciplines in the other years as well.

Funding

The Mellon Foundation committed significant funds for this study; these funds were necessary for a study as complex as this one to succeed.

Workshop design

The assessment workshops were well designed and allowed time for socializing and relaxation during the three days.

Limitations of the Study Design

Unequal disciplinary representation of student writing

Effort was made to collect papers from across the disciplines; however, many of the papers came from the humanities and, to a lesser extent, the social sciences. This outcome reflects the form and distribution of writing assignments across the disciplines. When designing the study, we chose to include only standard essays for evaluation, eliminating other types of student assignments such as laboratory reports, creative writing, etc.

To some extent, the unequal disciplinary representation in the study reflects the nature of the distribution of writing across the curriculum at Hamilton. All Hamilton students are required to take a minimum of three writing intensive (WI) courses in the first three years, and students take a mean of six WI courses. Each semester approximately 120 WI courses are offered from across the curriculum and across levels. In addition, many other, non-WI courses include writing assignments. For all of these reasons, we anticipated that more students would be writing papers in more departments than actually seemed to happen. We failed to anticipate that some students in some years would not have papers to submit that fit our criteria for submission. An additional limitation to paper collection is that it appears that less writing is assigned for the sophomore and junior years, particularly outside of the humanities.

Other difficulties with collection of papers

Other difficulties with paper collection included students studying off-campus, students leaving the college, lack of student response to requests for papers, poor photocopying, submission of the same paper for two categories, poor timing of requests for papers, and the submission of papers not fitting the study criteria (e.g., journal entries, film review). Of the 100 students originally selected as the Panel group, 82 students graduated from the college four years later. In the second year of the study, we discovered too late that 56 students had submitted two or more papers for multiple categories. In these cases, Panel submissions were kept and non-Panel submissions dropped. To some extent, the collectors of papers learned over time to avoid certain problems, and the collection process became more effective.

Sample size

The total number of students represented in the final data set was 541; the total number of papers collected was approximately 1,100.

The original design for the study was to collect across four years 400 Panel student papers, one paper per year for the 100 Panel students. Due to the difficulties in collecting papers described above, the final sample of Panel papers differed considerably from the original design. In the final sample, variation occurred in the number of papers of each type. For example, the final sample contained nineteen Panel papers for all five years (high school through college), 22

Panel papers for all four college years, 44 Panel papers for either three or four college years, and 52 Panel papers for three or four years with the high school essay included. Among the possible paired comparisons of Panel students by year, 51 Panel students submitted first year and senior year papers, the largest set of pairs for Panel students.

The limitations in the number of Panel papers were offset somewhat by papers collected from non-Panel students and especially by the collection of papers from pairs of years for the same student, Panel and non-Panel combined. In the end, it was possible to make same-student paired comparisons between freshmen to senior papers for 67 students; sophomore to senior papers for 90 students; junior to senior papers for 54 students; freshmen to junior papers for 45 students, and sophomore to junior papers for 58 students. The same-student paired comparisons increased the sample size and the statistical power of the study's findings.

A-rhetorical assessment

The a-rhetorical nature of the assessment process was a limitation as well as a strength of the writing study. The evaluators sometimes felt that it was more difficult to evaluate a paper not knowing its rhetorical context; this factor became more crucial with upper-level papers that involved more discipline-specific knowledge, analysis, and sources. The evaluators had to assume that the writer had followed the assignment, fully answered the question, used appropriate sources, etc. On the whole, the evaluators did not evaluate writing as a course instructor could and would.

Rubric limitations

Evaluation rubrics need to be sufficiently general to be useful for wide-ranging studies, but their usefulness decreases as student writers compose more specialized essays. At upper levels, students are not necessarily writing for a general audience in content or in form. Standards become increasingly more discipline-based, and general readers are less able to judge upper level assignments, e.g., judge the difference between a score of 5 and a score of 6 on a rubric item. When evaluating upper level papers, the need for greater knowledge of disciplinary conventions, of what counts as best evidence, and for a greater understanding of paper topics becomes far more important. Generalist evaluators maybe too forgiving, or too demanding, or look for qualities not central to the assignment. The evaluators expressed concern that their effectiveness as readers sometimes was compromised when reading upper level papers outside their field of study. Especially as they evaluated progressively more advanced papers, the evaluators became more aware of their dependence upon their professional instincts and their intuitive sense of the logic and coherence of good writing.

This difficulty in scoring higher-level papers may account for the lack of statistically significant difference between junior and senior year papers. The types of papers collected and evaluated for senior year, mostly short papers rather than senior theses, may also have affected this finding. Seniors may not put their best effort into shorter assignments.

A related concern was how accurately the rubric measured student improvement over time. There is the danger of over-simplification when trying to use a rating scale to measure the complex conceptual task of learning to write well. Related to this is the concern whether the rubric could account for students' efforts to meet more complex challenges as they advance through levels of study. Could an individual student's scores over time change little while, in fact, the student is achieving gains as a writer? As with the a-rhetorical nature of the evaluation, the bar for demonstration of writing excellence was set high, which may have suppressed finding some actual improvement in student writing.

Despite these real concerns about the elasticity of the rubric, the evaluators expressed confidence that that they were able to make reliable and valid judgments about student writing across time. They found that the rubric allowed for the evaluation of the student's ability both to compose correct, clear text and to meet the challenges of higher-level assignments. The earlier items on the rubric measured the surface features of writing, while later items measured intellectual reach and maturity.

The rubric should have a NA ("not applicable") option for each of the items. The rubric reflects assumptions about 'typical' papers that sometimes do not apply (e.g., item #5, "Author used evidence effectively," is not applicable to papers with no references to outside sources).

Findings of the study

See separate report for a summary of findings.

Acknowledgments

Many people contributed to the design and the implementation of this study. I would like to single out for special thanks the evaluators of the Writing Assessment Study for their enthusiasm and commitment: Margaret Darby, Colgate University; Holly Davis, Smith College; Katy Gottschalk, Cornell University; Peter Grudin, Williams College; Joyce Seligman, Bates College; and Beverly Wall, Trinity College. I would also like to thank Dan Chambliss, the Director of the Assessment Project, and Jennifer Borton, Assistant Professor of Psychology, for analysis of data.

2. Summary Results of the Writing Study

(Daniel F. Chambliss)

1. Our *overall conclusion* is that Hamilton students do, indeed, improve in their writing from high school through college until their senior year. The improvement seems to be greatest in the move from high school to college, but there are demonstrable improvements each year thereafter at least until the junior year, although the size of improvement for any individual year is not great. Within the college years the improvement is "soft", sometimes producing results that are not statistically significant for any particular year. From the junior to the senior year, we found no improvement.

These findings hold true across the entire range of measured items in our rubric (nine different factors), although the gains were greater in some areas than in others. The major gains seem to come early in the college career. We made comparisons in various ways, both longitudinal and cross-sectional, among various samples and configurations of our data, and in every case we found improvement, although in some cases change was not statistically significant.

In general, our conclusion, then, is that students exhibit noticeable improvements in writing from high school to college and over the course of their college career; the gains may not be huge, but they are clearly detectable even by outside "blind" readers who do not know what they are reading, nor the purpose for which the paper was written.

In effect, *very* roughly speaking, an educated outsider could be handed five papers by the same student, one paper from each of the past five years; and the reader could (typically) sort the papers into the correct sequence, except for the junior versus senior papers.

2. Our findings are based on several *different analyses* of the data: (a) Longitudinal comparisons using only random-sample panelists who submitted papers for all five years (including high school) of the study. There were only 18 such people. (b) Pair-wise comparisons for any individual who submitted papers in more than one year – such comparisons are therefore also longitudinal, comparing an individual with herself. These comparisons take advantage of the fact that we have such papers both from panel members and from other students whose papers were submitted *en masse* by professors teaching certain courses. The pair-wise comparison analyses were made possible by our long-term strategy of creating an archive of over 1,000 student papers in the course of our study. (c) Cross-sectional analyses comparing groups (for instance, seniors and first years) of papers written during the same academic year, that is, different individuals being compared in class-year groups; these analyses were done in the study's first year, when longitudinal data were not yet available.

3. Our *sampling assumptions* in this study were fairly conservative. First, we started with a random sample (the panel), and made an effort to collect papers from those students, individually solicited, in each of four years. Our collection methods improved over the course of

the study; by the fourth year, we were getting papers from a fairly high percentage (78%) of the panelists remaining at Hamilton. Second, in the early years of the study, when collecting high school papers for the entire Class of 2005 and many first year papers for that class as well, we leaned in the direction of getting the best possible papers: (a) The high school papers were submitted with college applications, and so presumably were the best papers the students had. (b) Many of the first year papers came from English 110 and English 150 classes, which are significantly focused on good writing, and in which the students' papers can reasonably be expected to be fairly carefully done. In later years of the study, many of the papers came from any class with high proportions of students in the Class of 2005, so that writing quality might be less of a focus for the students. This approach, which we took quite deliberately, was established to "make it difficult" for students to demonstrate improvement in their writing over the course of a career, at least for non-panel papers. Finally, in our data, some results are statistically significant and some are not, but even when improvements are not statistically significant they quite consistently fall in the direction of improvement. The only exception to this seems to be the "junior to senior year" findings, in which there is no improvement.

4. Based on comments from the writing evaluators, the *rubric* may become less reliable, in the technical sense of that word, with more advanced students. Evaluators commented that in highly specialized areas (papers done probably by more advanced students), the rubric, designed to evaluate general writing skills, became harder to apply; there were more ambiguities in evaluation for such papers. This could have blurred differences between the junior and senior year, but we have no way of knowing.

5. *Future analyses*: Now that the five-year longitudinal evaluation of students' writing has been completed, we would like to correlate these data with other information about these students. For instance, we conducted interviews with most panel students in most of their years at Hamilton, often asking them about writing: how it was taught, what they felt they had learned, and so on. We can now integrate that database with the database of the objective writing study, and discover how self-report relates to objective progress. We can also link these data to student transcript information, GPA, writing intensive courses taken, and the like. So far as we know, no such data – longitudinal, combining objective with subjective analysis, and drawn on a large sample – have previously been available at a college such as Hamilton. Lessons from the analysis of such data could be exceptionally valuable.

3. Scoring Rubric
Hamilton College

Below are nine statements that describe characteristics of effective writing.
A score of 1 indicates that a paper *completely fails* to meet the criterion.
A score of 7 indicates that a paper *completely meets* the criterion of evaluation.

Please evaluate the paper according to these criteria.

1. **Writer edited to correct misspellings and other obvious mechanical errors.**
(spelling, capitals, apostrophes, mechanics of documentation, punctuation of quotations, …)

1 *2* *3* *4* *5* *6* 7

2. ***Writer followed standard conventions of grammar and usage.***
(grammatical rules, general punctuation, possessives, tense, dangling modifiers, correct word choice, e.g., affect/effect, …)

1 *2* *3* *4* *5* *6* 7

3. ***Writer omitted needless words and chose words wisely.***
(concise expression, correct use of idioms, effective diction, appropriately constructed sentences, including appropriate integration of quotations,...)

1 *2* *3* *4* *5* *6* 7

4. Writer developed unified and coherent paragraphs.

1 *2* *3* *4* *5* *6* 7

5. Writer used evidence effectively.

(appropriate evidence, sufficient evidence, clear identification of sources, responsible attribution of sources, evidence analyzed, evidence and claims logically related, sound logic, multiple points of view considered if appropriate, …)

1 *2* *3* *4* *5* *6* 7

6. Writer clearly communicated the purpose, design, and major points of the paper.

1 *2* *3* *4* *5* *6* 7

7. Writer <u>developed</u> an interesting theme or argument.

1 *2* *3* *4* *5* *6* 7

8. Writer demonstrated complexity of intellectual reach.

(critical thinking, insight, originality, ambitiousness, …)

1 *2* *3* *4* *5* *6* 7

(Suggestion: add NA *to scale.)*