Not Again:

The Humanities Assessment

February 31, 2019

Harold Washington College
Assessment Committee

TABLE OF CONTENTS

Introduction

In the fall of 2016, Harold Washington College's Assessment Committee conducted a college wide General Education assessment of student attitudes and abilities in regard to learning outcomes associated with our Humanities appreciation objective. The use of the measure followed approximately four months of planning and revisions, integrating recommendations from the last Humanities assessment conducted in 2006 and committee procedural changes, resulting from committee learning about effective practices in the intervening ten years. A small sub-committee, led by Jeff Swigart and Erica McCormack, began the work in April by starting with the measurement model (*Hummmm)* originally developed in-house by faculty member Amanda Loos in concert with the committee in 2006. Revisions included the selection of a new set of artifacts, the reduction and integration of required essays, and the addition of a new philosophy option, as well as a new testing design and procedure.

The measure was piloted in July 2016 with summer school classes, and conducted as a full-scale assessment over 4 weeks in fall 2016 between weeks 12 and 16. Responses were rated through the spring (a process that was completed in fall 2017) by an interdisciplinary team of faculty members and the data was analyzed by Phil Vargas, with an additional contribution from Fernando Miranda-Mendoza, in accordance with questions posed, over the course of the following year. This report, a labor of love, tells the story of our efforts and our grand, spectacular failure. If it tells the story of a city boy born and raised in south Detroit, that is merely an accidental, but delightful, bonus of this glorious journey.

Key Findings

1. We should not do this assessment again (see findings 2, 5, and 6, below).
2. The sample was NOT representative of the general education students at HWC, and so the findings are not generalizable.
3. Ratings showed poor differentiation in the application of the rubric (p. 14).
4. Indirectly: Humanities General Education requirements are difficult to source and, when sourced, poorly explained or difficult to understand (see note 17, p. 17)
5. Students who participated* did not, generally, demonstrate mastery of the outcomes measured, even when the sample was restricted to students who had completed their A.A./A.S. Humanities General Education requirements; students scored best on "Clarity and Correctness," while doing the worst on Analysis and Development (p. 19).
6. Students who completed the survey* are more likely to identify as "Writers" than any other category of artistic disciplines, and in significant numbers (43%), even while producing less than proficient quality writing (p. 19).

   *To reiterate, because of Finding 2, Findings 5 and 6 are NOT generalizable to the HWC student body for reasons explained in this report, and so of limited interpretive value for general education assessment and potentially misleading. Other findings from the survey regarding students' attitudes and experiences with the arts (generally highly positive), have been omitted from the report for the same reason.

Primary Recommendations

1. We should not do this assessment again—never again;
2. It would be good to clearly explain Humanities General Education requirements and make the information readily available;
3. We should seek a different means of measuring student achievement, especially among Humanities requirement completers, to corroborate or refute the poor showing in this measurement.
4. We should radically rethink general education assessment.

Background & History

Back in 2006, the *Hummmm* assessment required that classes be brought to a reserved computer lab where students received numerically coded and matched surveys and test booklets. Students completed the survey on "Bubble forms" before turning their attention to a computer screen, where they clicked through a PowerPoint file that directed them to make a choice. Students were presented with a set of three themed artifacts, listing the title, name of the artist, and the artistic discipline, and they then selected one artwork to analyze and interpret. Students could choose a poem, an image of a visual artwork, or a musical piece, and, upon making their choice, began the direct assessment by answering each of four questions in brief essays handwritten in their test booklets. The four essay questions were the same for all students, regardless of the artifact being analyzed and corresponded to the four learning outcomes related to 1) articulation of an emotional response, followed by 2) the interpretation, 3) analysis, and 4) evaluation of the piece. Responses were then reviewed and matched up later to the data from the survey.

The report (completed 2007) made some recommendations related to the structure and content of the measure based on the observations of the raters, including "the chosen artifacts should either be equally famous or equally obscure, and the number of essays students are asked to write should be reduced."[1] The report also recommended "that the Assessment Committee aggressively pursue the automation of the Humanities Assessment Tool."[2] Other recommendations, such as one related to inter-rater reliability standards and another about specific data points, were moot points nine years later either due to changes in committee practices and personnel, especially the presence and expertise of our data analysts, or better access to data through Open Book.

The 2006 assessment proved more valuable as a learning experience for the committee than as an effective measure owing to some data collection problems in relation to our (then necessary) reliance on student self-reporting of their academic history. Student confusion about one of the questions and resulting inaccuracies in the data set led to subsequent analysis problems, severely constraining the usefulness of the data set, along with a fairly debilitating ignorance of some key practices of data analysis on the committee as a whole.

Student learning-related recommendations focused on student vocabulary issues, particularly related to student descriptions of their emotional responses and encouragement of student use of technical, disciplinary vocabulary, and developing informational materials related to encouraging encounters with and appreciation of the arts.[3] Multiple humanities faculty members' tenure projects and activities related to this recommendation, and the administration has continued, to a degree, its support for art installations, musical performances, guest speakers, and more; however, the intervening decade has presented funding, and other, challenges, that have significantly strained the college's capacity to support and encourage student encounters with the arts.

---

[1] *Hummmm Report*, p. 103.

[2] Ibid

[3] From page 104 of the *Hummmm Report*: "We note student shortcomings in the following specific areas: Student responses rely on a very narrow emotional vocabulary in discussing how works make them feel; Student responses often mistakenly equate the idea that aesthetic judgment is subjective with the idea that aesthetic judgment is impossible; Students seem to lack the technical vocabulary to adequately perform the more sophisticated tasks in the assessment; Student responses do not reflect adequate vocabulary for discussing the component pieces of a given work of art; Student responses often omit textual evidence that would lend support to their reasoning.

"In response to these shortcomings, we recommend that instructors at HWC spend additional time working with students to develop their vocabulary as it relates to affective and interpretive tasks; to increase student opportunities to provide evidential support for their reasoning; and to introduce students to a wider array of critical tools for explaining, interpreting, evaluating, and contextualizing humanistic artifacts. Beyond these core recommendations, we urge the humanities department to create a "humanities toolkit" of vocabulary words, critical tools, and examples of evidential support to be used by instructors in the department and across the curriculum when applicable to ensure that all of our students are more successful at achieving desired outcomes."

Methodology

Frame

The sub-committee's first task was to revisit and review the General Education Objective and associated learning outcomes, revising as appropriate. The committee affirmed that, in regard to Humanities and the Arts, HWC's General Education Objective is for students be able to demonstrate their ability "to understand and appreciate the arts, literature, history, and philosophical systems of major world cultures."

In order to attempt to measure students' abilities in this area, the committee revised and approved the following set of associated learning outcomes in May 2016:

> Upon completion of the General Education Requirements in arts and the humanities, students will demonstrate the ability to:
>
> 1. Analyze artifacts (i.e., artistic products of human creativity) across multiple cultures for evidence of historical period, relations to major artistic movements, and basic elements, techniques, and conventions of the related discipline(s);
> 2. Evaluate artifacts according to criteria provided by disciplinary experts as well as students' own invention;
> 3. Interpret artifacts for meaning, justifying their view using evidence from the "text" of the artifact (e.g., convention, technique, technical vocabulary, theoretical lens);
> 4. Communicate clearly and effectively an appreciation for artifacts.

The sub-committee then began work on developing a tool for the effective measurement of student learning and proficiency regarding these abilities with the hopes of learning answers to all of the following *initial* questions:

1. A) What are students able to demonstrate with respect to their ability to analyze, interpret, and communicate about creative artifacts from disciplines within the Humanities and their own understanding of those, and B) how does this ability manifest at various levels of educational attainment, and C) can students demonstrate appropriate proficiency at the completion of their General Education?
2. What are students' A) self-perceptions and B) attitudes, especially in regard to their engagement with the arts, in relation to various topics, and C) how do they change (if at all)?
3. What sorts of experiences with the arts have students had before, during, and after their time at HWC?
4. Are there strong correlations between the number or type of humanities (or other) classes/subjects and proficient performance once other factors are controlled?
5. Are there strong correlations between the number or type of humanities (or other) classes/subjects and student attitudes, once other factors are controlled?
6. Are there strong correlations between student self-perceptions/attitudes and proficient performance (again, controlling for other factors)?

Initial Tool Design

In the spring of 2016, committee members invited the Humanities department to provide assistance with the General Education Humanities Assessment, planned to occur in the fall semester. Initial plans included use of the same locally-developed measure that had been used in 2006 with some procedural changes, different artifacts, and minor question revisions.

In the subsequent months, Jeff Swigart, Erica McCormack and Dave Richardson worked together to develop an updated measure by first reviewing the report and its recommendations. The review led to their decision to combine the cognitive tasks into a request for a single response, rather than four separate responses, and guided the selection of artifacts. The team also engaged in extensive discussion about the costs and benefits of procedural changes related to

allowing students to complete the assessment remotely and using Google Docs for the delivery mechanism, which would be consistent with the committee's recent practice as well as recommendations from the 2007 report.

In light of their review and subsequent discussions, between May and July, which ultimately came down to Swigart's enthusiasm for Google Docs and the conveniences of digital deployment and remote access for students (and his deep antipathy for the intense logistics of managing class by class assessment) versus Richardson's vague misgivings about various potential effects of the changes that, really, even he could not express full confidence about and McCormack's thoughtful epistemological restraint, Swigart developed a "Google Docs" version of the measure, allowing for the digital deployment and collection of responses. The new version also flipped the order of the tasks so that the survey *followed* the written response, per current committee practice, though this was changed twice more (once for the pilot and then back for the actual assessment, see below). McCormack and Richardson worked on selecting artifacts, in consultation with Humanities Department members, and revising the direct-measure portion of the assessment tool. They also proposed adding a fourth artifact—a philosophical text excerpt—and drafted an initial set of questions for the data analysis (the proposal was later accepted by the committee). For reasons that no one can remember clearly, likely arbitrary and accidental—it's possible that someone was looking at a Chicago flag and either thought about the absence of the fifth star that would have marked the 2016 Olympics, had history gone a different way and/or noted that of the four stars on the flag, three marked tragic events, which would have been "downers" to select—they selected the "1893 Chicago World's Fair" as a thematic/period anchor for their selections.

Selections, approved by acclimation noted through an absence of resistance verified by email non-responsiveness, were intended to be roughly equally known/unknown, in pieces that may have been, but probably weren't taught or assigned in classes. The artifact options were:

- Fine Art (Visual): "Women of Plymouth" by Lucia Fairchild Fuller (1893)[4]
- Music: "Maple Leaf Rag" by Scott Joplin (1899)
- Philosophy: "The Ethics of Democracy" (Excerpt) by John Dewey (1888)
- Poetry: "By the Stream" by Paul Laurence Dunbar (1896)

There was some concern about the length of the philosophical excerpt, but after editing it down to a little over 1000 words, the team estimated the reading time to be 3 to 5 minutes for an average reader, especially given the relatively low lexile density of the first few paragraphs, which rate out at "Junior High" levels. The selection is not simple, though; the second to last paragraph rates as "upper high school," and the last as "college level." Consequently, it was estimated that it would take about as much time as the music selection (3 minutes, 14 seconds), and was thus appropriate.

The measure was designed so that the first page opened with a brief introduction and our standard consent statement. Students were prompted to announce their consent (if they so chose) and enter their student ID along with their answer to a question about their location (i.e., in class, on campus, not on campus). As stated above, Swigart's first design placed the survey after the analysis, in the hopes of avoiding any survey fatigue, and there was some discussion and consideration of effects, and in the end, Swigart re-placed the survey at the beginning, first since that's how the measure was originally designed, and second out of an abundance of caution on account of some concern raised by Richardson that student experiences doing the analysis might color their attitude survey responses and affect whatever comparability they may have to the previous iteration.

On the second slide, students began the attitude survey, though another significant change from 2006 was the elimination of a section related to student experiences of the arts before and during their time at HWC. Upon completing the survey, students came to a slide inviting them to choose one of four selections featuring descriptions of the artifacts (e.g., "A painting by…," "A Musical Composition by…"), in lieu of the titles, along with the creator's name

[4] Changed after the pilot, see "Fall Committee Revisions" below.

and the year of its creation. This particular design for the selections won out in order to emphasize the discipline over the subject matter. Once they made their choice, the next slide had the complete information, along with the artifact and the task instructions. After completing their analysis, students were directed to submit their responses.

According to TeachingCopyright.org[5] all works published before 1923 are public domain. So, all four of these artifacts are public domain and can be freely shared in our survey without concern. The initial design was finalized in mid-July.

Pilot Assessment

Jeff Swigart managed all aspects of the pilot, sending out invitations to Assessment Committee faculty members who were teaching summer school on July 19th. Seven faculty members agreed to participate, posting the invitation and link for 1 or 2 classes each. 120 students completed all or most of the measure, which was enough to get a sense of how it worked.

Jen Asimow said that her students described being surprised, "shocked" actually, by the analytical writing task. She also said that of the five who took it, all of them understood that they were supposed to choose once and without looking at the other options, but two of the five students looked at them all before choosing. The feedback from other students who took the assessment was broadly positive, though, with many expressing gratitude and a lingering interest in the works or disciplines they analyzed. Overall, the pilot was deemed to have gone well by all involved, revealing no significant unknown problems and a few possibilities for helpful adjustments.[6]

Fall Committee Revisions

Data analyst and committee member Phil Vargas proposed minor changes to select survey questions (adding a neutral category for numbers one and two and the sub questions) on the survey and the suggestion to change "I am more likely to" in the squib for numbers three and four to "I now." (He also bet a beer in an email sent on 10/8/16 on his related hypothesis that "attitude will be more easily shifted than behavior" and Richardson took him up on it and intended to order his in a bucket before learning that Phil's hypothesis was not refuted by the data, though neither was the null hypothesis. So there.) The order of the sections was also changed, to place the analytic task in front of, rather than after, the survey, and a question about student motivations for their selection was added as "Part Three" along with the button to submit the response.

One major revision from the pilot version was McCormack's decision, with the support of her colleagues, to switch out the Visual Arts artifact to "In the Nursery" by Mary Fairchild MacMonnies (1897-1898), in order to "give students more opportunity to discuss stylistic elements" and to "avoid some problems we encountered in the other by making the subject contemporary with the artwork's creation (rather that depict a past event)." The majority of revision discussion, which took place at committee meetings in late October 2016, focused on discussion about the wording of the instructions for the written response and the response space, with the intention of encouraging students to write more than they otherwise might be tempted to provide and think sufficient. In the 2007 *Hummmm Report*, the committee noted, "The space provided to write responses in seems to have in some cases dictated the length of the responses."[7]

Various proposals were made and attempted and some even partially adopted, but in the end, the best solution to this particular problem came when Swigart changed the form to the old version of Google Docs and the "line" for student response became a text box. (We are not sure this actually happened for the final version, though; no one can remember, and the author of this report no longer has access to the form.) Committee discussion also led to the

[5] Accessed by Swigart at some point in 2017.

[6] Unfortunately, one structural problem with the design of the measure went unnoticed during the pilot and survived into the actual assessment in the fall, only coming to light in the analysis phase when Phil Vargas noticed that students who completed the Poetry analysis, were sent directly to "submit" rather than to the attitude survey questions, leaving us with 540 survey data sets out of the 818 participants. D'oh!

[7] *Hummmm Report*, 2007, p.16.

addition of a rating option stating the submission was "not long enough to rate" and a pair of checkboxes for "Indication of Possible Use of Outside Sources," with one for "Complete" and one for "Partial." Finally, there was a proposal to include a question about whether students have encountered works from outside of their own culture/comfort zone in their Humanities or other classes; however, the team decided that including it might entail a restructuring of other questions and lead to further delays. Consequently, they agreed to do include it as a recommendation for the next Humanities Assessment (see Recommendations, below).

Work on the Rating Rubric had commenced in early September with some initial revisions made by McCormack and Richardson to the rubric used in 2006, in light of the reduced number of tasks on this assessment compared to the original. They attempted to apply the rubric to a few examples drawn from the Pilot, and those efforts led to a few more minor revisions before being presented to the full committee for review and discussion (along with the final version of the survey) on Wednesday October 26th. Approval was delayed briefly, however, by rich discussion, even amid the hub-bub and hoopla of the Cubs World Series run, but the finalized, approved version of the assessment was ready for deployment Tuesday, November 8th. The first invitations to participate went out to committee faculty on that same day, and to selected faculty ("Friends of the Committee") on the following Monday. On November 17th, one additional change to the language on the front page of the survey was required to address important concerns raised by faculty member Kristin Bivens in relation to clarifying students right to non-participation and the committees' use of the data as well as fixing wording related to the confidentiality of the data. With those final changes, the measure was complete. And there was much rejoicing.

Implementation

By the end of Week 12, faculty members across the college had emailed or posted (in Blackboard) an invitation and link to the measure. They were also sent a list of suggestions for maximizing student participation and effort that read:

You may choose how to do this, though here are a few options to consider:

- *Email students the link https://goo.gl/forms/z0xg8RCAHwO40YCG2 or post it on Blackboard, or both, and ask them to do it during their own time outside of class.*
- *Allow students to complete it in class if you are in a computer lab.*
- *Let us know if you'd like to have your class take it in a computer lab (we have one reserved for this purpose), along with the day and time, and we will make arrangements for you to do so.*
- *Give them a solid deadline. We'd like it done for sure by Friday, November 25th.*
- *Please emphasize to your students the importance of them giving their time and focus to this survey so that we can get valid data. Also emphasize the importance of not using outside sources, such as Wikipedia, as they type out their writing responses. Their survey responses will be analyzed totally anonymously, so there is no reason for them to feel like they need to feel stressed about doing well.*
- *Please tell them they'll be automatically entered into a raffle for five $20 Potbelly gift cards, which we'll draw at the end of the semester. You are also welcome to offer them your own incentives, such as extra credit, but that's your call. If you do offer extra credit, tell your students to screenshot, phone pic, or print off the confirmation page at the end to show you as confirmation.*

As the last one suggests, to encourage participation, Swigart ginned up a pair of raffles for participants, promising to award two faculty volunteers who willingly posted the link a chance to win $20 Potbelly gift cards, to be sent out in December 2016 and a separate raffle for participating students that would have five winners of $20 Potbelly gift cards. The committee's (ambitious) goal was to have 1000 responses by December.

Swigart resent the invitation to the faculty at large on November 30th and on December 1st suggested that committee members send out 'targeted invitations,' writing, "Our numbers are a bit low, still at only about 400 students." Multiple surveys, including a major assessment on student perceptions of online learning, competed for student attention and possibly contributed to survey fatigue and lower than anticipated participation rates, along with the rather demanding nature of the measure. Participation numbers were monitored through the end of the semester. There were 740

responses on Tuesday the 13th, and the survey was finally closed at the end of the semester, Saturday, December 17th. The final tally of participants was 818 responses for analysis--lower than our targeted amount, but higher than the number rated in 2006 (665)[8].

Rating the Written Responses

The new rubric was reviewed and approved along with the measure having been adapted from the one used in 2006 for the analysis and interpretation questions, but simplified for digitization and ease of use. Late additions included checkboxes for "Too short to rate" and suspected "Indication of Possible Use of Outside Sources" (one for "Entirety" (ended rating) and one for "Partially" (continued rating)).

As Swigart worked to drum up responses, McCormack and Richardson developed support materials and plans for the rating team, including a pre-rating norming process. Because she is a goddess of persuasion and so deeply respected by our administration team, Committee Chair Carrie Nepstad, a.k.a. "Carrie Poppins," was able to secure approval for 30 hours of release from advising (15 in the spring of 2017 and 15 more in the fall of 2017). Raters were recruited from the committee and invited volunteers from the college faculty at large and composed an interdisciplinary team of 12 raters from seven different departments and 10 different disciplines.

List of Raters

1. Marcy Henry (Humanities (Multi-Disciplinary))
2. Sarah Kakumanu (Math)
3. Yev Lapik (Biology)
4. Mick Laymon (Humanities (Music))
5. Erica McCormack (Humanities (Fine Arts))
6. Carrie Nepstad (Social & Applied Sciences (Child Development))
7. Dave Richardson (Humanities (Philosophy))
8. Amy Rosenquist (English)
9. Matt Shevitz (Humanities (Music))
10. Ray Tse (Physical Science)
11. Paul Wandless (Art)
12. Allan Wilson (Physical Science)

In keeping with a previously effective practice, developed by former committee member Lynnel Kiely for the Social Science assessment of 2010, and in response to some concerns voiced by potential volunteers about rating responses for unfamiliar disciplines, McCormack and Richardson drafted four artifact guides to help the raters understand the pieces and their technical aspects, providing suggestions for what to look for in the essays regarding music, poetry, art, and philosophy. The guides all featured sections presenting the Content/Subject Matter Summary and Formal Elements that would be found in a high quality answer, as well as relevant biographical/historical context. Some of the guides also included information about how answers might be expected to go wrong or what kinds of moves might be typical of novice/uninformed responses or the elements whose presence indicates the likely use of outside sources (see Appendix).

With the end of the semester upon us, and rating planned to occur over the winter break if possible, time was short for both rater norming and turning the data around for distribution to raters. Norming sessions were scheduled for two separate hour-long Wednesday meetings in Weeks 15 and 16; participating raters were required to attend one and invited and encouraged to attend both. McCormack and Richardson selected a few student responses from completed assessments that they deemed would be informative for understanding and talking through the use of the rubric and led both of the sessions. Raters were provided with the rubric and artifact guide ahead of time, and then, in session, with four responses, one for each artifact. Everyone then rated the response according to their own proclivities and understanding, and then those ratings were compiled on easel paper to show where there was consensus and variation. Discussion then ensued according to rater questions and interest, with additional selections rated and discussed according to the needs of the group.

As those were happening, Swigart, after learning somewhat tardily that his original plan for distribution would need some revision to ensure that each batch of essays graded by each instructor would then be redistributed for the second round of rating (rather than merely passed along as a group to a second rater), barricaded himself in a quiet basement closet, jacked up some emo music, turned on his disco ball and got to work. He spent a big part of finals week furiously reformatting and spread-sheeting and randomly distributing the student responses into electronic documents for

[8] *Hummmm Report*, 2007 (p. 4).

distribution to raters before they left town for the holiday break. He also spent time time he didn't have converting the rubric into a Google Form to ease and automate the data collection. Somewhat miraculously, he managed all of the challenges he faced that week and sent emails on December 28th with the student responses for rating in Word and Excel formats, along with instructions.

The rating process worked pretty much as planned—individual raters used various processes for data entry, but by and large there were no technical difficulties or major flaws with the process. As of January 25th, three of the raters had finished their work (all from the Humanities department—whoop, whoop!), and by March 22nd, everyone was done. Unfortunately, approximately 70 responses were rated only once (rather than the requisite two times by different raters), and so, to speed the completion of the first stage, those were split between a trio of volunteers and completed within a fortnight (plus one week).

Initial analysis showed that about 130 of the 818 student responses (15.9%) were rated differently enough that they required a third read, and so those were emailed out on May 13th to the ten raters who had not volunteered previously. Those ratings were completed sometime between May and November, for a grand total of 1636 ratings, and data analysis began in earnest.

Research questions

A primary concern of the planning sub-committee was to discover what *student abilities* are with respect to analyzing and interpreting creative artifacts and then arguing for their interpretation in writing. We also wanted to find out how those changed and, hopefully, developed, over the course of students' time at HWC and whether students are acquiring the abilities promised in the General Education learning outcomes.

Additionally, we inquired about what *student attitudes* are with respect to the arts and themselves, and their learning experiences, and whether they are shifting in positive directions over time. Finally, we hoped to learn about *student experiences of the arts* while at Harold Washington College and while studying humanities—in other words, whether students are directly experiencing the arts during their time with us.

Areas of interest that arose beyond those big three (really seven) include a question about student engagement and discipline selection (i.e., why did they pick the artifact they chose?) three proposed by our intrepid data analyst, Phil Vargas, who proposed a word frequency analysis, a search for class history and terminology use correlation, and a study of the ability of raters to correctly identify plagiarized responses, the last of which turned out to be a project for another day. A second, ultimately unexplored possibility was to explore the effect of location (e.g., online, classroom, home) on the scores and lengths of the student responses, to gather information about the possible effects of the change in method.

The final list of analysis questions reads as follows:

1. A) What are students able to demonstrate with respect to their ability to analyze, interpret, and communicate about creative artifacts from disciplines within the Humanities and their own understanding of those, and B) how does this ability manifest at various levels of educational attainment, and C) can students demonstrate appropriate proficiency at the completion of their General Education?
2. What are students' A) self-perceptions and B) attitudes, especially in regard to their engagement with the arts, in relation to various topics and C) how do they change (if at all)?
3. What sorts of experiences with the arts have students had before, during, and after their time at HWC?
4. Are there strong correlations between the number or type of humanities (or other) classes/subjects and proficient performance once other factors are controlled?
5. Are there strong correlations between the number or type of humanities (or other) classes/subjects and student attitudes, once other factors are controlled?

6. Are there strong correlations between student self-perceptions/attitudes and proficient performance (again, controlling for other factors)?
7. Why did students select an artifact from a particular discipline for interpretation and analysis?
8. Were there significant differences in performance or attitudes among students who made different choices?
9. Is there a correlation between specific word usage and course history?
10. Is there a correlation between specific words and scores?

Data analysis was delayed for nearly a year while awaiting the completion of the ratings (see "Rating the Written Responses" above), but the sub-committee received Phil Vargas' initial analysis on November 17, 2017, featuring a class specific history point biserial correlation with the survey affective questions. A second and third set of files was sent to Richardson and McCormack for interpretation on March 12 and 13th, 2018. Subsequent meetings, questions, and investigations led to multiple revisions, additions, and corrections—all borne by Phil with quiet dignity, grace, and inhuman patience so that the final data analysis, sent on April 24th, 2018 included a revised analysis of the scoring breakdown (separating cohorts by English completion as well as General Education Humanities requirement completion via course history), student attitude and arts experience responses, a comparison of mean attitude scores from 2006 to means from 2016, complete Attitude data from Question 4 of the measure, and behavior data from Question 5 of the measure, plagiarism numbers, and a revised (reversed?) word frequency analysis, working from a list of technical terms supplied by McCormack and Richardson (see Appendix). Richardson and McCormack worked to interpret the data during that summer and early fall, developing additional questions, particularly about the sample, whose answers proved elusive until well into the spring of 2019 when the painful (on account of the thousands of work hours devoted collectively to the project) conclusions were finally no longer avoidable. Unfortunately, due to problems with the sample, the only research questions that could be productively and usefully answered were 1A, 1B, 1C, with the caveat that the informativeness of these findings is strictly limited to the population of students who completed the assessment. Arguably, questions 4, and 6 (answered below in **Findings: Skills Related** (p. 15)) fall into this category as well, so they have been included. The others based on or related to Attitudes, Experiences and Behaviors, Student Choices, and Language, could be answered for the students who took the exam, but would be of merely historical and anecdotal interest and so are not detailed in this report for two reasons: first, so as not to create the impression of findings about our general student body or other confusions (see "Sample Validity" below (p.12)), and second, since the method used to collect that information (a likert scale survey) is not controversial nor of particular interest to the committee.[9]

[9] One interesting aspect of the survey for this assessment provided a result for future exploration. We added a "Neutral" category for the questions this time, whereas in 2006, we forced a choice on students by including only four possible answers (Strongly Agree, Somewhat Agree, Somewhat Disagree, and Strongly Disagree). Initial comparisons of the survey results showed *even more* positive attitudes on the part of students towards the arts and their experiences in the classes than the highly positive 2006 results. Initial analysis showed similar levels of response among the positive categories for the questions, with declines in the "Somewhat" negative answers, suggesting a possible interpretation that students in 2006, when feeling neutral but faced with a forced choice, defaulted to the more negative answer. This hypothesis is far from proven but offers some interesting possibilities for exploration and potential usefulness if replicated.

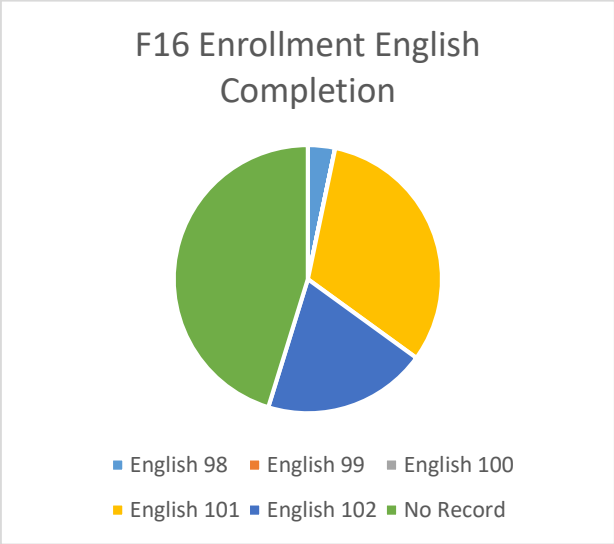
Findings: MEASURE RELATED

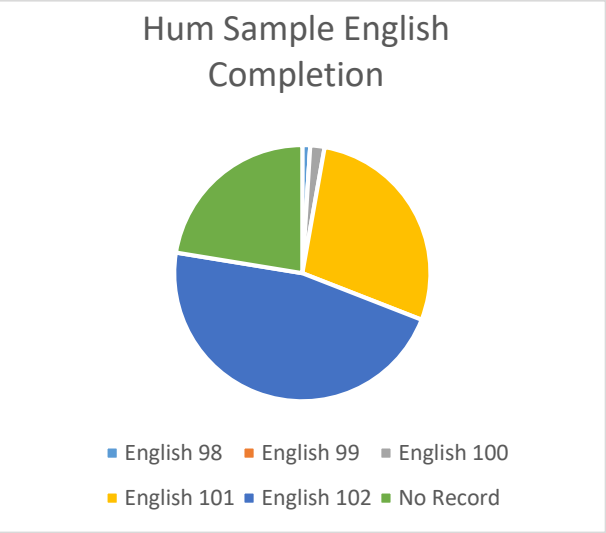
Sample Validity

Our method relied on self-selection, as described above in Methodology, and on initial review, our analyst found our sample to be in line with demographic expectations and of sufficient size to suggest sufficient representativeness to be generalizable. The specifics of the demographics are not available as of the writing of this report, but also not necessary on account of findings that only became apparent upon interpretation of the data.

It turns out that nearly half (46.6%) of the students who completed the assessment had, according to OpenBook data, successfully completed[10] English 102. That compares unfavorably to the general student body for the fall of 2016, of whom, less than 20% had successfully completed English 101, and just under half had no record of English completion, either because they had yet successfully completed an English course or had transferred credits from elsewhere.

	F16 Enrollm't	% of Total	Hum SAMPLE	% of Total
Total	8768	100.00%	605	100%
English 98	287	3.27%	6	.99%
English 99	2	0.02%	0	0.0%
English 100	n/a	0.0%	11	1.8%
English 101	2779	31.69%	170	28.1%
English 102	1733	19.77%	282	46.6%
No Comp/Record	3967	45.2%	136	22.4%





This disparity between the sample and the general population raises problems with generalizing from the abilities, attitudes, behaviors, and choices of the sample to the general HWC student body, and also highlights a significant data blind spot with respect to the general student body. In short, it seems that the self-selection built into the methodology led to a sample make-up that make any generalizations of interpretations of the data highly suspect. Thus, rather than report those suspect generalizations, and potentially mislead readers, in an act of noble and painful restraint after months of indulgence in "sunk-cost" and wishful thinking, they have been omitted from this report despite the pain of it.

[10] Defined as earning a grade of C or better in English 102 prior to the fall 2016 term.

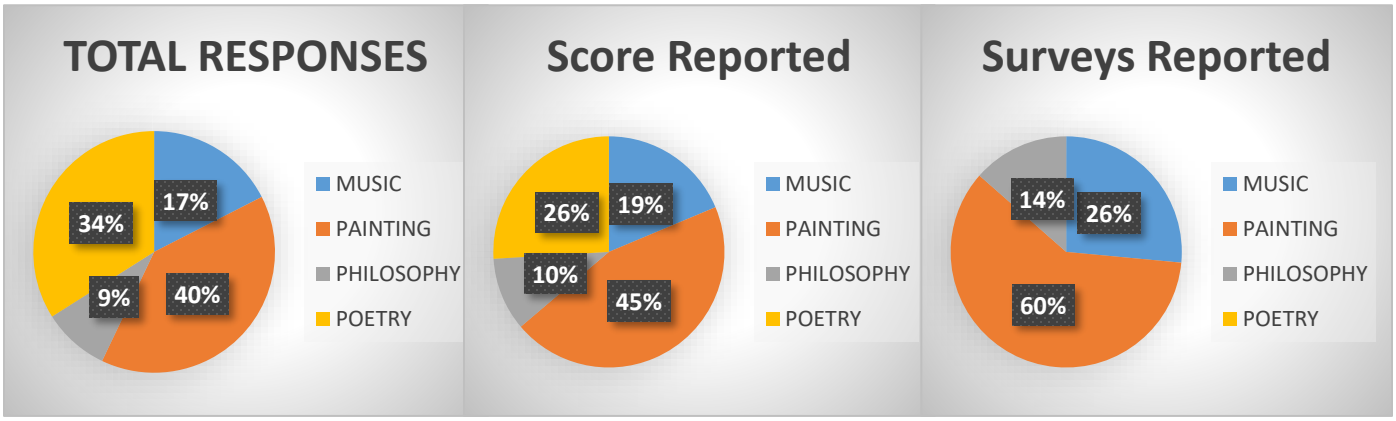
Response Details

Of the 818 student responses, 605 resulted in scoreable responses. Of the rest, most were too short, while nine of the 818 were identified as wholly plagiarized by both raters. There were 69 identifications of suspected plagiarism. Our analyst noted that there were disparities among raters in the frequency of such identifications and discounted some raters' suspicions, while prioritizing others, leading to case-by-case determinations and exclusions of some of the responses.

Those responses were divided by disciplines and included in the results as follows:

	MUSIC	PAINTING	PHILOSOPHY	POETRY	TOTALS
TOTAL RESPONSES	**143**	**324**	**73**	**278**	**818**
Score Reported	113	273	62	157	605
% Score-able	79%	84%	85%	56%	74%
Surveys Reported	143	324	73	0[11]	540
Mean Word Count*	80	194	110	102	122
Standard Deviation*	63	183	81	83	--
Maximum*	442	1180	528	519	667
*Data set includes non-scored essays (i.e., total responses)					

As shown in the numbers above and chart below, student response rates differed significantly, with students who selected "Painting" far outnumbering the other disciplines and outweighing the others in the data sets:



The mean word count is quite low given the tasks set for students, indicating, broadly speaking, a fairly low quality engagement in terms of effort and time on the part of the participating students. Unfortunately, we do not have these numbers from the 2006 assessment, so there is no way to compare whether there is a difference from the class-driven, hand-written procedure used then. On the other hand, there was little correlation between word count and scores. However, it is unclear what effect, if any, the inclusion of the 213 non-scored submissions had on these numbers.

Interrater Reliability

Data for this metric was not provided.

[11] See Note 6, above.

Rubric Effectiveness

Analysis of the results using the rubric (see chart below) shows extremely strong correlations among all of the categories, which implies that scores tended to travel together. This is not surprising, in and of itself—good writing tends to be good and bad writing tends to be bad, across the board—but it also suggests that raters had difficulty distinguishing among the categories and evaluating them independently of each other. Either that, or student skills in the various categories tended to be displayed simultaneously or not at all. However, this is a much less likely explanation when classroom experiences about student skill development are taken into consideration. In other words, the effectiveness of the rubric, in terms of distinguishing independent features of the writing submissions, was lacking, at least as used by the rating team. Consequently, discussion of student scores will concentrate on the overall score, rather than individual skill categories.

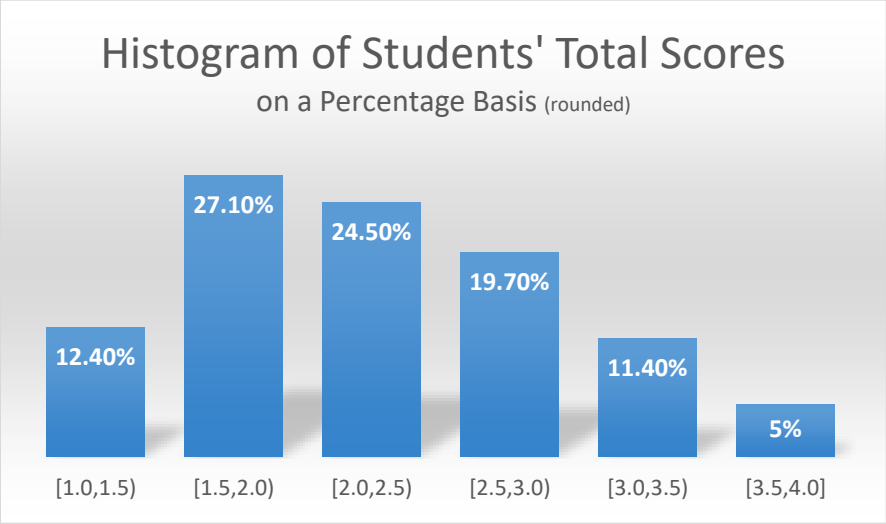
		Essay Scores				
		Focus & Organization	Support & Logic	Analysis & Development	Clarity & Correctness	Total
Essay Scores	Focus & Organization		0.7986203	0.78436296	0.75564653	0.922481
	Support & Logic	0.79862032		0.850807291	0.67900813	0.92289897
	Analysis & Development	0.78436296	0.8508073		0.67270077	0.91909127
	Clarity & Correctness	0.755646533	0.6790081	0.672700767		0.85209725
	TOTAL	0.922480998	0.922899	0.919091266	0.85209725	

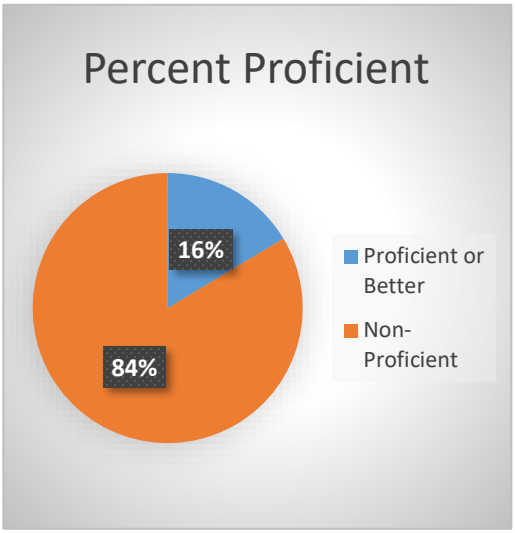
Post-assessment committee discussion by the raters suggested strong interest on the part of the committee for a more detailed, analytic rubric should the committee decide to conduct a future version of this assessment. Suggestions included developing separate rubrics for separate artifacts or the development of more descriptive analytic categories, or both. Significant challenges await such an effort, especially around useability; however, developing artifact-specific rubrics should resolve some of the potential complexity-related issues. While the rubric used by the committee may have held up better with a longer rater-norming calendar and procedure, the results clearly show that something needs to be different for any future iteration of the assessment, and the opinions and suggestions of the rating team should receive special weight.

Findings: SKILLS RELATED

Question 1A) What are students able to demonstrate with respect to their ability to analyze, interpret, and communicate about creative artifacts from disciplines within the Humanities and their own understanding of those?

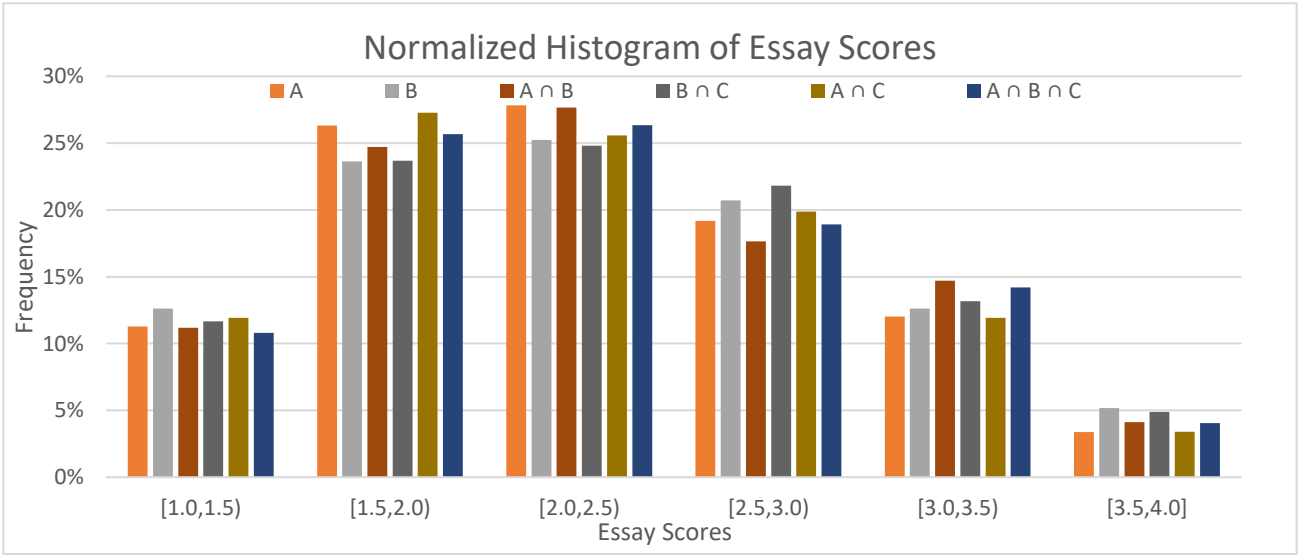
The overall student performance on the assessment was rather disappointing.





On the basis of the rubric, a score of 3.0 or higher would represent a student who demonstrated "Proficient" skill across all four analytic categories (Focus/Organization, Support/Logic, Analysis/Development, Clarity/Correctness) in their response, or at least three of the four, with "Accomplished" skill in one of those categories. These results suggest that the large majority of HWC students are not proficient in these skills (see pie chart, above right). This finding, on its own, is not all that surprising nor disappointing since it seems intuitively true that the large majority of HWC students have not completed their Humanities general education requirement, along with other things we know about our students' writing and analysis skills.

The normalized histogram of results, however, shows similar shapes, for <u>all</u> of the cohorts of students, including those who have completed one (whether A or B) or two (whether A and B, A and a second from the same category (here indicated as C), or B and a second from the same category (also indicated as C), or even three Humanities courses.



It is counter to our expectations, for example, to find out that the percentage of students scoring as "Proficient" or better for students who have completed their Humanities General Education requirement is about 18% of the population of students whose course history shows they have completed the three courses, which is close to the percentages of the cohort of students who have completed two classes (19%, 18%, and 15%, respectively), as well as the percentages for students who have completed only one class (15% and 18%) and closely mirrors the percentage for the sample as a whole (16%). Furthermore, it is disappointing to find so few students who have, apparently, completed their General Education Humanities requirement not demonstrating Proficient skill across the various aspects of this task.

So with respect to analysis question, *1A) What are students able to demonstrate with respect to their ability to analyze, interpret, and communicate about creative artifacts from disciplines within the Humanities and their own understanding of those?*—the assessment suggests "not very much" (but, here, it is important to recall that there is very likely error built into our measure, given the voluntary and ungraded nature of the assessment, as well as error in the application of measure (i.e., the ratings), and possibly error in our interpretation of results, especially given our inability to control for all the variables that we would have to control in order to separate student cohorts into meaningful groupings for this analysis, which brings us to our second analysis question.

Question 1B) How does this ability manifest at various levels of educational attainment?
A key challenge of assessment and analysis of student learning is making sense of the data in light of the wide variety of starting points that define our student body.[12] Per the committee's findings in the Natural Sciences assessment of 2015, our past practice, also used in analysis of the 2006 assessment, of grouping data sets by "completed credit hours" is an invalid approach for investigating differences in the manifestation of student skills "at various levels of educational attainment," and should be avoided.[13] Various alternatives were considered, including conducting an isolated analysis of IPEDS cohorts or an analysis of students within one semester of graduation of various degrees (AA, AS, and AGS, at least), as well as other possibilities.

After discussion, on advice of our analyst and in light of the rubric correlation analysis (see above), we proceeded by assuming that the quality of the writing in general had a primary effect on the score and would likely correlate with English Composition completion, per previous HWC Assessment.[14] After controlling for composition-level completion, scores were analyzed for differences in Humanities Gen Ed requirement fulfillment; however, Vargas stated that sample controlling for English completion and then comparing Humanities requirement completion (i.e., comparing students who have completed English 101 and zero or one or two Humanities[15] Gen Ed classes versus students who have completed English 101 and three Humanities Gen Ed classes) showed no differences among the cohorts, and their

---

[12] Eubanks, David. "A Guide for the Perplexed." *Intersection.* http://c.ymcdn.com/sites/www.aalhe.org/resource/resmgr/docs/Int/AAHLE_Fall_2017_Intersection.pdf (Association for the Assessment of Learning in Higher Education; Fall 2017), p. 10: "Recall that the mandated objective of assessment work is to place 'emphasis on assessing programs and institution (e.g. not individual courses).' But 'a program' is almost certainly a different experience for every student…with different instructors, different courses in a different sequence, different starting preparations, and many other important variables that can affect 'program outcomes.' Additionally, each student is different, with unique academic strengths and interests, and so on. Since none of these variables is usually accounted for, only large effects could possibly be detected, and even then we may fool ourselves as to the cause. This is a hopeless situation given the state of the actual data and inferential methods used. If an effect is large enough to show up under these conditions, the faculty almost certainly already know about it from their experiences with students. ***Even in cases where a curriculum is highly structured (e.g. cohort-based with a fixed course sequence), it is necessary to take into account student traits when trying to understand the cumulative effect of the curriculum."*** [emphasis, mine].

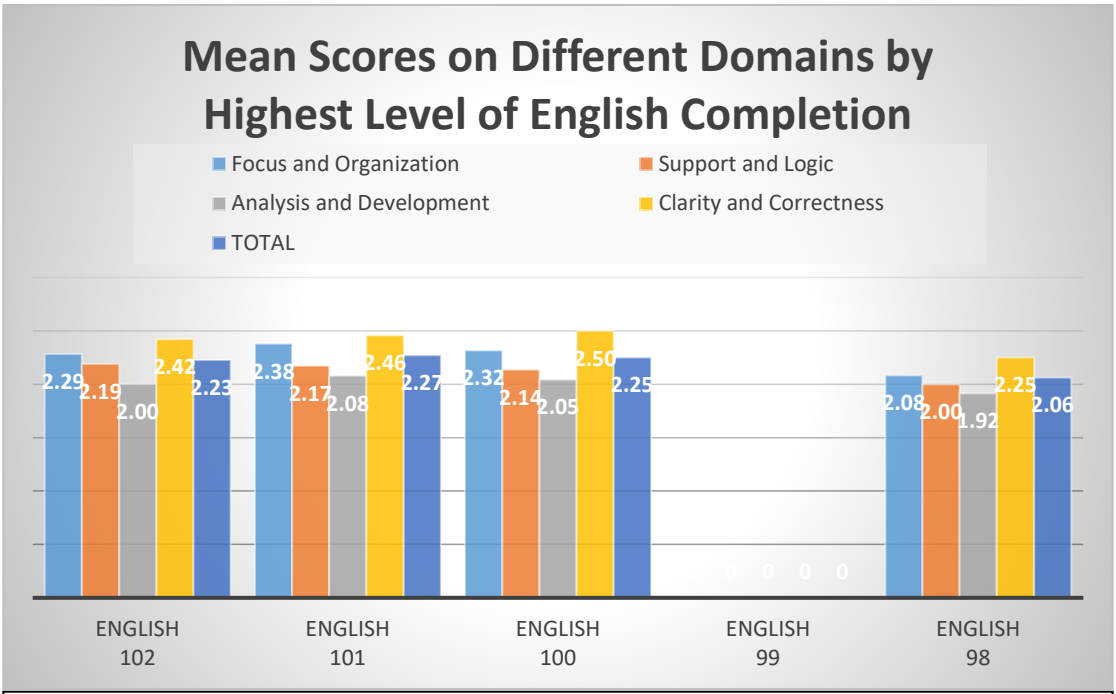
[13] Vargas, Phil. *Student Learning in the Natural Sciences*. Harold Washington Assessment Committee, p 7.

[14] Harold Washington College Assessment Committee. *Effective Writing at Harold Washington College.* (http://www.ccc.edu/colleges/washington/departments/Documents/HWCAC_Effective_Writing_Report_Spring_2013_02_14_Final_Approved.pdf), p. 20.

[15] Construed broadly and based on a course list that includes African-American Studies, Art, Fine Arts, Foreign Languages, Literature, Music, Philosophy, and Theater Arts classes.

sample sizes were reduced so drastically that it was unhelpful.[16] Thus, analysis was limited to scores bucketed by students' highest level of English class completion and, separately, by the category[17] and number of General Education Humanities classes completed.

Once again, as shown above, the results were disappointing with respect to General Education Humanities classes, and they were also surprising with respect to highest level of English class completion, giving us reason to question the efficacy of the measure with respect to students' skills:



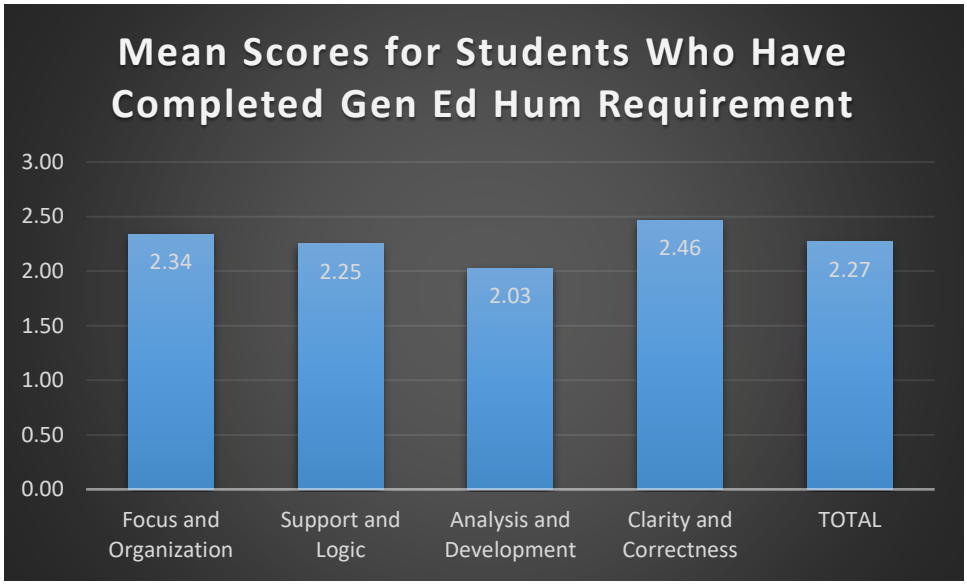
The total number of scores reported on this chart (469) differs from the number of scores because 136 students did not have any record of English completion (see Sample Validity, p. 13).

[16] Vargas, Phil. Email communication, April 24, 2018.

[17] The basic idea is that students have to take a text/writing-based (roughly) appreciation class (A column) and they have to take a performance/ non-written material (roughly) appreciation class and a third class, with an important caveat being that in our curriculum, Humanities classes are interdisciplinary, giving the student the required exposure to either, and so can be counted in either category. This requirement is often misunderstood, however, because of some category confusion and terminological equivocation. The stated requirement is one class from a list of classes described as "Humanities" classes, which we will call Group A (the set includes Philosophy, Literature, fourth-level Foreign Language classes, and more (see Note 15 above for a complete list)) as well as interdisciplinary classes, listed at HWC as Humanities classes), and one from a list of classes described as "Fine Arts" classes, which we will call Group B (the set includes HWC Fine Arts classes as well as others, such as Music and Dance Appreciation, as well as interdisciplinary classes, listed at HWC as Humanities classes. In other words, our Humanities classes fill either the A or B category requirement on account of their being interdisciplinary. To complete the requirement, students must fulfill a third class from the entirety of the list (A or B). So, would three Lit classes do it? No, because no "Fine Arts" (B Group) completion. Would three Fine Arts classes do it? No, because no "Humanities" (A Group) completion. Would three Hum classes do it? Yes (because Humanities classes are interdisciplinary and can fulfill both the A and B group requirement). Would two Philosophy and a Humanities class do it? Yes. Would two Philosophy and a Literature class do it? No, but a Literature, a Philosophy, and a Film class (e.g., Fine Arts 105) would. Get it? One important finding of this project was how hard it is for a non-expert both to find and understand the requirement.

Question 1C) Can students demonstrate appropriate proficiency at the completion of their General Education?
Strictly speaking, this question went unexamined. While it may have been possible to break out a cohort that was within, say, 12 hours of graduation or less, or selected the scores of students who graduated by the end of the Spring 2017 term, we did not do either of those. Instead, we limited ourselves to students who reportedly completed the Humanities General Education requirement, and based on the data from this assessment we would have to say, "No" in response to this question.

As the chart below shows, where 3 on the vertical axis would represent proficient, the mean score on each element and for the total was well below where we would hope to see it. In each case, "Proficient" measured roughly one standard deviation away from the mean, indicating that only about 25% of our students whose course history shows completion of the Humanities requirement were able to adequately demonstrate their analysis, interpretation, and writing skills on a measure of those that the students voluntarily completed.



Potential explanations for this result are myriad—perhaps students were distracted or uninterested or not sufficiently motivated to do their best work on it. It's possible that the visual format led to shorter, less developed responses, or that something went wrong with the ratings, or perhaps the questions were not as clear as the committee thought, or maybe the artifacts were not compelling in the eyes of the students. It may well be the case that scores would have been better if we looked at students who were close to the end of their full slate of General Education courses, under the theory that all of the courses contribute to the realization of the outcomes, but it's nonetheless clear that students did not demonstrate mastery of the skills they should have upon completion of what should be the largest contributor to these outcomes, i.e., their General Education Humanities requirement. That is a disappointment, to say the least.

The particularly low scores on "Analysis and Development" may be an indicator that student responses were more along the lines of an informal, surface level response that one might give to an acquaintance in response to a question, rather than a more formal, thoughtful, academic type of response, which may have been a function of the method or it could be that students were either not accessing their technical knowledge in their responses, if they had any about the particular discipline they were working in, and/or did not take the time to fully develop their answers the way they might if they were writing a response for their instructor or an interested party in a more formal situation. It's impossible to know, though, and so, it may be best to simply try a different measure next time to see if we can find a different way to get students to show what they know.

Question #4: Are there strong correlations between the number or type of humanities (or other) classes/subjects and proficient performance once other factors are controlled?

The analysis above seems to make pretty clear that, no, there were not strong correlations between proficient performance and the number of humanities classes, nor the type of humanities classes, nor even the level of English completion, all of which, as we say above, is rather surprising. Our data analyst ran a separate correlation of student scores with each of the four areas of the rubric as well as the overall score, yielding the following:

Rank	Essay Scores									
	Focus and Organization		Support and Logic		Analysis and Development		Clarity and Correctness		Total	
1	0.14	MUSIC-124	0.12	ECON-888	0.12	ECON-888	0.11	DMD-130	0.12	MUSIC-124
2	0.10	ECON-888	0.12	MATH-299	0.11	MUSIC-124	0.10	ENGLISH-197	0.11	ECON-888
3	0.09	SPEECH-100	0.11	DMD-168	0.10	DMD-168	0.09	DMD-168	0.11	DMD-168
4	0.09	SOC SCI-115	0.10	ART-103	0.10	VOC DA-989	0.09	VOC DA-989	0.10	VOC DA-989
5	0.09	INTCOM-99	0.10	ART-142	0.10	VOC DA-988	0.09	VOC DA-988	0.10	VOC DA-988
6	0.09	HISTORY-243	0.10	DMD-130	0.10	VOC DA-987	0.09	VOC DA-987	0.10	VOC DA-987
7	0.09	PSYCH-203	0.10	ART-132	0.10	VOC DA-5	0.09	VOC DA-5	0.10	VOC DA-5
8	0.09	ENGLISH-197	0.10	VOC DA-989	0.10	VOC DA-3217	0.09	VOC DA-3217	0.10	VOC DA-3217
9	0.08	MATH-299	0.10	VOC DA-988	0.10	VOC DA-3003	0.09	VOC DA-3003	0.10	VOC DA-3003
10	0.08	DMD-168	0.10	VOC DA-987	0.10	VOC DA-1007	0.09	VOC DA-1007	0.10	VOC DA-1007
11	0.08	HUM-212	0.10	VOC DA-5	0.10	SPEECH-100	0.09	INTCOMM-99	0.10	SPEECH-100
12	0.07	ART-888	0.10	VOC DA-3217	0.09	DMD-130	0.09	CE 014-626	0.10	DMD-130
13	0.07	ART-176	0.10	VOC DA-3003	0.09	CRM JUS-888	0.09	ABE GED-807	0.10	ENGLISH-197
14	0.07	ART-132	0.10	VOC DA-1007	0.09	ENGLISH-197	0.09	CIS-122	0.09	INTCOMM-99
15	0.07	ARCHITC-170	0.10	SPEECH-100	0.08	CIS-145	0.09	HISTORY-243	0.09	MATH-299
16	0.06	EMT-101	0.09	MUSIC-124	0.08	PSYCH-203	0.09	MUSIC-124	0.09	HISTORY-243
17	0.06	CRM JUS-888	0.08	ART-888	0.08	VOC HW-339	0.08	HUM-212	0.09	ART-132
18	0.06	CRM JUS-104	0.08	ART-176	0.08	THR ART-232	0.08	CHEM-121	0.08	ART-103
19	0.06	CE 014-625	0.08	BIOLOGY-120	0.08	INTCOMM-99	0.08	SOC-201	0.08	ART-888
20	0.06	CE 014-624	0.08	FIN ART-108	0.08	FAM DA-4308	0.07	CRM JUS-104	0.08	ART-176

As you can see, these are the strongest correlations between scores and courses students completed based on their Open Book course history. All of these correlations are extremely weak to the point of meaninglessness. Where a correlation (R value) of .3 to .7 suggests a 'moderate' effect, it's clear that all of these are closer to zero than they are to the minimum threshold of statistical relevance; these numbers suggest that something less than 2% of the variation is related.[18]

[18] By taking the r value and squaring it, then converting the result into a percentage, one can calculate this relationship. In this case, it is as follows: .14 x .14 = .0196 > 2%; https://www.surveysystem.com/correlation.htm

Question 6. Are there strong correlations between student self-perceptions/attitudes and proficient performance (again, controlling for other factors)?

The following table displays the results that our data analyst provided in response to this question:

		Essay Scores				
		Focus and Organization	**Support and Logic**	**Analysis and Development**	**Clarity and Correctness**	**Total**
I consider myself...	**Artist**	-0.0146	0.0268	0.0016	0.0291	0.0113
	Writer	-0.0184	0.0285	0.0152	-0.0396	-0.0028
	Musician	0.0396	-0.0077	-0.0019	0.0792	0.0284
	Actor or Performance	0.0257	0.0136	0.0256	-0.0083	0.0162
	Philosopher	0.0562	0.1095	0.0871	0.0658	0.0881

As you can see these are also extremely weak correlations, suggesting very little, if any, relation between the two variables. The strongest of these weak correlations runs through the philosophy respondents and the various tasks, but the effect is small, as is the size of the sample relative to the others. Granted, this table does not seem to correspond exactly to the question we asked (i.e., showing the correlation values for students who self-perceive in these ways and (exclusively) proficient scoring), but, we presume that Phil knows what he is doing and we do not understand something related to the question, made an error in the way we asked it, are confused about the means of measuring, or are confused about the table above. In any case, it does not seem that self-perception travelled together with student score, either on individual tasks or overall, in any meaningful way. This finding is unsurprising given the rest of what we've seen regarding student scoring, as there seems to have been some sort of major disconnect between our anticipations of student's ability to demonstrate their skills on these tasks and their actual performance, and it stands to reason that disconnect would cause interference in every other measurement we attempted related to student skills.

References

Hummm Report. (2007) Harold Washington College Assessment Committee.

TeachingCopyright.org. (2017). Public Domain Frequently Asked Questions. Available from https://www.teachingcopyright.org/handout/public-domain-faq.html

Creative Research Systems. “Correlation” (2018). Website: https://www.surveysystem.com/correlation.htm

Appendices

- Appendix A: Humanities Survey 2016
- Appendix B: Humanities Rubric 2016
- Appendix C: Rater Artifact Guide
- Appendix D: Key Terms List for Linguistic Analysis