America's Brightest **ORANGE**

Oklahoma State University
Committee for the Assessment of General Education
And
The Office of University Assessment and Testing
Annual Report, 2013

Committee for the Assessment of General Education:
Jon Comer (Chair), Geography
Melanie Bayles, Plant and Soil Sciences
Carol Beier, Nutritional Sciences
John Gelder, Chemistry
Bridget Miller, Applied Health & Educational Psychology
Greg Wilber, Civil & Environmental Engineering

Office of University Assessment & Testing:
Sarah R. Gordon, PhD, Interim Director
Lisa D. Cota, M.S., Statistical Analyst
uat@okstate.edu
(405)744-6687

Contents

List of Figures ... 3
List of Tables ... 3
Executive Summary ... 4
Overview ... 6
Introduction ... 6
Analytic Strategy ... 6
Key Findings, Scientific Reasoning: ... 9
Class Rank. ... 9
Year of Report. ... 10
GPA and ACT. ... 11
Academic College. ... 11
Subscale Scores. ... 12
Key Findings, Diversity: ... 14
Class Rank. ... 14
Year of Report. ... 15
GPA and ACT. ... 15
Academic College. ... 16
Other variables. ... 17
Use of Results and Future Plans ... 19
Assessment of Scientific Reasoning Skills ... 21
Scientific Reasoning Artifact Collection ... 21
Scoring Process and Reliability Estimation ... 21
Assessment of Diversity Learning Outcome ... 26
Diversity Artifact Collection ... 26
Scoring Process and Reliability Estimation ... 26

List of Figures

Figure 1. *Scientific Reasoning artifact scores by year of report.* 9
Figure 2. *Scientific Reasoning artifact scores by class rank.* 10
Figure 3. *Scientific Reasoning artifact scores by year of report.* 10
Figure 4. *Scientific Reasoning artifact scores by college.* 11
Figure 5. *Crossed pattern analysis.* 12
Figure 6. *Diversity artifact scores by year of report.* 14
Figure 7. *Diversity artifact scores by class rank.* 15
Figure 8. *Diversity artifact scores by GPA and ACT score (proportions).* 16
Figure 9. *Diversity artifact scores by college (proportions).* 16
Figure 10. *Diversity artifact scores by gender (proportions).* 17
Figure 11. *Diversity artifact scores by race/ethnicity (proportions).* 18
Figure 12. *Diversity artifact scores by General Education Designation (proportions)* 19

List of Tables

Table 1. *Scientific Reasoning Pattern Analysis: Original Model and Modeled Imprecision* 13
Table 2. *Diversity Crossed Pattern Analysis: Original and Modeled Imprecision* 17
Table 3. *2013 Collection of Scientific Reasoning Artifacts* 21
Table 4. *Inter-Rater Reliabilities(Scientific Reasoning)* 22
Table 5. *Student Demographics Associated with Scientific Reasoning Artifacts, 2007-2013* 23
Table 6. *Scientific Reasoning Artifact Scores, 2007-2013* 24
Table 7. *Scientific Reasoning Scores, 2013* 25
Table 8. *2013 Collection of Diversity Artifacts* 26
Table 9. *Inter-Rater Reliabilities(Diversity)* 27
Table 10. *Student Demographics Associated with Diversity Artifacts, 2007-2013* 28
Table 11. *Diversity Artifact Scores, 2007-2013*..29
Table 12. *Diversity Artifact Scores, 2013*...30

Executive Summary

In the summer of 2013, three teams of faculty raters scored 235 artifacts using the Scientific Reasoning rubric, and three teams of faculty raters scored 232 artifacts using the Diversity rubric. The purpose of general education assessment is to provide information on students' achievement of the objectives of the General Education program outcomes using an institutional portfolio process.

Key findings:

- Scientific Reasoning artifact scores have improved overall since 2009. In particular, the scores of juniors and seniors have increased markedly. Juniors had the highest Scientific Reasoning scores in 2013 out of all the academic classifications. More students received a score of 5 than in previous years.
- There are concerns about the extent to which the obtained sample of Scientific Reasoning artifacts accurately reflects the university population as a whole. Students from the College of Arts and Sciences were oversampled, and students from the Spears School of Business and the College of Human Services in particular were undersampled.
- Diversity artifact scores have improved slightly since 2010, with fewer students scoring a 1 than in previous years.
- The sample of Diversity artifacts does appear to adequately represent the student body in terms of race, gender, and academic college.
- There was no discernible effect of gender or ethnicity on Diversity artifact scores (in line with analyses in previous reports, these variables were not analyzed for Scientific Reasoning).
- There were distinct differences in the Diversity artifact scores when evaluating by College. For example, 12% of students from the College of Engineering, Architecture, and Technology scored a 1, as compared to 41% of students from the College of Agricultural Sciences and Natural Resources. No students in the College of Education scored above a 3.
- There were no differences in Scientific Reasoning artifact scores when evaluating by College.

Recommendations:

- Sampling procedures should be refined in order to ensure the obtained sample is more representative of the student body.
- Inter-rater reliabilities continue to be an issue. As this was a noted concern in last year's report, alternative scoring methods should be discussed. Inter-rater reliabilities for Scientific Reasoning are noticeably improved from last year's evaluation of Critical Thinking artifacts; however, inter-rater reliabilities on Diversity artifacts are poor.
- It is important to note that the purpose of assessment is not to examine individual faculty/instructors or courses; however, a closer evaluation of the assignment prompts for the Diversity and International Courses is strongly recommended.
- The Office of Institutional Research and Information Management and the Division of Institutional Diversity collect information on this campus regarding student ethnicity,

nationality, gender, and first generation student status. However, there are other markers of diversity which are not tracked by the University, including religion, culture, and ability (to name only a few). The fact that more students traditionally defined as minorities are attending this university is a positive. However, the attitudes towards diversity, as reflected in the Diversity artifact scores, do not reflect increased sensitivity to diversity among all students regardless of race or gender. A frank discussion about what exactly, constitutes diversity is warranted, as is a discussion regarding the types of assignments that best help facilitate understanding and respect of diversity in people, beliefs, and societies.

Assessment of general education is a critical aspect of our work to continuously improve our institution. We are fortunate that Oklahoma State University provides substantial resources to assess students' learning and to consider ways in which learning might be improved. Our challenge moving forward is clear: to make the most of this investment by using the results to make meaningful changes to our programs.

Thank you for your time and support of general education assessment. Please let us know if you have any additional questions or comments.

Sincerely,

Sarah R. Gordon, Ph.D.
Interim Director, University Assessment and Testing
Lisa D. Cota, M.S.
Statistical Analyst, University Assessment and Testing
Oklahoma State University
January, 2014

Overview

Introduction

General education at Oklahoma State University (OSU) is intended to:
A. Construct a broad foundation for the student's specialized course of study,
B. Develop the student's ability to read, observe, and listen with comprehension,
C. Enhance the student's skills in communicating effectively,
D. Expand the student's capacity for critical analysis and problem solving,
E. Assist the student in understanding and respecting diversity in people, beliefs, and societies, and
F. Develop the student's ability to appreciate and function in the human and natural environment.

Full details of the General Education program can be found at http://academicaffairs.okstate.edu/images/documents/gened/gened-criteriagoals.pdf

OSU has been involved in assessment of general education for more than 10 years. Three approaches are used to evaluate the general education program: institutional portfolios, review of general education course database, and college-, department-, and program-level approaches. This report focuses on OSU's use of institutional portfolios to assess the general education program. Institutional portfolios provide direct evidence of student achievement of the overall goals of general education. Institutional portfolios have been developed in five areas that represent the overall goals of the general education program (letters in parentheses map portfolios to the goals above):
1. Written communication (B and C)
2. Critical thinking (D)
3. Math problem solving (D)
4. Scientific Reasoning (D)
5. Diversity (E and F)

Recognizing that these goals cannot be achieved only through completion of courses with general education designations, student artifacts are collected from courses across campus that reveal students' achievement in each institutional portfolio area. These student artifacts are then assessed by a panel of faculty members using rubrics created by faculty members at OSU. Each rubric has a different number of categories used in the scoring process. All rubrics use a 1 to 5 scale where a 1 is low and a 5 is high. In 2013, portfolios were developed in the areas of Scientific Reasoning and Diversity.

Analytic Strategy

Data from 2013 were assessed statistically using Observation Oriented Modeling (OOM), rather than traditional inferential statistical methods. Data scored on a 1 to 5 scale, as is the case with the General Education data, are technically considered ordinal data; therefore, parametric statistics (e.g., t-tests, ANOVAs, linear regressions) are inappropriate analytic techniques. Nonparametric techniques are more appropriate but have the disadvantage of being difficult to interpret by individuals who are not well versed in statistical methodology. OOM provides an appropriate methodology for both parametric and nonparametric data alike, and it provides results that are transparent and comprehensible as well as free of common statistical assumptions. Results from analyses are both transparent and comprehensible. As such, it was the statistical method of choice employed for use with this data.

Rather than applying a series of statistical analyses to a sample dataset and then extrapolating to a population, OOM involves an analysis of the observed data only, without generalizing to a hypothetical population. The underlying philosophy is the Aristotelian definition of the *cause* of an occurrence. Modeling in OOM necessitates the researcher consider the cause/effect relationships of the variables in question.

The mathematical techniques underlying OOM also differ from null hypothesis statistical testing (NHST). Where traditional statistical methods such as *t*-tests or ANOVAs involve comparing the means of two or more groups in order to assess possible group differences, OOM assesses data at the level of the individual observation. Testing a hypothesis in OOM involves reducing a research question to a yes/no answer, then evaluating participants individually to assess whether or not a participant conformed to the hypothesis statement. Mathematically, the data are analyzed using a matrix algebra rotation called a binary Procrustes rotation.

Variables are referred to as observations, and can be considered either causal observations or target observations. Causal observations are conceptually similar to independent variables in NHST, and target observations are conceptually similar to dependent variables.

The objective of an OOM analysis is to conform the target observations to the causal observations. Mathematically, the observations are transposed into the binary system of zeroes and ones. This coding provides a matrix for both the causal and target observations, referred to in OOM as the deep structure. The deep structure matrix of the target observations is then rotated into the same number of units as the causal observations. The causal observations are then compared to the rotated deep structure matrix of the target observations in order to evaluate the percentage of observations classified correctly (PCC).

Results of the analyses are available in the form of frequency histograms. As with other statistical programs, the counts for the frequency histograms are derived from the number of participants a given category. For example, in the context of General Education Assessment, a histogram can be constructed to visually represent the distribution of scores on an assessment rubric based on the class rank of participants in a sample; the bars of the histogram will visually show the number of Freshmen who received a 1, number of Freshmen who received a 2, and so on. In OOM, the bars of the histograms are also color-coded based on the results of the matrix algebra rotation used in the analysis: Green bars represent correctly classified observations, and red bars represent incorrectly classified observations. The terms correctly classified and incorrectly classified must be considered by the researcher with a critical eye; in OOM, correct means that the classification conforms to the matrix algebra rotation but does not speak to the actual veracity of that classification. Researchers must evaluate critically both the shape of the distributions in the histograms, as well as the veracity of what is considered a correctly classified observation in the analysis.

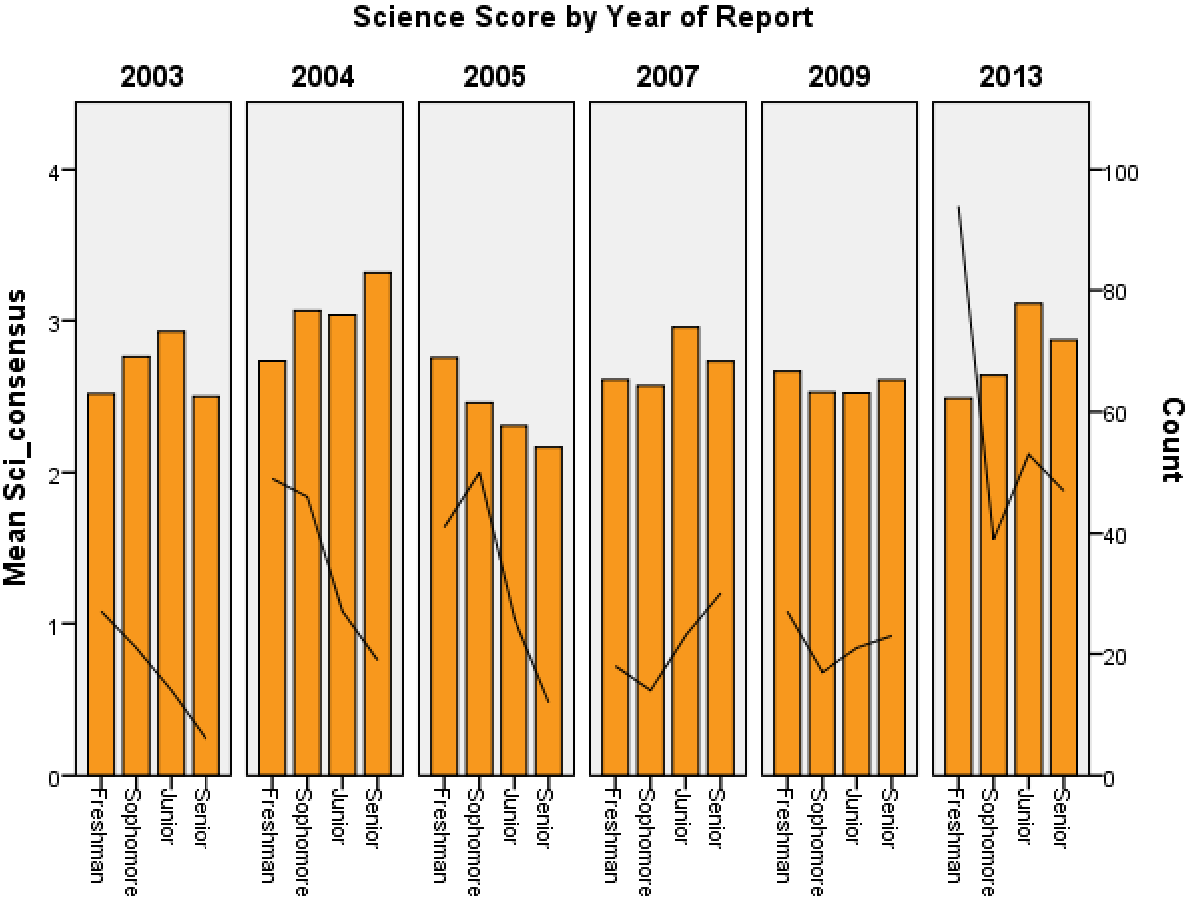
Researchers can then assess how often they might have arrived at their results by chance. This objective is accomplished through the use of randomization trials, the number of which is determined by the researcher. The randomization trials for the standard analysis involve shuffling the deep structure matrix of the target observations, performing the rotation, and then comparing the randomization results to the observed results. For all other tests, a random number generator is used to randomly assign values to the deep structure matrix of the target observations, and then the matrix is rotated and compared to the observed results. The percentage of trials classified correctly during the randomization trials are compared to the number classified correctly during the initial analyses. The resulting ratio is called the chance-value, or *c*-value.

A crucial difference between OOM and null-hypothesis statistical testing is that there is no ideal *c*-value, as opposed to the 0.05 *p*-value commonly encountered in traditional statistical analyses; it is left to the researcher to determine whether or not the obtained results are meaningful. For example, results of an OOM analysis may show a correct classification rate of 88%, and a *c*-value of 0.13. Were the same study analyzed using NHST methods, a *p*-value of 0.13 would almost certainly render the study un-publishable. However, as there is no acceptable *c*-level cut-off in OOM as there is using NHST, the researcher may decide that the correct classification rate of 88% is important, even if that classification rate occurred by chance 13% of the time during the randomization trials.

Key Findings, Scientific Reasoning:

In the summer of 2013, three teams of faculty raters scored 235 samples of student work using the Scientific Reasoning rubric. Though 235 artifacts were scored, only 225 were analyzed because little to no demographic information could be obtained for 10 students. Of the artifacts analyzed, 89 samples were written by freshmen, 37 samples were written by sophomores, 52 samples were written by juniors, and 47 samples were written by seniors. Of the 225 artifacts that were analyzed, 10 (4.4%) received a score of 1, 64 (28.4%) received a 2, 112 (49.8%) received a 3, 32 (14.2%) received a 4, and 7 (3.1%) received a 5. A comparison of scores by year of assessment is shown in Figure 1; more information across assessment years can also be found in Table 7 in the descriptive information at the end of this report.

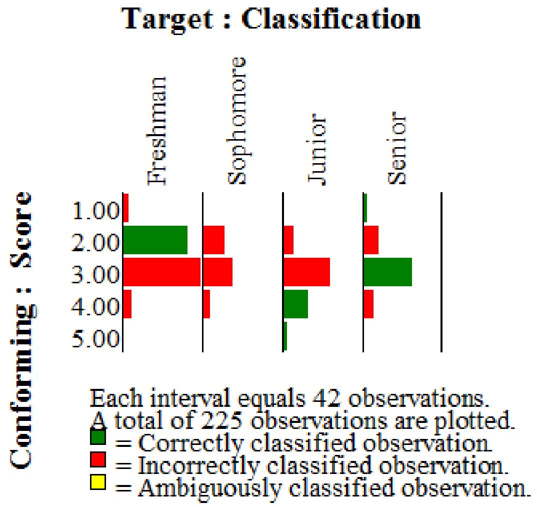
Figure 1. *Scientific Reasoning artifact scores by year of report.*



Class Rank.

Figure 2 is a frequency history histogram portraying the number of students in each academic classification that scored a 1, 2, 3, 4, or 5 on Scientific Reasoning (raw numbers and percentages of totals can be found in Table 7, further in this report). The shape of the distribution is similar across each category, with no discernible difference in the distribution of scores of Freshmen, Sophomores, Juniors, or Seniors; in other words, it is not possible to determine a student's score based on that student's academic classification.

Figure 2. *Scientific Reasoning artifact scores by class rank.*

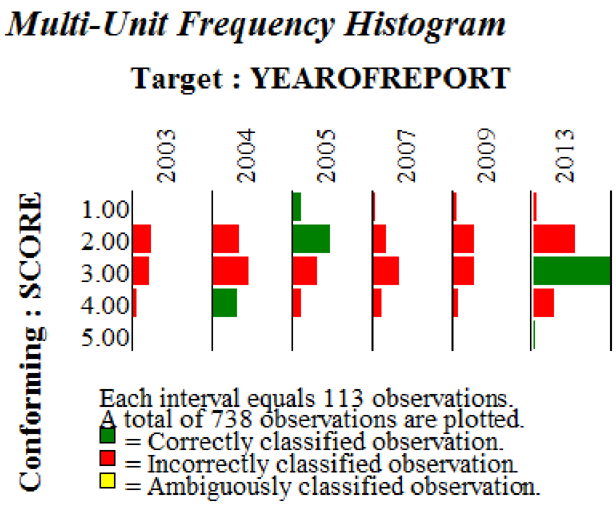


Mathematically, there was no discernible difference in the artifact scores based on class rank (PCC = 36.44%, c = 0.11; see Figure 2). Upon visual examination, Juniors had the highest scores; however, the distribution of those scores did not differ meaningfully from the score distributions of the other academic classifications, as reflected in the similarity of the shape of the score distributions in the histogram. Additionally, as stated in the introductions, the observations classified correctly refer to observations which conformed to the matrix algebra rotation; the term correctly classified does not speak to the actual veracity of those classifications. For example, Juniors who scored a 4 or 5 conformed in terms of the matrix algebra rotation; however, when critically evaluating those classifications, it makes no logical sense for Juniors to only be accurately classified as having scored a 4 or 5. This principle is more obvious when evaluating Seniors, who conformed to the matrix algebra rotation when scoring a 1 or a 3; however, it makes no logical sense that a Senior would only be classified accurately if that student scored a 1 or 3.

Year of Report.

Figure 3 shows a histogram portraying the distribution of scores based on year of evaluation. (raw numbers are summarized in Table 5, later in this report). Again, the shape of the distributions is similar from year to year. Scientific Reasoning scores continue to rise from a low point in 2005. The PCC rate for the analysis conforming Year of Report to Artifact Score was quite low (PCC = 31.30%); however, the c-value was less than 1/1000, indicating the distribution of the scores occurred randomly less than 1 time in 1000 trials. Furthermore, more artifacts received a score of 5 (n=7) in 2013 than in any other year (see Figure 3).

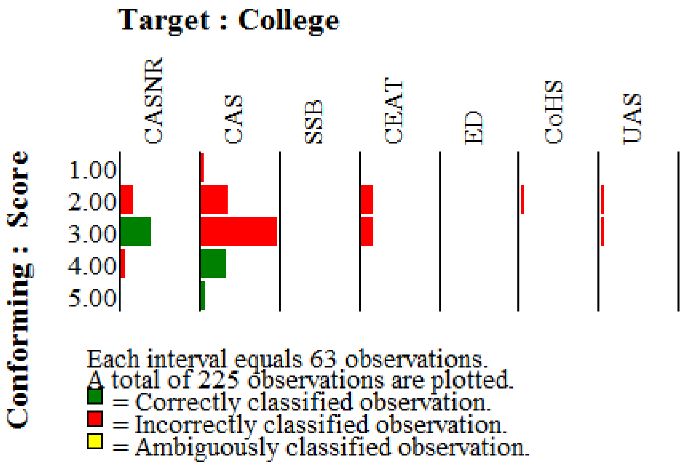
Figure 3. *Scientific Reasoning artifact scores by year of report.*



GPA and ACT.
There was no relationship between OSU GPA and artifact score (PCC = 24.44%, c = 0.38). There was no relationship between comprehensive ACT score and artifact score (PCC = 37.63, c = 0.11), nor was there a meaningful relationship between Science ACT score and artifact score (PCC = 22.68, c = 0.29).

Academic College.
Figure 4 shows a histogram portraying the distribution of scores based on academic college (raw numbers are summarized in Table 7, later in this report).

Figure 4. *Scientific Reasoning artifact scores by college.*



An analysis conforming academic college (i.e., Arts & Sciences, Spears School of Business, etc.) to artifact score yielded 26.67% of artifacts classified correctly, with a c-value of 0.001. An evaluation of the histogram indicated the obtained sample did not accurately reflect the proportions of students at the University. For example, the Spears School of Business is the second largest college at OSU; however, only two of the 225 artifacts were written by students enrolled in academic majors from the Spears School of Business; as such, the bar on the histogram representing the Spears School of Business is virtually invisible. Because the samples of students from the various colleges are so dissimilar (even proportionally), it is not appropriate to infer a relationship between academic college and scientific reasoning score.

Subscale Scores.

Each of the six subscales was evaluated separately to determine the extent to which scores on that subscale related to the overall artifact score. A Crossed Pattern Analysis was specified for each subscale. It was expected that lower scores on the given subscale would be associated with lower overall artifact scores, and higher subscale scores would be associated with higher artifact scores. An example, using the Problem subscale, is shown in Figure 5.

Figure 5. *Crossed pattern analysis.*

Original

A. Problem \ Score	1.00	2.00	3.00	4.00	5.00
1.00	🟩				
2.00		🟩			
3.00			🟩		
4.00				🟩	
5.00					🟩

Modeled Imprecision

A. Problem \ Score	1.00	2.00	3.00	4.00	5.00
1.00	🟩	🟩			
2.00	🟩	🟩	🟩		
3.00		🟩	🟩	🟩	
4.00			🟩	🟩	🟩
5.00				🟩	🟩

The subscales were also tested with a modeled imprecision, whereby the accepted hypothetical region was expanded by a region of one. Results of these analyses are shown in Table 1.

There was a distinct relationship between the subscales and the overall score. Results from this analysis should not be interpreted in terms of a linear relationship; however, this analysis indicates the overall score does appear to adequately reflect performance on the individual subscales.

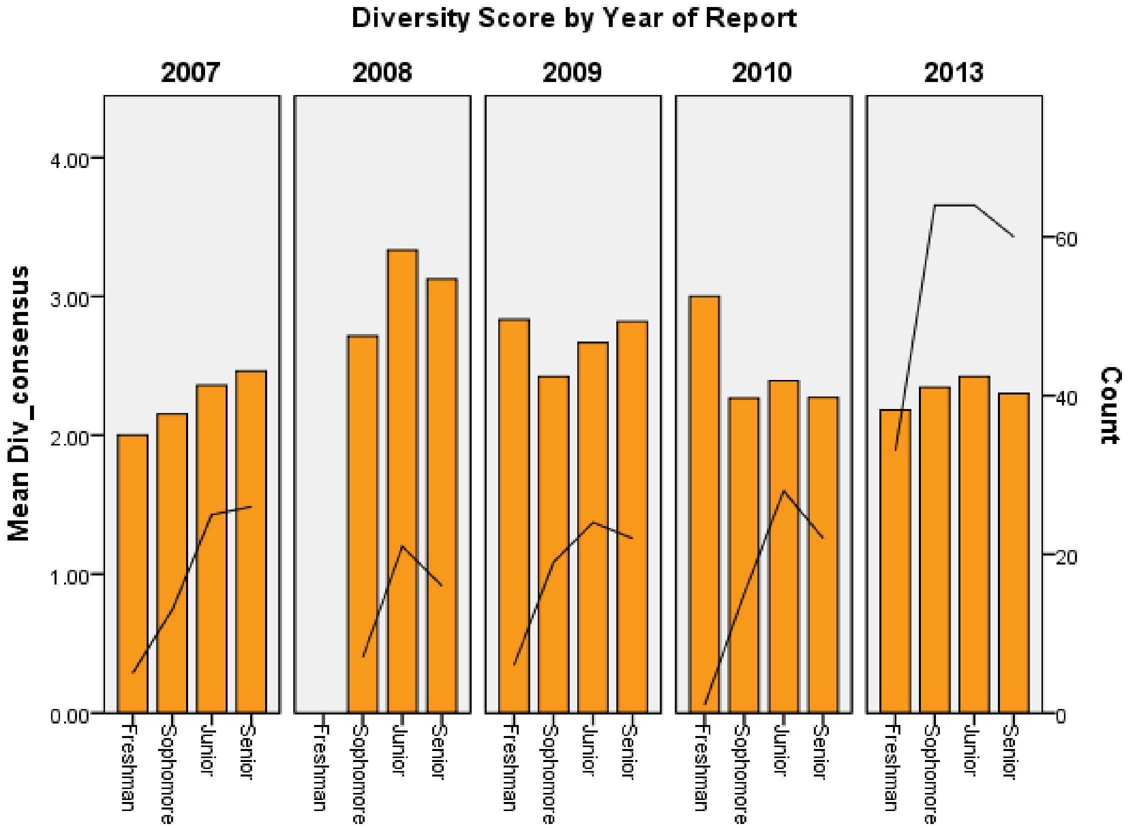
Table 1. *Scientific Reasoning Pattern Analysis: Original Model and Modeled Imprecision*

	Original Analysis		Modeled Imprecision	
Subscale	PCC	c	PCC	c
A. Understanding of problem	54.67	<0.001	96.89	<.001
B. Use of terms and symbols	58.22	<.001	97.78	<.001
C. Calculations and graphical data presentation (optional)	59.68	<.001	98.39	<.001
D. Solution and data interpretation	70.65	<.001	99.50	<.001
E. Answer and conclusions	70.22	<.001	99.56	<.001
F. Evidence of higher level thinking	65.33	<.001	96.44	<.001

Key Findings, Diversity:

Also in the summer of 2013, 232 samples of student work were evaluated using the Diversity rubric. Graduate students (n=2) and students for whom class rank information was not available (n=9) were excluded from further analyses, leaving 221 artifacts for analysis. Of the 221 artifacts that were analyzed, 33 were written by freshmen, 64 written by sophomores, 64 written by juniors, and 60 were written by seniors. Of the 221 samples, 45 (20.4%) received a score of 1, 81 (36.7%) received a 2, 73 (33.0%) received a 3, 21 (9.5%) received a 4, and one (0.5%) received a 5. A comparison of Diversity scores across assessment years is shown in Figure 6; more information across assessment years can also be found in Table 12 in the descriptive information at the end of this report.

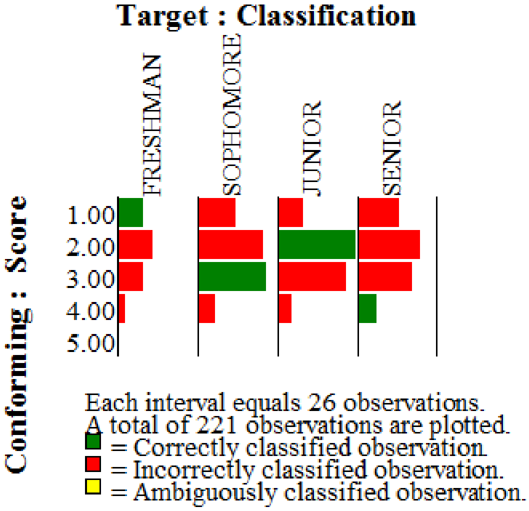
Figure 6. *Diversity artifact scores by year of report.*



Class Rank.

Figure 7 displays a histogram of Diversity score based on Classification. In this analysis, the proportions of scores are displayed on the Histogram itself, as the histogram of the raw participant counts are difficult to interpret. In 2013, the sample consisted of 33 Freshmen, 64 Sophomores, 64 Juniors, and 60 Seniors (see Table 12, later in this report).

Figure 7. *Diversity artifact scores by class rank.*



There was no effect of class rank on Diversity artifact score (PCC = 29.86%, c = 0.60). As was the case with the Scientific Reasoning data, it is not possible to determine a student's Diversity artifact score based on that student's class rank.

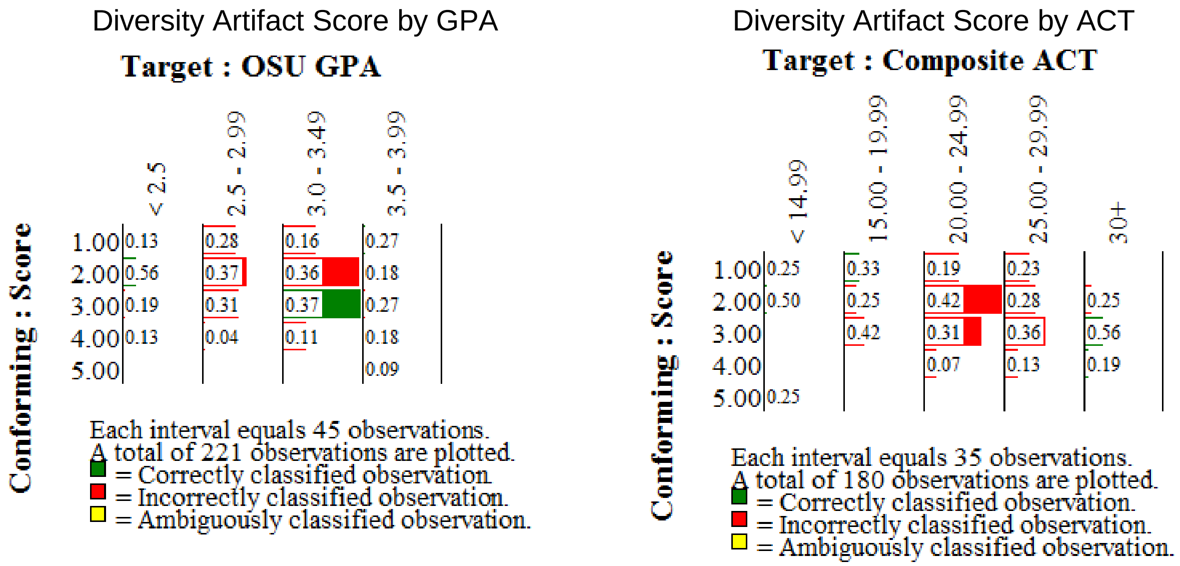
Year of Report.

Diversity artifact scores have improved slightly from 2010, although they are still proportionally not as good as they were in 2008. The PCC rate for the analysis conforming Year of Report to Artifact Score was quite low (PCC = 20.33%), and the c-value was 0.47, indicating the score distribution occurred randomly. As in 2009, one individual received a score of a 5 in 2013.

GPA and ACT.

Figure 8 displays two histograms: one shows the relationship between GPA and Diversity score, and the second shows the relationship between ACT and Diversity score. There was no discernible relationship between OSU GPA and artifact score (PCC = 27.15%, c = 0.47), nor was there a relationship between ACT score and artifact score (PCC = 12.78%, c = 0.86; see Figure 8).

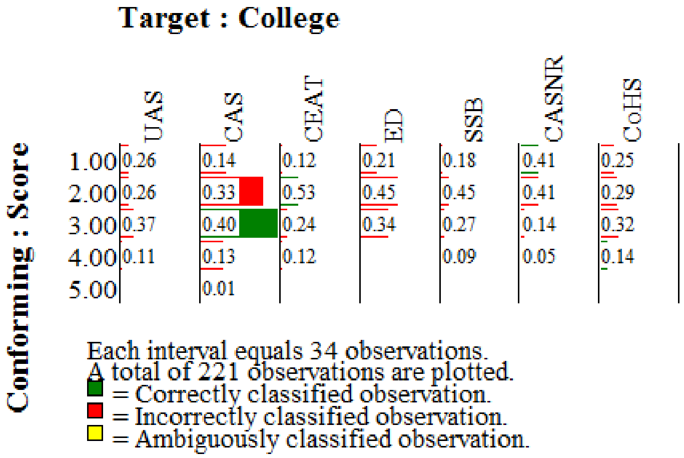
Figure 8. *Diversity artifact scores by GPA and ACT score (proportions).*



Academic College.

Score distributions as proportions based on academic college are shown in Figure 9. While the matrix algebra rotation scored a low percentage of the observations correctly, the distribution of scores did not occur randomly (PCC = 25.79%, c = 0.08); furthermore, the distributions of scores throughout those colleges are noteworthy. For example, no students in the College of Education scored above a 3. The proportions of students who scored a 1 ranged from 12% (College of Engineering, Architecture, and Technology) to 41% (College of Agricultural Sciences and Natural Resources).

Figure 9. *Diversity artifact scores by college (proportions).*



Subscale Scores.

Each of the six subscales was evaluated separately to determine the extent to which scores on that subscale related to the overall artifact score. A Crossed Pattern Analysis was specified for each subscale. It was expected that lower scores on the given subscale would be associated with lower overall artifact scores, and higher subscale scores would be associated with higher artifact scores. The subscales were also tested with a modeled imprecision, whereby the accepted hypothetical region was expanded by a region of one. Results of these analyses are shown in Table 2.

Table 2. *Diversity Crossed Pattern Analysis: Original and Modeled Imprecision*

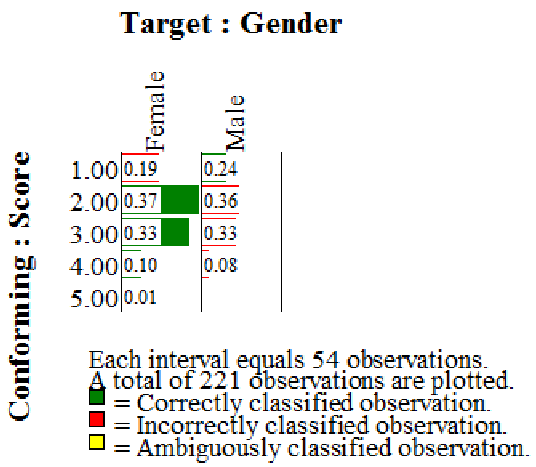
	Original Analysis		Modeled Imprecision	
Subscale	PCC	C	PCC	c
A. Conceptual understanding	72.40	<.001	100.00	<.001
B. Values diversity	66.06	<.001	99.10	<.001
C. Knowledge of historical context	47.06	<.001	90.95	<.001
D. Sources of understanding, value, and knowledge	63.80	<.001	96.38	<.001

There was a distinct relationship between the subscales and the overall score. Results from this analysis should not be interpreted in terms of a linear relationship; however, the overall score did appear to adequately reflect performance on the individual subscales. Even with the apparently low PCC of the Historical Context subscale, the scores did not occur randomly, and the PCC was improved when the imprecision was modeled.

Other variables.

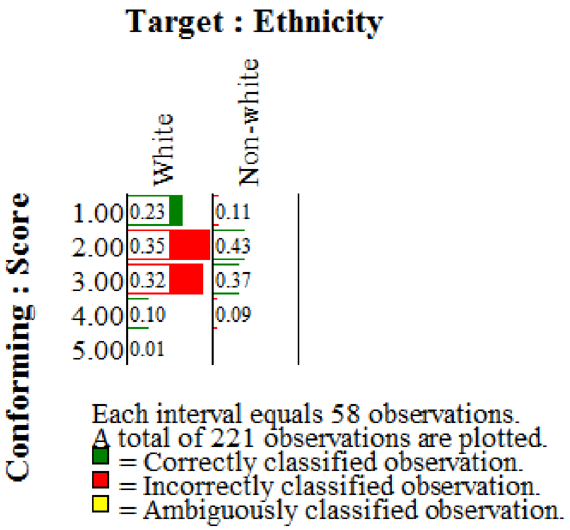
Gender: Figure 10 is a histogram showing the distribution of Diversity artifact scores by gender. The histogram shows proportions rather than raw numbers—for example, 19% of the females in this sample scored a 1, while 24% of males in this sample scored a 1(see Figure 10). Proportions were used rather than raw numbers because the sample consisted of nearly twice as many females (n=145) as males (n=76). As such, a visual comparison of the raw numbers is easily misinterpreted. Males and females did not differ noticeably in the distributions of their scores (PCC = 61.54, c = 0.10). While the PCC here is higher than in other analyses and the c-value is low, an examination of the histogram provides an explanation: the matrix algebra rotation correctly classified males when they scored a 1 and females when they scored a 2 through a 5. Again, correctly classified means that the target observations conformed to the causal variable, but does not speak to the veracity of that classification. Logically, it does not make sense for males to score only a 1, and not a 2 through 5. Furthermore, the distributions of scores are proportionally very similar for males and females (see Figure 10); when interpreting the histogram, the focus should be on the proportions.

Figure 10. *Diversity artifact scores by gender (proportions).*



Ethnicity:
Figure 11 is a histogram showing the distribution of Diversity artifact scores by ethnicity. The data on students' ethnicity was dichotomized into White and non-White for the sake of analysis. The histogram shows proportions rather than raw numbers—for example, 23% of the White students in this sample scored a 1, while 11% of the Non-White in this sample scored a 1(see Figure 11). Proportions were used rather than raw numbers because the sample included nearly three times as many White students (n=167) as Non-White students (n=54). As such, a visual comparison of the raw numbers is easily misinterpreted. As displayed in Figure 11, the proportions of the scores are quite similar; it is not possible to determine a student's score based on that student's ethnicity. (PCC = 44.80%, c = 0.89; see Figure 11).

Figure 11. *Diversity artifact scores by race/ethnicity (proportions).*

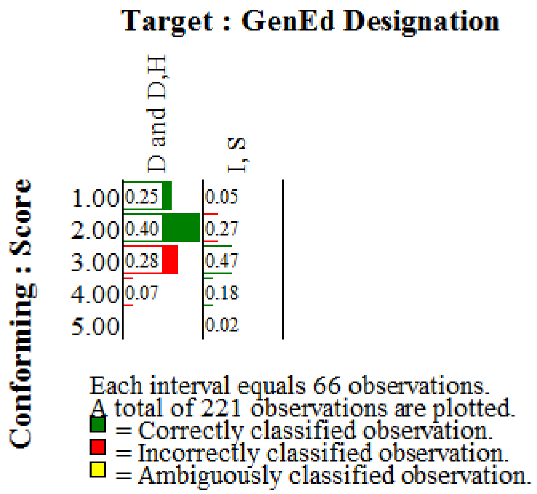


General Education Designation: There was a distinct difference in the scores of the courses with a D designation as compared to those courses with an I designation. I courses on the whole received higher scores than did the courses with a D designation (PCC = 65.61%, c = 0.08; see

Figure 12). These findings should be interpreted cautiously, as it is possible that the assignment prompt rather than the course designation influences the diversity score.

Figure 12 is a frequency histogram of Diversity artifact scores based on General Education Designation of the course from which the artifact was obtained.

Figure 12. *Diversity artifact scores by General Education Designation (proportions)*



Use of Results and Future Plans

There was a joint meeting on March 7, 2014, of the three committees or councils that share primary responsibility for the General Education program – Assessment and Academic Improvement Council (AAIC), General Education Advisory Council (GEAC), and the Committee for the Assessment of General Education (CAGE). The primary purpose of this meeting, which is held annually, is to discuss the contents of this annual report specifically and the broader implications and directions of assessment at OSU more generally.

The assessment of Diversity has historically generated much discussion on how well our students are performing, trends in scores over time, and the ability to (a) measure diversity through written artifacts and (b) draw meaningful conclusions from the results. This year, however, the discussion was particularly energetic and enthusiastic. A strongly stated view from at least some of the attendees was that OSU students are far more accomplished and progressive in their views and attitudes concerning diversity than is measured or revealed by our artifact/rubric evaluation process. While this opinion was largely based on anecdotal and personal observation, numerous members of the committees agreed with the basic concern that while evaluating written artifacts is a form of authentic assessment, the instrument (the rubric) and the method (sole authored, written papers for a grade in class) are both flawed due to the intrinsically personally, multidimensional, and sensitive issue of what constitutes diversity. Even the concepts of "appreciation" and "understanding" seem incredibly difficult to pin down compared to other skills that achieve wider agreement in the academic community, such as Writing or Scientific Reasoning. In fact, the Scientific Reasoning portion of this report garnered very little discussion, debate, or doubt as to the validity of those results, in part perhaps due to the suspiciously high level of inter-rater reliability found in two of the three Scientific Reasoning evaluation teams. Further, there was some agreement that the variety of different assignment prompts used to generate Diversity artifacts could produce an assortment of student reflections. Some writing prompts may illicit different reactions or require varying degrees of reflection on diversity issues, thus misrepresenting or failing to capture students' understanding and respect of diversity in people, beliefs, and societies.

Though no firm recommendations or plans of action came from the discussion, it seemed there was a critical mass of interested individuals who can hopefully be counted on to continue being involved in the discussion to explore other ways to measure Diversity learning goals, including co-curricular and extra-curricular activities, making use of individuals with expertise who are working in Student Affairs, Residential Life, Office of Multicultural Affairs, and other support units on campus. While the assessment of general education falls to CAGE, given that there are other portfolios to maintain and assess (including a huge effort in Writing and Critical Thinking in 2014 to meet Voluntary System of Accountability (VSA) standards) it seems logical that AAIC should consider appointing an ad hoc committee or task force to study ways to improve the assessment of Diversity on the OSU campus, drawn from AAIC, CAGE, and GEAC members as well as individuals from other offices identified who work with student populations as noted above. This task force could focus solely on researching and suggesting methods to improve not just student achievement of defined learning goals but also OSU’s ability to measure and evaluate this performance. The actual assessment thereof could then devolve to CAGE in its normal role.

With regard to the variety in assignment prompts, CAGE members acknowledged the importance of academic freedom for instructors to create their own assignments but suggested providing more information about the assessment process that might be beneficial for both students and instructors. CAGE members discussed the possibility of providing an informational packet that outlines the assessment process for Diversity to instructors who request the ‘D’ designation for their course through GEAC. Such a packet could provide instructors with more information about the assessment process (including the Diversity rubric), as well as suggestions for the types of writing prompts (based on previous assessment data) that seem to illicit more thoughtful reflection from students regarding their understanding and respect of diversity.

.

Assessment of Scientific Reasoning Skills

Scientific Reasoning Artifact Collection

Artifacts (embedded course assignments) were collected from faculty by direct request from a random sample of general education designated courses, as well as from faculty who attended the *Provost's Faculty Development Initiative: Focus on General Education*. The courses from which artifacts were sampled are shown in Table 3. Artifacts selected for the Institutional Portfolio were coded, and all identifying information was removed. Demographic data were collected separately from the Office of Institutional Research and Information Management (IRIM); these data were used for statistical analysis only, and cannot be used to identify individual students. Student demographic information was not shared with reviewers prior to scoring.

Table 3. *2013 Collection of Scientific Reasoning Artifacts*

Course No.	Course Name	General Education Designation (if any)	Number of Artifacts Submitted	Number of Artifacts Scored
BOT 1404	Plant Biology	N	108	57
CHEM 1314	General Chemistry	L, N	191	87
CIVE 3853	Environmental Engineering Laboratory		8	8
PSYC 3073	Neurobiological Psychology	N	50	50
ZOOL 3104	Invertebrate Zoology		33	33
Total Number of Scientific Reasoning Artifacts			382	235*

Note: *Though 235 artifacts were scored, only 225 were analyzed; 10 artifacts were removed from the analyses because sufficient demographic information on these students could not be obtained.

Scoring Process and Reliability Estimation

All reviewers met for a training session in the beginning of Summer 2013. After reviewing the Scientific Reasoning rubric, reviewers reviewed Scientific Reasoning artifacts from previous years. This provided raters with the opportunity to ask questions or discuss any concerns, as well as aligned raters' scores with each other.

Three teams, each composed of two raters, reviewed the artifacts independently. Each artifact received a score from 1 to 5, with 1 being the lowest possible score, and 5 being the highest possible score. Reviewers also scored the artifacts on six sub-scales: understanding of problem; use of terms and symbols; calculations and data (optional); solution and data interpretation; answer and conclusions; and evidence of higher level thinking.

After the teams rated the artifacts, the team captain reviewed the scores. Artifacts on which the reviewers differed by more than one point were discussed as a group. The team captain attempted to bring the reviewers to a consensus; absent that, the team captain scored the artifact in question. Estimates of inter-rater reliability are provided in Table 4.

Table 4. *Inter-Rater Reliabilities*[1]*(Scientific Reasoning)*

	Group 1			**Group 2**			**Group 3**		
Method	Value	SE	C.I.	Value	SE	C.I	Value	SE	C.I.
AC1	0.99	0.01	0.96 to 1.00	0.99	0.01	0.96 to 1.00	0.66	0.06	0.54 to 0.78
Kappa	0.98	0.02	0.95 to 1.00	0.98	0.02	0.95 to 1.00	0.56	0.08	0.41 to 0.71
PI	0.98	0.02	0.95 to 1.00	0.98	0.02	0.95 to 1.00	0.56	0.08	0.41 to 0.71
BP	0.99	0.01	0.96 to 1.00	0.99	0.01	0.96 to 1.00	0.64	0.06	0.52 to 0.77

There are numerous ways to evaluate the adequacy of reliability estimates. Guidelines proposed by Altman (1991) are provided below:

- < .20 = Slight Agreement
- .21 to .40 = Fair Agreement
- .41 to .60 = Moderate Agreement
- .61 to .80 = Good Agreement
- .81 to 1.00 = Very Good.

These guidelines indicate that two teams (1 and 2) had “very good” levels of agreement, and one team (3) had “moderate” levels of agreement.

Tables 5-7 represent descriptive statistics for the scientific reasoning artifacts and scores.

[1] Descriptions of reliability coefficient may be found at http://agreestat.com/research_papers.html. AC1 = variation of Kappa statistic and BP statistic that incorporates the conditional probability that two random rater will agree given no chance agreement; Kappa = omnibus measure of percent agreement among raters when corrected for chance agreement wherein chance is defined as the expected value if ratings were completely independent; PI = probability that a randomly selected rater will classify a randomly selected artifact into specific category. BP = Brennan-Prediger modification of Kappa statistic that incorporates a modification of marginal estimates so that chance is redefined to adjust for the number of possible categories.

Table 5. *Student Demographics Associated with Scientific Reasoning Artifacts, 2007-2013*

			2007-2010 # of artifacts (% of total)	2013 # of artifacts (% of total)	Combined # of artifacts (% of total)
Class	Freshman		162 (31.7%)	89 (39.6%)	251 (34.1%)
	Sophomore		148 (29%)	37 (16.4%)	185 (25.1%)
	Junior		111 (21.7%)	52 (23.1%)	163 (22.1%)
	Senior		90 (17.6%)	47 (20.9%)	137 (18.6%)
		Total	N=511	N=225	N=736
College	CAS		183 (35.8%)	121 (53.8%)	304 (41.3%)
	CASNR		127 (24.9%)	49 (21.8%)	176 (23.9%)
	SSB		53 (10.4%)	2 (0.9%)	55 (7.5%)
	COE		84 (16.4%)	6 (2.7%)	90 (12.2%)
	CEAT		31 (6.1%)	29 (12.9%)	60 (8.2%)
	CoHS		22 (4.3%)	7 (3.1%)	29 (3.9%)
	UAS		11 (2.2%)	11 (4.9%)	22 (3.0%)
		Total	N=511	N=225	N=736
Gender	Male		196 (38.4%)	94 (41.8%)	290 (39.4%)
	Female		315 (61.6%)	131 (58.2%)	446 (60.6%)
		Total	N=511	N=225	N=736
Admit Type	Regular (A, AR, L)		361 (70.1%)	163 (72.5%)	524 (71.2%)
	Alternative Admit (F)		19 (3.7%)	2 (0.9%)	21 (2.9%)
	Adult Admit (G)		-	-	-
	International (J)		7 (1.4%)	2 (0.9%)	9 (1.2%)
	Transfer (M, MR)		121 (23.7%)	55 (24.4%)	176 (23.9%)
	Other or Blank		3 (0.5%)	3 (1.3%)	6 (0.4%)
		Total	N=511	N=225	N=736
ACT	<22		127 (30.0%)	41 (21.1%)	168 (27.2%)
	22 to 24		125 (29.5%)	52 (26.8%)	177 (28.6%)
	25 to 27		100 (23.6%)	40 (20.6%)	140 (22.7%)
	28 to 30		49 (11.6%)	43 (22.2%)	92 (14.9%)
	>30		23 (5.4%)	18 (9.3%)	41 (6.6%)
		Total	N=424	N=194	N=618
OSU GPA	<2.0		34 (6.7%)	9 (4.0%)	43 (5.8%)
	2.0 to 2.49		77 (15.1%)	30 (13.3%)	107 (14.5%)
	2.50 to 2.99		129 (25.2%)	36 (16.0%)	165 (22.4%)
	3.00 to 3.49		125 (24.5%)	72 (32.0%)	197 (26.8%)
	3.50 to 4.00		146 (28.6%)	78 (34.7%)	224 (30.4%)
		Total	N=511	N=225	N=736

Note: The numbers presented in this table represent students for which demographic information was available. Sum totals for each category/column/row vary according to the information available.

Table 6. *Scientific Reasoning Artifact Scores, 2007-2013*

			SCORE n(%)			
	1	2	3	4	5	**N**
Overall	46 (6.3%)	247 (33.6%)	306 (41.6%)	121 (16.4%)	16 (2.2%)	736
Class						
Freshman	14 (5.6%)	97 (38.6%)	105 (41.8%)	30 (12.0%)	5 (2.0%)	251
Sophomore	14 (7.6%)	62 (33.5%)	73 (39.5%)	33 (17.8%)	3 (1.6%)	185
Junior	12 (7.4%)	45 (27.6%)	63 (38.7%)	37 (22.7%)	6 (3.7%)	163
Senior	6 (4.4%)	43 (31.4%)	65 (47.4%)	21 (15.3%)	2 (1.5%)	137
College						
CAS	19 (6.3%)	85 (28.0%)	131 (43.1%)	59 (19.4%)	10 (3.3%)	304
CASNR	10 (5.7%)	56 (31.8%)	77 (43.8%)	29 (16.5%)	4 (2.3%)	176
SSB	4 (7.3%)	22 (40.0%)	22 (40.0%)	7 (12.7%)	0 (0.0%)	55
COE	7 (7.8%)	30 (33.3%)	37 (41.1%)	16 (17.8%)	0 (0.0%)	90
CEAT	2 (3.3%)	28 (46.7%)	22 (36.7%)	6 (10.0%)	2 (3.3%)	60
CoHS	1 (3.4%)	14 (48.3%)	11 (37.9%)	3 (10.3%)	0 (0.0%)	29
UAS	3 (13.6%)	12 (54.5%)	6 (27.3%)	1 (4.5%)	0 (0.0%)	22
Gender						
Male	20 (6.9%)	93 (32.1%)	125 (43.1%)	47 (16.2%)	5 (1.7%)	290
Female	26 (5.8%)	154 (34.5%)	181 (40.6%)	74 (16.6%)	11 (2.5%)	446

Table 7. *Scientific Reasoning Scores, 2013*

	%		**SCORE n(%)**			
	1	2	3	4	5	**N**
Overall	10 (4.4%)	64 (28.4%)	112 (49.8%)	32 (14.2%)	7 (3.1%)	225
Class						
Freshman	4 (4.5%)	35 (39.3%)	42 (47.2%)	6 (6.7%)	2 (2.2%)	89
Sophomore	1 (2.7%)	13 (35.1%)	17 (45.9%)	5 (13.5%)	1 (2.7%)	37
Junior	2 (3.8%)	7 (13.5%)	26 (50.0%)	14 (26.9%)	3 (5.8%)	52
Senior	3 (6.4%)	9 (19.1%)	27 (57.4%)	7 (14.9%)	1 (2.1%)	47
College						
CAS	3 (6.1%)	13 (26.5%)	27 (55.1%)	6 (12.2%)	0 (0.0%)	49
CASNR	5 (4.1%)	24 (19.8%)	63 (52.1%)	23 (19.0%)	6 (5.0%)	121
SSB	0 (0.0%)	2 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	2
COE	2 (6.9%)	13 (44.8%)	12 (41.4%)	1 (3.4%)	1 (3.4%)	29
CEAT	0 (0.0%)	2 (33.3%)	3 (50.0%)	1 (16.7%)	0 (0.0%)	6
CoHS	0 (0.0%)	5 (71.0%)	2 (29.0%)	0 (0.0%)	0 (0.0%)	7
UAS	0 (0.0%)	5 (45.5%)	5 (45.5%)	1 (9.1%)	0 (0.0%)	11
Gender						
Male	7 (7.4%)	27 (28.7%)	44 (46.8%)	13 (13.8%)	3 (3.2%)	94
Female	3 (2.3%)	37 (28.2%)	68 (51.9%)	19 (14.5%)	4 (3.1%)	131

Assessment of Diversity Learning Outcome

Diversity Artifact Collection

Artifacts included in the Diversity portfolio were collected from faculty by direct request from a random sample of general education designated courses, as well as from faculty who attended the *Provost's Faculty Development Initiative: Focus on General Education*. The courses from which artifacts were sampled are shown in Table 8. Artifacts selected for the Institutional Portfolio were coded, and all identifying information was removed. Demographic data were collected separately from the Office of Institutional Research and Information Management (IRIM); these data were used for statistical analysis only, and cannot be used to identify individual students. Student demographic information was not shared with reviewers prior to scoring.

Table 8. *2013 Collection of Diversity Artifacts*

Course No.	Course Name	General Education Designation (if any)	Number of Artifacts Submitted	Number of Artifacts Scored
AGLE 2403	Agricultural Leadership in a Multicultural Society	D	45	29
ANTH 3353	Cultural Anthropology	I, S	33	32
ENGL 2413	Introduction to Literature	D, H	37	37
GWST 2123	Introduction to Gender Studies	D, H	42	29
SOC 4653	Gender and the Middle East	I, S	24	24
SCFD 3223	Role of Teacher in American Schools	D	65	62
TH 3633	Voices of Diversity	D, H	19	19
	Total Number of Diversity Artifacts		265	232*

Note: *Though 232 artifacts were scored, only 221 were analyzed; 11 artifacts were removed from the analyses because two were from graduate students and little to no demographic information could be obtained for nine students.

Scoring Process and Reliability Estimation

All reviewers met for a training session in the beginning of Summer 2013. After reviewing the Diversity rubric, reviewers reviewed Diversity artifacts from previous years. This provided raters with the opportunity to ask questions or discuss any concerns, as well as aligned raters' scores with each other.

Three teams, each composed of two raters, reviewed the artifacts independently. Each artifact received a score from 1 to 5, with 1 being the lowest possible score, and 5 being the highest possible score. Reviewers also scored the artifacts on four sub-scales: conceptual understanding; values diversity; knowledge of historical context; and sources of understanding, value, and knowledge. After the teams rated the artifacts, the team captain reviewed the scores. Artifacts on which the reviewers differed by more than one point were discussed as a group.

The team captain attempted to bring the reviewers to a consensus; absent that, the team captain scored the artifact in question. Estimates of inter-rater reliability are provided in Table 9.

Table 9. *Inter-Rater Reliabilities*[2]*(Diversity)*

	Group 1			**Group 2**			**Group 3**		
Method	Value	SE	C.I.	Value	SE	C.I	Value	SE	C.I.
AC1	0.33	0.07	0.20 to 0.47	0.36	0.07	0.21 to 0.50	0.34	0.07	0.20 to 0.47
Kappa	0.22	0.07	0.07 to 0.37	0.33	0.08	0.18 to 0.48	0.26	0.08	0.11 to 0.42
PI	0.22	0.07	0.07 to 0.36	0.32	0.08	0.17 to 0.48	0.25	0.08	0.10 to 0.40
BP	0.31	0.07	0.18 to 0.45	0.35	0.07	0.21 to 0.49	0.32	0.07	0.19 to 0.49

There are numerous ways to evaluate the adequacy of reliability estimates. Guidelines proposed by Altman (1991) are provided below:

- < .20 = Slight Agreement
- .21 to .40 = Fair Agreement
- .41 to .60 = Moderate Agreement
- .61 to .80 = Good Agreement
- .81 to 1.00 = Very Good.

These guidelines indicate that all three teams had “slight” levels of agreement. Use of scores when there is poor inter-rater reliability estimates, before corrected by a third rater, is problematic. However, most analyses use the scores after correction by a third rater. The extent to which this process “corrects” for score inconsistency across raters remains empirically difficult to define, as the “true” scores are subject to interpretation.

Tables 10-12 represent descriptive statistics for the diversity artifacts and scores.

[2] See Footnote 1 on Page 21

Table 10. *Student Demographics Associated with Diversity Artifacts, 2007-2013*

		2007-2010	2013	Combined
		# of	# of artifacts	# of artifacts
		(% of total)	(% of total)	(% of total)
Class	Freshman	12 (4.8%)	33 (14.9%)	45 (9.6%)
	Sophomore	54 (21.6%)	64 (29.0%)	118 (25.1%)
	Junior	98 (39.2%)	64 (29.0%)	162 (34.4%)
	Senior	86 (34.4%)	60 (27.1%)	146 (31.0%)
	Total	N=250	N=221	N=471
College	CAS	95 (38.0%)	86 (38.9%)	181 (38.4%)
	CASNR	6 (2.4%)	22 (10.0%)	28 (5.9%)
	SSB	17 (6.8%)	11 (5.0%)	28 (5.9%)
	COE	61 (24.4%)	38 (17.2%)	100 (20.7%)
	CEAT	33 (13.2%	17 (7.7%)	50 (10.6%)
	CoHS	22 (8.8%)	28 (12.7%)	51 (10.8%)
	UAS	16 (6.4%)	19 (8.6%)	35 (7.4%)
	Total	N=250	N=221	N=471
Gender	Male	140 (56.0%)	76 (34.4%)	216 (45.9%)
	Female	110 (44.0%)	145 (65.6%)	255 (54.1%)
	Total	N=250	N=221	N=471
Admit Type	Regular (A, AR, L)	141 (56.6%)	147 (66.5%)	288 (61.3%)
	Alternative Admit (F)	22 (8.8%)	9 (4.1%)	31 (6.6%)
	Adult Admit (G)	-	-	-
	International (J)	3 (1.2%)	2 (0.9%)	5 (1.1%)
	Transfer (M, MR)	83 (33.3%)	60 (27.1%)	143 (30.4%)
	Other or Blank	-	3 (1.4%)	14 (3.0%)
	Total	N=249	N=221	N=470
ACT	<22	63 (34.1%)	52 (29.4%)	115 (31.8%)
	22 to 24	52 (28.1%)	56 (31.6%)	108 (29.8%)
	25 to 27	30 (16.2%)	37 (20.9%)	67 (18.5%)
	28 to 30	22 (11.9%)	20 (11.3%)	42 (11.6%)
	>30	18 (9.7%)	12 (6.8%)	30 (8.3%)
	Total	N=185	N=177	N=362
OSU GPA	<2.0	12 (4.8%)	16 (7.2%)	28 (5.9%)
	2.0 to 2.49	44 (17.6%)	26 (11.8%)	70 (14.9%)
	2.50 to 2.99	73 (29.2%)	45 (20.4%)	118 (25.1%)
	3.00 to 3.49	56 (22.4%)	70 (31.7%)	126 (26.6%)
	3.50 to 4.00	64 (25.6%)	64 (29.0%)	130 (27.6%)
	Missing	1 (.004%)	0 (0.0%)	10 (2.1%)
	Total	N=250	N=221	N=471

Note: The numbers presented in this table represent students for which demographic information was available. Sum totals for each category/column/row vary according to the information available.

Table 11. *Diversity Artifact Scores, 2007-2013*

	SCORE n(%)					
	1	2	3	4	5	**N**
Overall	87 (18.5%)	156 (33.1%)	157 (33.3%)	67 (14.2%)	4 (0.8%)	471
Class						
Freshman	10 (22.2%)	16 (35.6%)	16 (35.6%)	3 (6.7%)	0 (0.0%)	45
Sophomore	26 (22.0%)	41 (34.7%)	36 (30.5%)	14 (11.9%)	1 (0.8%)	118
Junior	26 (16.0%)	50 (30.9%)	57 (35.2%)	27 (16.7%)	2 (1.2%)	162
Senior	25 (17.1%)	49 (33.6%)	48 (32.9%)	23 (15.8%)	1 (0.7%)	146
College						
CAS	23 (12.7%)	49 (27.1%)	73 (40.3%)	33 (18.2%)	3 (1.7%)	181
CASNR	11 (39.3%)	12 (42.9%)	4 (14.3%)	1 (3.6%)	0 (0.0%)	28
SSB	6 (21.4%)	7 (25.0%)	7 (25.0%)	7 (25.0%)	1 (3.6%)	28
COE	18 (18.2%)	42 (42.4%)	29 (29.3%)	10 (10.1%)	0 (0.0%)	99
CEAT	9 (18.0%)	16 (32.0%)	19 (38.0%)	6 (12.0%)	0 (0.0%)	50
CoHS	9 (18.0%)	17 (34.0%)	16 (32.0%)	8 (16.0%)	0 (0.0%)	50
UAS	11 (31.4%)	13 (37.1%)	9 (25.7%)	2 (5.7%)	0 (0.0%)	35
Gender						
Male	49 (22.7%)	73 (33.8%)	67 (31.0%)	26 (12.0%)	1 (0.5%)	216
Female	38 (14.9%)	83 (32.5%)	90 (35.3%)	41 (16.1%)	3 (1.2%)	255

Table 12. *Diversity Artifact Scores, 2013*

	SCORE n(%)					
	1	2	3	4	5	**N**
Overall	45 (20.4%)	81 (36.7%)	73 (33.0%)	21 (9.5%)	1 (0.5%)	221
Class						
Freshman	9 (27.3%)	12 (36.4%)	9 (27.3%)	3 (9.1%)	0 (0.0%)	33
Sophomore	13 (20.3%)	22 (34.4%)	23 (35.9%)	6 (9.4%)	0 (0.0%)	64
Junior	9 (14.1%)	26 (40.6%)	23 (35.9%)	5 (7.8%)	1 (1.6%)	64
Senior	14 (23.3%)	21 (35.0%)	18 (30.0%)	7 (11.7%)	0 (0.0%)	60
College						
CAS	12 (14.0%)	28 (32.6%)	34 (39.5%)	11 (12.8%)	1 (1.2%)	86
CASNR	9 (40.9%)	9 (40.9%)	3 (13.6%)	1 (4.5%)	0 (0.0%)	22
SSB	2 (18.2%)	5 (45.5%)	3 (27.3%)	1 (9.1%)	0 (0.0%)	11
COE	8 (21.1%)	17 (44.7%)	13 (34.2%)	0 (0.0%)	0 (0.0%)	38
CEAT	2 (11.8%)	9 (52.9%)	4 (23.5%)	2 (11.8%)	0 (0.0%)	17
CoHS	7 (25.0%)	8 (28.6%)	9 (32.1%)	4 (14.3%)	0 (0.0%)	28
UAS	5 (26.3%)	5 (26.3%)	7 (36.8%)	2 (10.5%)	0 (0.0%)	19
Gender						
Male	18 (23.7%)	27 (35.5%)	25 (32.9%)	6 (7.9%)	0 (0.0%)	76
Female	27 (18.6%)	54 (37.2%)	48 (33.1%)	15 (10.3%)	1 (0.7%)	145