TEXAS WOMAN'S UNIVERSITY
DENTON DALLAS HOUSTON

Core Assessment

Report on the 2015 Pilot Results

Dr. Gray Scott
Dept. of English, Speech, & Foreign Languages

For Undergraduate Studies & the Office of Academic Assessment

Goal: To Assess the Assessment

Our report summarizes findings from Spring 2015's pilot assessment of the new core curriculum at Texas Woman's University.

TWU adopted the new core in Fall 2014 in the Texas legislature's revision of rules governing the range of courses that Texas universities could require students to take for their general education. As part of the new law, Texas required that universities require 42 semester credit hours of general education (or core) instruction across eight broad subject areas: **Communication**; **Mathematics**; **Life & Physical Sciences**; **Creative Arts**; **Social & Behavioral Sciences**; **History**; **Government**; and **Language, Philosophy & Culture**. Moreover, each university had to develop a plan for assessing those "Foundational Component Areas" on six core objectives: **Communication**, **Critical Thinking**, **Empirical & Quantitative Skills**, **Personal Responsibility**, **Social Responsibility**, and **Teamwork**.

The following report assesses 1,530 student artifacts submitted by 67 faculty members. The artifacts were rated by our inaugural Assessment Academy members: 49 volunteer raters from faculty, staff, administration, and the student body, with artifacts rated twice each to bolster reliability.

However, it wouldn't be fair to describe this report as an assessment of *students*. Rather, it is one step in an **assessment of the assessment**.

You are the next step.

Table of Contents

Goal: … (Cont'd)

Data are meaningless without interpretation. For that reason, we would appreciate your input on

- how best to interpret the data that we are seeing
- how we might change our approach to make it more informative – or less intrusive
- what hypotheses you might want to see tested
- or which questions you would like us to try to answer with the data being collected.

Although we have no choice about *whether* to assess, we do have some control over *how* we assess, over whether the assessments produce meaningful information, and what we do with that information. The core assessment we employed during the pilot (and the version that we are using this year) are "rough draft" approaches. Like all drafts, it will require feedback and revision before it evolves into something we all value. We know this, and we look forward to reading or hearing your ideas.

Please send questions or feedback to Gray Scott, Michelle Buggs, or Terry Senne. We may also, from time to time, survey faculty, staff, or students to collect additional input. If you can take a few minutes to answer those surveys, we would be most grateful.

Thank you.

Gray Scott, Ph.D.
Assistant Director of Academic Assessment

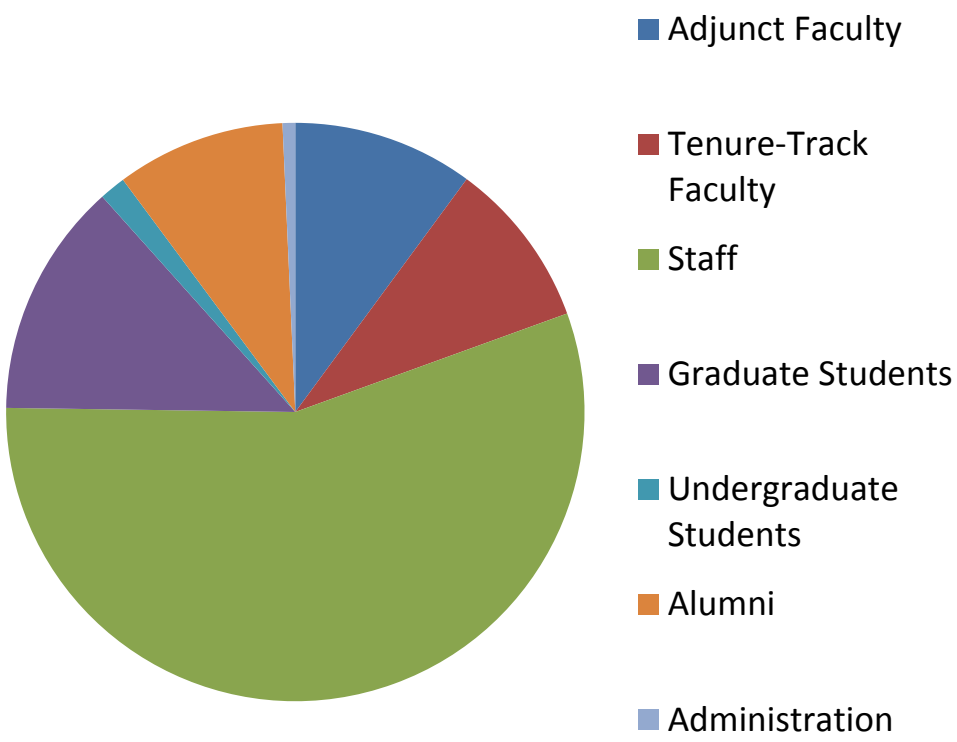
Our Raters

We would like to thank the following raters for their gracious assistance during the pilot.

Melissa Bittner
Valerie Borgfield
Renae Bruce
Nancy Casey
Katherine Coffey
Wylijanna Cole
Jennifer Danley-Scott
Svetlana Galuzinschii
Brandie Golleher
Diana Gomez
Liliana Grosso
Lisa Haynes
Amy Hicks
Garrett Isom
Vicki Jennings
Emily Johns
Rachelle Land
Angelica Landeros
Ileana Lopez-Jimenez
Mindy Menn
Sarah Merrill
Lauren Meyer
Jessica Murphy
Annita Owens
Chre Parnell
Sita Periathiruvadi
Annie Phillips
Ann Rathbun
Michelle Reeves
Rachel Reeves
Brittanie Romine
Sully Saucedo
René Scott
Ashley Spinozzi
Teresa Starrett
Abigail Tilton
Suzanne Townsdin
Akeitha Walton
Susan Whitmer
Danielle Woolery

Rater Participation

Figure 1. Proportion of Artifacts Rated by Each Rater Type



A diverse group of volunteer raters, each named on the previous page, rated artifacts for the pilot. Students, administrators, staff, and faculty came together to complete rater-training in early Spring 2015 and met to assess student work in late May and early June.

A handful of raters also agreed to assess **live student performances and speeches** in classrooms during the term. As a result, a broad cross-section of the TWU community got to see what a broad cross-section of our faculty and students are doing in the classroom—and it was routinely inspiring, encouraging, and enlightening. In disciplines as diverse as art, women's studies, and math, students are often grappling with well-designed intellectual challenges. And what we saw, consistently, was that when students were challenged to meet tough expectations, they rose to meet them.

As Figure 1 (this page) shows, most artifacts were rated by **staff**. **Undergraduate**, **graduate**, and **former students** comprised another large wedge. **Adjunct faculty**, **tenure-line faculty**, and **administrators** formed the smallest of the three major groupings.

Rater Participation, Cont'd

Figure 2. Average Rating by Type of Rater

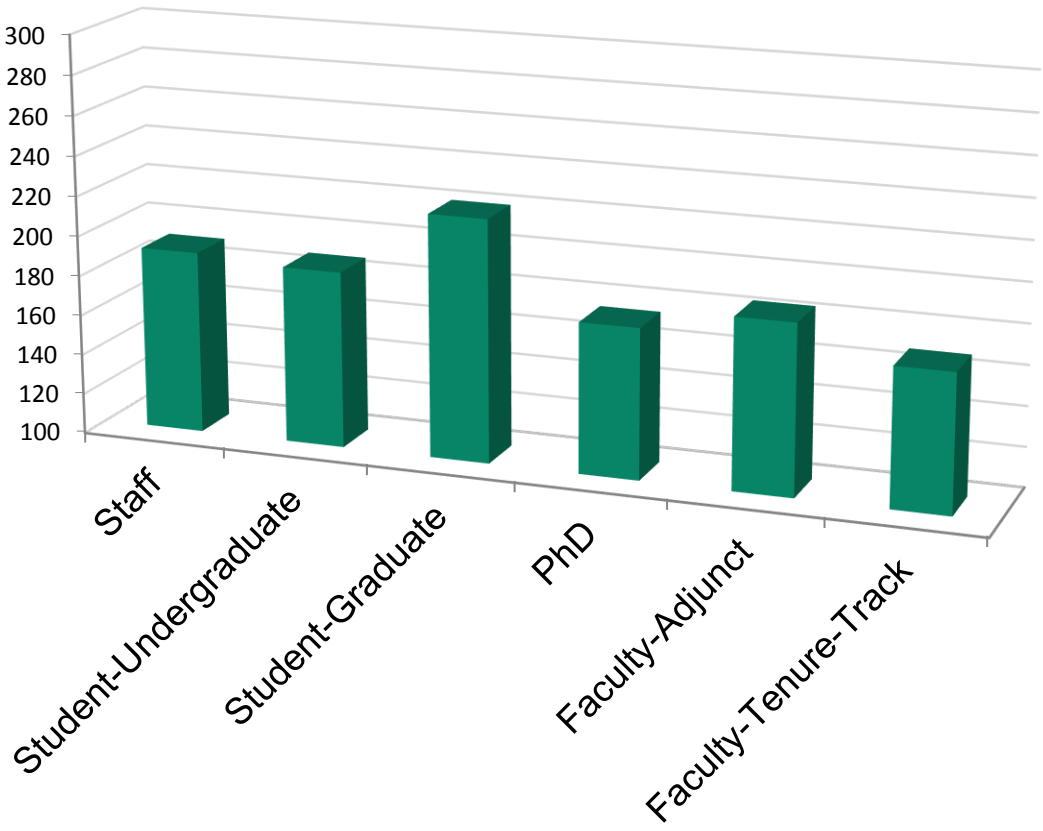


Figure 2 (left) displays the differences in how raters scored, providing the average for each type on a scale of 100 to 300. Although raters were trained for consistent scoring, rater reliability was (as we'll see later) quite low. Six objectives, 10 to 15 criteria for each rubric, and eight foundational areas with a wide array of classes and assignments made comprehensive training an unrealistic goal.

As a result, differences emerged in how tough or kind each type of rater was. Graduate student raters proved the most generous, followed by staff. Raters with PhDs, particularly those on the tenure-track, were tougher. This pattern suggests we may need to control for future changes in rater population. If more tenure-track faculty rate, for instance, scores might appear to drop even if students are learning more.

The next page (**Table 1**) displays what percentage of faculty in each area submitted artifacts online and how many artifacts from each area could be rated. **Figure 3** compares student participants to TWU's overall demographics.

Faculty Participation

Table 1. Participation by Foundational Component Area[1]

Foundational Component Area	% Faculty Submitting Artifacts	% Student Artifacts Submitted	% of Rated Entries Recorded as "No Artifact"	% of Criteria Rated "Not Applicable"
Communication	100%	100%	35%	6%
Mathematics	85%	89%	30%	7%
Life & Physical Sciences	63%	67%	36%	48%
Language, Philosophy, & Culture	99%	94%	56%	5%
Creative Arts	77%	76%	21%	19%
History	100%	99%	39%	0%
Government	100%	96%	15%	0%
Social & Behavioral Sciences	66%	73%	44%	23%

The 8 **Foundational Component Areas** of the Core Curriculum are listed in the left column.

% Student Artifacts Submitted refers to the percentage of sampled students in each area for whom faculty completed the online artifact-submission form or arranged for live rating.

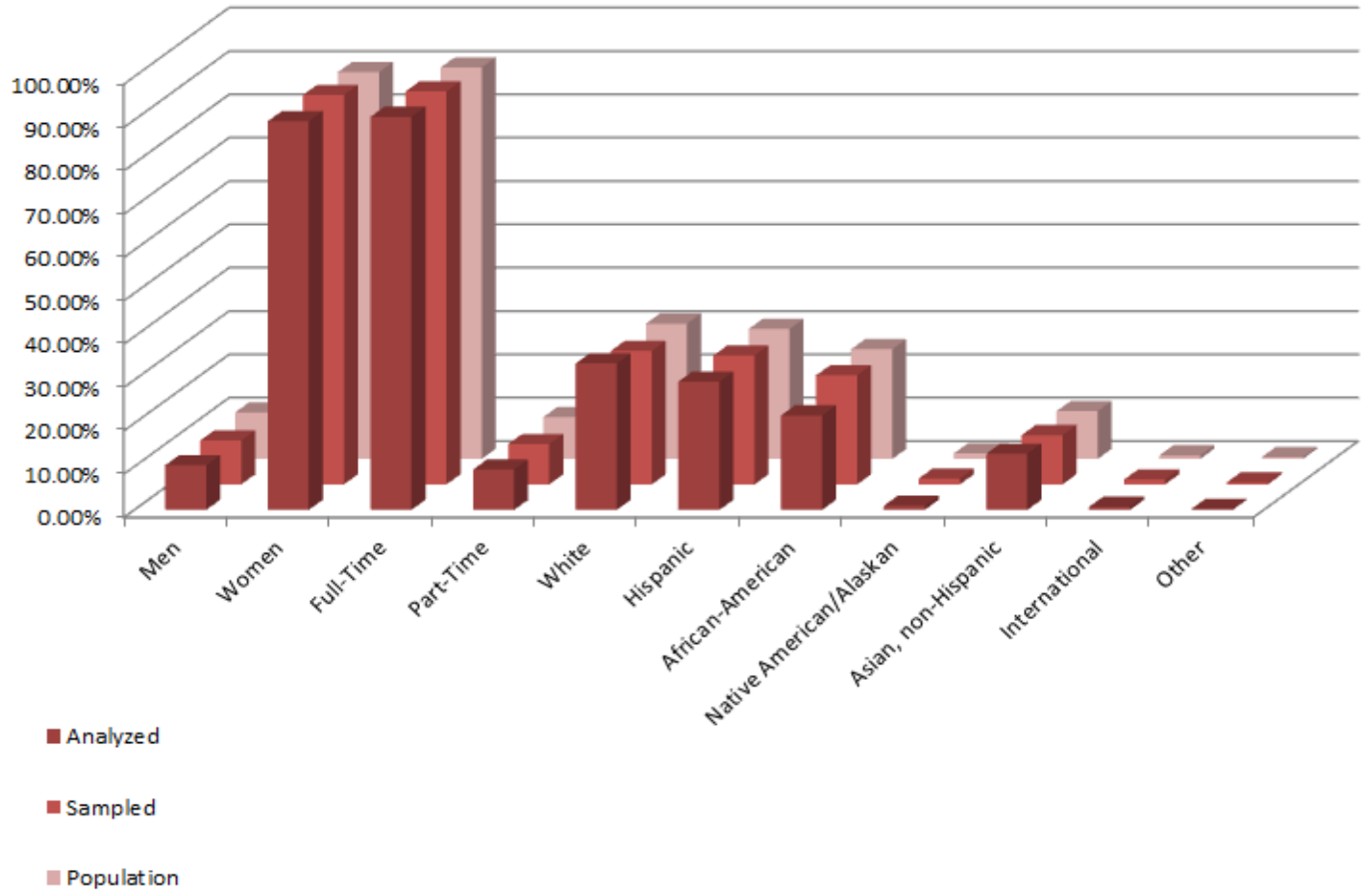
The **No Artifact** percentage indicates how often those faculty indicated students had not submitted artifacts to attach. (Average: 35%)

The **Not Applicable** column indicates how often raters reported that the artifact could not be assessed on the criteria selected by faculty. (Average: 13.5%)

[1] Not including hard-copy submissions and video uploads.

Student Participation

Figure 3. Sample Student Demographics, Compared with TWU Overall



Thanks to a sampling design created by Mark Hamner, Associate Provost for Institutional Research and Improvement, the **sampled** student population closely matched the **overall** demographics of students enrolled in core classes.

In fact, even though we were unable to rate some artifacts and other artifacts, though rated, are not included in Figure 1 because we were unable to match them to student IDs, the subset of students who were **analyzed** also closely matches the overall core population.

Benchmarks for Student Performance

The first question you might have – "How did students do?" – is one of the tougher questions to answer. We have no basis for comparison. We can't compare the pilot results to those of a previous year. Nor can we (as you'll see in Table 2 below) responsibly compare discipline to discipline. Some departments had students do fairly easy things in relaxed, take-home, or extra-credit conditions. Others submitted student artifacts that were challenging in-class, closed-note, written exams. What we can do with the pilot data is set some benchmarks for interpreting future results *within a given area*. The table below tells you the average for each core objective in each subject area, as well as the range within which roughly two-thirds of the students fell. As we conduct future assessments, look for gains or losses in these areas and use them to spark conversations about what might be causing those trends.

Table 2. Typical Scores by Strata (Core Area)

STRATA and OBJECTIVES	Average	Standard Deviation	Typical Range
COMMUNICATION	**183**	**69**	**From 115 to 252**
COMMUNICATION	194	65	From 128 to 259
CRITICAL THINKING	172	69	From 102 to 241
PERSONAL RESPONSIBILITY	208	67	From 141 to 275
TEAMWORK	157	62	From 094 to 219
MATHEMATICS	**191**	**82**	**From 108 to 273**
COMMUNICATION	186	80	From 106 to 265
CRITICAL THINKING	190	84	From 106 to 273
EMPIRICAL/QUANTITATIVE	195	82	From 113 to 277
LIFE & PHYSICAL SCIENCES	**201**	**74**	**From 126 to 275**
COMMUNICATION	206	71	From 134 to 277
CRITICAL THINKING	190	69	From 121 to 258
EMPIRICAL/QUANTITATIVE	198	78	From 120 to 276
TEAMWORK	229	84	From 144 to 300
LANGUAGE, PHILOSOPHY, & CULTURE	**186**	**72**	**From 114 to 258**
COMMUNICATION	202	69	From 133 to 270
CRITICAL THINKING	183	71	From 111 to 254
PERSONAL RESPONSIBILITY	184	74	From 110 to 258
SOCIAL RESPONSIBILITY	173	73	From 100 to 246

STRATA and OBJECTIVES	Average	Standard Deviation	Typical Range
CREATIVE ARTS	**207**	**72**	**From 136 to 279**
COMMUNICATION	223	67	From 156 to 291
CRITICAL THINKING	206	75	From 131 to 281
SOCIAL RESPONSIBILITY	184	65	From 119 to 250
TEAMWORK	208	73	From 136 to 281
HISTORY	**249**	**66**	**From 183 to 300**
COMMUNICATION	247	68	From 178 to 300
CRITICAL THINKING	251	69	From 182 to 300
PERSONAL RESPONSIBILITY	244	63	From 181 to 300
SOCIAL RESPONSIBILITY	253	62	From 190 to 300
GOVERNMENT	**236**	**73**	**From 163 to 300**
COMMUNICATION	238	73	From 165 to 300
CRITICAL THINKING	241	69	From 172 to 300
PERSONAL RESPONSIBILITY	225	77	From 148 to 300
SOCIAL RESPONSIBILITY	239	71	From 168 to 300
SOCIAL & BEHAVIORAL SCIENCES	**200**	**73**	**From 127 to 273**
COMMUNICATION	200	72	From 128 to 272
CRITICAL THINKING	206	72	From 134 to 278
EMPIRICAL/QUANTITATIVE	190	75	From 115 to 264
SOCIAL RESPONSIBILITY	175	50	From 125 to 225

Factors Affecting Success

Figure 5. How Each Class Level Performed:

♦Average Scores

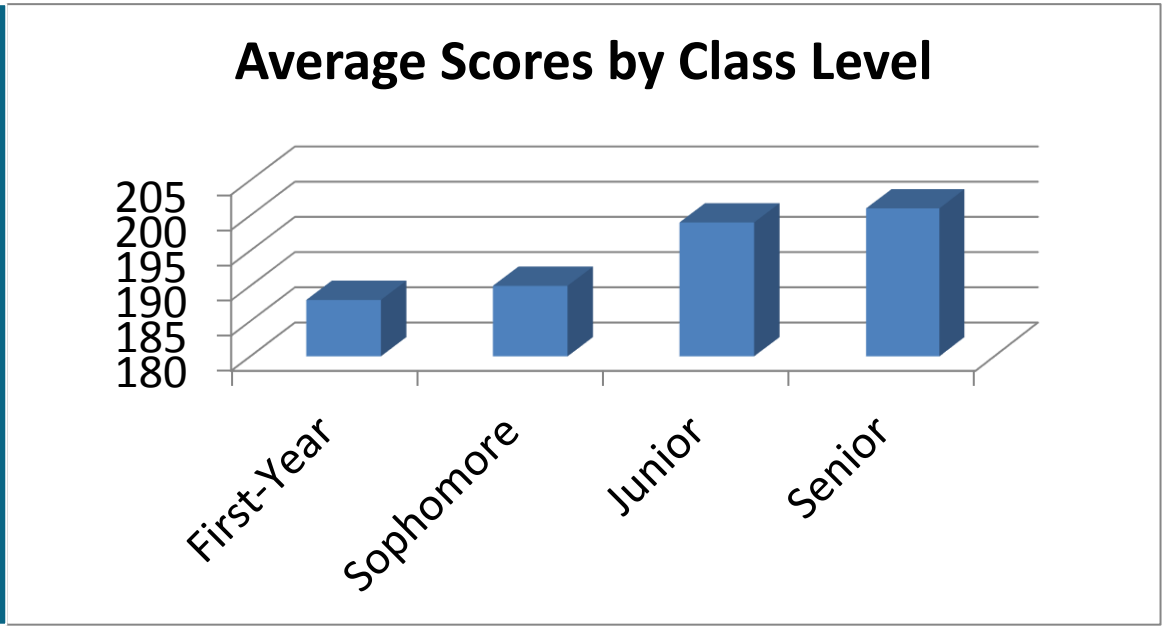
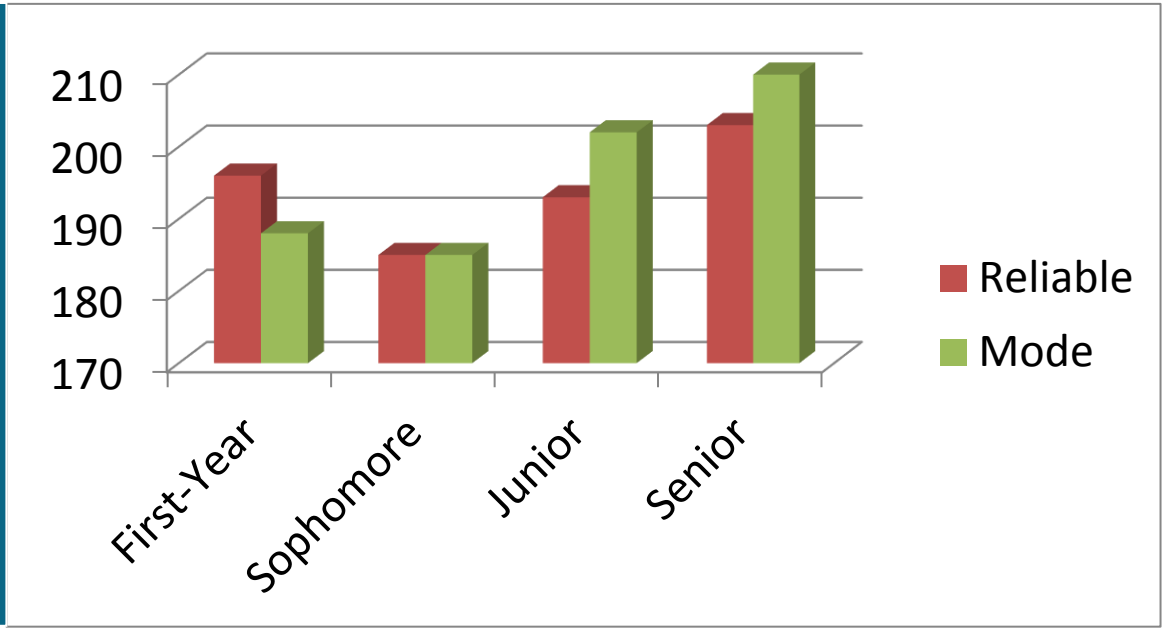


Figure 6. How Each Class Level Performed:

♦ Most Reliable Criteria (Reliable)

♦ Majority Rater Opinions (Mode)



Although our raters didn't know the class levels of the students they assessed, clear differences emerged among levels. At each class level, the **average** score improved, with a large jump from sophomore to junior year. We suspect student attrition may explain some of the jump from year 2 to year 3.

When we control better for rater variability, however, a puzzling drop appears between years 1 and 2. "Reliable" reflects results only on criteria with inter-rater reliability above 0.2.* Because most artifacts were rated multiple times, it was also possible to look only at artifacts about which raters—or a majority of raters—agreed on a rating ("Mode").

We are still exploring hypotheses for the year 2 dip.

** A 0.2 is low, but it was average for this pilot.*

Factors Affecting Success

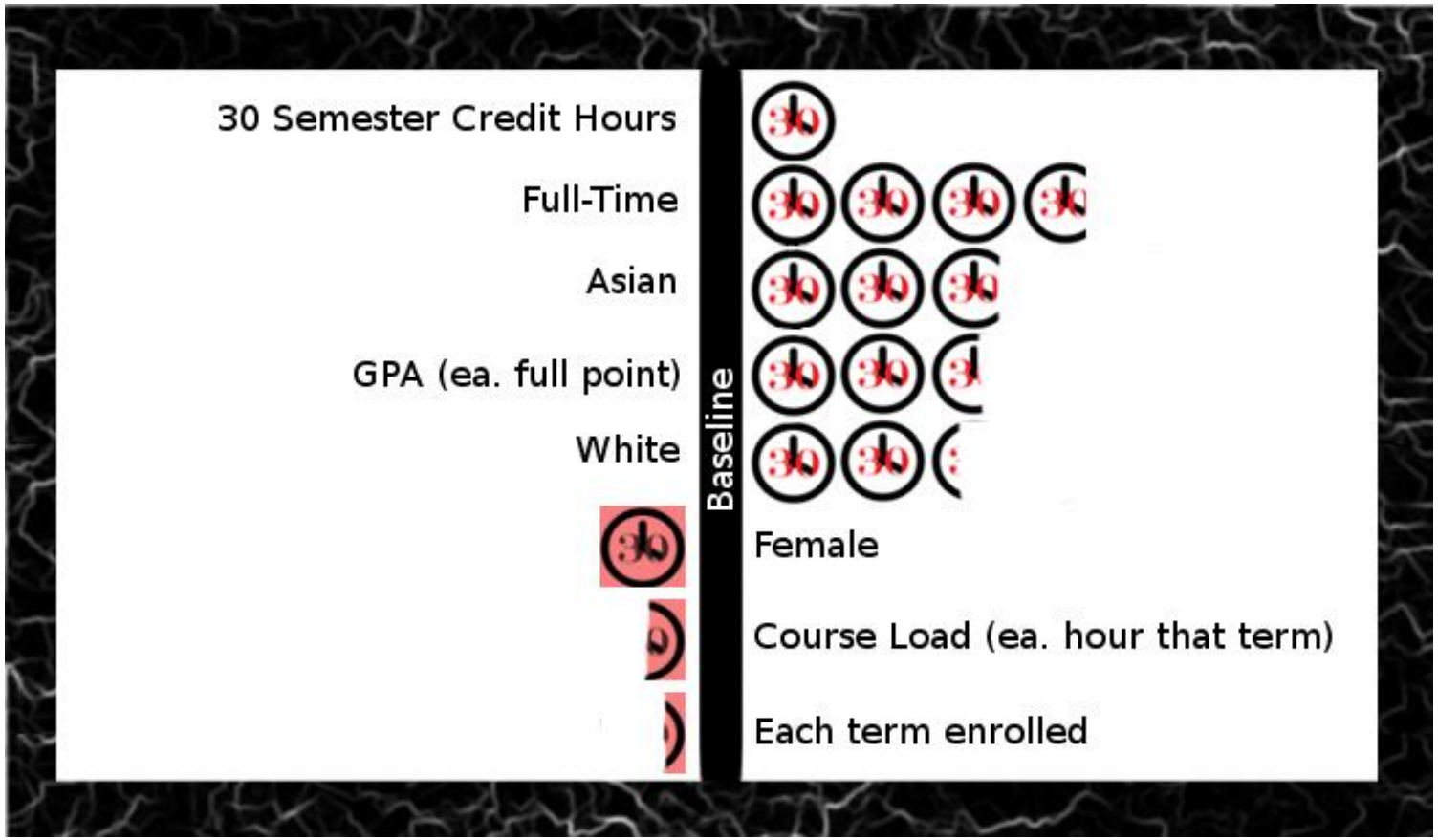


Figure 7. Measuring Variable Impacts

A Note on Gender Impacts. Red-shaded icons above indicate *negative* results. By far the variable that has provoked the most discussion is gender. In defiance of national trends, women students at TWU tended to score 30 hours behind men once GPA and other variables were controlled for. (Women had higher GPAs, however, with an average of 3.01 compared with 2.84 for men.) Selection bias may play a role in this result, in that men who select TWU may be different from college males elsewhere and may be more likely than women to choose TWU for specific programs (like nursing). Nevertheless, men's lower GPA suggests TWU men may be less likely than women to perform at their potential.

For every 30 semester credit hours a typical participating student had completed, the student added 8 points to a base score of 143 (on a 300-point scale).

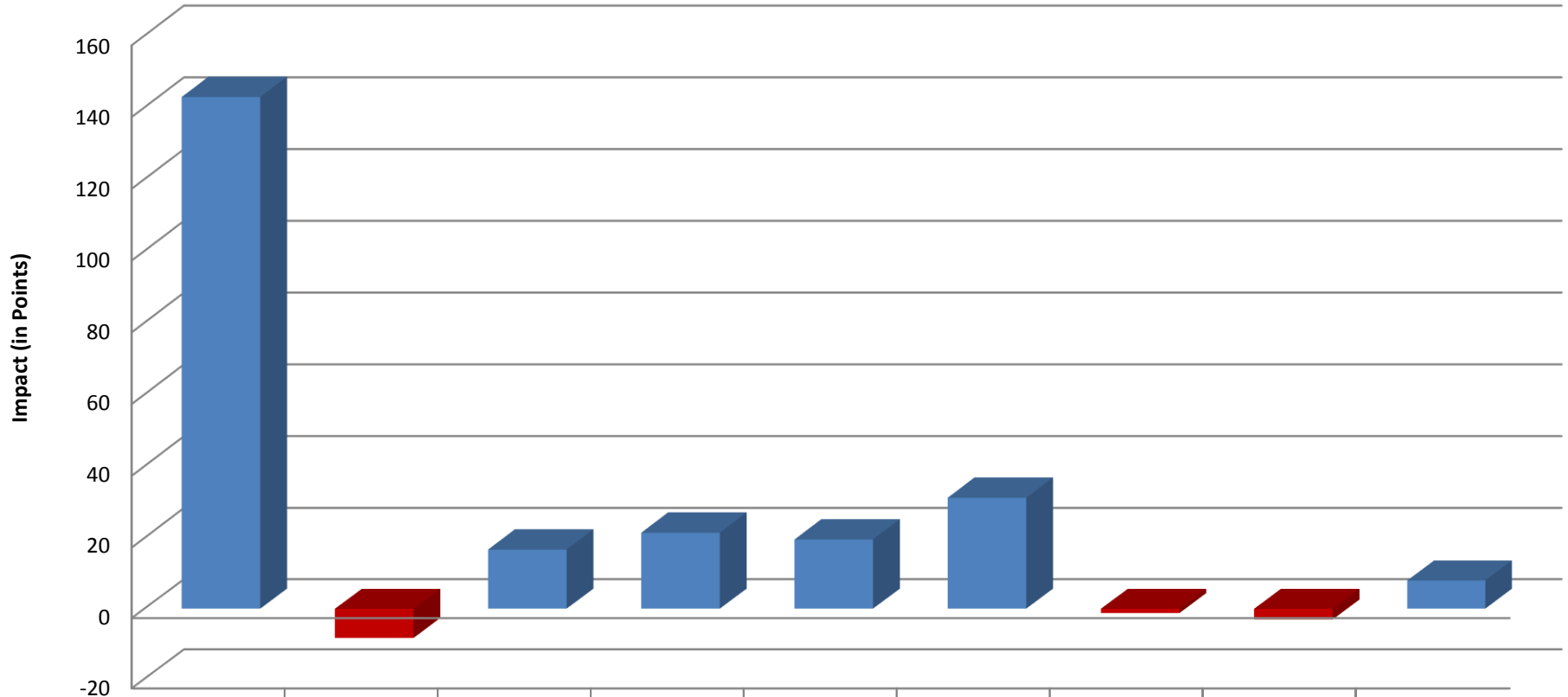
We can use that 30-hour impact (i.e., a year of full-time study) as a way to measure the importance of other variables. For instance, full-time students outperformed part-time students to a degree that was roughly equivalent to being 120 hours ahead in the curriculum.

Asian and White students performed more than 60 hours ahead of their peers. And GPA was highly associated with assessment performance, with each full point of GPA predicting a 60+ hour difference.

** Results in Figure 7 emerged from multiple regression analysis. All findings in this figure were significant at better than p<0.01.*

Factors Affecting Success

Figure 8. Significant Variables and Their Effects



	Base	Female	White	Asian	Cumulative GPA	Full-Time	# of Terms Enrolled	Course Load that Term	Effect of 30 SCH
Variable Impacts	142.896	-8.318	16.761	21.466	19.57	31.199	-1.225	-3.055	8.022

Figure 8 (above) offers another way to visualize the regression data from Figure 7. To predict a first-year student's likely score, start with a base of about 143 points and adjust as noted above for race, sex, full-time/part-time status, GPA, and so forth. For instance, the above model would predict that an Hispanic, female sophomore who has completed 60 semester-credit hours at a 3.0 GPA, and who has been enrolled for three terms with 15 SCH in the term being assessed, would end up with an average assessment score of 190.4 on a 300-point scale.

Assessing the Assessment

Evaluating Criteria & Their Rating

The remaining pages of this report evaluate the criteria that we have been using, the accuracy with which faculty selected criteria, and the consistency with which raters scored those criteria.

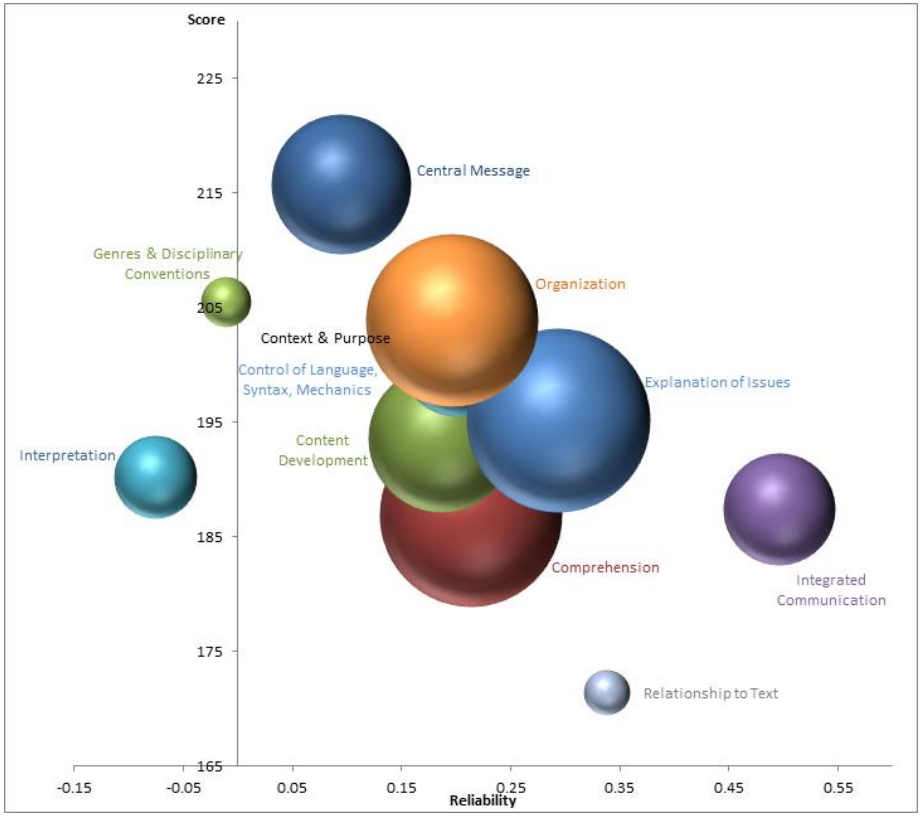
The goal of these pages is to help us assess our assessments. By honing our instruments and processes, we hope to improve the accuracy of our system while minimizing intrusions into pedagogy.

Although our long-term goal is to use the results to steer conversations on how to better educate our students, those conversations will only be as good as the data that spark them.

This pilot report gives us some starting benchmarks and direction for future improvements to the system.

The pages that follow deal in turn with each of the six core objectives: Communication, Critical Thinking, Empirical & Quantitative Skills, Personal Responsibility, Social Responsibility, and Teamwork.

The first chart for each objective is a **criteria map**.



A small-scale criteria map appears above, but those later will be large enough to read. Each criteria map shows how students did, how often faculty used each criterion, and how much raters agreed with each other.

Following each map is a **table** listing the criteria for each objective and all associated data. The next page defines the headers and columns for those tables and explains the color-codes used in them.

Decoding Headers and Color-Codes in Tables 3 through 8

In the following pages, each Core Objective table features these columns:

- **Reliability** (Krippendorff's alpha). A 0 represents no agreement beyond chance. A 1.0 represents perfect agreement beyond chance.
- **Score**. Average student performance for the criterion on a scale of 100-300. A 200 meets our assessment target. (The overall average was196.)
- **Count**. How many times we rated students on that criterion. Indirectly suggests how often faculty selected each criterion.
- **N/A**. How often raters selected "N/A" (not applicable). Indicates the criterion might not mean what faculty choosing it thought it meant.
- **Flagged**. How often raters, in a survey, indicated the criterion might need revision or rethinking or re-naming.
- **Fmode**. On a 1 to 5 scale (5 being highly applicable to their assignments, 1 indicating they picked the criterion only because they had to pick 4), how did most surveyed faculty rate the criterion? (Figure represents the most common answer.)
- **Favg**. As Fmode, but provides the average (mean) faculty response, rather than the most common one.
- **Fstdev**. As Fmode, but presents the standard deviation for faculty responses. Roughly two-thirds of faculty gave ratings within 1 standard deviation of the mean.

Highlighted Criteria Names: **Green** indicates popular criteria that students do well on. **Red** indicates popular criteria that students struggle with. **Black** indicates unpopular, problematic criteria.

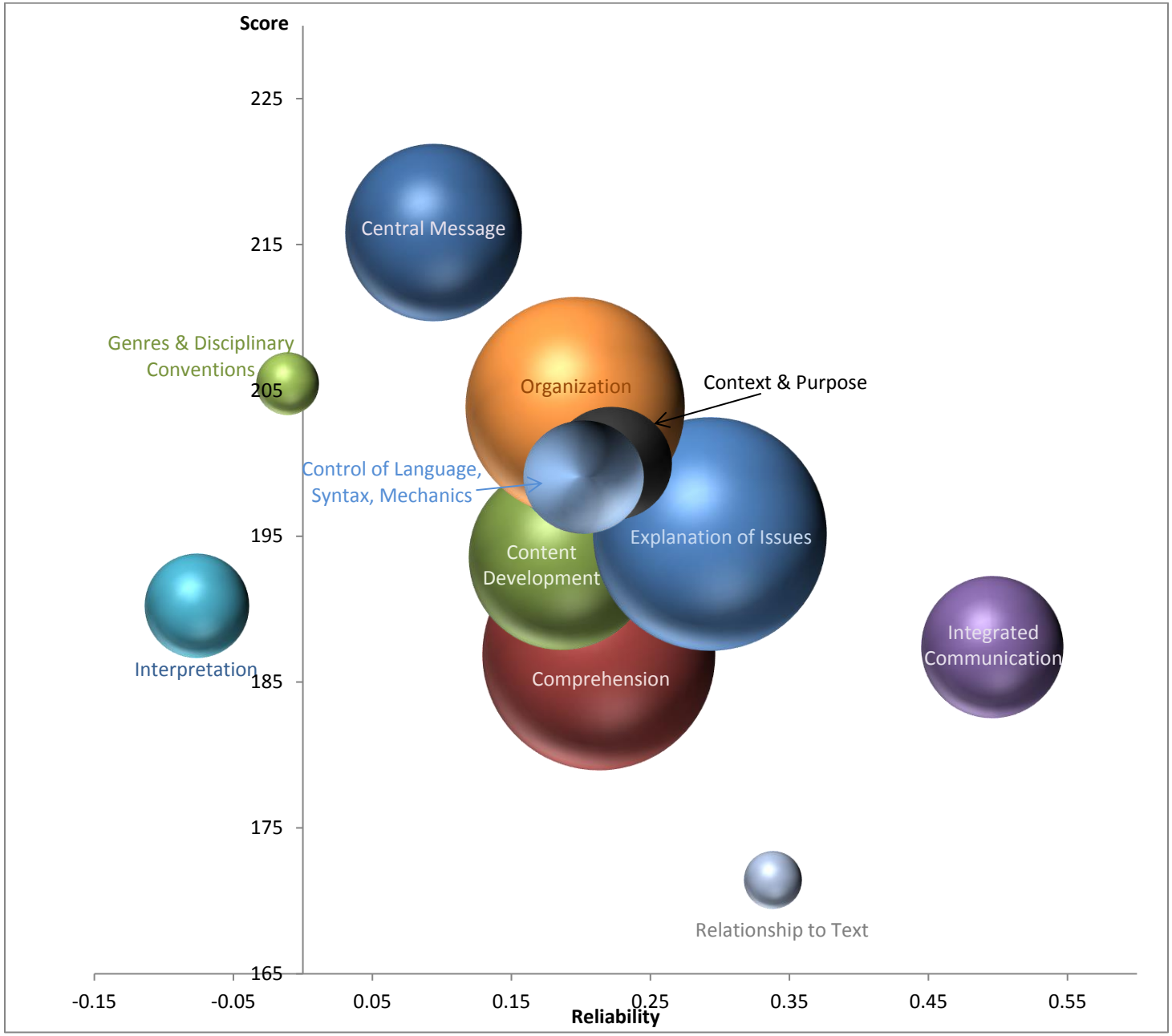
Highlighted Values: The more green the value, the better it compares with other criteria. The more red it is, the worse it compares to other criteria.

Reliability Issues

Our greatest challenge was **reliability**. Raters disagreed with each other enough that ratings for many criteria appear indistinguishable from chance. Some criteria, like **Implement Solution**, showed relatively high agreement—not at levels expected for peer-reviewed research, but at levels higher than expected for uncontrolled scenarios like customer reviews and assessments. Other criteria, like **Embracing Contradiction**, showed staggering *disagreement*. Several factors likely contributed to low reliability:

1. **We asked faculty to identify four criteria per rubric**. Struggling to find four that applied, some faculty picked criteria more or less at random. (**Genre**, for instance, seems to have been picked mostly out of desperation.) When raters rated, they sometimes gave these instances "N/A" ratings, but experience has showed us that often raters will attempt to rate *anyway*, at which point the ratings also become random. *Solution:* This term we've reduced the requirement to three criteria per objective and have cross-listed some criteria (as, for instance, both Empirical and Critical Thinking), to open up more opportunities to find suitable choices.
2. **Hurried faculty sometimes only read the name of the criteria**, and thus picked criteria inapplicable to what they had students do. **Delivery** is one example: It's an oral presentation criterion, but some faculty picked it for their written work. (**Interpretation (COM)** had a similar problem.) *Solution*: As we hold more workshops and gradually provide feedback, and as we rename some of the criteria, this factor may decline in influence.
3. **It's tough to maintain rater proficiency on this many criteria**. In an attempt to make these assessments accessible to a wide array of departments and assignments, we included many VALUE criteria. However, to train raters appropriately on all of them would take an estimated 112 hours *per rating session*. (Raters fall out of sync quickly – human nature.) For many criteria, our raters never received training. Instead, we trained on a sample of criteria. *Solution:* Over time, we expect faculty will cluster around a more limited range of common criteria, while others fall out of use enough that we might simply remove them from the list. Those black-coded on the tables above are first in line to consolidate, remove, or rewrite.

Criteria Map I. Communication



The **sizes of the spheres** correspond with how frequently those criteria were selected by faculty—the larger a criterion's bubble, the more often faculty picked it.

Score represents how well students did on the criterion (with scores ranging from 100 to 300 and averaging 192 for all criteria).

Reliability indicates roughly how often raters agreed when rating artifacts aligned with each criterion—the larger the value, the more faith we can have that the scores are accurate.

TABLE 3. COMMUNICATION CRITERIA

COMMUNICATION	Reliability	Score	Count	N/A	Flagged	Fmode	Favg	Fstdev
Central Message	0.094	216	471	5%	6.30%	5	3.6	1.6
Comprehension	0.213	187	815	11%	0%	5	4.1	1.2
Content Development	0.186	194	519	16%	12.50%	5	4	1
Context & Purpose	0.206	201	259	17%	6.30%	4	3.9	0.9
Control of Language, Syntax, Mechanics	0.2	200	295	4%	6.30%	4	3	1.25
Delivery		200	116	99%	18.80%	3	3.1	0.9
Explanation of Issues	0.293	195	819	7%	0%	5	4.4	1
Genres		180	5	0%	6.30%	1	2.3	1.6
Genres & Disciplinary Conventions	-0.011	205	59	7%	6.30%	1	2.8	1.7
Integrated Communication	0.496	187	300	15%	6.30%	5	3.5	1.2
Interpretation (COM)	-0.076	190	163	12%	12.50%	4	3.9	1.1
Organization	0.196	204	722	4%	0%	4	3.5	1.1
Relationship to Text	0.338	171	49	0%	12.50%	5	3.8	1.4

For a color key and explanation of column headers, see the Decoding page.

EVALUATING COMMUNICATIONS CRITERIA

Among Communications criteria, **Explanation of Issues** rules the roost. Highly ranked in faculty and rater surveys, it was also the most commonly assessed criterion. Although student scores in this area were average, it lands among our more reliable criteria.

Close behind, **Organization, Central Message,** and **Content Development** all proved to be popular choices among faculty. Students tended to do well on them. Although Content Development's N/A rate was well above the mean, faculty seem interested enough in that criterion that any time spent refining it will be well-spent. The other criteria named in this paragraph seemed to encounter few problems. Focusing rater training on these criteria could raise our reliability on the criteria that are most often encountered.

We red-coded **Comprehension**, **Delivery**, and **Integrated Communications**. All three of these criteria seem to be worth keeping: Comprehension was popular faculty choice; Delivery and Integrated Communications cover forms of communication that aren't writing. However, all three ran into N/A problems and students scored poorly

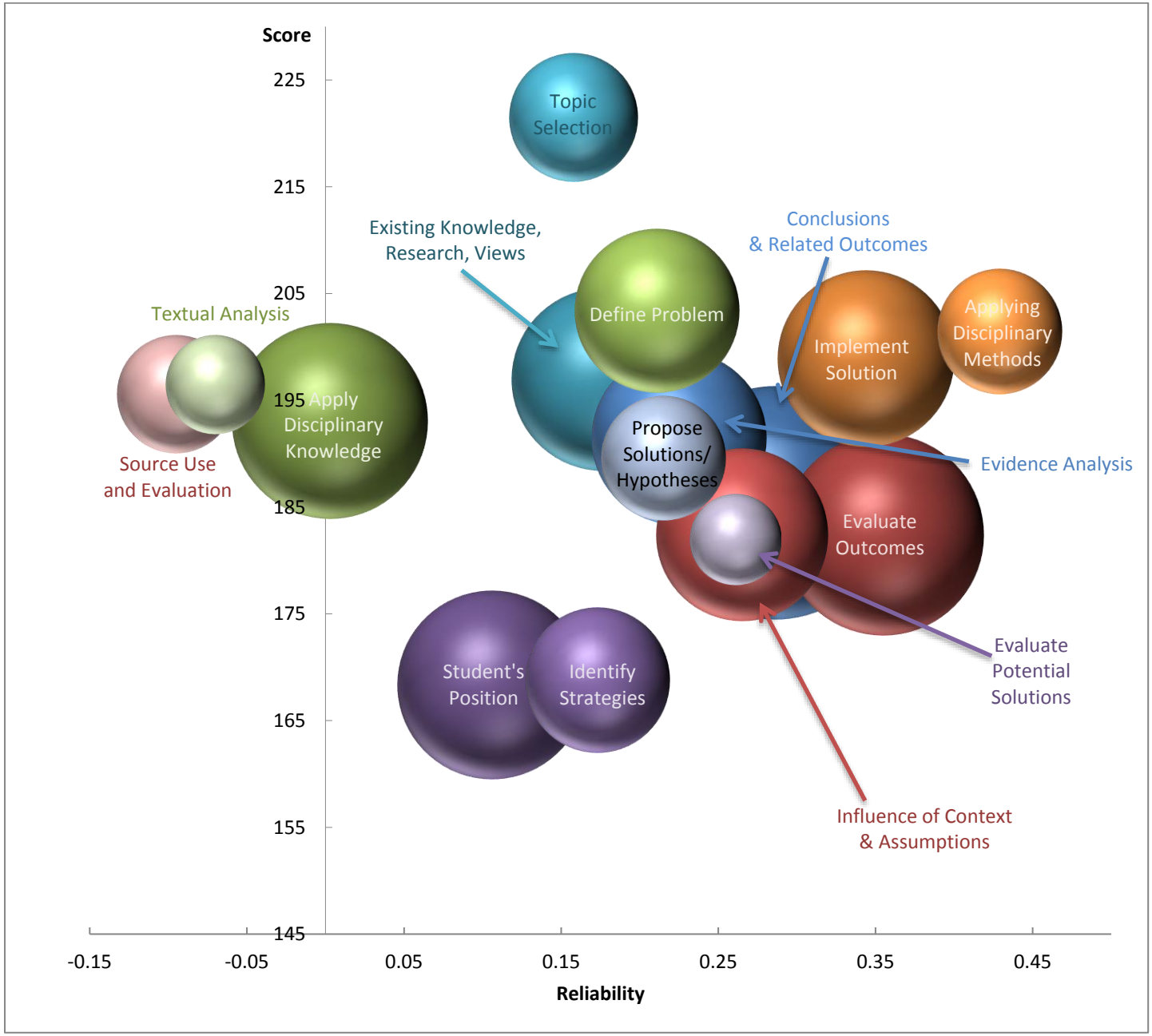
on two of them. For **Comprehension**, the issue was a simple one: raters couldn't assess a student's reading comprehension unless the faculty member had also submitted the text that the student was attempting to comprehend. Most faculty did so, but some didn't. (Comprehension may be most appropriate for tasks in which students respond to a single, assigned text, rather than for research papers, simply because raters won't be able to look up every source that a student has cited in order to check for comprehension.)

The N/A issues for Delivery and Integrated Communications, meanwhile, suggest confusion about what those criteria cover. (**Delivery** is for oral presentations, recorded or viewed live. **Integrated Communications** is for assignments that require students to blend writing with images, tables, and other non-word elements.) Improving our guidance to faculty—by including short descriptions rather than just criteria names on the assessment upload site—may mitigate these sorts of issues in the future.

We have black-coded **Genres**, **Genre & Disciplinary Conventions**, **Interpretation**, and **Relationship to Text**. These criteria were, at once, both rarely selected and highly problematic. Clarifying what the criteria are might help, but given how seldom these were selected, we're interested to see how much interest in them drops off now that we're only asking for three criteria per objective.

If faculty were selecting these only because they needed a fourth criterion (as the surveys suggest in some cases), then use of them might drop enough that we could simply remove them from the list. ♦

Criteria Map II. Critical Thinking



The **sizes of the spheres** correspond with how frequently those criteria were selected by faculty—the larger a criterion's bubble, the more often faculty picked it.

Score represents how well students did on the criterion (with scores ranging from 100 to 300 and averaging 192 for all criteria).

Reliability indicates roughly how often raters agreed when rating artifacts aligned with each criterion—the larger the value, the more faith we can have that the scores are accurate.

TABLE 4. CRITICAL THINKING CRITERIA

CRITICAL THINKING	Reliability	Score	Count	N/A	Flagged	Fmode	Favg	Fstdev
Apply Disciplinary Knowledge	0.003	193	578	28%	6.30%	5	4.2	1.3
Applying Disciplinary Methods	0.429	201	236	11%	0%	5	3.7	1.5
Conclusions & Related Outcomes	0.286	185	825	25%	6.30%	5	4.1	1.2
Define Problem	0.211	203	406	6%	6.30%	5	4.2	1.3
Evaluate Outcomes	0.355	182	612	12%	0%	5	4.4	0.8
Evaluate Potential Solutions	0.261	182	125	11%	0%	5	3.7	1.5
Evidence Analysis	0.225	191	460	8%	0%	4	3.8	1
Existing Knowledge, Research, Views	0.177	197	509	29%	6.30%	4	3.8	1.1
Identify Strategies	0.173	169	317	11%	0%	4	4	1.1
Implement Solution	0.344	199	469	23%	6.30%	5	3.7	1.4
Influence of Context and Assumptions	0.265	182	448	9%	0%	4	3.7	1.1
Propose Solutions/Hypotheses	0.215	190	236	6%	0%	5	3.8	1.3
Source Use & Evaluation	-0.095	196	211	3%	12.50%	4	3.1	1.5
Student's Position	0.106	168	538	14%	6.30%	5	3.4	1.6
Textual Analysis	-0.07	196	147	23%	18.80%	5	4.3	0.9
Topic Selection	0.158	221	249	8%	12.50%	1	3.2	1.6

For a color key and explanation of column headers, see the Decoding page.

EVALUATING CRITICAL THINKING CRITERIA

Four Critical Thinking criteria stood out positively in this analysis: **Define Problem**; **Evidence Analysis**; **Existing Knowledge, Research, Views**; and **Implement Solution**. All were frequently selected by faculty, had relatively high reliabilities, and scored close to average or better. Of these, **Evidence Analysis** was weaker on student performance while **Existing Knowledge** had more trouble with N/A ratings—raters and faculty didn't always seem to agree on what the criterion meant.

Our six red-coded criteria were typically more popular among faculty, but they often suffered from weaker scores, weaker reliability, or higher N/A ratings. Due to high demand, fine-tuning makes more sense than deletion.

Conclusions & Related Outcomes, for instance, might trigger less confusion if renamed to "Findings & Outcomes." **Evaluate Outcomes** might need to be renamed "Evaluate Attempted Solution." **Apply Disciplinary Knowledge** runs into reliability problems because it is unclear from the AAC&U language whether sources need

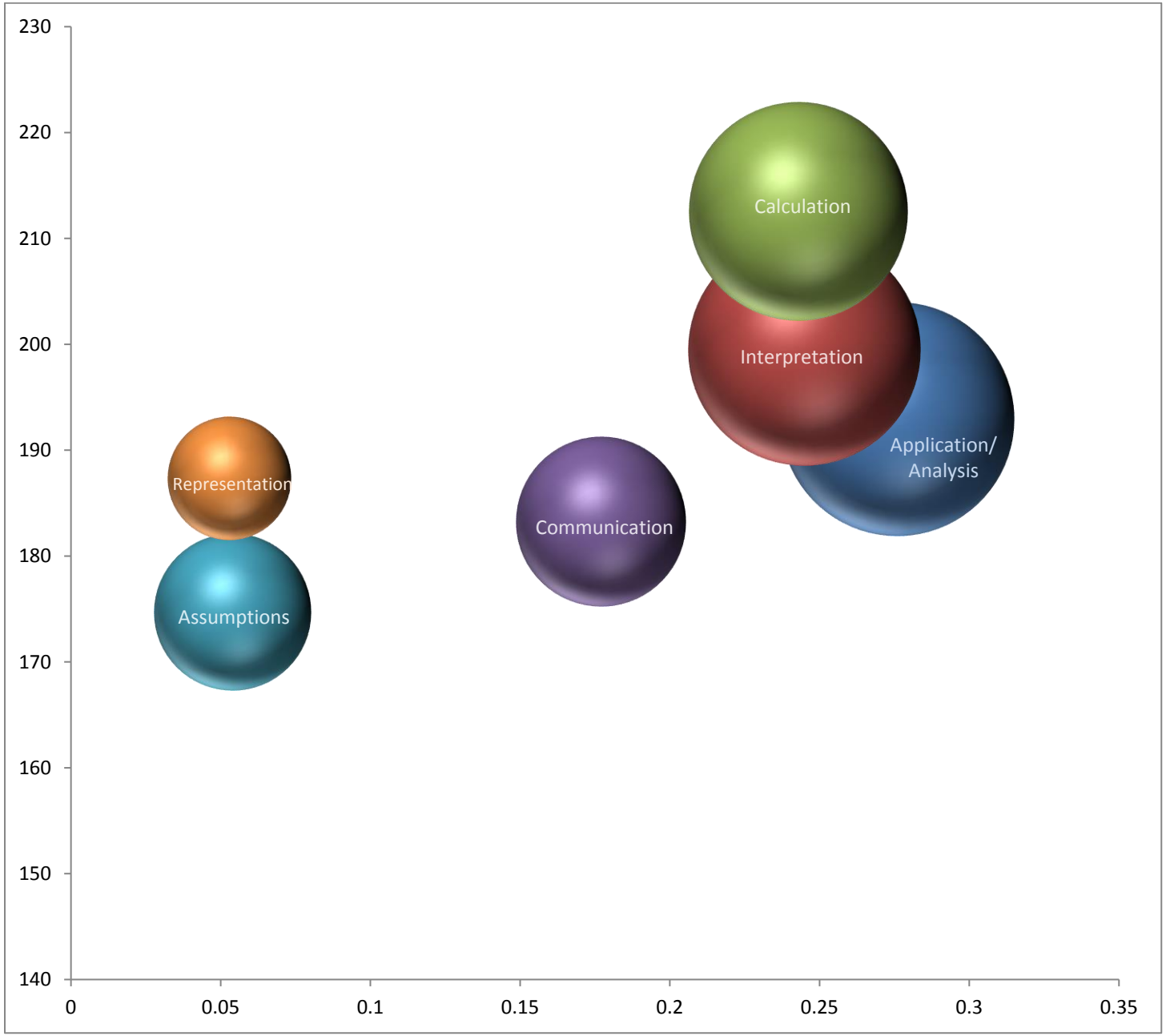
to be cited for that criterion. **Student's Position** should probably be consolidated with **Central Message** from the Communication rubric and cross-coded as applying to both objectives.

Unique among the red-coded criteria is **Influence of Contexts and Assumptions**, which performs relatively well on everything but average score – students may simply need more practice and instruction in this area.

Textual Analysis has been black-coded due to poor reliability, low use, and high N/A rates – most of which stems from faculty and rater confusion over what it means. (To clarify: One performs *Textual Analysis* when analyzing a speech for rhetorical features or analyzing poetry for poetical features.) It may be best to fold textual analysis under **Evidence Analaysis** as simply one kind of evidence among many.

Finally, the high-scoring **Topic Selection** was flagged by surveyed faculty as not being terribly helpful. Its high score may be deceptive, as it is among the least demanding criteria on the list. ♦

Criteria Map III. Empirical & Quantitative Skills



The **sizes of the spheres** correspond with how frequently those criteria were selected by faculty—the larger a criterion's bubble, the more often faculty picked it.

Score represents how well students did on the criterion (with scores ranging from 100 to 300 and averaging 192 for all criteria).

Reliability indicates roughly how often raters agreed when rating artifacts aligned with each criterion—the larger the value, the more faith we can have that the scores are accurate.

TABLE 5. EMPIRICAL/QUANTITATIVE CRITERIA

EMPIRICAL/QUANTITATIVE	Reliability	Score	Count	N/A	Flagged	Fmode	Favg	Fstdev
Application/Analysis	0.276	193	1038	22%	6.30%	5	4.4	0.7
Assumptions	0.054	175	468	48%	6.30%	5	4.1	0.8
Calculation	0.243	213	911	22%	0%	4	4.2	0.9
Communication	0.177	183	545	10%	6.30%	4	3.8	1.3
Interpretation (EQS)	0.245	199	1023	22%	0%	5	4.1	0.9
Representation	0.053	187	288	34%	6.30%	5	4	1

For a color key and explanation of column headers, see the Decoding page.

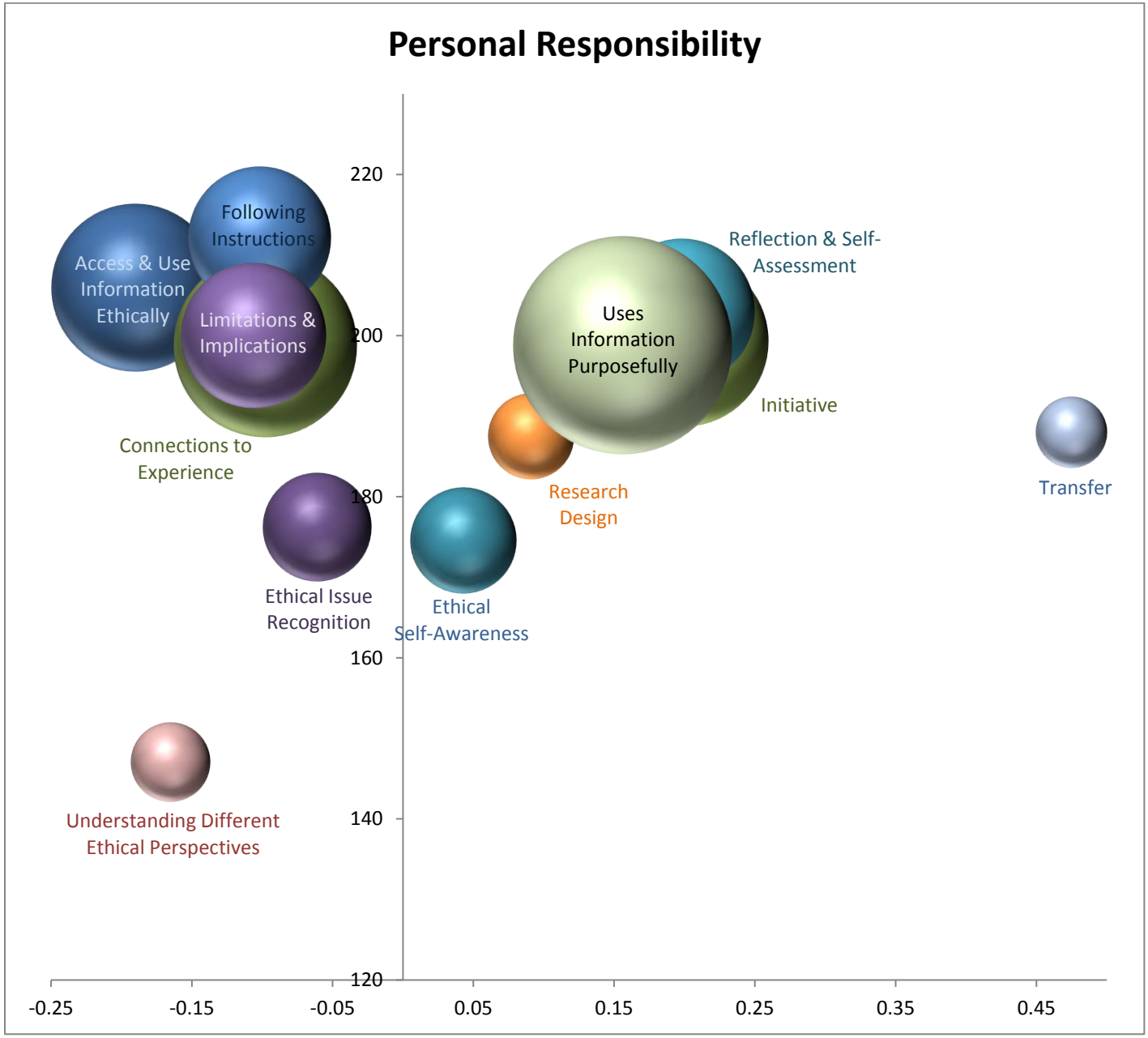
EVALUATING EMPIRICAL/QUANTITATIVE CRITERIA

The three most promising criteria for the Empirical/Quantitative objective are **Application/Analysis**, **Calculation**, and **Interpretation (EQS)**. All three were frequently selected by faculty, relatively reliable, and high-scoring.

They did, however, have high N/A rates. All of the Empirical criteria did. Raters often couldn't assess the artifacts that were submitted on the criteria that faculty picked because the criteria require written discussion or explanation of the student's thinking, and a lot of the submitted student work displayed only calculations or final answers.

The remaining three criteria ran into other trouble. **Assumptions** and **Representation** both had low reliability and particularly high N/A ratings, as well as low scores. Because surveyed faculty indicated the criteria were valuable – and we agree – these two may require some fine-tuning, or else we need to better coach faculty on how to set up assignments so that these can be assessed better. **Communication** also scored poorly. However, since it was not rated as highly by surveyed faculty and seems redundant with criteria on the Communications objective rubric, it may in the long run be better to remove it from the list. ♦

Criteria Map IV. Personal Responsibility



The **sizes of the spheres** correspond with how frequently those criteria were selected by faculty—the larger a criterion's bubble, the more often faculty picked it.

Score represents how well students did on the criterion (with scores ranging from 100 to 300 and averaging 192 for all criteria).

Reliability indicates roughly how often raters agreed when rating artifacts aligned with each criterion—the larger the value, the more faith we can have that the scores are accurate.

TABLE 6. PERSONAL RESPONSIBILITY

PERSONAL RESPONSIBILITY	Reliability	Score	Count	N/A	Flagged	Fmode	Favg	Fstdev
Access and Use Information Ethically and Legally	-0.19	206	153	1%	0%	1	2.5	1.7
Application of Ethical Perspectives/Concepts			1	100%	0%	1	3	2
Connections to Experience	-0.098	199	182	10%	6.30%	1	2.7	1.6
Ethical Issue Recognition	-0.061	176	64	2%	0%	4	4.5	0.5
Ethical Self-Awareness	0.043	175	61	3%	12.50%	5	5	0
Evaluate Information and Its Sources Critically		129	16	13%	0%	1	2.7	1.3
Following Instructions	-0.102	212	110	3%	0%			
Independence		100	5	80%	6.30%	1	3.3	1.7
Initiative	0.199	199	158	8%	6.30%	3	3.3	1.3
Limitations and Implications	-0.106	200	114	27%	0%	1	2.3	1.3
Reflection and Self-Assessment	0.198	203	116	12%	6.30%	5	4.3	0.9
Research Design	0.091	188	40	40%	0%	1	2.7	1.3
Transfer	0.475	188	27	7%	12.50%	4	3.3	1.3
Understanding Different Ethical Perspectives, etc.	-0.165	147	34	0%	6.30%	5	3.7	1.9
Uses Information Purposefully	0.156	199	259	5%	6.30%	1	3.2	1.7

For a color key and explanation of column headers, see the Decoding page.

EVALUATING PERSONAL RESPONSIBILITY CRITERIA

The **Personal Responsibility** objective had very uneven results.

Initiative, **Reflection and Self-Assessment**, and **Uses Information Purposefully** all had relatively high reliability, above-average scores, healthy usage frequencies, and low N/A rates. The remaining four green-coded criteria above had good usage and relatively high sores, but poor reliability. Better training of raters on such criteria, or improving the criteria descriptions, may help bring the reliabilities up.

However, many other criteria proved problematic. Four are black-coded because they had so few cases that we couldn't calculate reliability for them:

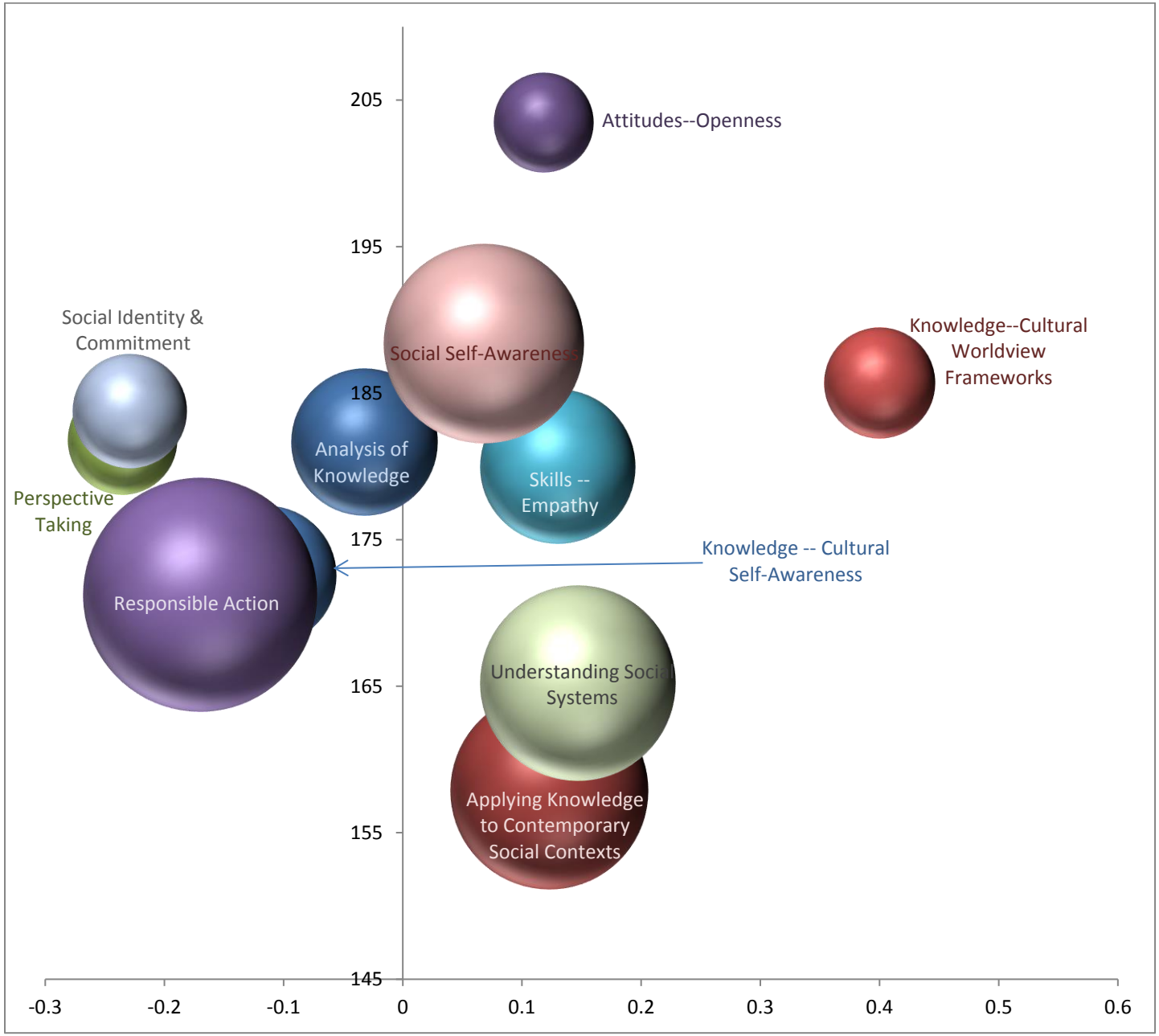
1. **Evaluate Information and Its Sources Critically** may be easily deleted because a similar criterion in the Critical Thinking set is now cross-coded as Personal Responsibility.
2. **Independence** may need to be deleted because of the difficulty involved in assessing it properly.
3. **Understanding Different Ethical Perspectives** is black-coded because of its low average score of 147,

low-usage, and low reliability.

4. **Application of Ethial Perspectives/Concepts** has no reliability rating at all because almost no one used it.

On a final note, **Research Design**, though not color-coded, has a relatively high N/A rating of 40%. We suspect this is because the criterion requires that students describe their methodology. When assignments didn't require students to describe their own methods, the resulting student artifacts couldn't be assessed. ♦

Criteria Map V. Social Responsibility



The **sizes of the spheres** correspond with how frequently those criteria were selected by faculty—the larger a criterion's bubble, the more often faculty picked it.

Score represents how well students did on the criterion (with scores ranging from 100 to 300 and averaging 192 for all criteria).

Reliability indicates roughly how often raters agreed when rating artifacts aligned with each criterion—the larger the value, the more faith we can have that the scores are accurate.

TABLE 7. SOCIAL RESPONSIBILITY

SOCIAL RESPONSIBILITY	Reliability	Score	Count	N/A	Flagged	Fmode	Favg	Fstdev
Analysis of Knowledge	-0.032	182	63	5%	0%	5	3.9	1.5
Applying Knowledge to Contemporary Social Contexts	0.123	158	115	1%	0%	5	3.6	1.5
Attitudes -- Curiosity		193	15	7%	0%	1	3.3	1.7
Attitudes -- Openness	0.118	203	29	0%	0%	5	3.5	1.8
Cultural Diversity		183	12	0%	6.30%	4	3.9	1.2
Diversity of Communities and Cultures						5	3.9	1.4
Knowledge -- Cultural Self-Awareness	-0.114	173	57	11%	6.30%	5	4.5	0.8
Knowledge -- Cultural Worldview Frameworks	0.4	186	36	3%	6.30%	5	4.4	0.7
Perspective Taking	-0.235	182	35	6%	6.30%	4	3.7	1.3
Responsible Action	-0.17	171	160	4%	0%	1	2.8	1.6
Skills -- Empathy	0.13	180	70	0%	6.30%	5	3.8	1.2
Skills -- Verbal and Nonverbal Communication		186	23	39%	0%	4	4.2	0.8
Social Identity and Commitment	-0.229	184	38	3%	0%	1	3.3	1.7
Social Self-Awareness	0.068	188	117	4%	6.30%	5	3.9	1.4
Understanding Social Systems	0.147	165	112	0%	0%	5	3.8	1.6

For a color key and explanation of column headers, see the Decoding page.

EVALUATING SOCIAL RESPONSIBILITY CRITERIA

The Social Responsibility criteria that we borrowed from the AAC&U VALUE rubrics polled well with faculty and, for the most part, had low N/A rates. Nevertheless, conferences with departments associated with this objective may be required to fix some of the problems that we're seeing here.

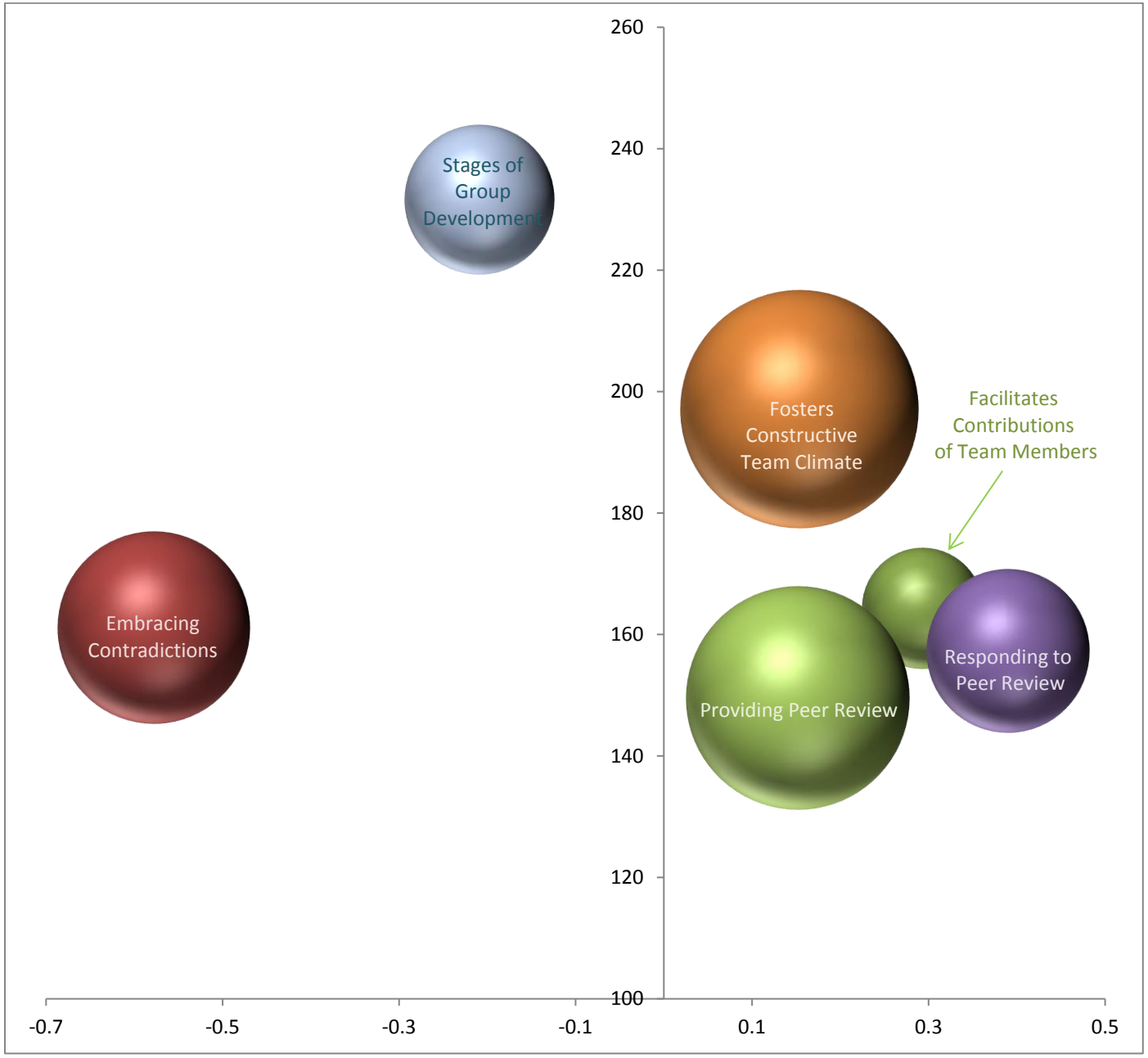
Only **Social Self-Awareness** emerged from this analysis with relatively high performances in reliability, score, and count – but even for that criterion, those numbers lag behind those of criteria for other objectives.

Otherwise, Social Responsibility had the lowest reliabilities, scores, and counts of any core objective. **Diversity of Communities and Cultures** was never rated or selected. **Attitudes-Curiosity** and **Cultural Diversity** both had 15 or fewer occurrences. Raters often couldn't rate artifacts on **Skills – Verbal and Nonverbal**, which was rarely

selected by faculty anyway.

Revising or offering faculty better training on the red-coded criteria – **Applying Knowledge to Contemporary Social Contexts**, **Responsible Action,** and **Understanding Social Systems** – could help faculty find a few viable options in this set. Otherwise, we may need to broaden the language of, or replace, some of the black-coded criteria to better accommodate faculty. ♦

Criteria Map VI. Teamwork



The **sizes of the spheres** correspond with how frequently those criteria were selected by faculty—the larger a criterion's bubble, the more often faculty picked it.

Score represents how well students did on the criterion (with scores ranging from 100 to 300 and averaging 192 for all criteria).

Reliability indicates roughly how often raters agreed when rating artifacts aligned with each criterion—the larger the value, the more faith we can have that the scores are accurate.

TABLE 8. TEAMWORK CRITERIA

TEAMWORK	Reliability	Score	Count	N/A	Flagged	Fmode	Favg	Fstdev
Contributes to Team Meetings		281	194	92%	18.80%	5	4.7	0.5
Embracing Contradictions	-0.578	161	177	42%	0%	5	3.9	1.4
Facilitates Contributions of Team Members	0.293	164	70	80%	18.80%	5	3.5	1.7
Follows Directions of Leader			1	100%	6.30%			
Follows Instructions		100	3	67%	0%			
Fosters Constructive Team Climate	0.154	197	272	62%	18.80%	5	3.6	1.5
Handles or Sets-Up Shared Property			55	100%	6.30%			
Individual Contributions Outside of Team Meetings		232	187	80%	18.80%	5	3.9	1.3
Providing Peer Review	0.152	150	240	52%	18.80%	5	4.4	1
Responding to Peer Review	0.39	157	128	25%	12.50%	4	4.4	0.5
Responds to Conflict			58	100%	12.50%	5	4.3	0.8
Responds to Director Feedback			5	100%	6.30%			
Stages of Group Development	-0.209	232	108	82%	6.30%	5	4.4	0.7

For a color key and explanation of column headers, see the Decoding page.

EVALUATING TEAMWORK CRITERIA

Like Social Responsibility, Teamwork has been a challenge. Chiefly this is because it is impossible to assess students on most of the listed criteria *without seeing the collaborative behavior*.

If, for instance, a faculty member simply submits group projects without any team minutes, records of team meetings, or the like, most of the criteria above cannot be rated. If the artifact is a video of a presentation but the students don't identify themselves on camera, the raters cannot identify which student is the one to assess.

For these reasons, Teamwork had by far the highest frequencies of N/A ratings, sometimes reaching 100% in the computerized (non-live) scoring sessions.

Even when criteria could be rated, there were other problems: **Embracing Contradictions** was, for instance, low-scoring and had a reliability rating so low (-.58) that it appears raters were disagreeing with each other at rates far beyond chance. Other criteria – like **Facilitates Contributions of Team Members** – scored well below average. Only **Fosters Constructive Team Climate** was coded green, with average reliability and scores, high usage, and a better-than-average N/A rate (for this objective).

Because Teamwork will be the first among the more problematic objectives to be assessed officially (in AY 2016-2017), our efforts have so far focused on ...

1) **Meeting with affected faculty** to coordinate better kinds of assignments for assessment. For instance, a cadre of faculty are working with the assessment office to pilot online, recorded, out-of-class collaborative tasks through Blackboard Collaborate.
2) **Revising criteria**. For instance, we have replaced the two Peer Review criteria with what we think will be a better set of three: **Applying Criteria through Peer Review**, **Constructive Framing of Peer Review**, and **Clarity of Peer Review**, all of which can be assessed from just a written peer review and the work being reviewed. ♦

COMMUNICATIONS & CRITICAL THINKING
SUGGESTIONS FOR COMMON ASSIGNMENTS

Because Communication and Critical Thinking apply across the core, it might be helpful to highlight some criteria from the objectives that fit well together and which apply to common tasks:

PROBLEM-SOLVING POWERPOINT PRESENTATION

Required: Video recording with student(s) clearly identified, or raters present for live rating

- **Communications**: Delivery, Integrated Communication, Organization
- **Critical Thinking**: Define Problem; Existing Knowledge, Research, Views; Propose Solutions/Hypotheses

ESSAY RESPONDING TO ASSIGNED TEXT

Required: Both student response and a copy of assigned text

- **Communications**: Comprehension, Explanation of Issues, Content Development
- **Critical Thinking**: Student's Position, Textual Analysis, Influence of Context & Assumptions

RESEARCH PAPER

- **Communications**: Genre and Disciplinary Conventions, Organization, Explanation of Issues
- **Critical Thinking**: Existing Knowledge, Research, Views; Source Use & Evaluation; Apply Disciplinary Methods.

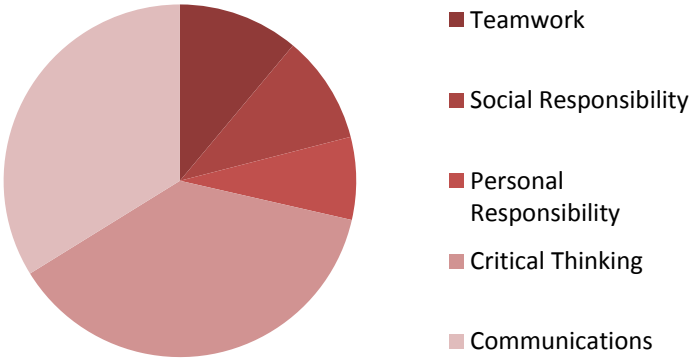
COMPARING SCIENTIFIC ARTICLE TO NEWS REPORT ON SAME DISCOVERY

Required: Student comparison, as well as both articles being compared

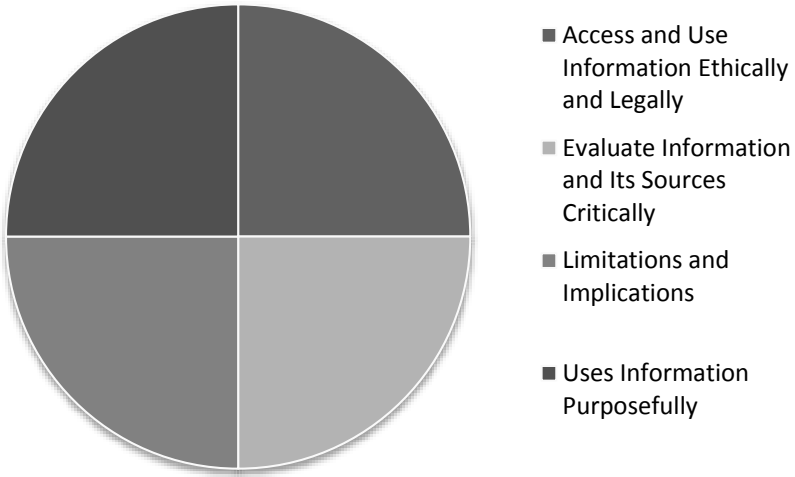
- **Communications**: Comprehension; Organization; Explanation of Issues
- **Critical Thinking**: Textual Analysis, Applying Disciplinary Knowledge, Conclusions & Related Outcomes

Live Rating Results

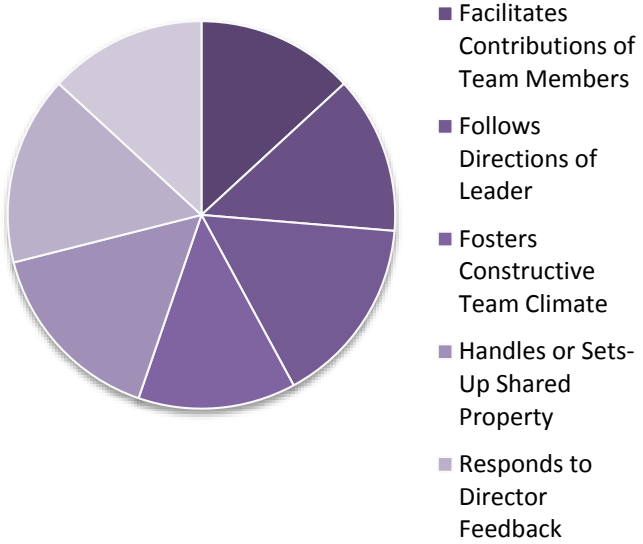
Live-Ratings by Core Objective



Personal Responsibility Criteria



Teamwork Criteria

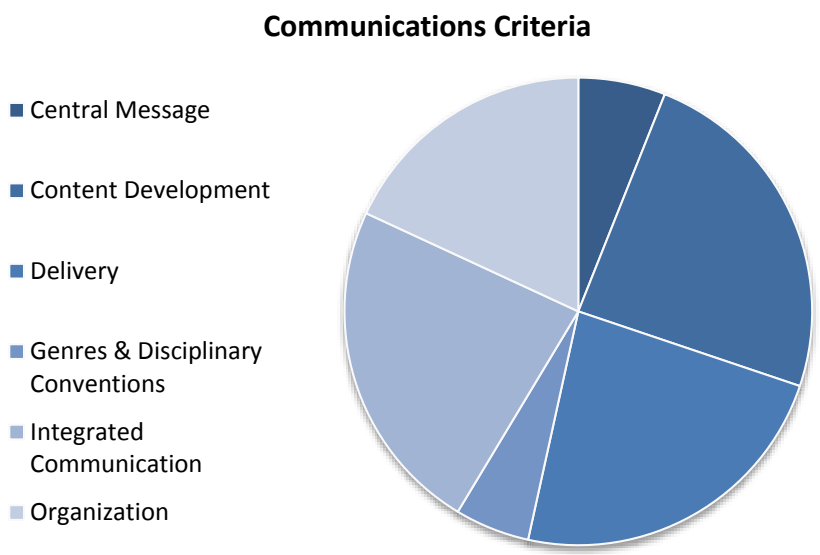


We also conducted live ratings during the pilot. Raters visited classes taught by seven TWU faculty and rated 44 live student activities, from speeches and team presentations to musical performances.

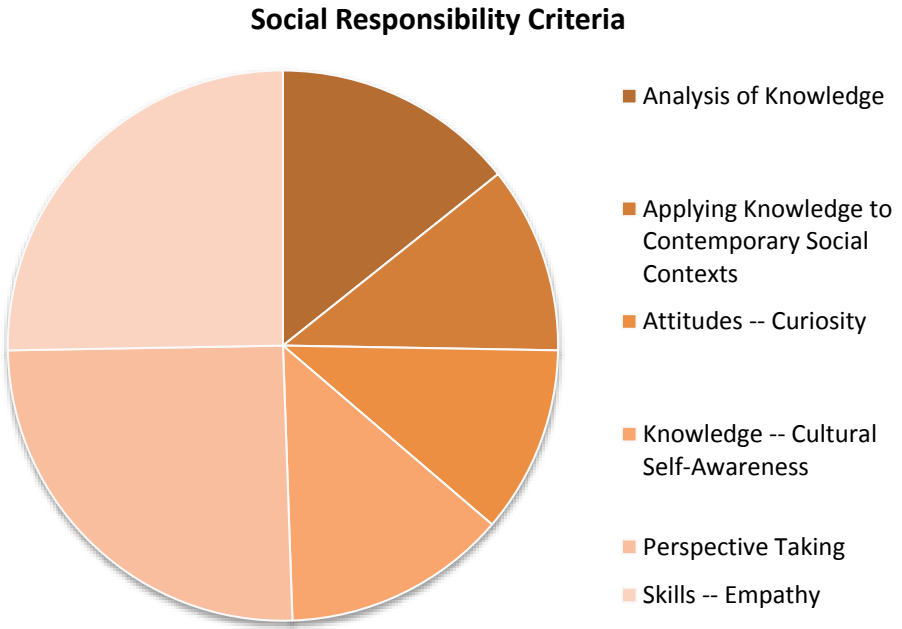
For this report, we've segregated the data for the live rating. This page identifies which criteria and objectives were assessed.

Empirical/Quantitative Skills were never assessed in live ratings. And four Teamwork criteria were *only* assessed live: Follows Directions of Leader, Fosters Constructive Team Climate, Handles or Sets up Shared Property, and Responds to Director Feedback.

Live Rating Results, Cont'd



The pie charts and other results in this section exclude criteria for which fewer than five students were assessed.



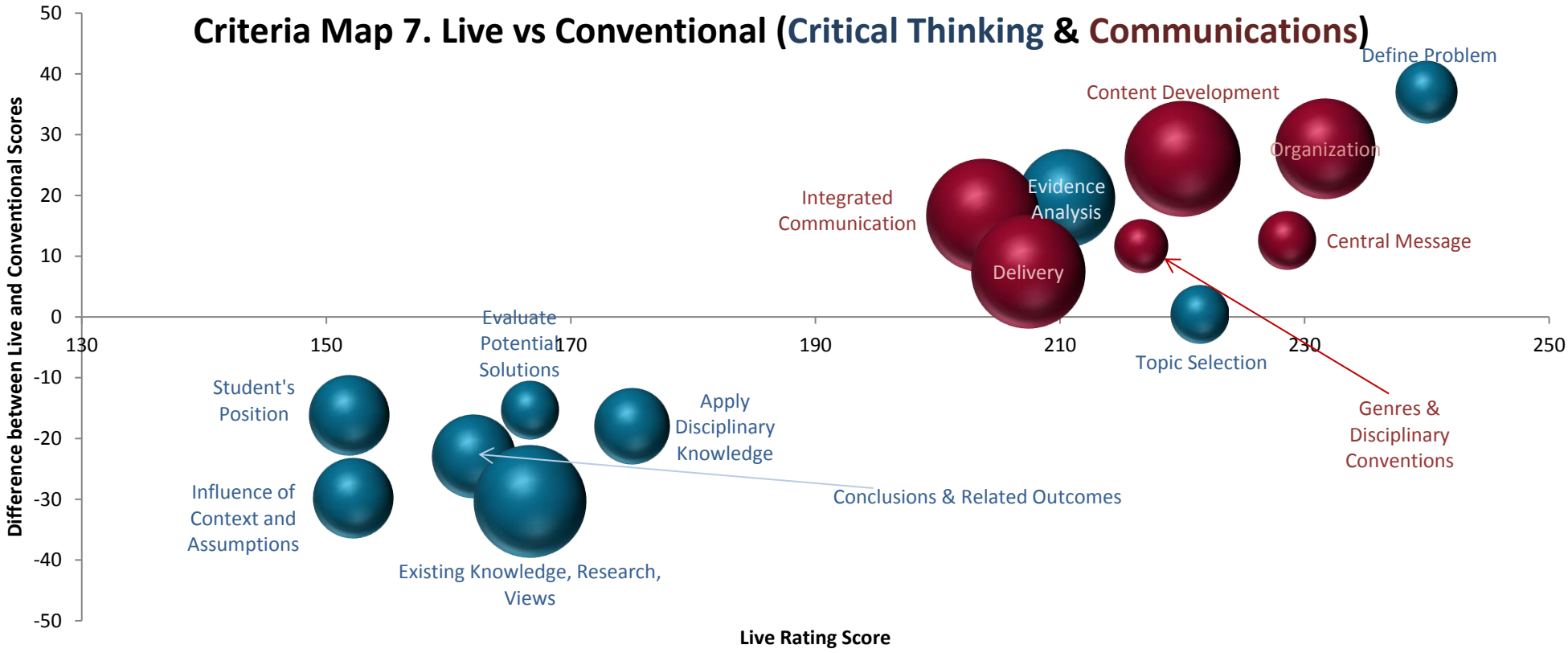
Live Rating Results, Cont'd

Compared to the earlier ratings of static artifacts, live rating seemed to result in more reliable scores. Reliability here (Table 9, using Krippendorff's alpha) is judged from a sample of two common criteria, across several rater pairings.

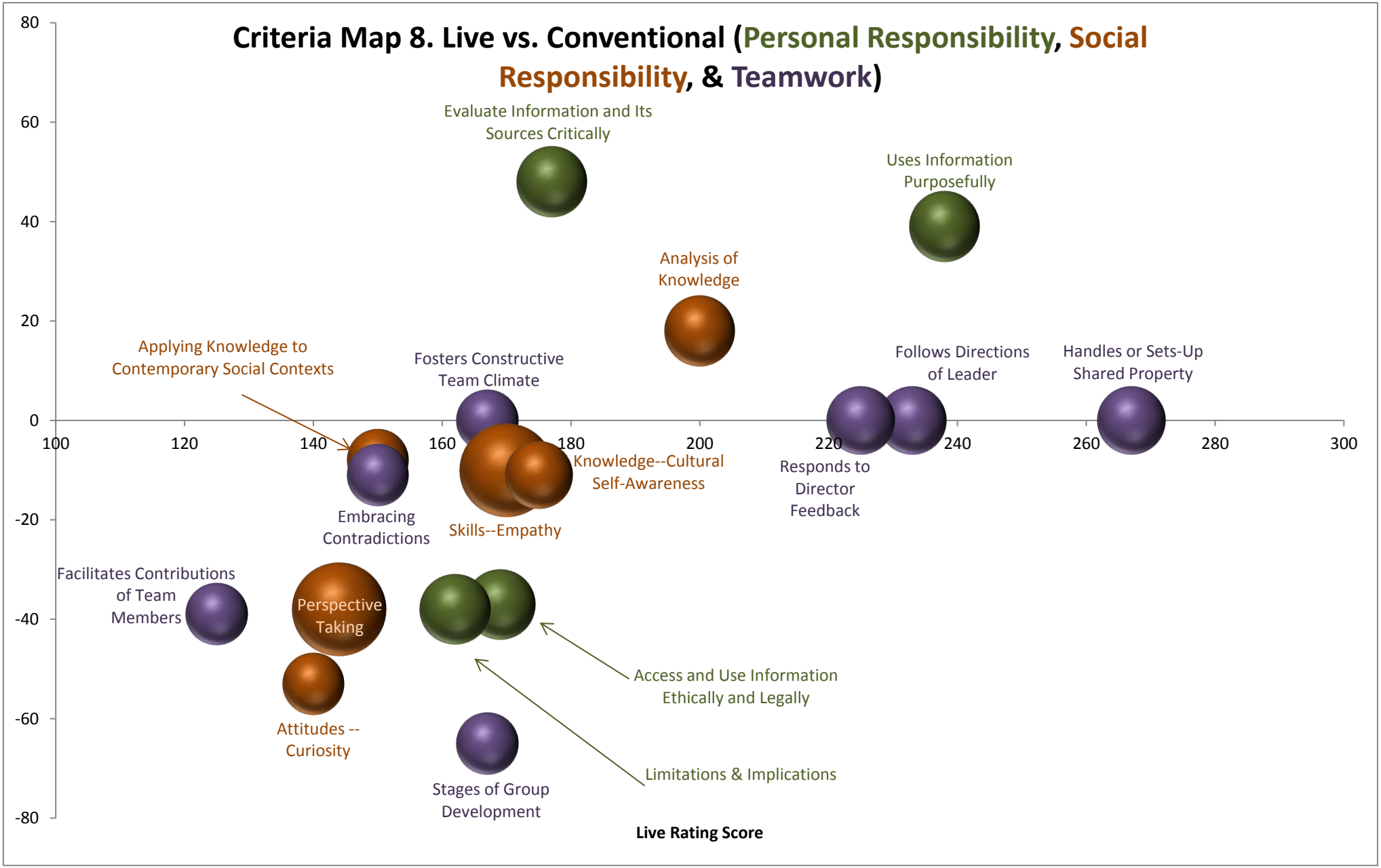
Criteria Maps 7 & 8 don't chart reliability, although again the size of the bubble indicates how often the criterion was rated. The X-Axis indicates the average live-rating score by criterion, while the Y-Axis (vertical) indicates the difference between live and conventional scores for those criteria. For Communication criteria (red, in Criteria Map 7), live ratings tended to fare better. For every other objective, they tended to fare worse.

Table 9. Live Rater Reliability

Sampled Criteria	Rater Pair 1	Rater Pair 2	Rater Pair 3
Content Development (Communciations)	.25	.45	-.22
Conclusions and Related Outcomes (Critical Thinking)	n/a	.25	.37



Live Rating Results, Cont'd



Multiple-Choice Strata

Table 10. Multiple-Choice Strata OBJECTIVE Criteria	AVERAGE SCORE		STANDARD DEVIATION	
	MC STRATA	ALL OTHERS	MC STRATA	ALL OTHERS
COMMUNICATION	**241**	**195**	**71**	**75**
Comprehension	229	187	74	75
Explanation of Issues	263	195	65	74
Interpretation	251	190	70	72
Organization	222	204	69	76
CRITICAL THINKING	**245**	**193**	**69**	**73**
Define Problem	235	203	70	69
Evidence Analysis	243	191	68	72
Existing Knowledge, Research, Views	249	197	69	68
Influence of Context & Assumptions	254	182	69	78
PERSONAL RESPONSIBILITY	**232**	**193**	**72**	**67**
Ethical Issue Recognition	238	176	74	73
Evaluate Information & Its Sources Critically	235	129	72	47
Limitations & Implications	235	200	71	56
Uses Information Purposefully	222	199	73	67
SOCIAL RESPONSIBILITY	**244**	**178**	**68**	**68**
Analysis of Knowledge	248	182	66	70
Attitudes -- Openness	246	203	66	73
Perspective Taking	248	182	67	46
Responsible Action	233	171	73	69
OVERALL AVERAGE	**241**	**188**	**70**	**68**

During the term, two strata piloted a common multiple-choice test, following these guidelines for converting multiple-choice responses to rubric-equivalent scores.

Both strata shared exactly the same assessment, using the test for all of their courses. Students completed the tests online for extra credit, and could take the test as many times as they wished.

The resulting scores were significantly higher ($p<.0001$) than those for the overall population. Since the same students were often assessed at lower values on the same criteria in other classes, it seems reasonable to conclude that the .77 effect size obtained by the strata was mostly a side effect of test conditions. From the data, we can draw no conclusions about the validity of multiple-choice assessments in general.

Going Forward

Suggestions for improving core assessment – both its practice and its results – appear below, organized by role.

ASSESSMENT LEADERS

- Improve assessment site to provide faculty with short descriptions, in addition to names, of criteria.
- Rename criteria whose labels are causing confusion. (Example: “Communication” from the Empirical Rubric.)
- Eliminate criteria weakly supported by faculty in affected disciplines.
- Consolidate cross-listed criteria (like Student’s Position and Central Message) that appear to be redundant.
- Work with faculty in affected disciplines to revise popular criteria with low reliability and/or high N/A rates.
- Focus rater training on commonly used criteria.
- Continue providing faculty workshops on assignment and assessment design, focusing on sets of criteria that fit well together and are applicable to common assignment genres.
- Work with faculty to develop and pilot-test new assignments or assessments, particularly for Teamwork, Personal Responsibility, and Social Responsibility.

FACULTY

- Check existing assignments for black-listed criteria (see criteria evaluation section above) that might soon be removed from the assessment rubrics. If such criteria are critical, please inform Gray Scott.
- Review full descriptions – including performance descriptors for each level of accomplishment – for the criteria that you plan to use.
- Doublecheck your criteria for high “N/A” ratings above. If they have high N/A ratings, it’s usually because those criteria require elements that faculty aren’t always including.
- Consider working with colleagues to develop (or even share) common assessments or assignments for core classes in the department. Common assignments need not be mandatory, but the more similar assignments are from class to class, the more visible gains in learning are likely to become.
- Volunteer to rate (if you haven’t already) even if it’s just for one session. Faculty who rate artifacts bring valuable perspectives to the room—and often come away with ideas for new assignment approaches.