Spring 2011 (and brief Fall 2010)GEAR Artifact Assessment Results

August 30, 2011

Introduction

- This report is organized as follows:
  - Summary of reviewer comments (from both fall and spring artifacts).
  - Executive summary of main findings.
  - Possible discussion points as we consider modifying the process.
  - Detailed results of spring artifact analysis.
  - Brief results of fall artifact analysis.

Reviewers' Comments and Suggestions(Fall and Spring Artifacts): General Themes

- Some artifacts were not tagged correctly – the metacognitive domain was especially mentioned.
- Reviewers were sometimes not able to download artifacts. It was suggested that artifacts be uploaded in specified formats, perhaps in PDF.
- The inability to download artifacts was especially problematic when the artifact was connected to an external link. We need to determine how to solve this problem.
- It might be more efficient, and yield more accurate results, if we calculate the mean between two or among three reviewers rather than calculating final scores based on at least two agreements.
- Rubric descriptions lean heavily to "process," which cannot be observed in an artifact. It might help reviewers if process papers or assignment instructions were included with the artifacts. Based on artifacts alone, it also is impossible to assess the student's level of independence in completing it.
- The artifacts contained too many spelling and grammar problems. A suggestion was made that writing mechanics should be addressed in FYS classes. Another suggestion was that students should upload only major projects in their final state of completion.

Reviewers' Comments and Suggestions(Fall and Spring Artifacts): General Themes Continued:

- There was too great a diversity of assignments and this contributed to the challenge of using the rubrics.
- It was sometimes not possible for students to achieve at higher rubric levels because the nature of the assignments did not give them the opportunity to do so.
- The rubrics need to be more descriptive so that evaluators have a better idea of what to look for in the artifacts. This lack of description made level assignments difficult.
- Training or discussion regarding using the rubrics prior to conducting the evaluations would be helpful.
- There needs to be a better definition of "metacognitive" reflection and the rubric needs to be clearer as to level descriptions.
- For science artifacts it might be helpful to have someone from the College of Science explain what should be included in an artifact that is uploaded as either "Scientific Contexts" or "Scientific Experimentation." There were several uploaded as "Scientific Experimentation," but there was no experiment proposed. Even for "Scientific Contexts," students should explain "why" things happen, not just present a series of facts they've looked up on the internet.
- It might be helpful to have meetings among raters when there are significant discrepancies. This would allow each to discuss the reasons for his/her scores and could help to resolve the discrepancies.

Executive Summary

- Below are what I consider to be important findings from the assessment of spring artifacts. However, there were a lot of data to sift through, so I invite you to add important findings I may have missed. ☺

  - In general, interrater reliability was poor. Of the 280 artifacts assessed, two independent raters agreed on 85 artifacts (30%) during the first review. Scores for an additional 97 artifacts differed by one point. Total agreement or a one-point difference accounted for 182 (65%) of the artifacts. However, the fact that scores differed by more than one point on 35% of the artifacts suggests that further analysis of rubrics, examination of the types of artifacts submitted, consideration of submission of assignment instructions with artifacts, and addressing the need for rater instruction prior to scoring are warranted.
  - Scores of "0," suggesting that reviewers felt that artifacts did not minimally meet the rubric specifications, were high. Thirty-four of the 280 artifacts assessed (12%) received two scores of "0" and therefore did not receive a third review. An additional 33 artifacts received one score of "0" plus a higher score, so were assigned to a third reviewer, who assigned a second score of "0." Potentially 24% of the artifacts submitted did not meet rubric specifications. This may have been due, at least in part, to inappropriate tagging. This suggests the need to re-examine how artifacts are selected and tagged.
  - Artifacts from the domains most closely aligned to FYS outcomes (Communication, Metacognitive Reflection, Information/Technical Literacy, and Multicultural/International Thinking) were uploaded most frequently. There were also a fair number of artifacts uploaded from the Ethical/Social/Historical Thinking Domain.

Executive Summary Continued:

- Of the domains, Information/Technical Literacy is most in need of further analysis. Only 35 out of 63 artifacts in this domain had at least two agreements (56%) and 24 of those (69%) were at level "0". This suggests that either students did not correctly tag these artifacts or they did not show evidence of achieving even minimal competence in this domain.
- The Multicultural/International Domain also is in need of further analysis. Although there was fairly high agreement between two out of three raters on these artifacts, 19 out of the 25 artifacts on which two scorers agreed were rated at level "0, " for a percentage of 76%. Again, this suggests that either students did not correctly tag these artifacts or they did not show evidence of achieving even minimal competence in this domain.
- Metacogntive Reflection also needs a more thorough look. Thirty-one percent of artifacts on which at least two reviewers agreed received scores of "0."
- After eliminating scores of "0," when final scores (scores on which there were at least two agreements) were analyzed, the majority of students scored at levels 1 and 2. The only exception to this was in the Aesthetic/Artistic Domain, but there were only six artifacts that received usable scores other than "0."
- When examining means for each domain based on all scores submitted and when considering means based on agreement between at least two scorers, students had the highest scores in the domains of Aesthetic/Artistic Thinking, Communication, and Ethical/Social/Historical Thinking.

Thoughts for Consideration

- Below I suggest some issues to consider (and wish to acknowledge Dr. Chris Green for some ideas), but I'm sure you have many more, which we can discuss at our meeting.

  - Consider having students in FYS upload only high-stakes course projects.
  - Consider focusing the assessment of FYS on artifacts that demonstrate competence with course outcomes. These would most closely align to the domains of Communication, Multicultural/International Thinking, Information/Technical Literacy, and Metacognitive Reflection.
  - Since students are introduced to the Domains of Thinking in FYS, consider assessing an artifact that demonstrates "Integrative Thinking." One of the AAC & U Value Rubrics could be used for this purpose.
  - Since one of the FYS outcomes is "reasoning," since "critical thinking" is supposed to be at the core of the other Domains of Thinking, and since Marshall currently uses the Collegiate Learning Assessment as one measure of student performance in analytic reasoning and problem solving, consider developing a series of CLA performance type tasks that could be used as part of FYS instruction and for purposes of assessment. This would allow us to align one part of GEAR assessment with a national benchmark. CLA has developed a rubric to assess analytic reasoning and evaluation and problem solving.
  - Further examine rubrics for domains where interrater reliability was lowest or in which students scored at level "0." Start with Information/Technical Literacy, Multicultural/International Thinking, and Metacognitive Reflection.
  - Consider conducting an experiment during our next round of artifact assessment to compare outcomes using Marshall's rubrics versus AAC & U's rubrics. The rubrics from AAC & U do not perfectly match our domains of thinking, but they come close. They have the advantage (or disadvantage, depending on your point of view) of being more detailed. Additionally, they have been field tested across the country. Their reliability is currently being determined.
  - Develop more detailed tagging specifications.
  - Find a solution for the problem of assessors not being able to open artifacts.
  - Consider either training or some face-to-face meetings among assessors.
  - Other ideas????

Spring Artifact Statistics

- 416 artifacts assigned
  - 6 were not reviewed.
  - 6 were not able to be opened – (5 with two scores and 1 with one score).
  - 125 had only one score – assigned reviewers did not complete task.
  - 34 were scored “0” by the first two independent raters.
  - An additional 33 were tagged “0” by two out of three raters.
- 280 artifacts had usable scores – we included scores of “0” as usable.
- Initial Review
  - Two independent reviewers agreed on 85 artifacts (30%).
- Third Party Reviews
  - 195 artifacts were assigned to a third reviewer.
  - Of these artifacts, 90 (46%)had two scores that agreed.
  - 105 (54%)had three different scores.
  - Out of 280 artifacts, 175 (62.5%) had agreement from two independent raters.

Number of Artifacts and Their Distribution by Domain

Domain	# Uploaded	# Assigned	% Assigned	# with Usable Scores	% with Usable Scores
Aesthetic/Artistic	144	29	20%	14	10%
Communication	481	96	20%	52	11%
Ethical/Social/Historical	237	47	20%	47	20%
Information/Technical	337	67	20%	63	19%
Abstract/Mathematical	55	11	20%	11	20%
Metacognitive Reflection	427	85	20%	54	13%
Multicultural/International	291	58	20%	28	10%
Scientific	117	23	20%	11	9%
Total	2,089	416	20%	280	13%

Number of Artifacts with Usable Scores and Their Distribution by Outcome

Domain	Outcome	Total	% of Total
Aesthetic/Artistic	Aesthetic Interpretation	3	1.1
	Aesthetic Creation	11	3.9
Communication	Communicating	31	11.1
	Interpreting Communication	21	7.5
Ethical/Social/Historical	Social Problems	42	15.0
	Social Science Methodology	5	1.8
Information/Technical	Information Literacy	35	12.5
	Technical Literacy	28	10.0
Abstract/Mathematical	Mathematical and Abstract Reasoning	7	2.5
	Mathematical Problem Solving	1	0.4
	Mathematical Applications to other Disciplines	3	1.1
Metacognitive Reflection	Reflecting on the Learning Experience	54	19.3
Multicultural/International	Intercultural Communication	6	2.1
	Intercultural Appreciation	9	3.2
	Global Awareness	13	4.6
Scientific	Scientific Contexts	8	2.9
	Scientific Experimentation	3	1.1
Total		280	100

Analysis of Possible Mistagging “Scores of 0”

- 34 artifacts scored “0” by first two independent reviewers, so did not have a third reviewer.
  - Aesthetic Creation = 1
  - Communicating = 1
  - Information Literacy = 10
  - Technical Literacy = 5
  - Mathematical and Abstract Representation = 1
  - Metacognitive Reflection = 6
  - Intercultural Communication = 1
  - Intercultural Appreciation = 1
  - Global Awareness = 6
  - Scientific Experimentation = 2
- 21 artifacts had two scores of “0” and one score of “1”
  - Communicating = 1
  - Interpreting Communication = 1
  - Social Problems = 1
  - Information Literacy = 4
  - Technical Literacy = 4
  - Metacognitive Reflection = 3
  - Intercultural Communication = 1
  - Intercultural Appreciation = 2
  - Global Awareness = 4
- 7 artifacts had two scores of “0” and one score of “2”
  - Social Problems = 2
  - Technical Literacy = 1
  - Metacognitive Reflection = 2
  - Intercultural Communication = 2
- 3 artifacts had two scores of “0” and one score of “3”
  - Social Science Methodology = 1
  - Mathematical Application to Other Disciplines = 1
  - Intercultural Appreciation = 1
- 2 artifacts had two scores of “0” and one score of “4”
  - Metacognitive Reflection = 1
  - Intercultural Communication = 1

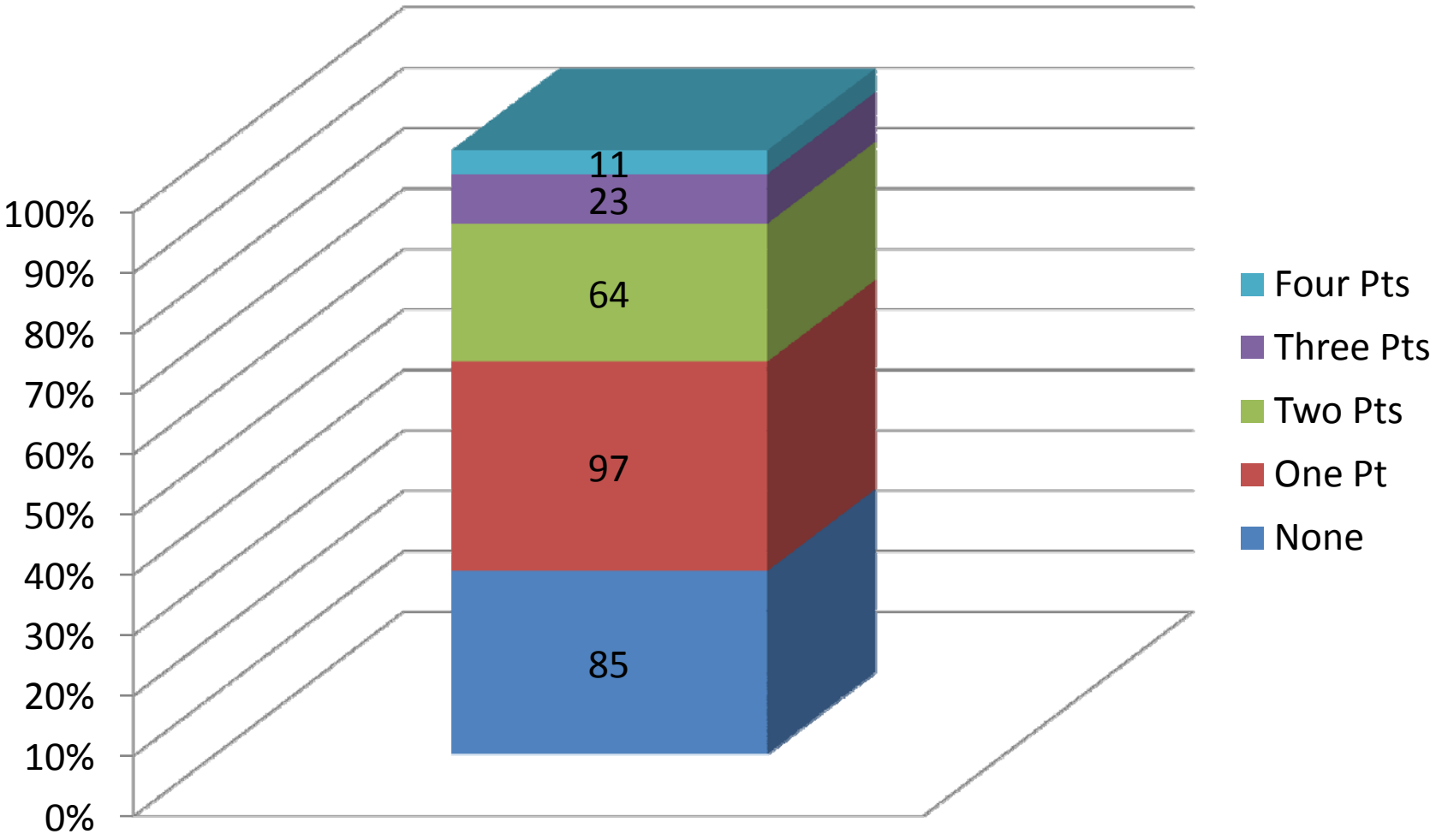
Number of Artifacts with Usable Scores and Their Distribution by Outcome with information regarding # and % with at least one score of "0"

Domain	Outcome	Total	Number with at least one score of "0"	% with one score of "0"
Aesthetic/Artistic	Aesthetic Interpretation	3	0	0
	Aesthetic Creation	11	1	9.1
Communication	Communicating	31	2	6.5
	Interpreting Communication	21	1	4.8
Ethical/Social/Historical	Social Problems	42	3	7.1
	Social Science Methodology	5	1	20.0
Information/Technical	Information Literacy	35	14	40.0
	Technical Literacy	28	10	35.7
Abstract/Mathematical	Mathematical and Abstract Reasoning	7	1	14.3
	Mathematical Problem Solving	1	0	0
	Mathematical Applications to other Disciplines	3	1	33.3
Metacognitive Reflection	Reflecting on the Learning Experience	54	12	22.2
Multicultural/International	Intercultural Communication	6	5	83.3
	Intercultural Appreciation	9	4	44.4
	Global Awareness	13	10	76.9
Scientific	Scientific Contexts	8	0	0
	Scientific Experimentation	3	2	66.7
Total		280	67	23.9

More In-Depth Analysis of Rater Agreement

- Initial Review
  - Two independent reviewers agreed on 85 artifacts (30%).
    - Level 0 = 34
    - Level 1 = 29
    - Level 2 = 14
    - Level 3 = 5
    - Level 4 = 3
  - Two independent reviewers differed by only one point on 97 artifacts (35%):
    - Level 0-1 = 45
    - Level 1-2 = 25
    - Level 2-3 = 15
    - Level 3-4 = 12
  - Two independent reviewers differed by two points on 64 artifacts (23%):
    - Level 0-2 = 23
    - Level 1-3 = 17
    - Level 2-4 = 24
  - Two independent raters differed by three points on 23 artifacts (8%):
    - Level 0-3 = 7
    - Level 1-4 = 16
  - Two independent raters differed by four points on 11 artifacts (4%):
    - Level 0-4 = 11

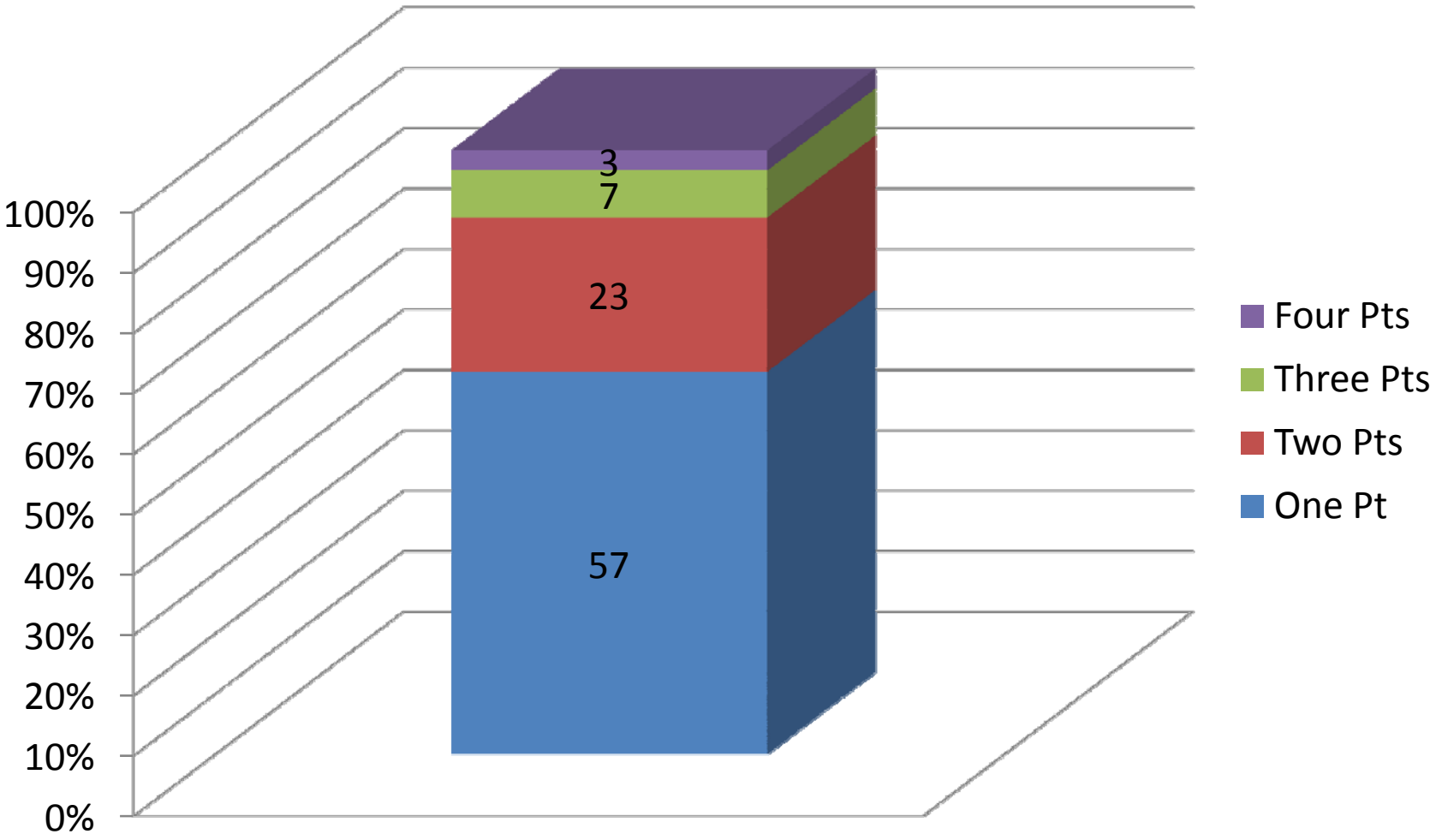
Point Differences Between First Two Independent Raters



More In-Depth Analysis of Rater Agreement

- Third Party Review: Agreement Between Two Reviewers (90 out of 195 artifacts)
  - For 57 reviews, two reviewers agreed and the third reviewer differed by only one point
    - Level 0-0-1 = 21
    - Level 0-1-1 = 13
    - Level 1-1-2 = 4
    - Level 1-2-2 = 7
    - Level 2-2-3 = 2
    - Level 2-3-3 = 4
    - Level 3-3-4 = 4
    - Level 3-4-4 = 2
  - For 23 reviews, two reviewers agreed and the third reviewer differed by two points
    - Level 0-0-2 = 7
    - Level 1-1-3 = 3
    - Level 0-2-2 = 5
    - Level 2-2-4 = 6
    - Level 1-3-3 = 2
    - Level 2-4-4 = 0
  - For 7 reviews, two reviewers agreed and the third reviewer differed by three points
    - Level 0-0-3 = 3
    - Level 1-1-4 = 4
    - Level 0-3-3 = 0
    - Level 1-4-4 = 0
  - For 3 reviews, two reviewers agreed and the third reviewer differed by four points
    - Level 0-0-4 = 2
    - Level 0-4-4 = 1

When Two our of Three Reviewers Agreed, Point Difference Between Them and Third Reviewer



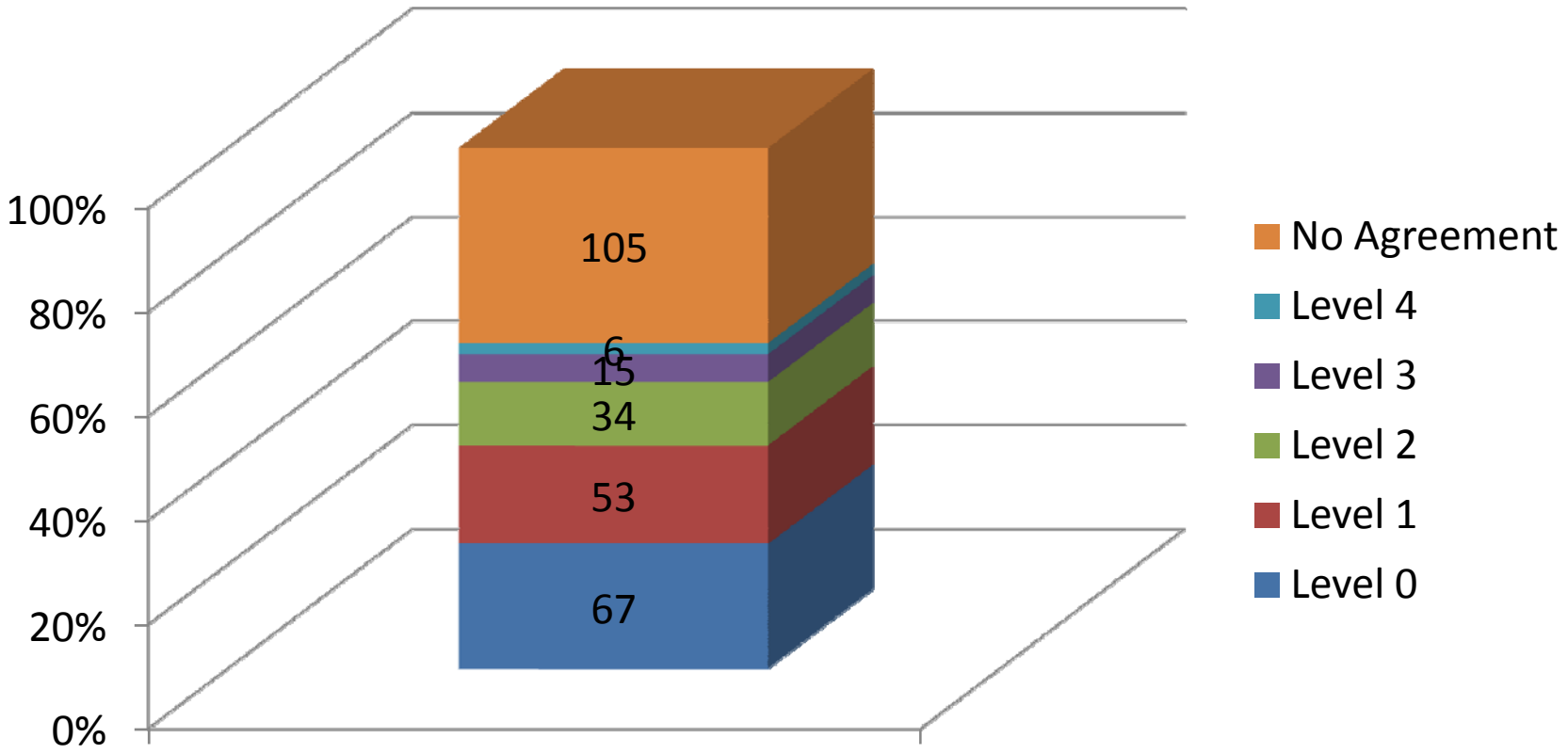
More In-Depth Analysis of Rater Agreement

- Third Party Review (Analysis of remaining 105 artifacts):
  - For 43 reviews, the three reviews were stairsteps:
    - Level 0-1-2 = 23
    - Level 1-2-3 = 12
    - Level 2-3-4 = 8
  - For 55 reviews, two reviewers were within a point of each other, while the third jumped 2 to 3 points:
    - Level 0-1-3 = 10
    - Level 0-1-4 = 11
    - Level 0-2-3 = 5
    - Level 0-3-4 = 3
    - Level 1-2-4 = 19
    - Level 1-3-4 = 7
  - For the remaining 7 artifacts, each review was a 2-point spread
    - Level 0-2-4 = 7

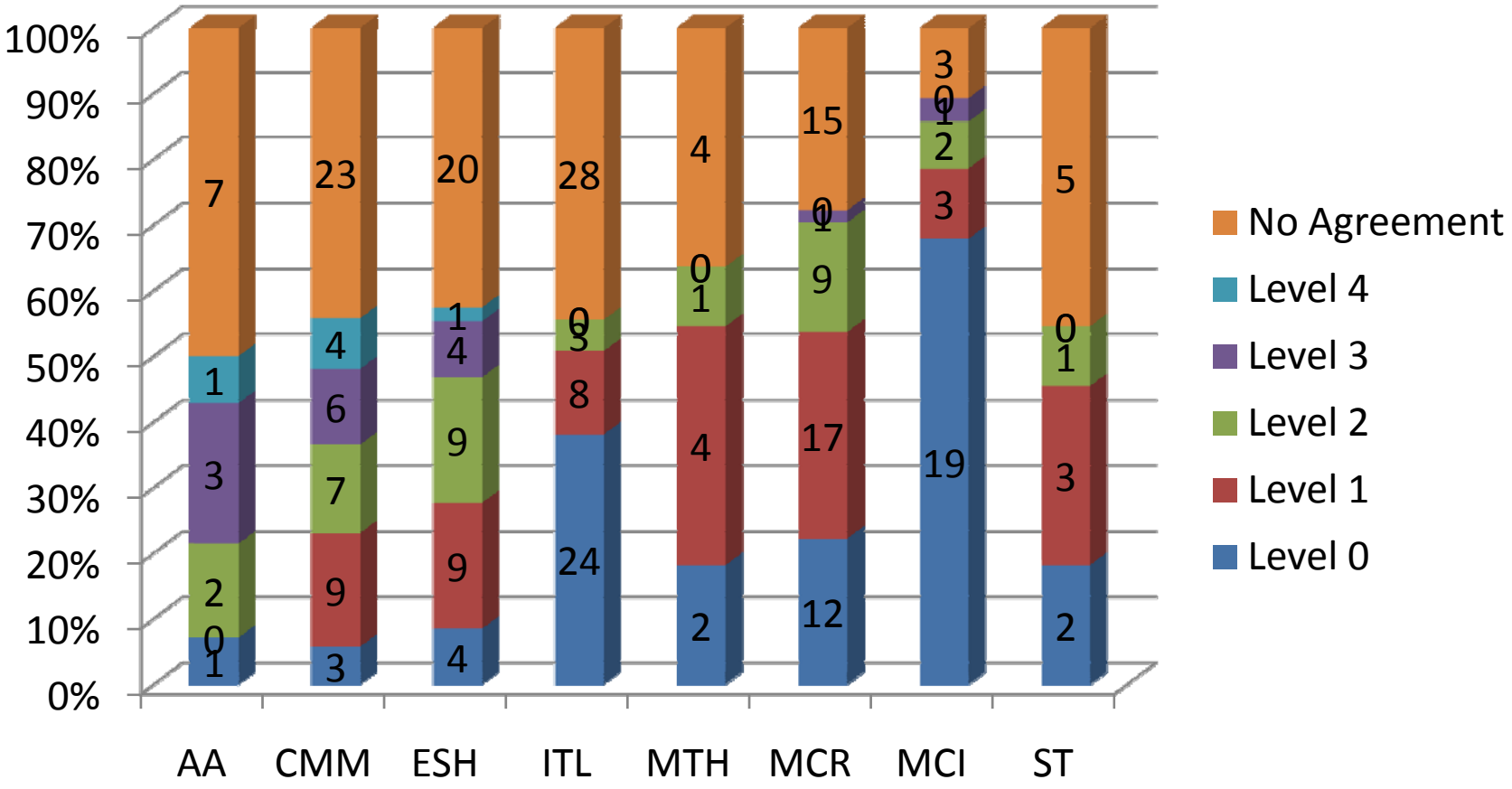
This Chart Shows the scores (by domain) at each level where there was agreement between at least two assessors and the number of artifacts for which there was no agreement.

Domain	Level 0	Level 1	Level 2	Level 3	Level 4	No Agreement	Total # Artifacts
Aesthetic/Artistic	1	0	2	3	1	7	14
Communication	3	9	7	6	4	23	52
Ethical/Social/Historical	4	9	9	4	1	20	47
Information/Technical	24	8	3	0	0	28	63
Abstract/Mathematical	2	4	1	0	0	4	11
Metacognitive Reflection	12	17	9	1	0	15	54
Multicultural/International	19	3	2	1	0	3	28
Scientific	2	3	1	0	0	5	11
Total	67	53	34	15	6	105	280

This Chart Shows the scores at each level where there was agreement between at least two assessors and the number of artifacts for which there was no agreement.



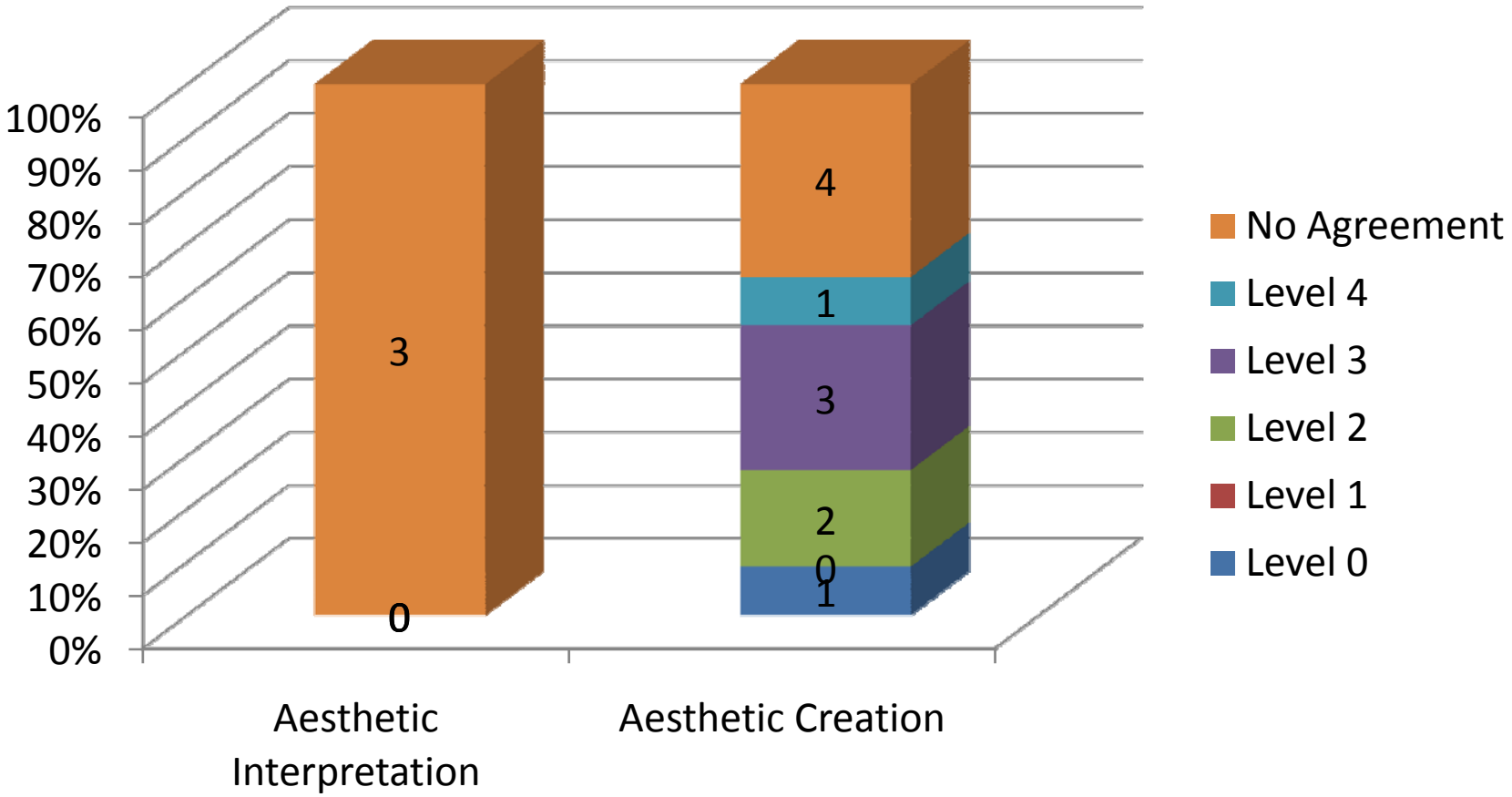
This Chart Shows the scores (by domain) at each level where there was agreement between at least two assessors and the number of artifacts for which there was no agreement.



This Chart Shows the scores (by outcomes) for the Aesthetic/Artistic Domain at each level where there was agreement between at least two assessors and the number of artifacts for which there was no agreement.

Outcome	Level 0	Level 1	Level 2	Level 3	Level 4	No Agreement	Total # Artifacts
Aesthetic Interpretation	0	0	0	0	0	3	3
Aesthetic Creation	1	0	2	3	1	4	11
Total	1	0	2	3	1	7	14

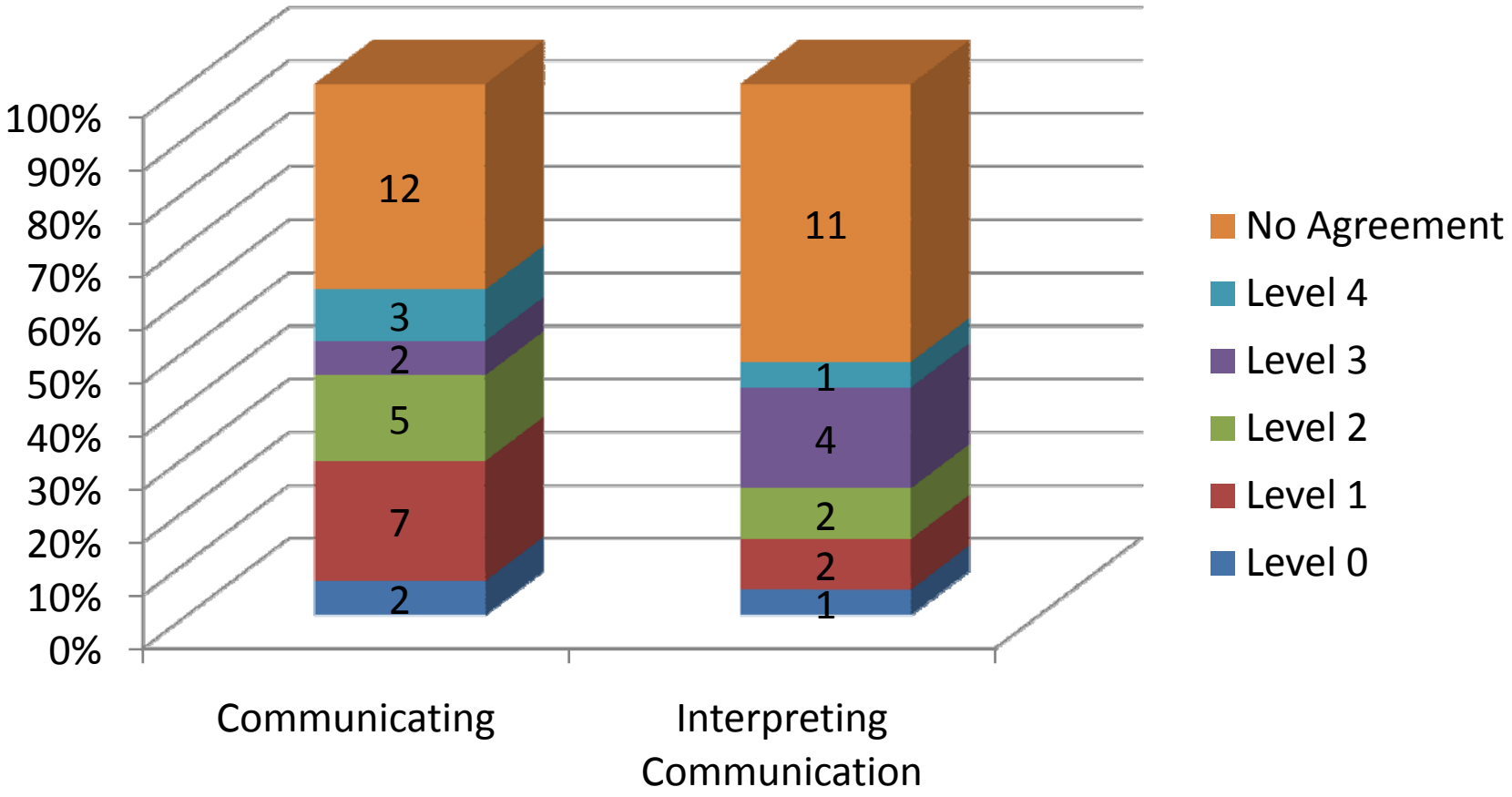
This Chart Shows the scores (by outcomes) for the Aesthetic/Artistic Domain at each level where there was agreement between at least two assessors and the number of artifacts for which there was no agreement.



This Chart Shows the scores (by outcomes) for the Communication Domain at each level where there was agreement between at least two assessors and the number of artifacts for which there was no agreement.

Outcome	Level 0	Level 1	Level 2	Level 3	Level 4	No Agreement	Total # Artifacts
Communicating	2	7	5	2	3	12	31
Interpreting Communication	1	2	2	4	1	11	21
Total	3	9	7	6	4	23	52

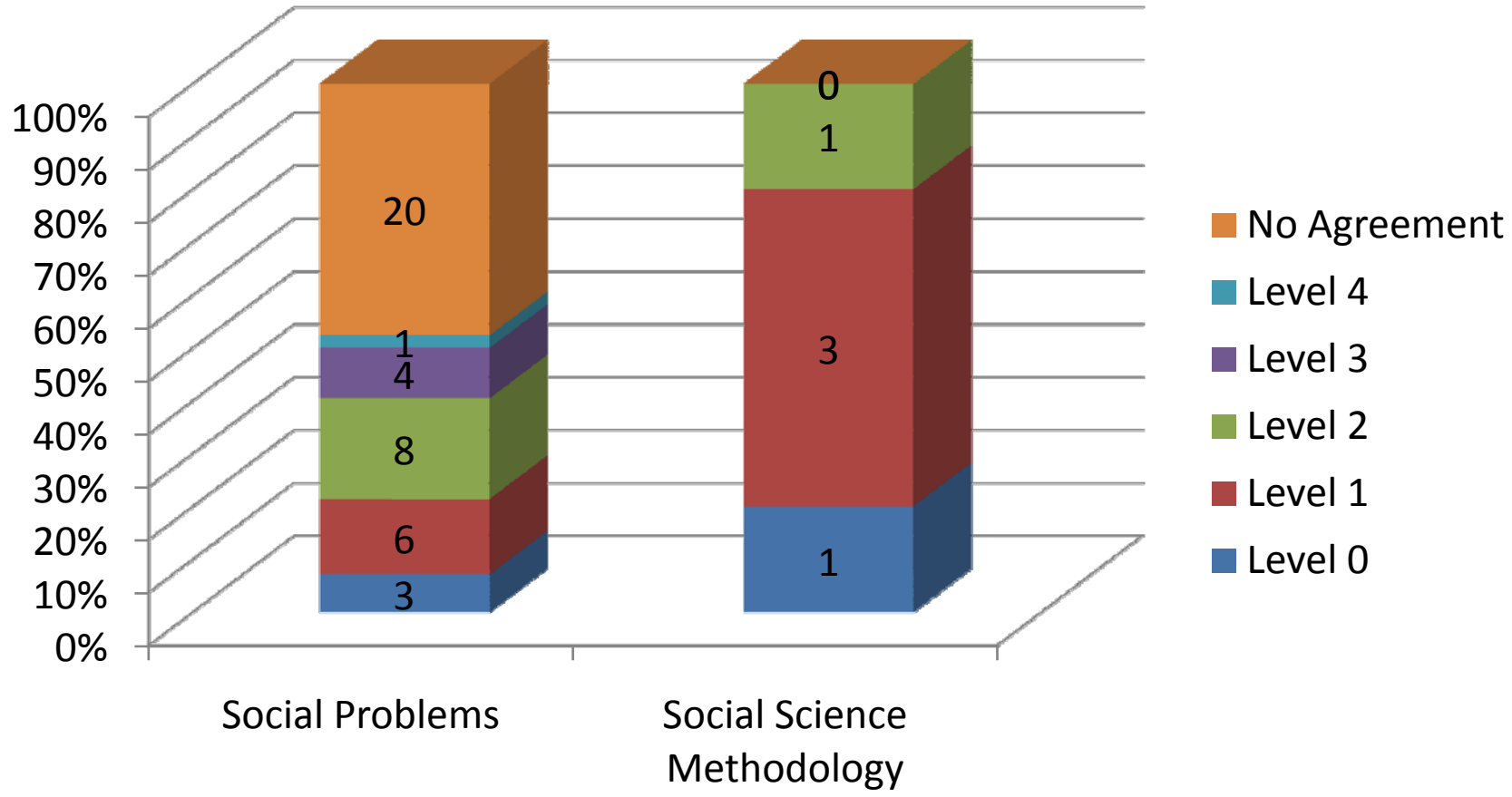
This Chart Shows the scores (by outcomes) for the Communication Domain at each level where there was agreement between at least two assessors and the number of artifacts for which there was no agreement.



This Chart Shows the scores (by outcomes) for the Ethical, Social, Historical Domain at each level where there was agreement between at least two assessors and the number of artifacts for which there was no agreement.

Outcome	Level 0	Level 1	Level 2	Level 3	Level 4	No Agreement	Total # Artifacts
Social Problems	3	6	8	4	1	20	42
Social Science Methodology	1	3	1	0	0	0	5
Total	4	9	9	4	1	20	47

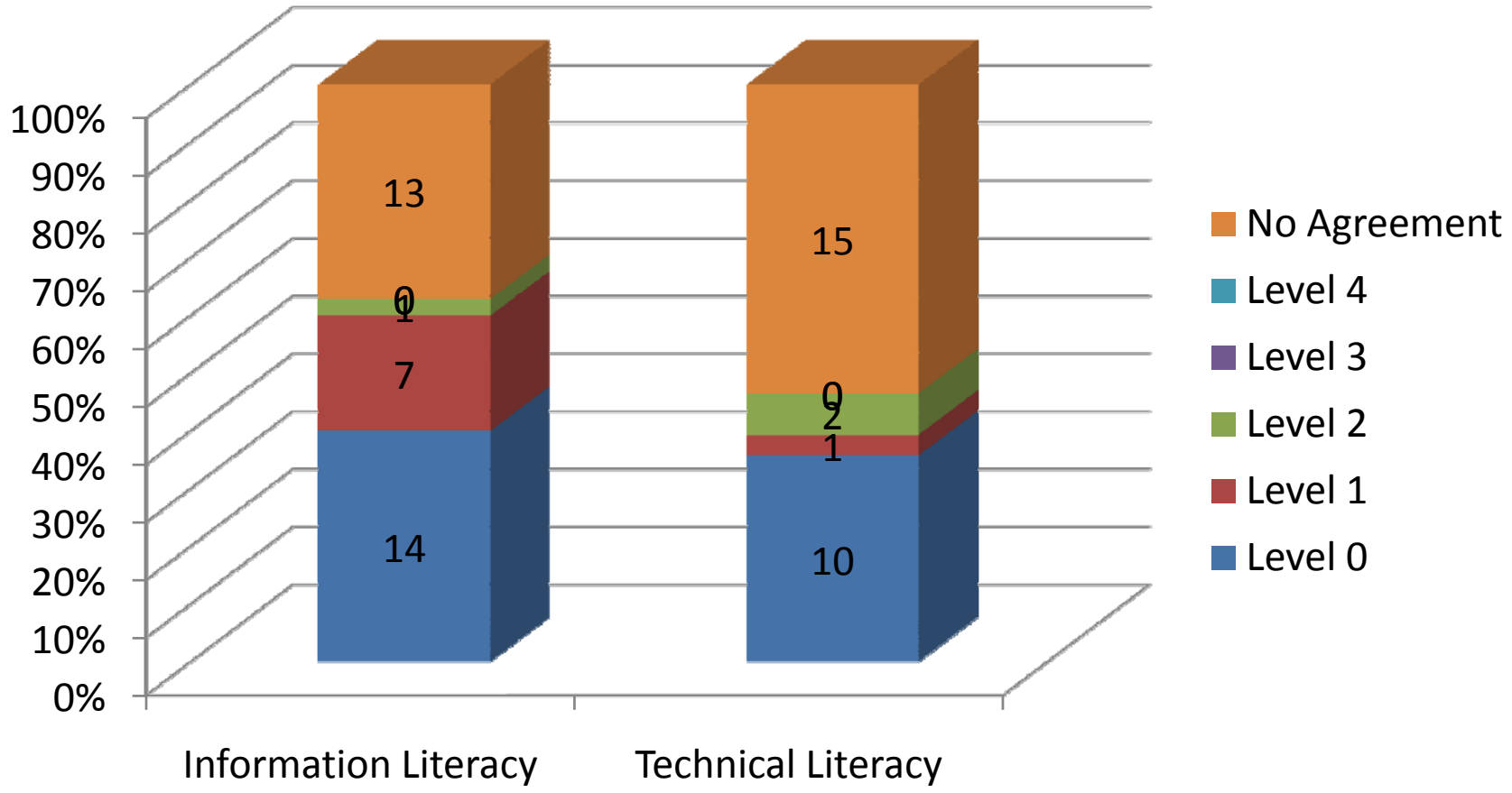
This Chart Shows the scores (by outcomes) for the Ethical, Social, Historical Domain at each level where there was agreement between at least two assessors and the number of artifacts for which there was no agreement.



This Chart Shows the scores (by outcomes) for the Information/Technical Literacy Domain at each level where there was agreement between at least two assessors and the number of artifacts for which there was no agreement.

Outcome	Level 0	Level 1	Level 2	Level 3	Level 4	No Agreement	Total # Artifacts
Information Literacy	14	7	1	0	0	13	35
Technical Literacy	10	1	2	0	0	15	28
Total	24	8	3	0	0	28	63

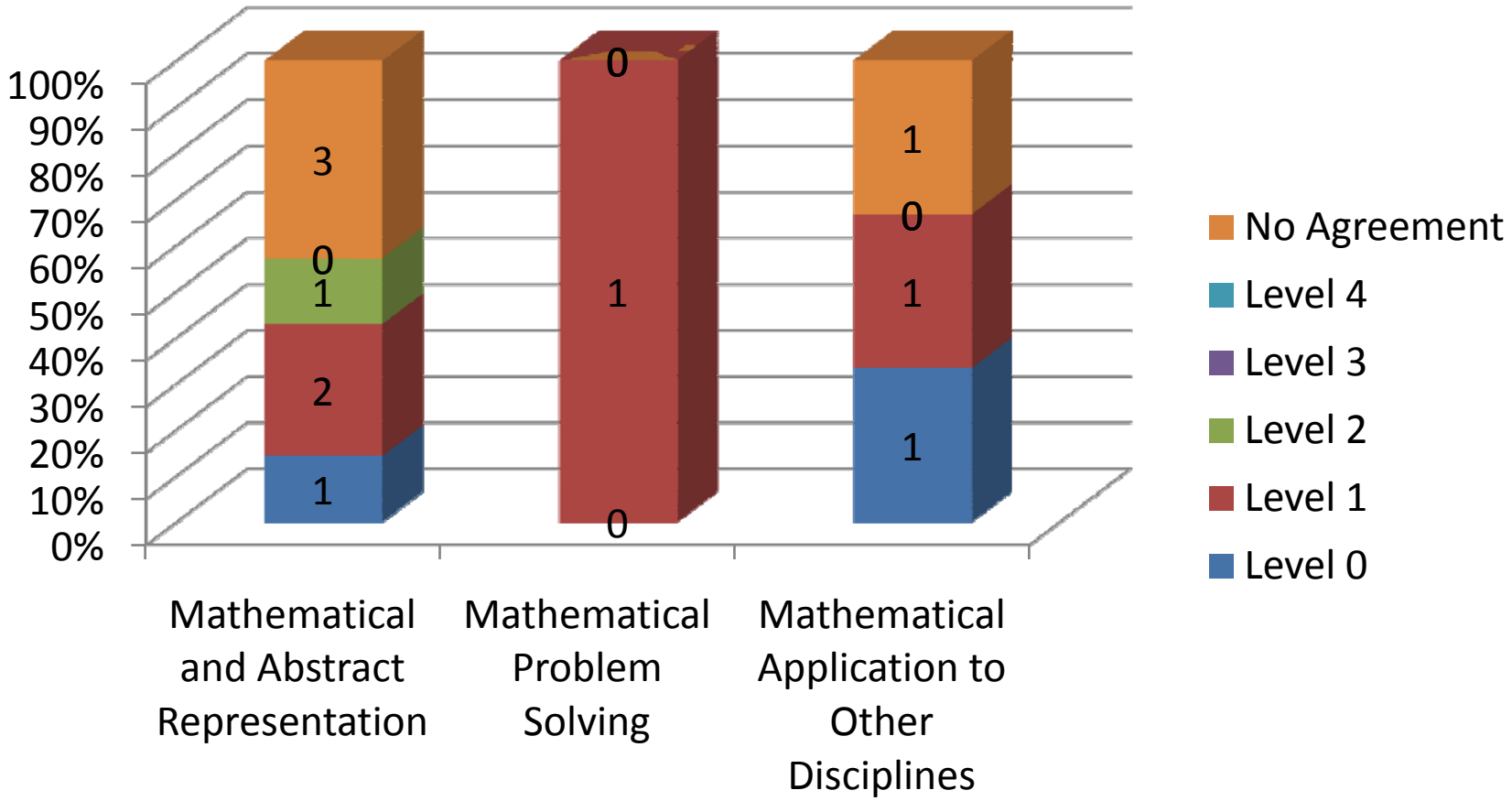
This Chart Shows the scores (by outcomes) for the Information/Technical Literacy Domain at each level where there was agreement between at least two assessors and the number of artifacts for which there was no agreement.



This Chart Shows the scores (by outcomes) for the Abstract and Mathematical Thinking Domain at each level where there was agreement between at least two assessors and the number of artifacts for which there was no agreement.

Outcome	Level 0	Level 1	Level 2	Level 3	Level 4	No Agreement	Total # Artifacts
Mathematical Abstract Representation	1	2	1	0	0	3	7
Mathematical Problem Solving	0	1	0	0	0	0	1
Mathematical Application to Other Disciplines	1	1	0	0	0	1	3
Total	2	4	1	0	0	4	11

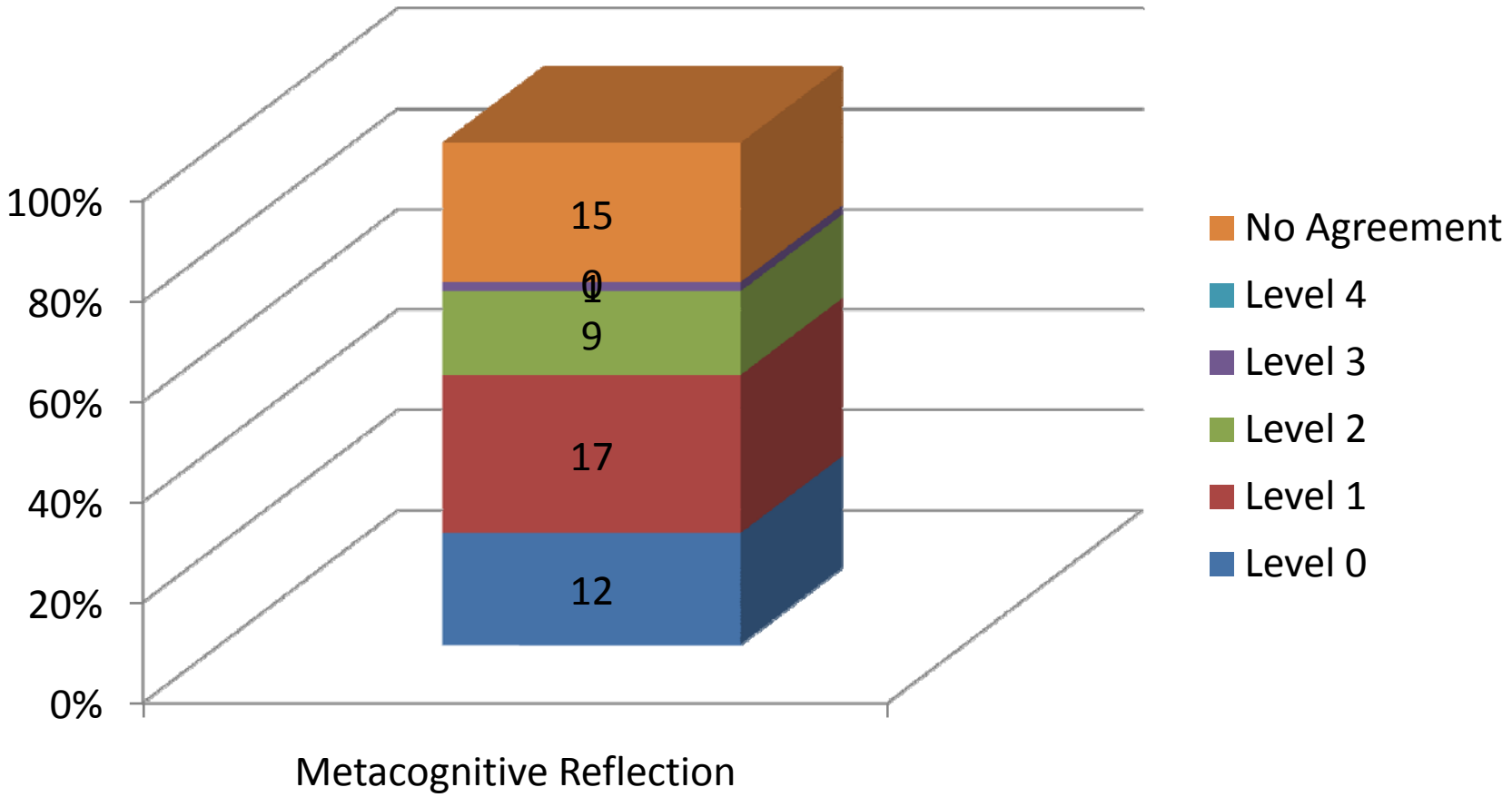
This Chart Shows the scores (by outcomes) for the Abstract and Mathematical Thinking Domain at each level where there was agreement between at least two assessors and the number of artifacts for which there was no agreement.



This Chart Shows the scores (by outcomes) for the Metacognitive Reflection Domain at each level where there was agreement between at least two assessors and the number of artifacts for which there was no agreement.

Outcome	Level 0	Level 1	Level 2	Level 3	Level 4	No Agreement	Total # Artifacts
Metacognitive Reflection	12	17	9	1	0	15	54
Total	12	17	9	1	0	15	54

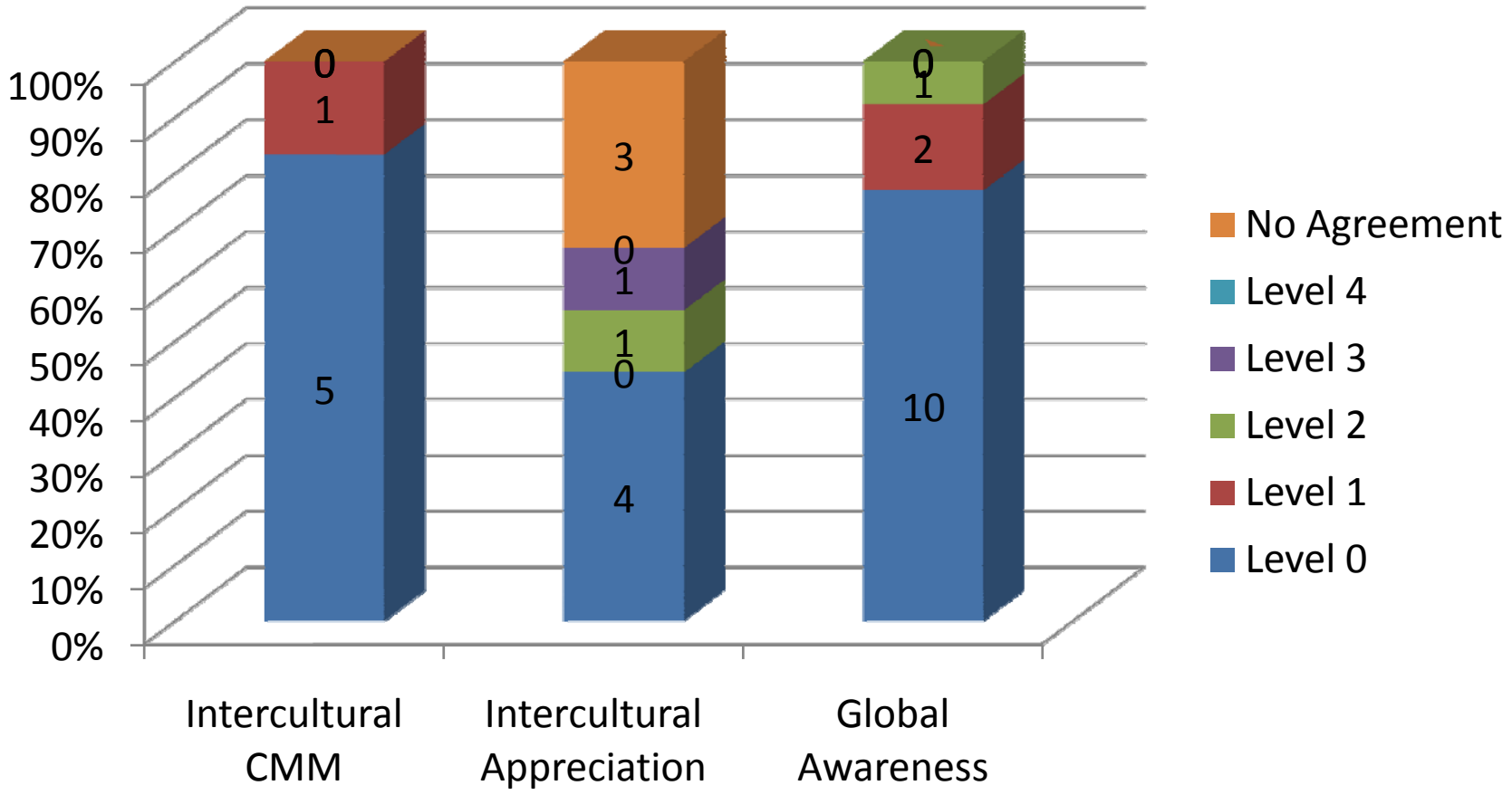
This Chart Shows the scores (by outcomes) for the Metacognitive Reflection Domain at each level where there was agreement between at least two assessors and the number of artifacts for which there was no agreement.



This Chart Shows the scores (by outcomes) for the Multicultural/InternationalThinking Domain at each level where there was agreement between at least two assessors and the number of artifacts for which there was no agreement.

Outcome	Level 0	Level 1	Level 2	Level 3	Level 4	No Agreement	Total # Artifacts
Intercultural Communication	5	1	0	0	0	0	6
Intercultural Appreciation	4	0	1	1	0	3	9
Global Awareness	10	2	1	0	0	0	13
Total	19	3	2	1	0	3	28

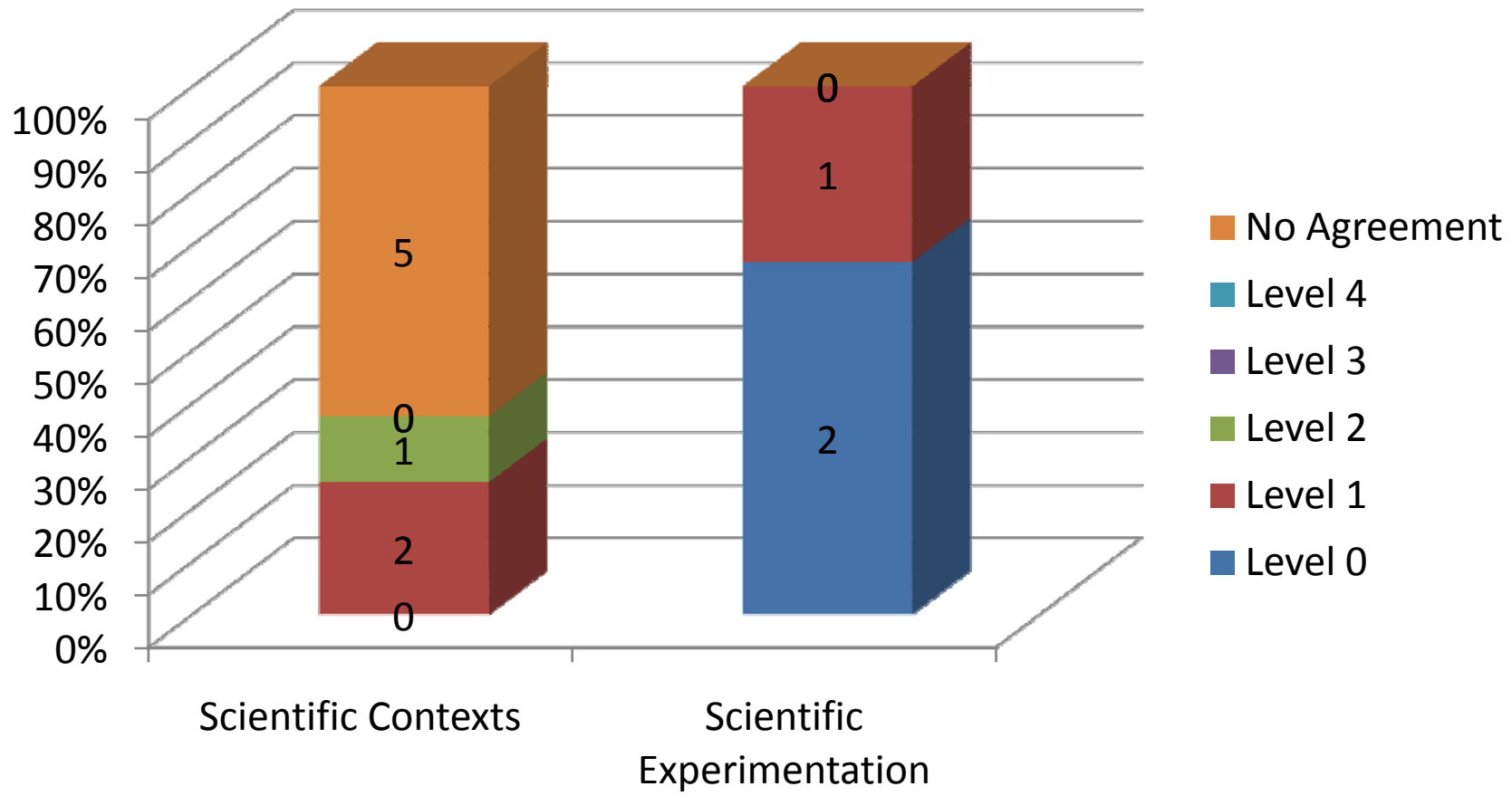
This Chart Shows the scores (by outcomes) for the Multicultural/InternationalThinking Domain at each level where there was agreement between at least two assessors and the number of artifacts for which there was no agreement.



This Chart Shows the scores (by outcomes) for the Scientific Thinking Domain at each level where there was agreement between at least two assessors and the number of artifacts for which there was no agreement.

Outcome	Level 0	Level 1	Level 2	Level 3	Level 4	No Agreement	Total # Artifacts
Scientific Contexts	0	2	1	0	0	5	8
Scientific Experimentation	2	1	0	0	0	0	3
Total	2	3	1	0	0	5	11

This Chart Shows the scores (by outcomes) for the Scientific Thinking Domain at each level where there was agreement between at least two assessors and the number of artifacts for which there was no agreement.



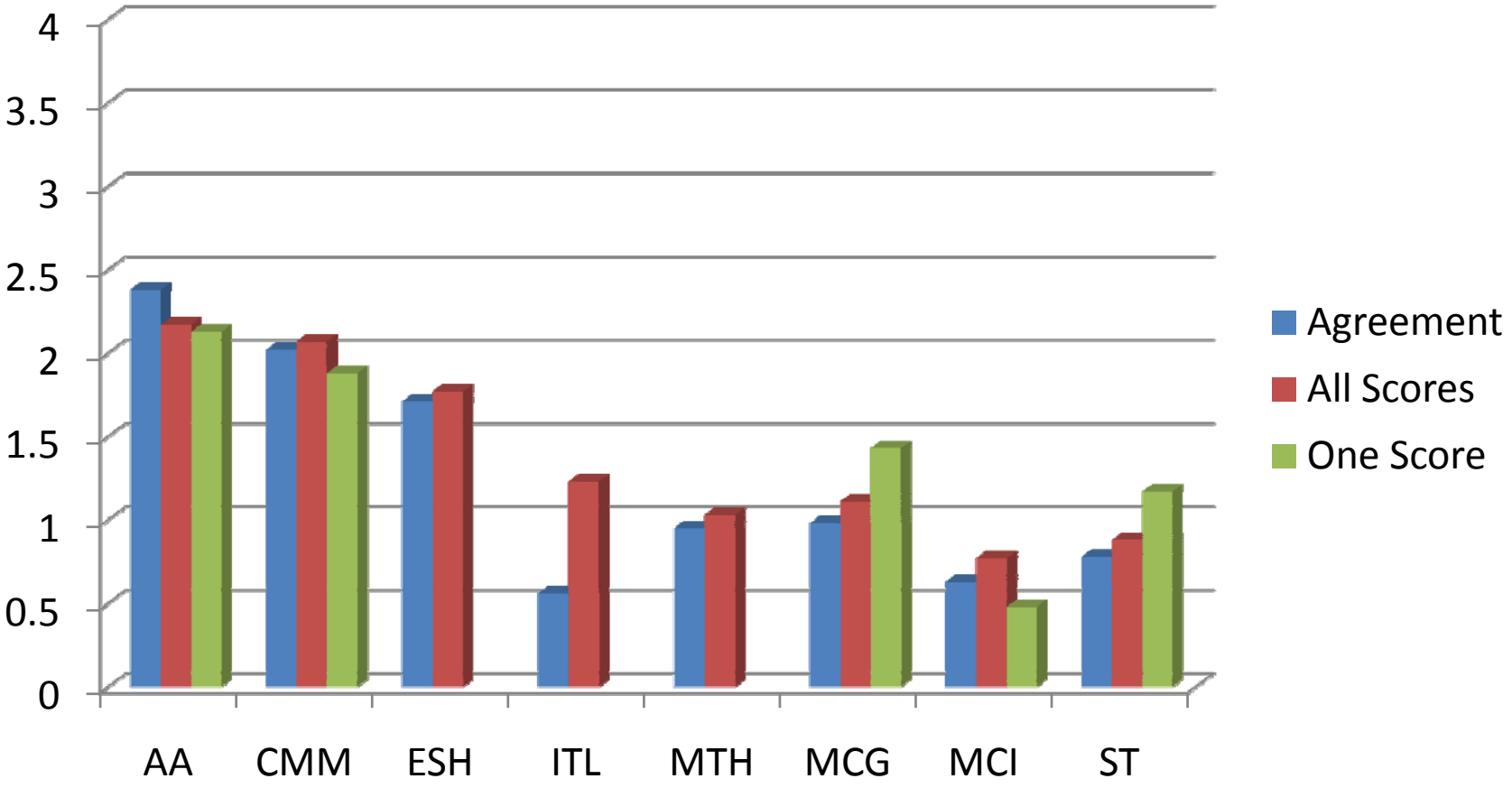
Mean Scores for final score (2 agreements) and mean of all scores submitted (2 or 3)

Domain	*N of artifacts*	Final Mean	*SD*	*N of artifacts*	Mean of all Scores submitted	*SD*
Aesthetic/Artistic	7	2.38	1.42	14	2.17	1.12
Communication	29	2.02	1.24	52	2.07	0.90
Ethical/Social/ Historical	27	1.71	0.84	47	1.77	0.76
Information/ Technical	35	0.56	0.73	63	1.23	0.98
Abstract/ Mathematical	7	0.95	0.52	11	1.03	0.46
Metacognitive Reflection	39	0.98	0.71	54	1.11	0.71
Multicultural/ International	25	0.63	0.77	28	0.77	0.86
Scientific	6	0.78	0.86	11	0.88	0.62
Total	175	1.18	1.03	280	1.43	0.96

Mean Scores for each domain using artifacts that received only one score (*n* = 124)

Domain	*N of artifacts*	Mean	*Standard Deviation*	*Low to High Score*
Aesthetic/Artistic	15	2.13	1.41	0 – 4
Communication	40	1.88	1.24	0 – 4
Metacognitive Reflection	28	1.43	1.29	0 – 4
Multicultural/ International	29	0.48	0.63	0 – 2
Scientific	12	1.17	1.12	0 – 3
Total	124	1.40	1.28	0 – 4

Means for the following categories: Agreement (at least two reviewers agreed on a score), mean of all scores (2 or 3) submitted, and mean of the 124 artifacts that had a single score.



Less Elegant Analysis of Fall 2010 FYS Artifacts

- 100 Artifacts were reviewed (16 Aesthetic/Artistic; 32 Communication; 25 Ethical/Social/Historical; 14 Information/Technical; 13 Multicultural/International)
  - Agreement between two independent reviewers on 35% of artifacts following first review
  - Agreement between two of three independent reviewers on 35% of artifacts
  - No agreement among three independent reviewers on 30% of artifacts
- <u>Artifact Levels</u>
  - 25% rated at level 0
  - 16% rated at level 1
  - 16% rated at level 2
  - 4% rated at level 3
  - 9% rated at level 4
  - 30%: no agreement among three independent reviewers
- *Comments*
  - Level “0” most prevalent in Information/Technical Literacy Domain.
  - No agreement statistics were as follows:
  - Aesthetic/Artistic = 44%
  - Communication = 25%
  - MC/I = 39%
  - Information/Technical Literacy = 7%
  - Ethical/Social/Historical Thinking = 36%
  - Aesthetic/Artistic, Communication, and MC/I had some artifacts at level 4.