



Nebraska Student-Centered Assessment System (NSCAS) Grades 5 and 8 Science Standard Setting

Technical Report

August 11, 2022

Submitted By:
Susan Davis-Becker, Ph.D.
Kelley Wheeler, M.S.

Contents

- Overview..... 3
- NSCAS Assessment 3
- Standard Setting Workshop Design..... 3
- Results 7
- Validity Evidence and Next Steps 11
- References 12
- Appendix A: Panelist Agenda..... 13
- Appendix B: Study Materials 14
- Appendix C: Achievement Level Descriptors..... 15

Overview

The Nebraska Department of Education and Early Development (NDE) has partnered with NWEA and ACS Ventures, LLC (ACS) to establish cut scores for the Nebraska Student-Centered Assessment System (NSCAS) Science assessments administered in grades 5 and 8. The first operational administration of the new NSCAS Science assessments was in the spring of 2022. As part of the implementation, it was necessary to establish cut scores for the NSCAS Science assessments which: (a) reflect the current Nebraska state standards, (b) link students' scores on the assessments to the state's expectations for students in each performance level, and (c) are articulated between grades 5 and 8. On July 6 and 7, 2022, NWEA and ACS worked with subject matter expert panelists from Nebraska to formulate recommended cut scores. The Extended Angoff methodology (Hambleton & Plake, 1995) was used to establish the recommendations for two cut scores for each assessment. The cut scores are intended to guide interpretation of test scores in alignment with the Achievement Level Descriptors (ALDs).

NSCAS Assessment

The NSCAS Science assessment is formed around tasks, which are collections of prompts focused on a single phenomenon or problem. Tasks may use elements from the DCIs, SEPs, and CCCs in any combination. In the test design, elements are differentiated by the frequency they are to be assessed. Frequent elements will most likely be assessed every year, infrequent elements will be presented at least every two years, and rare elements will be presented at least every three years. The assessment was developed following the *SIPS Complexity Framework* (see Appendix B).

Each test is a series of tasks with associated items. Grade 5 includes 4 tasks with 21 total items and grade 8 includes 8 tasks with 27 items. There are three types of items that may appear on a test form:

- Single multiple-choice: These items are presented with the stimulus and are scored as correct (1) or incorrect (0)
- Composite multiple-choice: These items are presented as a pair and are scored as both correct (1) or incorrect (0)
- Polytomous: These items may be hot text and gap match types in which the student needs to make multiple hot text selections or drag multiple objects into gaps. These are scored as correct (2), partially correct (1) or incorrect (0)

Student performance on each test will be classified into one of three achievement levels: *Developing*, *On-Track*, or *Advanced*. Therefore, two cut scores are required to interpret test scores – one to distinguish *On-Track* performance from *Developing* performance and one to distinguish *Advanced* performance from *On-Track* performance.

Standard Setting Workshop Design

ACS worked with NWEA and NDE to design the standard setting process for the NSCAS Science assessment. The standard setting meeting was organized to occur over two days (July 6 & 7, 2022) in Lincoln, NE. The agenda for the meeting is included in Appendix A. The meeting was facilitated by Dr. Susan Davis-Becker and Ms. Kelley Wheeler from ACS.

Panelists

NDE recruited subject matter experts to serve as panelists on one of two panels (grade 5, grade 8).

Table 1. Panelist Demographic Information

	Grade 5	Grade 8
Panelists	8	8
Position		
Science Teacher	5	5
Curriculum / Instructional Leader	3	2
University Professor	0	1
Years of Experience (average)	16.9	12.4

Two weeks before the standard setting, ACS provided the panelists with pre-meeting information that included the agenda for the standard setting along with a brief description of expectations for their participation. Panelists were asked to complete an online demographic survey as well as a security agreement indicating they agree to keeping the meeting discussions and results confidential.

General Training

The first part of the meeting began with a large-group general session with a welcome and introductions from NDE and ACS. During this general session, ACS provided training on the purpose of the standard setting, their role in it, the meaning of the achievement level descriptors (ALDs), the content of the assessments, and the procedures they were to follow in recommending cut scores. The full set of training materials for the study is included in Appendix B.

After the large-group general session, panelists transitioned to their grade-level panels to continue their work.

Grade-Level Training

The first activity in the grade-level session was for participants to learn more about the NSCAS Science assessments and review a form of the test online. This served as an opportunity for panelists to better understand the student experience as they engaged with the online platform and various item types.

After experiencing the test, panelists had a chance to review the draft Range ALDs associated with their test and work in small groups to identify the expectations that differentiated each achievement level from the one below. These expectations represented the Threshold achievement levels including *Just Barely On-Track* and *Just Barely Advanced*. Threshold performance is the location where a student's level of knowledge, skills, and abilities is "just barely" past the point-of-entry for a given achievement level. Each small group had the opportunity to share the results of their brainstorming and the ACS facilitator. Panelists were provided a copy of the final descriptors for use in making their standard setting judgments (see Appendix C).

Participants were then provided training on how to make and record their Extended Angoff judgments. This training included instructions as to how panelists should evaluate each item, make their judgments, and record their judgments. After the training, the facilitator led panelists in a practice exercise for a sample of items to ensure they understood the judgmental task. At the end of this training session, participants completed an evaluation to gather information on the general session training, the Threshold ALD development process, the practice exercise, and to gauge their understanding of the standard setting process.

Standard Setting Judgments

After the training activities, panelists were asked to work individually to consider the knowledge, skills, and abilities measured by each task and subsumed items and compare these against the Threshold ALDs. Panelists made a judgment for each item within each task as to how well they believe the *Just Barely On-Track* and *Just Barely Advanced* students will perform (i.e., what score they will earn). The specific instructions provided for this task and a sample rating form are included in Figure 1.

Panelists completed the first day of the standard setting meeting by making their initial judgments (Round 1) and a second mid-process evaluation form.

To start the second day of the standard setting, panelists were provided with the following information about the Round 1 standard setting judgments:

- Overall and individual recommendations for each cut score
- Distribution of panelist recommendations for each cut score
- Empirical estimates of item and task difficulty
- Estimated impact (i.e., percent of 2022 students within each achievement level)

The ACS facilitators led a discussion of a sample of items and panelists had the opportunity to discuss their judgments including their rationale grounded in the Threshold ALDs and how the item difficulty information may impact the revision of these judgments in Round 2. After this discussion, panelists completed their second round of Extended Angoff judgments and a third mid-process evaluation form.

Vertical Articulation

All panelists participated in a vertical articulation meeting. During this meeting, they were asked to consider the impact of the grade 5 and grade 8 Round 2 recommendations along with the impact from the 2021 ACT Science assessment to determine whether the recommended cut scores represented a reasonable set of expectations and impact between grades 5 and 8. The vertical articulation process was based on two underlying principles:

- Achievement level expectations should be coherent between grades and tests.
- Judgments of standard setting panels should be honored, unless doing so would clearly violate the above principle.

The ACS facilitator explained the purpose and process of vertical articulation to the panelists and that their task was to determine whether the magnitude of any differences between the estimated impact at each grade match the magnitude of the shifts and expectations from the academic content standards as well as a curricular and instructional perspective. The vertical articulation meeting operated on consensus where panelists were able to suggest and discuss any potential changes to cut scores with their colleagues to come to consensus. At the end of the discussion panelists were asked to complete a fourth evaluation.

Results

The results of each round of standard setting judgment are shown in Tables 2 and 3 for grades 5 and 8, respectively including the range of cut scores (minimum, maximum), the median recommended cut scores, the standard error of the recommended cut scores, the recommended range (median \pm 2 standard error), and the estimated impact of the median recommended cut score (% of students in each achievement level). Overall, the results show the panel had some variability in their recommendations but yielded reasonable standard errors. The panelists adjusted their recommendations slightly between rounds.

Table 2. Grade 5 Standard Setting Results

	Round 1			Round 2		
	Developing	On-Track	Advanced	Developing	On-Track	Advanced
Minimum	--	10	16	--	10	16
Maximum	--	13	19	--	13	22
Median	--	12	18	--	11	19
Std Error (SE)	--	0.32	0.44	--	0.40	0.76
Range (Median \pm 2 SE)	--	11 - 13	17 - 19	--	10 - 12	17 - 21
Estimated Impact	42%	48%	10%	36%	60%	4%

Table 3. Grade 8 Standard Setting Results

	Round 1			Round 2		
	Developing	On-Track	Advanced	Developing	On-Track	Advanced
Minimum	--	12	26	--	9	20
Maximum	--	18	33	--	13	30
Median	--	14	29	--	11	24
Std Error (SE)	--	0.63	0.81	--	0.49	1.21
Range (Median \pm 2 SE)	--	13 - 15	27 - 31	--	10 - 12	22 - 26
Estimated Impact	64%	36%	<1%	48%	49%	3%

The detailed student performance distributions for grades 5 and 8 are shown in Tables 4 and 5 below with the recommended range for each cut score.

Table 4. Grade 5 Performance Distribution

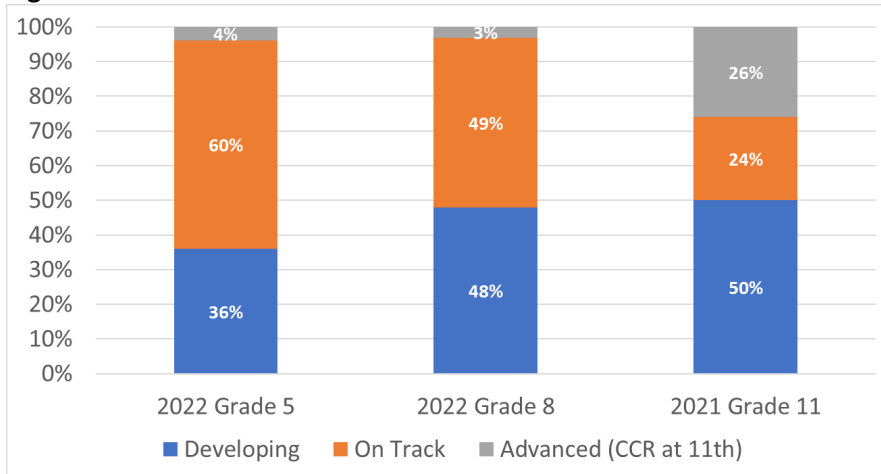
Score	Frequency	% at/above	Results
0	3	100%	
1	45	100%	
2	142	99%	
3	282	98%	
4	490	96%	
5	664	93%	
6	846	89%	
7	898	85%	
8	980	80%	
9	1142	75%	
10	1191	70%	On-Track Range
11	1287	64%	
12	1399	58%	
13	1599	51%	
14	1659	43%	
15	1800	35%	
16	1905	26%	
17	1896	18%	Advanced Range
18	1713	10%	
19	1314	4%	
20	687	1%	
21	242	0%	
22	0	0%	

Table 5. Grade 8 Performance Distribution

Score	Frequency	% at/above	Results
0	33	99.9%	
1	195	99.0%	
2	457	97.1%	
3	787	93.7%	
4	1021	89.4%	
5	1141	84.6%	
6	1225	79.4%	
7	1247	74.1%	
8	1287	68.6%	
9	1312	63.0%	
10	1327	57.4%	On-Track Range
11	1244	52.1%	
12	1282	46.7%	
13	1238	41.4%	
14	1245	36.1%	
15	1085	31.5%	
16	1042	27.1%	
17	1022	22.8%	
18	929	18.8%	
19	873	15.1%	
20	708	12.1%	
21	656	9.3%	
22	596	6.8%	Advanced Range
23	508	4.6%	
24	367	3.1%	
25	290	1.8%	
26	215	0.9%	
27	138	0.3%	
28	47	0.1%	
29	28	0.0%	
30	6	0.0%	
31	0	0.0%	
32	0	0.0%	
33	0	0.0%	

Figure 2 shows the results presented at the vertical articulation session including the impact of the Round 2 recommendations and the 2021 spring administration of the high school Science assessment (ACT). This chart shows the percent of students at the *Developing, On-Track, and Advanced* achievement levels (*College and Career Ready* at grade 11).

Figure 2. Cross-Grade Vertical Articulation Review



The panelists did not have any specific recommendations for adjustments to the cut scores based on the vertical articulation results. A couple of the 8th grade panelists did suggest that the percentage of students in the *Developing* achievement level at 5th grade seemed a bit low but also noted that there could be several factors influencing this difference between grades 5 and 8 including number of standards to cover and availability of curriculum resources.

Workshop Evaluations

Panelists completed four evaluations during the standard setting meeting. The results are presented in Tables 6-9. The results in Table 6 summarize the panelist perceptions of the panel on the various training activities. Largely, the panelists indicated the different aspects of training (general orientation, test review, Threshold ALD development, practice activity) were useful in preparing for the standard setting activities with only minor exceptions. In addition, most panelists felt the right amount of time was dedicated to these activities and all panelists felt prepared to make their standard setting judgments.

The results in Table 7 were captured after each round of standard setting judgments. Panelists indicated that they felt there was sufficient time to make their judgments in each round and indicated confidence in their judgments with higher levels of confidence in their Round 2 judgments (compared to Round 1).

The results in Table 8 provide more specifics as to what aspects of the Round 1 feedback influenced their Round 2 judgments. Panelists indicated the strongest influence from the item difficulty information (average score) and the Round 1 results.

Finally, the results in Table 9 indicate that panelists felt there was sufficient time to consider the results in the vertical articulation and that they had confidence in the results.

Table 6. Training Evaluation

Utility	Not Very Useful		Useful		Very Useful	
	N	%	N	%	N	%
General Orientation	1	6.7%	3	20.0%	11	73.3%
Test Review	1	6.7%	3	20.0%	11	73.3%
Threshold ALD Development	1	6.7%	3	20.0%	11	73.3%

Practice	1	6.7%	1	6.7%	13	86.7%
	Not enough time		Right amount of time		More than enough time	
Time	N	%	N	%	N	%
General Orientation	0	0.0%	12	80.0%	3	20.0%
Test Review	0	0.0%	11	73.3%	4	26.7%
Threshold ALD Development	1	6.7%	10	66.7%	4	26.7%
Practice	0	0.0%	13	86.7%	2	13.3%
	No		Yes			
	N	%	N	%		
Do you feel prepared to make Round 1 judgments?	0	0.0%	15	100.0%		

Table 7. Post-Judgment Evaluation

	Round 1		Round 2	
Time to Make Judgments	N	%	N	%
Sufficient Time	16	100.0%	16	100.0%
Insufficient Time	0	0.0%	0	0.0%
	Round 1		Round 2	
Confidence in Results	N	%	N	%
Confident	5	31.3%	11	68.8%
Somewhat confident	11	68.8%	2	12.5%
Not confident	0	0.0%	0	0.0%

Table 8. Round 2 Judgment Evaluation

	Somewhat Considered		Considered		Strongly Considered	
	N	%	N	%	N	%
Threshold ALDs	0	0.0%	7	46.7%	9	60.0%
Individual Round 1 Judgments	0	0.0%	8	53.3%	8	53.3%
Round 1 Panel Results	0	0.0%	5	33.3%	10	66.7%
Round 1 Impact Data	0	0.0%	7	46.7%	9	60.0%
Item Difficulty Information	0	0.0%	3	20.0%	13	86.7%

Table 9. Vertical Articulation Evaluation

	Insufficient		Sufficient					
	N	%	N	%				
Time to Consider Results	0	0%	15	100%				
	Not Confident		Somewhat Confident		Confident		Very confident	
	N	%	N	%	N	%	N	%
Confidence in Final Recommendations	0	0%	2	13%	8	53%	5	33%

Validity Evidence and Next Steps

This report summarizes the standard setting process developed and applied to achieve recommended cut scores for the NSCAS Science assessment at grades 5 and 8. This information can also be evaluated to identify the validity evidence in support of the use of these cut scores for interpreting test scores from the Spring of 2022 (and future exams). Kane (2001) identified three areas of validity evidence: procedural, internal, and external.

Procedural

This category of validity evidence focuses on the standard setting process including the selection of panelists, the methodology selected, and the process by which the methodology is applied. The procedural validity evidence from this study includes:

- The panelists selected for each panel represented a range of school districts from across the state and included science teachers, curriculum and instructional leaders from different parts of the state, and one university processor.
- The Extended Angoff methodology was well suited for this standard setting given the organization of the test (multiple items linked to the same set of stimuli) and the nature of the items (some multi-point items with partial scoring)
- The panel had the opportunity to be trained on the methodology, practice making such judgments, and indicated sufficient time and confidence in their judgments.

Internal

This category of validity evidence focuses on the level of agreement among the panelists indicating they had similar expectations for each achievement level. The internal validity evidence from this study includes:

- Panelists were provided an opportunity to review the detailed Range ALDs from NDE and then discuss and develop the threshold expectations for the two transition points. The results of this discussion were available throughout the judgmental process.
- Panelists indicated a general level of agreement in their expectations as evidenced by the standard error of the judgments from each round. Although some standard errors increased between rounds indicating panelists reacted differently to the Round 1 feedback, the general level of agreement is acceptable.

External

This category of validity evidence focuses on the reasonableness of the standard setting results by stakeholder groups and/or in comparison to external sources of information. The external validity evidence from this study includes:

- Panelists were provided the opportunity to review the results from their grade-level panel against those from the other grade (5 or 8) as well as the results from high school. This review included a discussion about expectations and influence on student performance from this school years. The results indicated these results were reasonable.

At the conclusion of all workshop activities, the cut score recommendations were provided to NDE for their review. Based upon the evidence collected and the review of the performance of panelists, it does appear that the cut scores recommendations recommend appropriate cut scores for the NSCAS Science.

References

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*, 41–55.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum.

Appendix A: Panelist Agenda

Day/Time	Key activities	Materials
Pre-Meeting	ACS will provide a pre-meeting document for panelists describing the process and their expectations for participation. Panelists will also be asked to complete an online demographic survey and an online security agreement.	Pre-Meeting Document Pre-Meeting Non-Disclosure Form Demographic Form
Day 1		
9:00 – 10:00	General Orientation - Overview of the Assessment - Overview of the standard setting activities	Orientation and Training PowerPoint
10:00 – 10:15	<i>Transition to grade level panels</i>	
10:15 – 10:45	Review the Assessment	Operational test form
10:45 – 11:45	Draft Threshold ALDs	Range ALDs Threshold ALD template
11:45 – 12:30	Lunch	
12:30 – 1:15	Finalize Threshold ALDs	
1:15 – 2:30	Practice judgments Evaluation #1	Field test items/tasks Practice rating form Evaluation #1
2:30 – 2:45	Break	
2:45 – 4:00	Round 1 judgments Evaluation #2	Operational test form Operational rating form Evaluation #2
Day 2		
9:00-10:00	Review and discuss Round 1 judgments	Round 1 Results Assessment data
10:00-11:00	Complete Round 2 judgments	Operational rating form
11:-11:30	Review and discuss Round 2 judgments Evaluation #3	Round 2 results Evaluation #3
11:30-12:15	Break and Lunch	
12:15-1:30	Vertical articulation discussion Evaluation #4	Vertical articulation orientation PowerPoint Evaluation #4

Appendix B: Study Materials

Training Materials



1 - Panelist Advance Information2.docx



4 - Orientation and Training Slides.pptx



5 - SIPS Complexity Framework Draft_Jun

Standard Setting Materials



9 - Grade 5 Practice Rating Form.docx



9 - Grade 8 Practice Rating Form.docx



13 - Grade 5 Operational Rating Fc



13 - Grade 8 Operational Rating Fc

Vertical Articulation Discussion



17 - Vertical Articulation Slides.ppt

Appendix C: Achievement Level Descriptors

Grade 5

RALDs	<i>Just Barely On-Track</i>	<i>Just Barely Advanced</i>
1A – Asking Questions	<p>Ask simple relevant questions about the phenomena (“what”). Start to make connections to prior experience (see applications)</p> <p>Identify variables (things we can change) that are part of a cause and effect relationship</p>	<p>Ask questions to generate evidence</p> <p>Ask a variety of questions that demonstrate they understand causality</p> <p>Make explicit connections to other areas or phenomena.</p>
1B – Defining Problems	<p>Able to identify the problem but unable to design all the steps to the solution</p> <p>Ask simple questions about the design problem, starting to determine the constraints on the design</p>	<p>Able to organize their thinking.</p> <p>Apply scientific ideas to the design problem (clear connection)</p> <p>Can answer questions about the design</p> <p>Considering more than one criteria/constraint to the design problem</p>
2 – Developing and Using Models	<p>Starting to connect a given model to prior experiences</p> <p>Explain some cause and effect relationships</p> <p>Develop simple models that represent a system</p>	<p>Compare, contrast, and connect different models to explain causes of phenomena</p> <p>Develop different types of models including complex models and making the implicit explicit</p>
3 - Planning and Carrying Out Investigations	<p>Start to use collected data to predict and explain phenomena</p> <p>Seeing purpose of data collected</p>	<p>Organize data to help evaluate and predict</p>
4 - Analyzing and Interpreting Data	<p>Comparing and contrasting patterns to interpret (make sense) of patterns</p> <p>Create graphical ways to represent patterns</p> <p>Beginning to identify how changes might affect the design</p>	<p>Making connections between patterns and causes</p> <p>Beginning to make predictions and explain based on evidence</p> <p>Identify changes needed to improve design solution</p>
5 - Using Mathematics and	<p>Organize simple data to identify patterns</p> <p>Identify relevant data to create graphical evidence</p>	<p>Organize complex data to make predictions based on patterns</p>

Computational Thinking	Complete and/or modify simple graphs or charts	Find data, graph, and compare it to identify patterns for evidence Create graphs and charts, make simple comparisons
6A – Constructing Explanations	Select some relevant information to construct explanations Identify and use relevant evidence to construct or support explanations	Looking carefully at multiple explanations to explain relationships
6B – Designing Solutions	Beginning to generate some solutions to solve a design problem Using evidence to create solutions	Implement a testable solution within the criteria and constraints
7 - Engaging in Argument from Evidence	Beginning to use evidence to refine arguments Beginning to use research findings to explain phenomenon	Beginning to evaluate quality research data to construct an argument to explain the cause and effect relationship Progressing towards the ability to generate data Starting to explain why an argument needs to be modified
8 - Obtaining, Evaluating, and Communicating Information	Beginning to understand, comprehend, and compare information from grade-level text, graphic organizers, and charts Sharing obtained information verbally or in written form	Independently reading or listening to analyze more complex grade-level text, graphic organizers, and charts Sharing obtained information in different forms (e.g., media, diagrams, tables, charts)

Grade 8

RALDs	<i>Just Barely On-Track</i>	<i>Just Barely Advanced</i>
1A – Asking Questions	Likely to use empirical evidence from the data set (qualitative or quantitative) Relationships in models Do seek additional info AND clarify	Asking questions instead of just evaluating questions Use empirical evidence from the data set (quantitative required) to support
1B – Defining Problems	Defines a problem that includes a criterion and/or constraint	Problem <u>must</u> involve the development of a process / system Identifying interaction between the process / system
2 – Developing and Using Models	<u>Develops</u> a model that includes evidence	Evaluate merits and limitations of a model Revise models based on evidence Use models to test unobservable mechanisms / scales in the model Use model to generate data
3 - Planning and Carrying Out Investigations	Work individually Understand cause and effect relationship between variables Identifying and organizing data Conduct an investigation	Identify multiple relationships between independent and dependent variables Conduct and evaluate experimental design Evaluate data
4 - Analyzing and Interpreting Data	Demonstrates 2 of 3 (construct, analyze, interpret) Identify quantitative relationship Identify limitations in data OR try to improve accuracy of data	Demonstrates all three (construct, analyze, interpret) Identify AND <u>apply</u> quantitative relationships Analyze the data to find <u>optimal</u> range Identify limitations in data AND try to improve accuracy of data
5 - Using Mathematics and Computational Thinking	Use quantitative data <u>Order</u> the steps to solve a problem Apply concept to solve problem	Evaluate qualitative AND quantitative data <u>Evaluate</u> steps to solve a problem Compare quantitative data
6A – Constructing Explanations	Construct an explanation that predicts relationships between variables that describe a phenomena	Construct an explanation that predicts AND describes relationships between variables

		Construct an explanation that critically evaluates investigation
6B – Designing Solutions	Apply scientific ideas to design, construct, and/or test a solution Makes best choices to optimize performance of solution / design	Apply scientific ideas to design, construct, AND test a solution that meets specific criteria and constraints
7 - Engaging in Argument from Evidence	Can compare AND critique two arguments Can provide critiques by cite relevant evidence Individually construct and use arguments Construct an argument that refutes a device / system Evaluates design solutions	Compare and critique <u>multiple</u> arguments Can provide and receive critiques by citing relevant evidence AND posing / responding to questions Individually construct, use, and present arguments Constructs an argument that supports AND refutes a device / system
8 - Obtaining, Evaluating, and Communicating Information	Integrates quantitative information Analyzes evidence from multiple given appropriate sources Evaluates data and hypotheses in scientific texts Communicates in writing	Reads <u>multiple</u> scientific texts Integrates qualitative AND quantitative information Selects appropriate sources to analyze evidence Evaluates data, hypotheses, AND conclusions in scientific texts Communicates oral AND digital