



Chemistry  
Education Research  
and Practice

**Visualizing Chemistry Teachers' Enacted Assessment Design Practices to Better Understand Barriers to "Best Practices"**

Journal:	<i>Chemistry Education Research and Practice</i>
Manuscript ID	RP-ART-06-2020-000179.R2
Article Type:	Paper
Date Submitted by the Author:	14-Jan-2021
Complete List of Authors:	Schafer, Adam; Miami University, Chemistry & Biochemistry Borland, Victoria; Miami University, Chemistry & Biochemistry Yeziarski, Ellen; Miami University, Chemistry & Biochemistry

SCHOLARONE™  
Manuscripts

---

# Visualizing Chemistry Teachers' Enacted Assessment Design Practices to Better Understand Barriers to "Best Practices"

Adam G. L. Schafer, Victoria M. Borland, Ellen J. Yeziarski\*

Department of Chemistry and Biochemistry, Miami University, Oxford, Ohio 45056, United States

## ABSTRACT

Even when chemistry teachers' beliefs about assessment design align with literature-cited best practices, barriers can prevent teachers from enacting those beliefs when developing day-to-day assessments. In this paper, the relationship between high school chemistry teachers' self-generated "best practices" for developing formative assessments and the assessments they implement in their courses are examined. Results from a detailed evaluation of several high school chemistry formative assessments, learning goals, and learning activities reveal that assessment items are often developed to require well-articulated tasks but lack either alignment regarding representational level or employ only one representational level for nearly all assessment items. Implications for the development of a chemistry-specific method for evaluating alignment are presented as well as implications for high school chemistry assessment design.

## KEYWORDS

High School, Assessment, Teacher Professional Development, Chemical Education Research

Assessment is a complex process involving several interrelated decisions that impact a teacher's ability to draw inferences about the teaching and learning process. Several literature resources are available to high school teachers to assist with the complex decisions that go into designing assessments and evaluating assessment quality (Bell and Cowie, 2001; Martone and Sireci, 2009; Towndrow *et al.*, 2010; Ruiz-Primo *et al.*, 2012; Towns, 2014a; Harshman and Yeziarski, 2017; Dini *et al.*, 2020). Often, a teacher's personal goals for assessment design align well to the guidelines present in the literature (Sandlin *et al.*, 2015; Schafer and Yeziarski, 2020a). However, enacting personal goals and literature-suggested guidelines during assessment development can be difficult for high school chemistry teachers (Black and Wiliam, 1998; Mandinach *et al.*, 2006; Sandlin *et al.*, 2015). The barriers hindering enactment can cause misalignment between a teacher's goals and the assessments they generate. In this work, a "barrier" is defined as something preventing a teacher from enacting their personal goals. For example, a "barrier" could be a gap in knowledge, lack of awareness, or even a mismatch of resources. A better understanding of the relationship between chemistry teacher assessment goals and assessment products is necessary for helping teachers better design formative assessments.

## FRAMEWORK FOR STUDY

Establishing goals is often considered the first step in initiating a cycle of formative assessment. Goals could be, for example, learning goals to describe expectations for student learning or even goals that establish practices the teacher will incorporate into formative assessment design. Essentially, the goals a teacher sets should guide the design of opportunities to collect evidence about the progress of student learning and the success of learning activities (Harshman and Yeziarski, 2017). The evidence gathered can then inform new goals as teachers craft an instructional response, repeating the formative assessment cycle. This iterative formative assessment cycle has been given many names in the literature; here we are calling the iterative process data-driven inquiry (Harshman and Yeziarski, 2017).

The study described herein follows a cohort of high school chemistry teachers participating in a professional development designed to follow the process of data driven inquiry (Figure 1). The teachers started this professional development by generating a set of “best practices” for designing and interpreting formative assessments (Schafer and Yeziarski, 2020a) before employing these

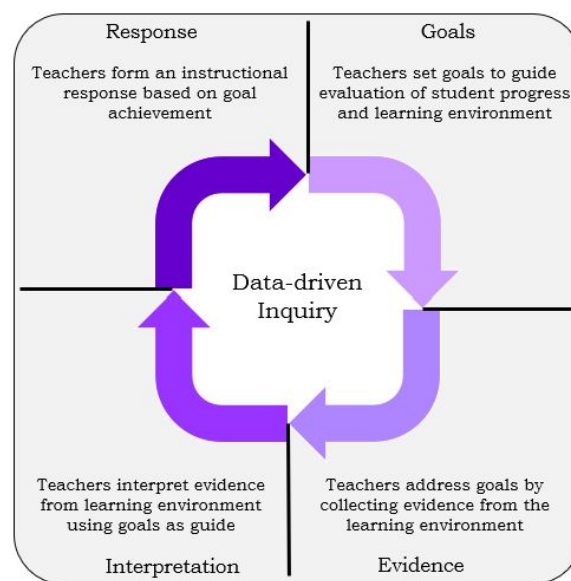


Figure 1. Data-driven Inquiry Cycle

“best practices” to guide the design of planned, formative assessments for their classrooms. Here, “best practices” are in quotations to reflect that these practices are the participating teachers’ proposed ideas about how to best design and interpret formative assessments, not necessarily evidence-based best practices from the literature, although the teacher-generated “best practices” were found to align with relevant literature for formative assessment design (Schafer and Yeziarski, 2020a). This investigation uses the teacher-generated “best practices” to better understand the relationship between chemistry teacher assessment goals and design practices. Following data-driven inquiry, the “best practices” specified by the teachers should inform the design of formative assessments (*i.e.*, the tools used to collect evidence) alongside any lesson-specific learning goals. What follows is a brief primer on formative assessments from the literature, a summary of the “best practices” generated by teachers in the earlier work (Schafer and Yeziarski, 2020a), and literature that contextualizes and/or reinforces the high school chemistry teachers’ ideas.

## FORMATIVE ASSESSMENT BEST PRACTICES

Assessments that are used to inform continued instruction are considered formative assessments (Black and Wiliam, 1998; Irons, 2008; Clinchot *et al.*, 2017). Formative assessments can range from spontaneous in-class questioning to planned

1  
2  
3 quizzes. Planned, formative assessments allow the teacher to consider items that students will respond to in the future. When  
4  
5 designing and evaluating the quality of planned, formative assessments, a teacher must enact a practical knowledge of the  
6  
7 students and classroom alongside content knowledge to effectively generate an opportunity to judge student learning and the  
8  
9 learning activity (Bell and Cowie, 2001; Remesal, 2011; American Chemical Society, 2012; DeLuca *et al.*, 2018). How this  
10  
11 knowledge is enacted by a teacher is informed by the beliefs they hold about assessment design. Arguably, a teacher will  
12  
13 attempt to design assessment items in a way that agrees with what they believe to be the “best practices” to employ. A  
14  
15 teacher’s perception of what counts as a “best practice” is informed by several sources (e.g., peers, current literature, prior  
16  
17 experiences, training, and content knowledge). However, many sources informing chemistry teachers’ perceptions of “best  
18  
19 practices” create barriers for enacting beliefs about how an assessment should be designed by requiring a significant amount  
20  
21 of interpretation by the teacher (Black and Wiliam, 1998). Although several studies propose best practices for assessment  
22  
23 design, few have examined how literature-based practices that align with chemistry teachers’ personal assessment beliefs  
24  
25 about best practices inform the decisions teachers make while generating planned, formative assessments.

26  
27 When Schafer and Yeziarski (2020a) examined the discourse between a group of high school chemistry teachers  
28  
29 generating a set of “best practices” for formative assessment design, they learned that the teachers collaboratively proposed  
30  
31 and revised practices as critical friends (i.e., peers who were comfortable and had experiences collegially contributing to  
32  
33 discourse and resolving disagreements) (Curry, 2008; Loughran and Brubaker, 2015; Schafer and Yeziarski, 2020a). Among  
34  
35 the teachers’ “best practices” were the considerations that an assessment should:

- 36 • articulate clearly what the student should do when responding to the assessment item,
- 37
- 38 • address a variety of conceptual and representational levels, and
- 39
- 40 • align assessment items to the instructional materials (*e.g.*, learning goals).

41  
42 Although this is not an exhaustive list of the “best practices” generated by the teachers, the stated practices represent those  
43  
44 specific to barriers high school chemistry teachers may face when designing planned, formative assessments. The same  
45  
46 chemistry teachers who generated the practices in the bulleted list from Schafer and Yeziarski are the participants in the study  
47  
48 described herein. While generating the “best practices”, teachers had multiple opportunities to contribute, revise, and remove  
49  
50 practices before agreeing on a final draft. As such, the “best practices” represent consensus ideas from the participating  
51  
52 teachers. Employing the “best practices” generated by these teachers presents a unique opportunity to examine the  
53  
54 relationship between chemistry teacher beliefs and practices about planned, formative assessment design. Before  
55  
56  
57  
58

1  
2  
3 investigating the relationship between the chemistry teacher's beliefs about best practices and the practices they enact when  
4  
5 85 designing planned, formative assessments, each of the teacher-generated "best practices" are described in further detail.

6  
7 Assessments Should Articulate Clearly What the Student Should Do When Responding to the Assessment Item of a Task  
8 The teacher-generated "best practice" of using clear articulations of what the student should do matches guidelines  
9  
10 proposed in educational literature (National Research Council, 1999, 2001, 2014; Bell and Cowie, 2001; Stiggins, 2001;  
11  
12 Gibbs and Simpson, 2004; Dwyer, 2007; Lyon, 2011). What the student is required to do is often referred to as the "task" of  
13  
14 90 an assessment item (McDonald, 1964; Hoffman and Medsker, 1983; Jonassen *et al.*, 1999; Merrill, 2007; Tomanek *et al.*,  
15  
16 2008). Establishing an appropriate task is a key process teachers undergo when designing or selecting planned, formative  
17  
18 assessment items, and literature guidance provides teachers with specific considerations in how to productively design tasks  
19  
20 for assessments (Bell and Cowie, 2001; Tomanek *et al.*, 2008; Kang *et al.*, 2016; Schafer and Yeziarski, 2020b). For  
21  
22 example, teachers may ask students to calculate a value, explain a relationship, or select an appropriate answer from a set of  
23  
24 95 choices. In these examples, the tasks are to *calculate*, *explain*, and *select*, respectively. Numerous methods for evaluating  
25  
26 assessment quality employ task as a criteria, further highlighting the importance of establishing a specific, expected action to  
27  
28 be performed by the student when responding to the assessment item (Webb, 1997; Porter and Smithson, 2001; Rothman *et*  
29  
30 *al.*, 2002; Martone and Sireci, 2009). Any task designed into an assessment item is inevitably informed by a teacher's goals  
31  
32 for a given formative assessment cycle (Harshman and Yeziarski, 2017).

33  
34 100 Assessments Should Address a Variety of Conceptual and Representational Levels

35 The National Research Council encourages teachers to provide students with opportunities to perform tasks that go  
36  
37 beyond simple recall (National Research Council, 2014). Teachers can move past recall of facts by requiring students to  
38  
39 apply their conceptual knowledge. However, a common pitfall teachers face is to include mainly recall tasks during  
40  
41 assessment, potentially resulting in inadequate information about student understanding (Stiggins, 2001; Towns, 2014b;  
42  
43 105 Schafer and Yeziarski, 2020a). When designing assessments, items incorporating recall tasks should not be perceived as  
44  
45 "bad" formative assessment items, although including only recall items may limit the data available to inform continued  
46  
47 instruction. Many sources agree that collecting data from a variety of tasks can help teachers better judge student  
48  
49 understanding (Bell and Cowie, 2001; Gearhart *et al.*, 2006; National Research Council, 2014). If possible, teachers should  
50  
51 collect multiple measurements, since the additional data can provide the teacher and student with insights about student-  
52  
53 110 specific challenges and knowledge gained (Black and Wiliam, 1998; Bell and Cowie, 2001; National Research Council,  
54  
55 2001; Stiggins, 2001; Cizek, 2009). Arguably, an appropriate task is one that corresponds to similar items in the instructional  
56  
57 materials and fulfills the goals(s) of the assessment (Harshman and Yeziarski, 2017).

1  
2  
3 Assessing student knowledge within a variety of representational levels was a “best practice” generated by the chemistry  
4 teachers in the study by Schafer and Yeziarski, but is not a commonly cited practice in domain-general literature (Schafer and  
5 Yeziarski, 2020a). Establishing an appropriate representational level has been found as a practice chemistry teachers engage  
6  
7 115 in when developing formative assessment items (Schafer and Yeziarski, 2020b). Although some teachers explicitly focus on  
8  
9 representational level when designing assessment items, several investigations have documented student difficulties  
10  
11 navigating between different representational levels (Gabel *et al.*, 1987; Nakhleh, 1992; Russell *et al.*, 1997; Gkitzia *et al.*,  
12  
13 2020). As such, literature guidance suggests carefully scaffolding the number of representational levels per item when the  
14  
15 goal is to assess student knowledge and abilities about representational level (Taber, 2013).  
16  
17 120

18 While representational levels have been accepted within chemistry as important descriptors of chemical information,  
19  
20 other disciplines likely do not have the same need for describing chemical phenomena (Johnstone, 1991; Taber, 2013;  
21  
22 Vilardo *et al.*, 2017). Inevitably, many assessment practices are domain-general, while others are more discipline specific  
23  
24 (Coffey *et al.*, 2011). Few investigations focus on the role representational level plays in conjunction with other assessment  
25  
26 125 item components (such as task); however, better understanding as to how these components are developed in unison could  
27  
28 improve understanding of chemistry teacher assessment design practices.  
29

### 30 Assessments Should Align Assessment Items to the Instructional Materials

31 Alignment refers to the degree to which learning goals, learning activities, and assessments are in agreement and  
32  
33 mutually support students in learning what they are expected to know and do (Tyler, 1949; Webb, 1997; Martone and Sireci,  
34  
35 130 2009), although there is some disagreement as to what “alignment” means in studies regarding current science standards  
36  
37 (Fulmer *et al.*, 2018). In their review of studies about aligning to the Next Generation Science Standards (NGSS), Fulmer *et*  
38  
39 *al.* (2018) acknowledge that the two main questions regarding alignment involve (1) What aspect of the NGSS are being  
40  
41 considered? and (2) How is alignment being judged? Addressing these questions presents challenges unique to the discipline  
42  
43 of chemistry that require teachers receive chemistry-specific support and guidance for developing skills aligning assessment  
44  
45 135 and instruction. Unfortunately, a review of chemistry education literature reveals that few chemistry-specific methods exist  
46  
47 for evaluating the alignment of an assessment to instruction; however, alignment methods are prevalent within domain-  
48  
49 general education literature (e.g., Porter and Smithson, 2001; Rothman *et al.*, 2002; Webb and Herman, 2006; Martone and  
50  
51 Sireci, 2009; Kaderavek *et al.*, 2015; Fulmer *et al.*, 2018; Young *et al.*, 2019). One instrument high school chemistry teachers  
52  
53 may consider when evaluating lesson alignment is the *EQuIP Rubric for Lessons and Units: Science* (Achieve, 2016). The  
54  
55 140 EQuIP rubric evaluates lessons sequences and units for NGSS alignment using three domains: a three-dimensional design,  
56  
57  
58

1  
2  
3 instructional supports, and monitoring student progress (Achieve, 2016). The rubric is extensive in its descriptions for NGSS-  
4 alignment; however, the comprehensiveness of the rubric can be challenging to use, especially for teachers new to the rubric  
5 (Fulmer *et al.*, 2018). In response to teachers' struggles using the EQUiP rubric, Achieve released a reduced version of the  
6 rubric (Achieve, 2016), but the reduced version is not meant to be used as a complete evaluator and requires a follow-up with  
7 the full EQUiP. Overall, the lack of chemistry-specific guidelines and methods burdens high school chemistry teachers with  
8 translating domain-general practices to their own context or employing comprehensive instruments that can be challenging to  
9 apply.  
10

11  
12  
13  
14  
15  
16 Since assessment and instruction do not exist in isolation, improvements in learning are dependent upon the quality of  
17 alignment among assessments, curriculum, and instruction (Bell and Cowie, 2001; National Research Council, 2001, 2014;  
18 Broman *et al.*, 2015). However, criteria for alignment specific to the discipline of chemistry (such as representational level)  
19 can increase the difficulty of designing and implementing formative assessment tasks that are tightly aligned to instructional  
20 materials. The interrelatedness between assessment and instruction implies that teachers cannot make valid inferences from  
21 assessment data without alignment to instruction (Datnow *et al.*, 2007; Hamilton *et al.*, 2009; Sandlin *et al.*, 2015; Harshman  
22 and Yeziarski, 2017). Although several investigations recognize the importance of alignment, there is not a consensus on the  
23 number of aligned criteria necessary for adequate measurement of student competency. Typically, studies suggest that  
24 between six and eight observations of students performing a task is sufficient for a reliable measurement (Webb, 2006;  
25 Martone and Sireci, 2009; Praetorius *et al.*, 2014; Briggs and Alzen, 2019). Although the study presented herein does not  
26 seek to confirm or refute these values, we recognize that teachers should consider how many items are necessary to evidence  
27 a given learning goal when designing a planned, formative assessment. Inevitably, what is considered a sufficient number of  
28 ways and instances of assessing a learning goal must be decided by the assessment designer based on the purpose of the  
29 assessment (Kane, 2006; Webb, 2006; American Educational Research Association *et al.*, 2014; Harshman and Yeziarski,  
30 2017). Processes such as data-driven inquiry can support teachers' considerations of alignment by establishing goals which  
31 inform the design of tools to collect evidence and the conclusions drawn from the evidence collected (Harshman and  
32 Yeziarski, 2017). Better understanding the relationship between teachers' perceived "best practices" (such as alignment to  
33 instruction) and the assessments they design is essential for helping teachers improve formative assessment design practices.  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## RESEARCH QUESTIONS

The purpose of this study is to investigate the nature of the relationship between high school chemistry teachers' perceptions of "best practices" for formative assessment and the planned, formative assessments generated for their classrooms. The research questions guiding this study are:

- 1) How can the tasks and representational levels of planned, formative assessment items be diagrammed to allow for comparison with corresponding instructional materials (i.e., learning goals and learning activities)?
- 2) What is the nature of the relationship between teachers' self-generated "best practices" for formative assessments and the assessments they generate?

## METHODS

To address the research questions, the assessments, learning activities, and learning goals of a group of high school chemistry teachers were collected. This research was approved by the university's Institutional Review Board as an investigation into the alignment between high school chemistry teachers' practices and beliefs about assessment. All methods were in compliance with the university's policies on ethics. Informed consent was obtained for all participants prior to participation.

### Sample

Five teachers participated in a long-term professional development program focused on improving assessment practices during the Spring/Summer of 2018. As part of the professional development program, all teachers generated an assessment for an inquiry activity implemented in their own classrooms. However, one teacher was unable to complete the professional development and another decided to generate a summative assessment. As this investigation focuses on planned, formative assessments, these two teachers were removed from the study. Demographic information for the three teachers included in this study is presented in Table 1.

**Table 1. Teacher Demographic Information**

Participant	Bachelor's Degree Major	Highest Degree Earned <sup>a</sup>
Celine	Biology Education	MEd, MS
Claude	Chemistry	MEd
Emmerson	Chemistry/Earth Science Education	MEd

<sup>a</sup>MEd = Master of Education, MS = Master of Science

### Data Sources

As part of the professional development, the chemistry teachers were asked to provide the planned, formative assessment for a lesson as well as any learning goals and learning activities associated to the lesson to serve as artifacts for the study described herein. In this study, an *artifact* is defined as a singular document, such as the learning goals, a learning activity, or



1  
2  
3 195 an assessment instrument, whereas an *item* is any singular task requested or question asked of the student within an artifact.  
4  
5 Further information about the implementation of the lesson can be found in the results. For this work, the learning goals are  
6  
7 the student expectations set by the teachers at the beginning of the lesson. All teachers recorded the learning goals on the  
8  
9 materials provided to the students at the beginning of the lesson. Learning activities are any instructional materials used by  
10  
11 the teachers throughout the lesson. Claude and Emmerson separately implemented laboratory activities while Celine  
12  
13 200 implemented a lab along with two group work activities.

#### 14 15 Artifact Analysis

16 Following the process of data-driven inquiry, the “best practices” generated by the teachers represent goals for designing  
17  
18 formative assessment items. Although other goals likely contributed to their assessment design, the “best practices” should be  
19  
20 evident in the tools that teachers designed and implemented to collect evidence. As such, the “best practices” generated by  
21  
22 205 the teachers were distilled into distinct analysis categories. The analysis categories synthesized were *task* (referring to the  
23  
24 practice of “clearly articulate what the student should do”), *representational level* (referring to the practice of “address a  
25  
26 variety of conceptual and representational levels”), and *chemistry content* (referring to the practice of “align assessment items  
27  
28 to the instructional materials”). The synthesis of each of these analysis categories is presented in further detail below.

29  
30 **Task.** The tasks within teacher artifacts were coded to address the teacher-generated “best practices” of “clear  
31  
32 210 articulation of a task” and “assessment of a variety of conceptual levels.” Codes within the task category were inspired by  
33  
34 “The New Taxonomy” by Marzano and Kendall (Marzano and Kendall, 2008). This taxonomy was modified from previous  
35  
36 taxonomies (*e.g.*, Bloom’s Taxonomy (Bloom *et al.*, 1956)) to describe specific actions reflective of individual knowledge  
37  
38 categories (Marzano and Kendall, 2008). The task codes (Table 2) for this study similarly reflect the process students must  
39  
40 carry out to demonstrate knowledge within each item.

41 Table 2. Example Tasks for Each Task Code

42 Task Code	43 Example Tasks
44 Retrieval	Identify, recognize, calculate, complete, apply, demonstrate
45 Explanation	Explain, summarize
46 Representation	Draw, use models, represent, show
47 Analysis	Sort, categorize, differentiate, assess, critique, evaluate, diagnose
48 Knowledge Utilization	Test, how would you determine, generate and test

49 215 **Representational Level.** To address the teacher-generated “best practice” of “assessing at a variety of representational  
50  
51 levels” the representational level of each item was coded. Johnstone’s representational levels were used to generate codes for  
52  
53 the representation category (Johnstone, 1991). The representational level codes describe how information is to be represented  
54  
55  
56  
57  
58

in the student response, if specified. Table 3 presents descriptions of the three representational levels, including the “Ambiguous” code for when no representational level is communicated.

Table 3. Code Descriptions for Each Representational Level

Representational Level Code <sup>a</sup>	Representational Level Descriptions
Macroscopic	Representation of species/events on a visible scale to communicate chemical ideas/events/species
Symbolic	The use of descriptive words, symbols, or values to communicate chemical ideas/events/species
Particle-level	Representation of species/events on an invisible scale to communicate chemical ideas/events/species
Ambiguous	No representational level communicated

<sup>a</sup>Mixtures of representational level codes are possible and are written as the two representational levels present in the item (e.g., Macroscopic/Symbolic)

**Chemistry Content.** To provide a consistent means of addressing the teacher-generated “best practice” of “assessment is aligned with instructional materials” each item was inductively coded for the chemistry content embedded within the item using constant comparative analysis to ensure independence across content themes (Maxwell, 2013). The complete codebooks for all coding schemes are available in Appendices A, B, and C. Although other studies have investigated alignment *via* comparison of the assessment and/or learning activity items to a learning goal (or state standard), addressing research question one required the generation of a diagram that was descriptive of items misaligned with the learning goal while precisely illustrating items that were aligned with learning goals (Rothman *et al.*, 2002; Martone and Sireci, 2009; Polikoff and Porter, 2014).

It is essential to establish trustworthiness of coding for any qualitative investigation (Patton, 2002). As such, evidence of trustworthiness was established by addressing the credibility, dependability, and transferability of the findings. All investigators have previous experience teaching at the high school level and facilitating professional development for high school chemistry teachers, granting credibility to the codes generated. To establish evidence of dependability, interrater agreement was conducted as well as frequent debriefings. Interrater agreement of task and representational level codes was established by having two researchers independently code items from one learning activity. Code applications were compared, with an agreement of 79%. Disagreements in code application were negotiated, code descriptions were collaboratively revised, and the codes were reapplied to the data set once complete agreement was established. Codes for chemistry content were generated by an individual researcher and were independent for each group of artifacts. Once chemistry content codes were generated, the descriptions were shared with other members of the research team and collaboratively revised until agreement was reached. Debriefings between the authors were held weekly. In addition, monthly debriefings were held with several graduate students and another chemistry education research faculty member who were not involved in the data collection. Transferability of the findings presented are limited but carefully defined; as a small group of high school chemistry teachers with several years of professional development. More generally, findings have the capacity to

generate discussion and avenues for future studies as well as guidance for high school chemistry teachers and education researchers.

Assessment items were compared to items in the corresponding learning activity and learning goal to investigate for code co-occurrence. This comparison was facilitated through the use of a novel diagram called an “alignment plot.” These alignment plots illustrated each of the codes applied to the items within teacher artifacts. For this work, alignment is defined as the presence of code co-occurrence among all artifacts with regard to task and/or representation level within a particular chemistry content category. The alignment plots served to visualize the teachers’ enacted practices when designing planned, formative assessments. By organizing the features of the artifacts in the alignment plot, we were able to characterize how teachers designed their assessments as compared to how they thought assessments *should* be designed.

## RESULTS AND DISCUSSION

### Overall Code Occurrences

The first research question seeks to establish a means of diagramming teacher artifacts to illustrate how teachers enact their “best practices” during assessment design. To address this research question, a table quantifying the tasks and representational levels of all the items present in teacher artifacts was generated. The table helps to reveal code occurrence across learning goals, assessment items, and learning activity items.

Table 4 communicates the total number of items that include each task, representational level, and chemistry content topic for Claude’s artifacts. The learning goals for Claude’s lesson were that the student would be able to:

1. Identify a redox reaction based on symbolic representations.
2. Represent particulate level representations of redox reactions.
3. Predict products of redox reactions.

Claude generated one assessment and one learning activity to address these learning goals. Students in Claude’s classes were provided one day to complete the learning activity, taking the assessment the following day.

Table 4. Counts of Each Code Occurrence for Claude’s Artifacts

Category	Code	Learning Goals	Assessment	Learning Activity	Total
Task	Retrieval	1	13	22	26
	Explanation	-	1	13	14
	Representation	1	2	5	8
	Analysis	1	-	2	3
	Knowledge Utilization	-	-	-	0
Representational Level	Symbolic	1	14	26	41
	Symbolic/Particle	-	-	-	0
	Particle	1	1	3	5
	Particle/Macroscopic	-	-	-	0
	Macroscopic	-	1	11	12

	Macroscopic/Symbolic	-	-	1	1
	Ambiguous	1	-	1	2
	Static Chemical System	3	7	28	38
	Chemical Phenomena	-	4	7	11
	Electron Count and Movement	-	5	2	7
	Observations	-	-	4	4
	All Content Areas	-	-	1	1
	Total	3	16	42	

Claude's artifacts addressed three chemistry topics within oxidation and reduction reactions with an additional code for items that call for in-lab observations. One item in Claude's learning activity asked students to summarize what happens during a redox reaction. This item addressed all content codes and thus was given its own content code (shown in Table 4 as "All Content Areas"). Example items from Claude's artifacts are shown in Figure 3. The tasks articulated in Claude's items were mainly retrieval (26 total) with fewer explanation (14 total), representation (8 total), and analysis (3 total) tasks. Additionally, Claude's artifacts addressed a variety of representational levels; however, the symbolic representational level was emphasized much more than others with 41 items out of 61 total. The items in Claude's artifacts meet the self-imposed requirements of articulating specific tasks to complete while also incorporating a variety of conceptual and representational levels, although the symbolic representational level was disproportionately emphasized.

Table 5 communicates the total number of items that include each task, representational level, and chemistry content topic found in Celine's artifacts. The learning goals for Celine's lesson were that students will be able to:

1. successfully employ ratios and proportions to obtain relative mass for particles of imaginary elements.
2. successfully explain how Avogadro's law allows scientists to assign mass to particles as tiny as atoms.
3. use ratio relationships and reasoning to assign a relative mass to an unknown particle, based on given information.

Celine generated one assessment and three learning activities to address these learning goals. Students in Celine's classes were provided one day to complete each learning activity, taking the assessment the day after the third learning activity.

Table 5. Number of Each Code Occurrence for Celine's Artifacts

Category	Code	Learning Goals	Assessment	Learning Activity	Total
Task	Retrieval	2	6	40	48
	Explanation	1	-	3	4
	Representation	-	1	-	1
	Analysis	-	-	-	0
	Knowledge Utilization	-	-	-	0
Representational Level	Symbolic	3	5	42	50
	Symbolic/Particle	-	-	-	0
	Particle	-	-	-	0
	Particle/Macroscopic	-	1	-	1
	Macroscopic	-	-	1	1
	Macroscopic/Symbolic	-	1	-	1
Chemistry Content	Ambiguous	-	-	-	0
	Element and Number Relation	3	3	27	33
	Element Comparison	-	4	2	6

	Periodic Table	-	-	7	7
	Math Knowledge	-	-	6	6
	Observations	-	-	1	1
	Total	3	7	43	

Celine's artifacts addressed four chemistry topics within a lesson about isotopes and atomic mass, with an additional code for items about recording observations. Celine has multiple learning activities because she made in-class decisions about student progress, deciding that her students needed more in-class instruction before taking the assessment. Example items from Celine's artifacts are shown in Figure 4. The tasks articulated in Celine's items were nearly all retrieval (48 total) with far fewer explanation tasks (4 total) and only one representation task. Similarly, nearly all items in Celine's artifacts incorporated the symbolic representational level (50 out of 53 total items). Celine's items were disproportionately retrieval tasks at the symbolic representational level, meaning that the self-imposed requirements of addressing a variety of conceptual and representational levels was not met.

Table 6 communicates the total number of items that include each task, representational level, and chemistry content topic for Emmerson's artifacts. The learning goals for Emmerson's lesson were that the learner will be able to:

1. classify reactions as a synthesis, decomposition, double displacement, single displacement, or combustion reaction.
2. interpret symbolic representations of equations to make predictions of observable behaviors to link the macroscopic and symbolic levels of understanding according to Johnstone's triangle of chemical levels of thinking.
3. analyze and develop sub-microscopic representations of reactions.

Emmerson generated one assessment and one learning activity to address these learning goals. Students in Emmerson's classes were provided two days to complete the learning activity, taking the assessment electronically the day after the learning activity.

Table 6. Number of Each Code Occurrence for Emmerson's artifacts

Category	Code	Learning Goals	Assessment	Learning Activity	Total
Task	Retrieval	1	10	43	54
	Explanation	-	-	29	29
	Representation	1	-	-	1
	Analysis	2	-	5	7
	Knowledge Utilization	-	-	6	6
Representational Level	Symbolic	-	3	6	9
	Symbolic/Particle	-	2	-	2
	Particle	2	3	5	10
	Particle/Macroscopic	-	-	1	1
	Macroscopic	-	-	54	54
	Macroscopic/Symbolic	1	2	12	15
Chemistry Content	Ambiguous	1	-	5	6
	Reaction Type	1	6	17	24
	Reaction Representation	3	4	-	7
	Lab Knowledge	-	-	1	1
Total	Observations	-	-	65	65
Total		4	10	83	

Emmerson's artifacts included two chemistry content codes for a lesson about types of chemical reactions, with additional codes to represent items that asked students to make observations or employ their laboratory knowledge. Example items from Emmerson's artifacts are shown in Figure 6. The tasks articulated in Emmerson's items were largely retrieval (54 total) with at least 1 assessment or learning activity item incorporating explanation, analysis, and knowledge utilization tasks. Although the learning activity incorporated a variety of tasks, the assessment included only retrieval tasks. The emphasis of retrieval tasks is likely a result of the number of observations in the learning activity, which also likely contributes to the emphasis on the macroscopic representational level (54 items). Even though the macroscopic representational level was disproportionately emphasized, the learning activity included at least one item from all other possible representational levels (excluding items including both symbolic and particle levels) and the assessment spanned a variety of representational levels. Emmerson's items met the self-imposed requirements of articulating a task for each item and incorporating a variety of conceptual and representational levels, although retrieval items at the macroscopic level were emphasized.

Although investigating individual code occurrences provided insight about overall trends in task and representational level use, the complex nature of the items is not illustrated. For example, an item coded as a retrieval task may include a different representational level than other retrieval tasks within the same chemistry content. Highlighting the co-occurrence of the tasks and representational levels of items within each chemistry content code allows for a more thorough characterization of alignment to instructional materials which is described in the following section.

#### Construction of the Alignment Plots

A two-dimensional diagram (called an alignment plot) was generated to allow for a synchronous characterization of the alignment of tasks and representational levels of items within teacher items from her/his artifacts. To compare individual items among assessment and instructional artifacts, every action within each learning goal, learning activity item, and assessment item was assigned a shape based on its task code (Table 7) and a color based on its representation level (Figure 2).

Table 7. Task Code Key for Alignment Plot

Task Code	Shape on Alignment Plot
Retrieval	Circle
Explanation	Triangle
Representation	Square
Analysis	Pentagon
Knowledge Utilization	Star

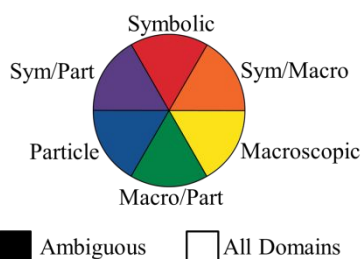





Figure 2. Representational Level Codes

Example items from Claude's artifacts are shown in Figure 3. The items in Figure 3 align with the "Static Chemical System" chemistry content code, because they ask the student about a system that is assumed to be static (i.e., unmoving). The learning goal's task is depicted by a circle to show that students are asked to *identify*. The learning goal sets the requirement that, to demonstrate competency, students need to recognize the correct answer when critical information is provided. The circle is colored red to show the symbolic representational level communicated in the learning goal. The assessment item's task is depicted by a triangle to show that the item asks the student to explain. When responding to this assessment item, students are provided with the information that a portion of a magnesium strip is placed in a solution of chromium (III) iodide. From this information, the student would need to discern critical information from noncritical information to demonstrate that they understand where the reaction occurred and what observations would be evident. Since the item asks for a macroscopic description, the item's shape is colored yellow. The task of the learning activity item in Figure 3 is depicted as a square because the student is asked to draw. To demonstrate competency for this item, the student would need to generate a representation. The color of the shape is blue, since a particle-level representation is to be generated by the study. Example items for each task and representational level can be found in Appendix A and B.

Artifact	Item	Codes	Shape
Learning Goal	Identify a redox reaction based on symbolic representations.	Content: <i>Static Chemical System</i>  Task: <i>Retrieval</i>  Representational Level: <i>Symbolic</i>	
Assessment	The student group removes the metal strip from the solution after the reaction has completed. Describe what the students would see at EACH END of the strip.  <div style="display: flex; justify-content: space-around; align-items: center;"> <div style="border: 1px solid black; padding: 5px; width: 150px; text-align: center;">This end was in solution.</div> <div style="border: 1px solid black; padding: 5px; width: 150px; text-align: center;">This end was not in solution.</div> </div> 	Content: <i>Static Chemical System</i>  Task: <i>Explanation</i>  Representational Level: <i>Macroscopic</i>	


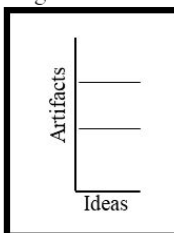
Learning Activity	Draw a particulate level representation of zinc metal.	Content: <i>Static Chemical System</i>  Task: <i>Representation</i>  Representational Level: <i>Particulate</i>	
-------------------	--	--	---

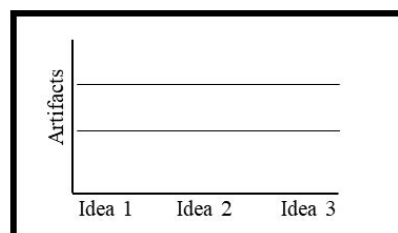
Figure 3: Example items, codes, and alignment plot identifiers from Claude's artifacts

Inspection for matching shapes and colors within content categories of the alignment plots (vertically) allows for a more thorough investigation of alignment than comparison of tasks and representational levels separately. A guide for reading the alignment plot is provided in Figure 4.

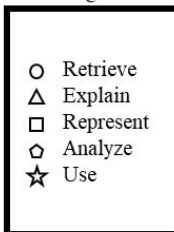
## Alignment Plot Structure



The alignment plot is organized using two dimensions. The vertical dimension describes the classroom artifacts. The horizontal dimension describes the chemistry ideas.

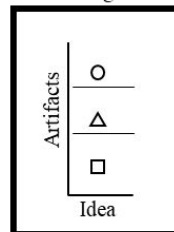


## Task Alignment: Shape

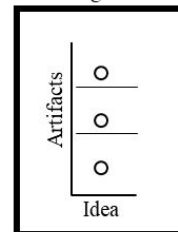


Shapes describe the task to be performed by the student. Task alignment is evaluated by inspecting for similar shapes across all artifacts

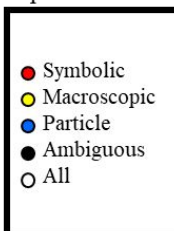
## Misaligned



## Aligned

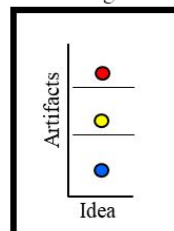


## Representational Level Alignment: Color

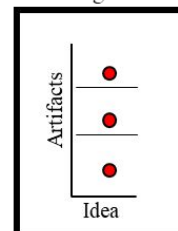


Each shape is colored to communicate the representational level requested in the student response. Representational level alignment is evaluated by inspecting for similar colors across all artifacts

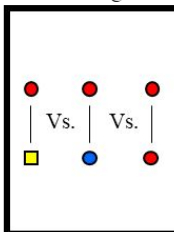
## Misaligned



## Aligned

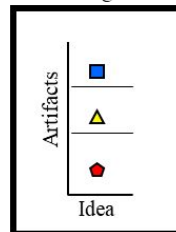


## Overall Alignment: Shape and Color

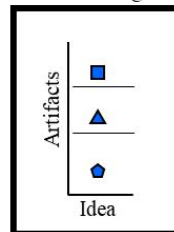


At minimum, teachers should observe a one-to-one match for each shape and color across all artifacts within a single chemistry idea. Multiple items can align to a single item from another artifact.

## Misaligned



## Task-Misaligned



## Synchronously Aligned

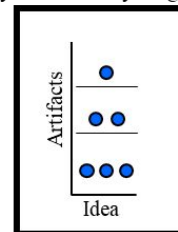


Figure 4. Guide for reading and interpreting the alignment plots.



## Alignment Plots

**Claude's Alignment Plot.** Claude's alignment plot is shown in Figure 5. The alignment plot provides a succinct and comprehensive diagram of the nature of what students were asked to do in the learning goals, learning activity items, and assessment items within a lesson about redox.

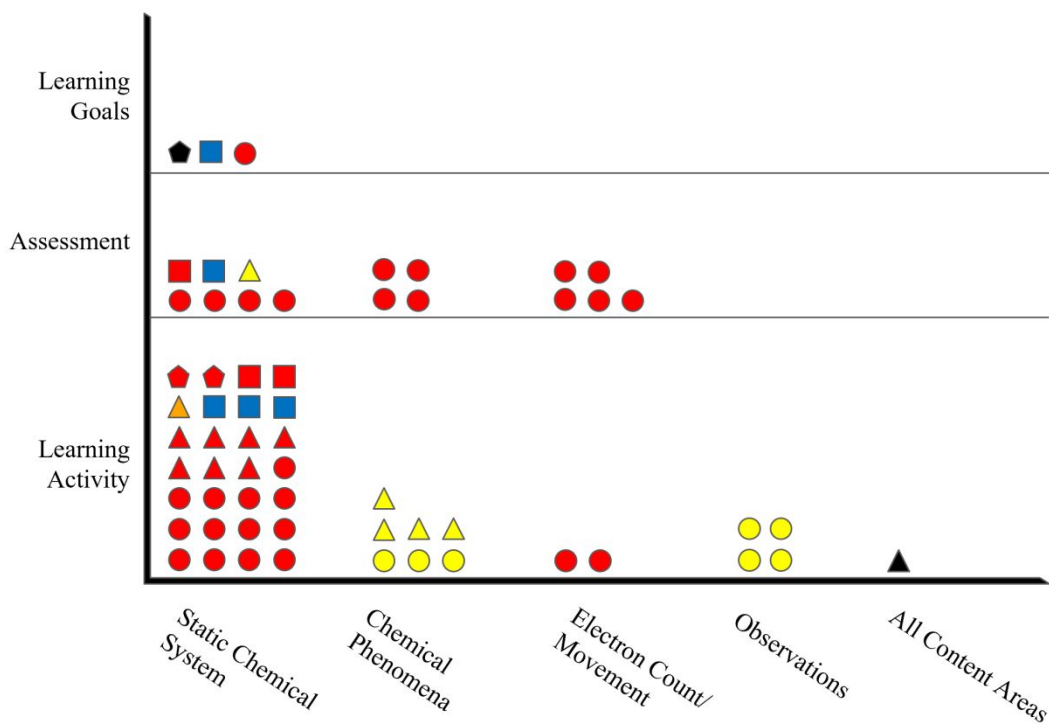


Figure 5: Claude's Alignment Plot

Inspecting only the shapes within Claude's alignment plot (Figure 5) reveals that the artifacts exhibit instances of both alignment and misalignment regarding assessment tasks. The "Static Chemical System" content in Figure 5 shows that retrieval (circles) and representation (squares) tasks are present in all artifacts, demonstrating alignment of tasks. However, analysis (pentagons) and explanation (triangles) tasks are not present in all artifacts, indicating misalignment of tasks. The other chemistry content areas shown in Figure 5 do not contain learning goals and thus cannot have alignment among assessment and instructional materials.

A similar evaluation can be performed for the representational level alignment by comparing colors present in Claude's artifacts. Figure 5 shows two learning goals with specified representational levels, one particle-level (blue color) and the other symbolic (red color), within the "Static Chemical System" content. The items are aligned with items in the assessment and learning activity, as indicated by the presence of red and blue shapes in both the assessment and learning activity for that content. Alternatively, the lone macroscopic-level (yellow) item in Claude's assessment for the "Static Chemical System"

1  
2  
3 content is not matched with any learning goals or items in the learning activity for that content, indicating misalignment of  
4 representational level. Again, the lack of learning goals for the other content areas means that there is not an opportunity for  
5 alignment among the assessment and instructional materials.  
6  
7 360

8  
9 Although evaluating task and representational level separately reveals valuable insights about Claude's assessment  
10 design practices, the alignment plot allows for these criteria to be evaluated synchronously. For example, one of Claude's  
11 learning goals is a symbolic/retrieval item (red/circle). This learning goal is exactly matched with four assessment items and  
12 13 learning activity items within the same content. Similarly, another of Claude's learning goals is a particle-  
13 level/representation item (blue/square) that is exactly matched with one assessment item and three learning activity items.  
14  
15 These exact matches represent the highest degree of alignment within Claude's artifacts.  
16  
17 365

18  
19  
20 **Celine's Alignment Plot.** Example items from Celine's artifacts are shown in Figure 6 and Celine's alignment plot is  
21 provided in Figure 7. Celine included three learning activities as part of her lesson, which are numbered chronologically for  
22 the alignment plot (i.e., Learning Activity 1, Learning Activity 2, Learning Activity 3).  
23  
24  
25

Artifact	Item <sup>a</sup>	Codes	Shape
Learning Goal	Students will be able to successfully employ ratios and proportions to obtain relative mass for particles of imaginary elements.	Content: <i>Element and Number Math Relation</i>  Task: <i>Retrieval</i>  Representational Level: <i>Symbolic</i>	●
Assessment	One particle of tinium actually weighs 0.12g. Use this mass and your discovered relative masses to calculate the mass of each individual new atom.	Content: <i>Element and Number Math Relation</i>  Task: <i>Retrieval</i>  Representational Level: <i>Symbolic</i>	●
Learning Activity 3	Jessie has 89g of Rd. How many moles does she have?	Content: <i>Element and Number Math Relation</i>  Task: <i>Retrieval</i>  Representational Level: <i>Symbolic</i>	●

<sup>a</sup>Celine's artifacts included hypothetical elements (such as tinium).

Figure 6: Example items, codes, and alignment plot identifiers from Celine's artifacts

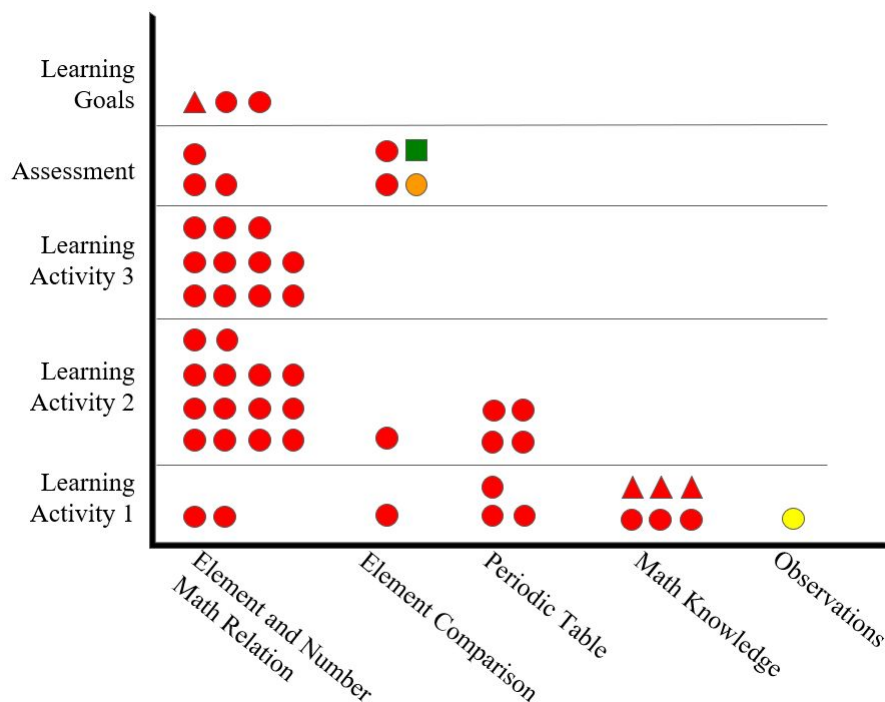


Figure 7: Celine's Alignment Plot

Beginning with the tasks (shapes) of Celine's artifacts in Figure 7, nearly all of the items require a retrieval (circle) task. Although retrieval tasks (circles) are present throughout all content categories, alignment is only observed for the "Element and Number Math Relation" content, since this is the only content with learning goals. Additionally, the lone explanation (triangle) learning goal is misaligned with items in the assessment and learning activity since there are no corresponding shapes within the same chemistry content. Evaluating the representational level (color) of the artifacts in Figure 7 reveals a similar trend. Nearly all the items are at the symbolic level (red color). However, since the learning goals are only within the "Element and Number Math Relation" content, only these are considered aligned.

Again, the alignment plot's true value stems from the ability to compare task (shape) and representational level (color) synchronously. Celine's artifacts show complete alignment within the "Element and Number Math Relation" content for items that involve symbolic/retrieval (red/circles) tasks as these items are present throughout all artifacts for this content.

**Emmerson's Alignment Plot.** Example items from Emmerson's artifacts are shown in Figure 8, and his alignment plot is provided in Figure 9.


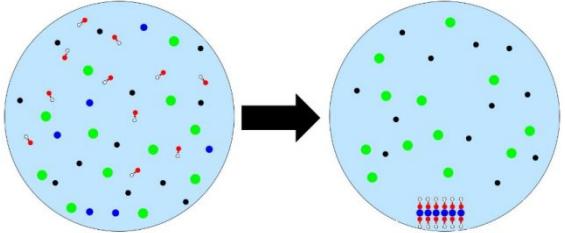


Artifact	Item	Codes	Shape
Learning Goal	Learner will be able to classify reactions as synthesis, decomposition, double displacement, single displacement, or combustion reaction.	Content: <i>Reaction Type</i>  Task: <i>Retrieval</i>  Representational Level: <i>Ambiguous</i>	
Assessment	Classify each reactions a. through e. as synthesis (S), decomposition (D), single replacement (SR), double displacement (DD), or combustion (C). 	Content: <i>Reaction Type</i>  Task: <i>Retrieval</i>  Representational Level: <i>Particulate</i>	
Learning Activity	$Zn(s) + I_2(s) \rightarrow ZnI_2(aq)$ $Mg(s) + O_2(g) \rightarrow 2MgO(s)$ $CaO(s) + CO_2(g) \rightarrow CaCO_3(s)$  Each of these reactions are classified as synthesis reactions, also sometimes called "addition" reactions. Explain why this name fits these reactions.	Content: <i>Reaction Type</i>  Task: <i>Explanation</i>  Representational Level: <i>Symbolic</i>	

Figure 8: Example items, codes, and alignment plot identifiers from Emmerson's artifacts

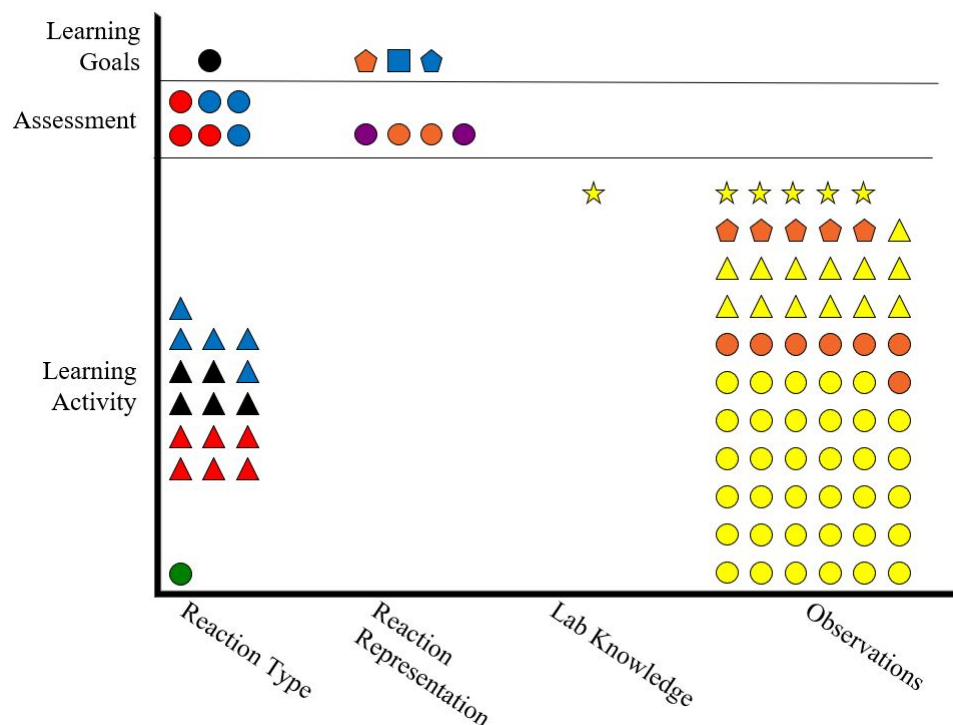


Figure 9: Emmerson's Alignment Plot

Figure 9 shows the tasks (shapes) in Emmerson's artifacts only exhibit alignment within the chemistry content "Reaction Type" for retrieval (circles) tasks, since all artifacts in that content include a circle. Emmerson's learning goals also incorporate representation (squares) and analysis (pentagons) tasks within the "Reaction Representation" content; however, the assessment and learning activity do not include corresponding shapes, indicating misalignment of tasks. Although Emmerson's artifacts incorporate a variety of representational levels (colors), the colors are not consistent throughout the artifacts, leading to no instances of representational level alignment within any chemistry content categories (i.e., one color is not present in each the learning goals, assessment, and learning activity for any particular content category). Without any instances of alignment in regard to representational level, Emmerson's artifacts do not exhibit any instances of matching task and representational level (shape and color) for the evaluated lesson.

The alignment plots allow for a synchronous evaluation of the items across instructional and assessment artifacts; however, it is unrealistic for a teacher to generate an alignment plot when designing every planned, formative assessment. Nonetheless, these alignment plots may be useful for teachers who wish to evaluate the quality of select planned, formative assessments.

---

### Relationship Between “Best Practices” and Assessment Design

The second research question guided the investigation of the relationship between the “best practices” for formative assessments generated by these teachers and the actual assessments they generated. The alignment plots illustrated teachers’ enacted design practices for their planned, formative assessments. Comparing these enacted practices to the teachers’ “best practices” allows for a characterization of the relationship between the tools teachers use to collect classroom evidence and their beliefs about how planned, formative assessments *should* be designed. In a research context, the alignment plots configure the codes from teachers’ artifacts into a format that allows a comparison among the enacted practices, “best practices,” and the literature. This comparison reveals the barriers that teachers encounter when designing planned, formative assessments. As such, the rest of the manuscript leverages assessment design practices illustrated in the alignment plots to uncover the relationship between teachers’ enacted practices and perceived “best practices”.

The first teacher-generated “best practice” states that an assessment item should clearly articulate a task for the student to perform. Other investigations have found teachers often consider the specific action to be performed by students when designing and selecting assessment items (Tomanek *et al.*, 2008; Kang *et al.*, 2016; Schafer and Yeziarski, 2020b). Teachers have a significant body of literature to draw from when considering how to design productive tasks for their learning environment (e.g., Harris *et al.*, 2016). A task was able to be assigned for each assessment item within the assessments generated by this group of teachers. Additionally, each learning goal and learning activity item was also able to have a specific task assigned. As such, few barriers seem to exist preventing teachers from enacting the “best practice” of stating a specific task for students to perform.

The second listed “best practice” for formative assessments by these teachers was to assess a variety of conceptual and representational levels. Assessing content in a variety of ways can help the teacher triangulate student competency (Sadler, 1989; National Research Council, 2001; Means *et al.*, 2011). Of the teachers in this study, Claude’s assessment included the greatest variety of tasks, which were considered synonymous with “conceptual levels” for the purpose of this study. Claude included retrieval, explanation, and representation tasks in his assessments, as shown in Figure 5 and Table 4. Celine included retrieval and representation tasks in her assessment, as shown in Figure 7 and Table 5. Emmerson included only retrieval tasks in his assessment, as shown in Figure 9 and Table 6. It is important to recognize that any particular task is not necessarily more conceptually difficult or challenging than other by default (National Research Council, 1999, 2014; Wauters *et al.*, 2012). For example, students may have opportunities to consistently practice generating representations for a chemical phenomenon without being able to reason about the atomic-level events that explain why the phenomenon occurs. Additionally, the assessments provided were formative, meaning that each serves the purpose to inform continued instruction.

As such, the assessments may be situated near the beginning of a unit of instruction, limiting the opportunity to include several conceptually challenging items. Either way, the greater variety of tasks on Claude's assessment more likely addresses a range of conceptual difficulties and potentially better equips him to judge the upper limit of his students' understanding.

Also included in the second "best practice" is the suggestion to include a variety of representational levels. Several investigations document student struggles navigating among representational levels (Gabel *et al.*, 1987; Nakhleh, 1992; Russell *et al.*, 1997; Gkitzia *et al.*, 2020). Thus, when triangulating student understanding, teachers could benefit from assessing competency using a variety of ways of perceiving chemical information (National Research Council, 1999).

However, literature guidance suggests that carefully scaffolding items assessing representational level and that limiting the number of representational levels per task may be beneficial (Taber, 2013). Table 6 and Figure 9 reveal that Emmerson addressed the greatest variety of representational levels by including four separate levels in his assessment. However, Claude and Celine both included three separate representational levels throughout their assessment items. All teachers seemed to meet the self-generated "best practice" of addressing a variety of representational levels in their assessments. Few barriers seem to exist preventing teachers from enacting the "best practice" of incorporating a variety of representational and conceptual levels. However, incorporating a variety of representational levels requires teachers to consider their students' prior knowledge and experience, making enactment a more complex process compared to "clearly state a task." Such considerations are commensurate with the process of data-driven inquiry, as employing data-driven methods must consider the needs their students and learning environment when designing tools for gathering evidence that are aligned with their goals (Harshman and Yeziarski, 2017).

The final relationship evaluated to address the second research question was that of the "best practice" of aligning assessment and instructional materials. Assessment and instruction do not exist in isolation and are intertwined throughout the learning process, especially formative assessments. Assessing tasks and concepts not addressed during instruction can unjustly increase the cognitive demand of an item (National Research Council, 2014; Kang *et al.*, 2016). While assessments in this study exhibited instances of alignment and misalignment regarding task or representational level, none were completely aligned or misaligned regarding task or representational level. The results from evaluating alignment of each teachers' assessment are shown in Table 8.

Table 8. Results of Assessment Alignment Evaluations

Teacher	Tasks Aligned	Representational Levels Aligned	Tasks and Representational Levels Synchronously Aligned (Shown as Representational Level/Task)
Claude	Retrieval, Representation	Symbolic, Particle	Symbolic/Retrieval, Particle/Representation
Celine	Retrieval	Symbolic	Symbolic/Retrieval

---

Emmerson Retrieval

For this investigation, an assessment was considered “aligned” if one instance of a specified criterion was present in each of the included artifacts. This one-to-one definition of alignment is commonly employed by methods used to evaluate alignment for state-level artifacts (Martone and Sireci, 2009), but may not be suitable for evaluating formative assessments. Indeed, several works recognize that a teacher may need to make between six and eight observations of student behavior to reliably judge student competency (Webb, 2006; Martone and Sireci, 2009; Praetorius *et al.*, 2014; Briggs and Alzen, 2019). The recommended six to eight observations of student observations come from investigations of assessment and instructional tasks, but not both. So, teachers may need to exercise caution when considering the number of assessment items necessary to reliably judge the number of items employed to judge student ability. Teachers in this study generally included several assessment items to evaluate a single learning goal and even more learning activity items to address the assessment items. For example, Figure 5 shows that Claude’s artifacts included one learning goal that was symbolic/retrieval and matched to four symbolic/retrieval assessment items and 13 symbolic/retrieval learning activity items. There are currently no literature-based guidelines on the ratio of learning goals-to-assessment items-to-learning activity items recommended for a reliable judgement of student learning, and a specific ratio is unlikely to be generalizable across all learning environments. However, the previously mentioned guidelines about the number of student observations suggests that a structure similar to Claude’s is more favorable than a 1-learning goal to 1-assessment item to 1-learning activity item ratio of co-occurrence.

Missing from the chemistry education literature is how representational level factors into evaluations of alignment. Results of synchronously evaluating task and representational level are included in Table 8. While Claude’s assessment was aligned along the synchronous criteria of both symbolic-level/retrieval tasks and particle-level/representation tasks, Celine’s assessment showed synchronous alignment regarding only symbolic-level/retrieval tasks and Emmerson’s assessment showed no synchronous alignment of representational level and task. Considering students’ documented struggles navigating among representational levels (Gabel *et al.*, 1987; Nakhleh, 1992; Russell *et al.*, 1997; Gkitzia *et al.*, 2020), developing a method for evaluating the alignment of chemistry formative assessments that incorporates task and representational level synchronously may be beneficial for research, teacher education, and teacher professional development. As such, significant barriers exist for chemistry teachers when evaluating alignment between assessment and instruction in terms of both 1) an appropriate ratio of learning goals-to-assessment items-to-learning activity items that can support reliable inferences from evidence; and 2) how to incorporate alignment criteria (such as task and representational level).



---

## CONCLUSIONS AND RESEARCH IMPLICATIONS

The first research question asks how tasks and representational levels of planned, formative assessment items can be diagrammed to allow for a synchronous evaluation of alignment between assessment and instruction. To address this question multiple representations were considered, ultimately resulting in the generation of an alignment plot for one lesson submitted by three high school chemistry teachers (three total lessons). The alignment plot is able to reveal the formative assessment design practices of the participating teachers by synchronously visualizing the assessment features of each assessment. In this way, the alignment plots may serve as a valuable tool to teachers seeking to evaluate the alignment between their assessment and instruction. Although the alignment plots are still extensive, teachers may find it less challenging to employ than other alignment evaluation tools, such as the EQuIP (Achieve, 2016; Fulmer *et al.*, 2018). The ability of the alignment plot to reveal instances of alignment and misalignment within teacher assessments shows that the diagram may serve as a tool teachers may use to evaluate the alignment between their assessments and instructional materials.

Findings from the second research question revealed relationships between high school chemistry teachers' stated "best practices" for formative assessment design and the assessments they generated. The "best practices" developed by these teachers state that a formative assessment should articulate a specific task to be performed by the student, address a variety of conceptual and representational levels, and align assessment items to instructional materials. Results showed that teachers met the practices of articulating a task and including a variety of representational levels. However, only Claude's assessment incorporated a variety of conceptual levels (i.e., tasks). Teachers' ability to consistently meet their "best practices" of articulating a task implies that there are few barriers to enacting this goal. Direct literature guidelines are available for designing learning activity and assessment tasks for science classrooms (Harris *et al.*, 2016; Laverty *et al.*, 2016; Penuel *et al.*, 2019). Although teachers met the "best practice" of incorporating a variety of representational levels, the emphasis on symbolic items could indicate that barriers still hinder enactment of this goal. Overall, assessment items generally required retrieval tasks at symbolic representational levels, potentially limiting the amount of information about student competency available to the teacher for effectively interpreting student understanding (Stiggins, 2001; Towns, 2014b; Schafer and Yeziarski, 2020a).

The third "best practice" generated by teachers required evaluating the extent that assessments aligned to instruction. For this study, a one-to-one ratio between items in instructional artifacts and assessment artifacts was used. While teachers were generally able to align some assessment tasks to instructional materials, representational levels of assessment items were frequently misaligned with instruction or only one representational level was employed for nearly all items. Synchronous alignment of tasks and representational levels to instructional materials was infrequently observed; however, Claude's

1  
2  
3 assessment included several items that synchronously aligned task and representation level to instructional materials (while  
4  
5 510 still employing a variety of tasks and representational levels). The instances of alignment revealed by the alignment plots  
6  
7 indicate that even experienced teachers with several years attending professional development can still encounter significant  
8  
9 barriers to aligning assessments to instructional materials. Although there is some literature-based guidance available to  
10  
11 teachers for aligning assessment and instruction, most studies result in tools for teachers to use for evaluating alignment (e.g.,  
12  
13 Webb, 2007; Kaderavek *et al.*, 2015; Achieve, 2016) which can be challenging for teachers to employ, given variations in  
14  
15 515 learning environments and teacher goals. Additionally, literature about evaluating alignment has mixed guidance on what  
16  
17 qualifies as “aligned” (Fulmer *et al.*, 2018), variations in the criteria to use when considering alignment (Martone and Sireci,  
18  
19 2009; Fulmer *et al.*, 2018), and disagreements in how many aligned criteria are necessary to assume a suitably reliable  
20  
21 interpretation of student knowledge (Webb, 2006; Martone and Sireci, 2009; Praetorius *et al.*, 2014; Briggs and Alzen, 2019).  
22  
23 The existing barriers these teachers face for enacting their “best practice” of aligning assessment to instruction implies that  
24  
25 520 teachers need more than tools for evaluating alignment. Chemistry teachers need guidance understanding the methodological  
26  
27 and conceptual underpinnings of the available tools, interpreting the results from the tools employed, and appropriately  
28  
29 adjusting classroom materials to better evaluate the success of the learning environment.

### 30 31 **LIMITATIONS AND FUTURE WORK**

32 Although the findings presented are useful for better understanding high school chemistry teacher design practices, there  
33  
34 525 are several limitations that bound the claims presented. To begin, the study described herein closely examines the physical  
35  
36 artifacts from three individual lessons and extracted from teachers’ day-to-day practices. The focus on physical artifacts is not  
37  
38 meant to imply that learning is paused during assessment. and we recognize that chemistry teachers evaluating the alignment  
39  
40 of their assessments will likely have access to more data, such as additional assessments, learning activities, classroom  
41  
42 interactions, and knowledge of students. However, investigating a defined set of artifacts allowed for a closer inspection of  
43  
44 530 the proposed measures, providing guidance for chemistry teachers as they interpret the alignment of their assessments. When  
45  
46 interpreting alignment in this work, the reader is cautioned against perceiving misaligned items as “bad” items, and  
47  
48 interpretation of the alignment plot should consider the overall structure of the instrument. For example, Claude included  
49  
50 items with retrieval tasks to scaffold or support complementary tasks, such as developing a representation. However, well-  
51  
52 designed, planned, formative assessments should include items that evaluate the stated goals. Future studies could expand  
53  
54 535 upon this investigation to include more teachers and more data sources to better understand how teachers incorporate these  
55  
56 many data sources into their evaluation of assessment quality.

Certain limitations exist within the measures, as well. For example, Marzano's and Kendall's original framework sets knowledge categories as a hierarchy of knowledge levels (Marzano and Kendall, 2008). To exemplify the complexities of item difficulty, this study does not employ the hierarchical use of knowledge tasks. Future studies could investigate the extent to which individual knowledge categories are representative of item difficulty and depth of understanding.

A limitation of the analysis described herein is that each learning goal, learning activity, and assessment item were investigated according to what is asked of the student in the artifact, not the student's actual response to the item. This purposeful bounding of the investigation around the artifacts does not include response process validation from students. Additionally, the items generated by these teachers may not be deemed "high-quality" items by chemistry content experts. The items generated by the teachers were part of an ongoing professional development and reflect what was implemented in chemistry teacher classrooms during early sessions. During later professional development sessions, teachers collaboratively interpreted student responses to the assessment items presented and posited potential changes to planned, formative assessment design practices. Future studies may investigate the tasks that are asked of the student versus the task the student performs.

### **IMPLICATIONS FOR TEACHERS**

Although this work holds many implications for research, it may also serve to benefit teachers looking to improve their classroom practices. For example, chemistry teachers should consider the representations used to teach and assess knowledge in their classrooms. Overemphasis on a single representational level or misalignment between representations on assessment and instructional materials can impact the quality of interpretations of student knowledge. Teachers may also consider how tasks and representations are aligned between assessment and instruction. It is likely the one-to-one ratio between assessment and instructional items employed in this study is not sufficient and that teachers may benefit from using a ratio more like what is found in Claude's artifacts (1 learning goal : 4 assessment items : 13 learning activity items). The exact numbers are likely less important than having one learning goal, assessed multiple times, with even more student learning opportunities. The alignment plot can be employed in chemistry teacher education (focused on formative assessment practices) and chemistry teacher professional development as a tool to examine alignment in the discipline-specific manner.

The "best practices" examined in this work were generated by high school chemistry teachers and align to high quality practices stated in relevant literature. The connection between these chemistry teachers' goals and literature-backed practices implies that high school chemistry teachers may benefit from incorporating the stated "best practices" when developing formative assessments. The "best practices" were generated as goals for designing and interpreting formative

assessments by a group of teachers knowingly using the process of data-driven inquiry (Schafer and Yeziarski, 2020a). Chemistry teachers may similarly benefit by using processes like data-driven inquiry to scaffold their progression from designing goals, collecting classroom evidence, and forming an evidence-based instructional response. Additionally, we hope chemistry teachers are inspired to reflect on the alignment of the purpose and design of the assessments employed in their own classrooms and employ formal measures such as the alignment plot in their practice.

## APPENDICES

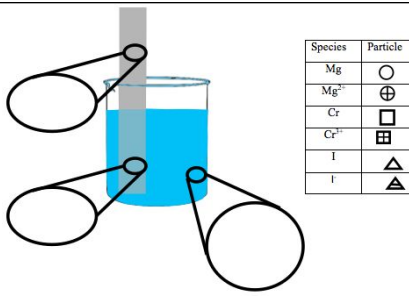
Appendix A – task codebook and example items

Table 9. Task Codebook

Code	Description (Demonstration of competency requires...)	Example Task	Plot Depiction
Retrieval	identification or stating of standalone information without synthesis or analysis.	Identify, recognize, math functions, complete, apply, demonstrate	circle (Shape)
Explanation	communication of critical/essential information from noncritical/nonessential information.	Explain, summarize	triangle (Shape)
Representation	generation of a depiction of a chemical idea, phenomenon, or relationship	Draw, use models, represent, show	square (Shape)
Analysis	processes that involve examining knowledge/content/observations with the intent of generating new conclusions (extending knowledge).	Sort, categorize, differentiate, assess, critique, evaluate, diagnose	pentagon (Shape)
Knowledge Utilization	processes that require the application or use of knowledge in a novel situation.	Test, how would you determine, generate and test	star (Shape)

Table 10. Example Items for Each Task Code

Code	Example	Example Source
Retrieval	Students will be able to successfully employ ratios and proportions to obtain relative mass for particles of imaginary elements.	Celine: Learning Goal
Explanation	$\text{Zn}(s) + \text{I}_2(s) \rightarrow \text{ZnI}_2(aq)$ $\text{Mg}(s) + \text{O}_2(g) \rightarrow 2\text{MgO}(s)$ $\text{CaO}(s) + \text{CO}_2(g) \rightarrow \text{CaCO}_3(s)$ <p>Each of these reactions are classified as synthesis reactions, also sometimes called “addition” reactions. Explain why this name fits these reactions.</p>	Emmerson: Assessment Item
Representation	<p>A student group performs an experiment where a strip of magnesium of placed into a solution of chromium (III) iodide. A reaction occurs according to the following equation.</p> <p>Draw a particulate representation of the products in the space that follows. Use the symbols in the key provided. (Note that the zoom-out in the <i>liquid</i> represents species that are <i>dissolved</i> in the liquid).</p>	Claude: Assessment Item

	 <table border="1" data-bbox="860 220 974 378"> <thead> <tr> <th>Species</th> <th>Particle</th> </tr> </thead> <tbody> <tr> <td>Mg</td> <td>○</td> </tr> <tr> <td>Mg<sup>2+</sup></td> <td>⊕</td> </tr> <tr> <td>Cr</td> <td>□</td> </tr> <tr> <td>Cr<sup>3+</sup></td> <td>⊠</td> </tr> <tr> <td>I<sup>-</sup></td> <td>△</td> </tr> <tr> <td>I<sub>2</sub></td> <td>▲</td> </tr> </tbody> </table>	Species	Particle	Mg	○	Mg <sup>2+</sup>	⊕	Cr	□	Cr <sup>3+</sup>	⊠	I <sup>-</sup>	△	I <sub>2</sub>	▲	
Species	Particle															
Mg	○															
Mg <sup>2+</sup>	⊕															
Cr	□															
Cr <sup>3+</sup>	⊠															
I <sup>-</sup>	△															
I <sub>2</sub>	▲															
Analysis	Predict products of redox reactions	Claude: Learning Goal														
Knowledge Utilization	<p>Calcium chloride (CaCl<sub>2</sub>) and sodium carbonate (Na<sub>2</sub>CO<sub>3</sub>) are ionic compounds that dissolve in water. When dissolved, these substances react as expressed by the following equation:</p> $\text{CaCl}_2(aq) + \text{Na}_2\text{CO}_3(aq) \rightarrow \text{CaCO}_3(s) + 2\text{NaCl}(aq)$ <p>Suggest a technique or a combination of techniques that would allow you to “recover” and observe the other product if you do not currently have evidence for its formation.</p>	Emmerson: Learning Activity														

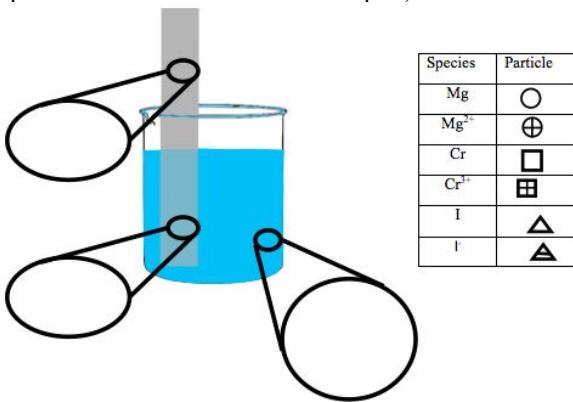
## Appendix B – representational level codebook and example items

Table 11. Representational Level Codebook

Code	Description	Example	Plot Depiction
Symbolic	The use of descriptive words, symbols, or values to communicate chemical ideas/events/species	Chemical Equations	Red (Shape)
Macroscopic	Representation of species/events on a visible scale to communicate chemical ideas/events/species	Observations	Yellow (Shape)
Particulate	Representation of species/events on an invisible scale to communicate chemical ideas/events/species	Atomic-Level Particles in Solution	Blue (Shape)
Symbolic/ Macroscopic		Equation paired with observation	Orange (Shape)
Symbolic/ Particulate	Communication of chemical ideas/events/species includes combinations of individual representational levels	Equation paired with atomic-level particle model	Purple (Shape)
Macroscopic/ Particulate		atomic-level particles paired with observation	Green (Shape)
All	Communication of chemical ideas/events/species occurs at all representational levels	Answer incorporates all levels	White (Shape)

Ambiguous	No representational level communicated	Representational level not specified	Black (Shape)
-----------	--	--------------------------------------	---------------

Table 12. Example Items for Each Representational Level

Code	Example	Example Source														
Symbolic	Identify a redox reaction based on symbolic representations.	Claude: Learning Goals														
Macroscopic	<p>Sodium carbonate (<math>\text{NaHCO}_3</math>) is a familiar household compound commonly referred to as baking soda. When heated, it reacts according to the following equation:</p> $2\text{NaHCO}_3(s) \rightarrow \text{Na}_2\text{CO}_3(s) + \text{H}_2\text{O}(g) + \text{CO}_2(g)$ <p>Summarize your observations below. Be complete when you record these observations. This includes any observations obtained by sight, feel, odor, or sound.</p>	Emmerson: Learning Activity														
Particulate	<p>A student group performs an experiment where a strip of magnesium of placed into a solution of chromium (III) iodide. A reaction occurs according to the following equation.</p> <p>Draw a particulate representation of the products in the space that follows. Use the symbols in the key provided. (Note that the zoom-out in the <i>liquid</i> represents species that are <i>dissolved</i> in the liquid).</p>  <table border="1" data-bbox="857 1098 1015 1312"> <thead> <tr> <th>Species</th> <th>Particle</th> </tr> </thead> <tbody> <tr> <td>Mg</td> <td>○</td> </tr> <tr> <td><math>\text{Mg}^{2+}</math></td> <td>⊕</td> </tr> <tr> <td>Cr</td> <td>□</td> </tr> <tr> <td><math>\text{Cr}^{3+}</math></td> <td>⊗</td> </tr> <tr> <td>I</td> <td>△</td> </tr> <tr> <td><math>\text{I}^-</math></td> <td>▲</td> </tr> </tbody> </table>	Species	Particle	Mg	○	$\text{Mg}^{2+}$	⊕	Cr	□	$\text{Cr}^{3+}$	⊗	I	△	$\text{I}^-$	▲	Claude: Assessment Item
Species	Particle															
Mg	○															
$\text{Mg}^{2+}$	⊕															
Cr	□															
$\text{Cr}^{3+}$	⊗															
I	△															
$\text{I}^-$	▲															
Symbolic/ Macroscopic	Which of the following choices best shows the mass relationship between 1 mole of Blonko (Bk) and 1 mole of Copperium (Cp)?	Celine: Assessment														

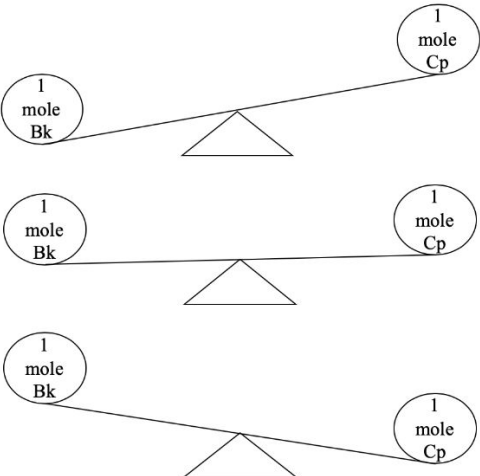
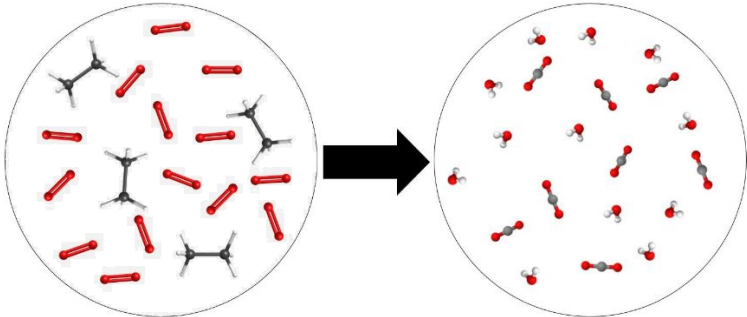
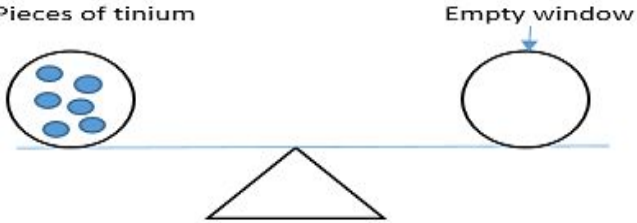
		
Symbolic/ Particulate	<p>Which equation best represents the following particulate representation?</p>  <p>a. <math>2 \text{C}_2\text{H}_6(\text{g}) + 7 \text{O}_2(\text{g}) \rightarrow 6 \text{H}_2\text{O}(\text{g}) + 4 \text{CO}_2(\text{g})</math>  b. <math>2 \text{C}_2\text{H}_6(\text{l}) + 7 \text{O}_2(\text{g}) \rightarrow 6 \text{H}_2\text{O}(\text{l}) + 4 \text{CO}_2(\text{g})</math>  c. <math>\text{C}_2\text{H}_6(\text{g}) + \text{O}_2(\text{g}) \rightarrow \text{H}_2\text{O}(\text{g}) + \text{CO}_2(\text{g})</math></p>	Emmerson: Assessment
Macroscopic/ Particulate	<p>Copperium has a mass approximately 3 times heavier than tinium. In the diagram below the window on the left has 6 particles of tinium (TN). Draw how many pieces of copperium would be in the empty window on the right so that the mass would balance.</p> <p>Pieces of tinium                      Empty window</p> 	Celine: Assessment
All	Did not occur in teacher artifacts.	N/A
Ambiguous	Predict products of redox reactions	Claude: Learning Goals

Table 13. Content Codebook for Claude's Artifacts

Code	Description
Static Chemical System	Content includes information about what atoms/molecules/ions present in a system, their features, and location within the system
Chemical Phenomena	Content includes information about changes that occur between provided species in a chemical system (including hypothetical, or predicted changes)
Electron Count and Movement	Content includes information regarding electrons in a specified system
Observations	Content is dependent upon an in-lab observation

Table 14. Content Codebook for Celine's Artifacts

Code	Description
Mathematical Relation between an Element and a Number	Content involves mathematical relationships and operations between values and their meanings in a chemistry context
Element Comparison	Content requires the consideration of features/ideas/information of multiple elements (includes hypothetical elements)
Periodic Table Information	Content is sourced from general periodic table knowledge
Mathematical Knowledge	Content includes knowledge about mathematical operations and principles
Observations	Content is dependent upon an in-lab observation

Table 15. Content Codebook for Emmerson's Artifacts

Code	Description
Reaction type	Content includes tasks specific to the type of reaction included in the item
Reaction Representation	Item requires student to use the information embedded within a representation
Laboratory Knowledge	Content includes general laboratory knowledge at the high school level
Observations	Content is dependent upon an in-lab observation

### AUTHOR INFORMATION

Corresponding Author

\*E-mail: yeziere@miamioh.edu

### CONFLICTS OF INTEREST

There are no conflicts to declare.

### ACKNOWLEDGMENTS

We thank the high school chemistry teachers for participating in this project. We also thank the Yezierski and Bretz research groups at Miami University for their feedback and guidance. This material is based upon work supported by the U.S. National Science Foundation under Grant No. DRL-1118749.



---

**REFERENCES**

1. Achieve, (2016), EQuIP rubric for lessons and units: Science. *NGSS*.
2. American Chemical Society, (2012), ACS Guidelines and Recommendations for the Teaching of High School Chemistry, Washington, DC: American Chemical Society.
3. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, (2014), Standards for Educational And Psychological Testing, Washington, DC: American Educational Research Association.
4. Bell B. and Cowie B., (2001), The Characteristics of Formative Assessment in Science Education. *Sci. Educ.*, **85**(5), 536–553.
5. Black P. and Wiliam D., (1998), Inside the Black Box: Raising Standards Through Classroom Assessment. *Phi Delta Kappan*, **80**(2), 139–148.
6. Bloom B. S., Engelhart M. D., Furst E. J., Hill, Walker H., and Krathwohl D. R. eds., (1956), Taxonomy of Educational Objectives: The Classification of Educational Goals, New York, NY: David McKay Company, INC.
7. Briggs D. C. and Alzen J. L., (2019), Making Inferences About Teacher Observation Scores Over Time. *Educ. Psychol. Meas.*, **79**(4), 636–664.
8. Broman K., Bernholt S., and Parchmann I., (2015), Analysing task design and students' responses to context-based problems through different analytical frameworks. *Res. Sci. Technol. Educ.*, **33**(2), 143–161.
9. Cizek G. J., (2009), Reliability and validity of information about student achievement: Comparing large-scale and classroom testing contexts. *Theory Pract.*, **48**(1), 63–71.
10. Clinchot M., Ngai C., Huie R., Talanquer V., Banks G., Weinrich M., et al., (2017), Better Formative Assessment: Making formative assessment more responsive to student needs. *Sci. Teach.*, **84**(3), 69–75.
11. Coffey J. E., Hammer D., Levin D. M., and Grant T., (2011), The missing disciplinary substance of formative assessment. *J. Res. Sci. Teach.*, **48**(10), 1109–1136.
12. Curry M. W., (2008), Critical Friends Groups: The Possibilities and Limitations Embedded in Teacher Professional Communities Aimed at Instructional Improvement and School Reform. *Teach. Coll. Rec.*, **110**(4), 733–774.
13. Datnow A., Park V., and Wohlstetter P., (2007), Achieving with Data: How high-performing school systems use data to improve instruction for elementary students, Los Angeles, CA: Center on Educational Governance.
14. DeLuca C., Valiquette A., Coombs A., LaPointe-McEwan D., and Luhanga U., (2018), Teachers' Approaches to Classroom Assessment: A Large-Scale Survey. *Assess. Educ. Princ. Policy Pract.*, **25**(4), 355–375.

- 
- 1  
2  
3 15. Dini V., Sevia H., Caushi K., and Orduña Picón R., (2020), Characterizing the formative assessment enactment of  
4 experienced science teachers. *Sci. Educ.*, **104**(2), 290–325.  
5  
6  
7 16. Dwyer C. A., (2007), Assessment and Classroom Learning: theory and practice. *Assess. Educ. Princ. Policy Pract.*, **5**(1),  
8 131–137.  
9  
10  
11 630 17. Fulmer G. W., Tanas J., and Weiss K. A., (2018), The challenges of alignment for the Next Generation Science  
12 Standards. *J. Res. Sci. Teach.*, **55**(7), 1076–1100.  
13  
14  
15 18. Gabel D. L., Samuel K. V., and Hunn D., (1987), Understanding the particulate nature of matter. *J. Chem. Educ.*, **64**(8),  
16 695.  
17  
18  
19 19. Gearhart M., Nagashima S., Pfothner J., Clark S., Schwab C., Vendlinski T., et al., (2006), Developing Expertise With  
20 Classroom Assessment in K-12 Science: Learning to Interpret Student Work Interim Findings From a 2-Year Study,  
21 635 Los Angeles, CA: Center for the Assessment and Evaluation of Student Learning (CAESL).  
22  
23  
24 20. Gibbs G. and Simpson C., (2004), Conditions Under Which Assessment Supports Students' Learning. *Learn. Teach.*  
25 *High. Educ.*, **1**(1), 3–31.  
26  
27  
28 21. Gkitzia V., Salta K., and Tzougraki C., (2020), Students' competence in translating between different types of chemical  
29 representations. *Chem. Educ. Res. Pract.*, **21**(1), 307–330.  
30 640  
31  
32 22. Hamilton L., Halverson R., Jackson S., Mandinach E., Supovitz J. A., Wayman J. C., et al., (2009), Using Student  
33 Achievement Data to Support Instructional Decision Making, Washington, DC: National Center for Education  
34 Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.  
35  
36  
37 23. Harris C. J., Krajcik J. S., Pellegrino J. W., Mcelhaney K. W., DeBarger A. H., Dahsah C., et al., (2016), Constructing  
38 Assessment Tasks that Blend Disciplinary Core Ideas, Crosscutting Concepts, and Science Practices for Classroom  
39 645 Formative Applications Center for Technology in Learning, Menlo Park, CA: SRI International.  
40  
41  
42  
43 24. Harshman J. and Yeziarski E., (2017), Assessment Data-driven Inquiry: A Review of How to Use Assessment Results to  
44 Inform Chemistry Teaching. *Sci. Educ.*, **25**(2), 97–107.  
45  
46  
47 25. Hoffman C. K. and Medsker K. L., (1983), Instructional analysis: The missing link between task analysis and objectives.  
48 *J. Instr. Dev.*, **6**(4).  
49 650  
50  
51 26. Irons A., (2008), Enhancing Learning Through Formative Assessment and Feedback, New York, NY: Routledge.  
52  
53 27. Johnstone A. H., (1991), Why is science difficult to learn? Things are seldom what they seem. *J. Comput. Assist. Learn.*,  
54 **7**(2), 75–83.  
55  
56  
57  
58
-

- 
- 1  
2  
3 28. Jonassen D. H., Tessmer M., and Hannum W. H., (1999), *Task Analysis Methods for Instructional Design*, Mahwah, New  
4 Jersey: Lawrence Erlbaum Associates.
- 5 655  
6  
7 29. Kaderavek J. N., North T., Rotshtein R., Dao H., Liber N., Milewski G., et al., (2015), SCIENCE: The creation and pilot  
8 implementation of an NGSS-based instrument to evaluate early childhood science teaching. *Stud. Educ. Eval.*, **45**, 27–  
9 36.
- 10  
11  
12 30. Kane M., (2006), Content-Related Validity Evidence in Test Development. in Downing S. M. and Haladyna T. M. (eds.),  
13 *Handbook of Test Development*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc., pp. 131–154.
- 14 660  
15  
16 31. Kang H., Windschitl M., Stroupe D., and Thompson J., (2016), Designing, launching, and implementing high quality  
17 learning opportunities for students that advance scientific thinking. *J. Res. Sci. Teach.*, **53**(9), 1316–1340.
- 18  
19  
20 32. Lavery J. T., Underwood S. M., Matz R. L., Posey L. A., Carmel J. H., Caballero M. D., et al., (2016), Characterizing  
21 college science assessments: The three-dimensional learning assessment protocol. *PLoS One*, **11**(9), 1–21.
- 22  
23  
24 665 33. Loughran J. and Brubaker N., (2015), Working with a Critical Friend: A Self-study of Executive Coaching. *Stud. Teach.*  
25 *Educ.*, **11**(3), 255–271.
- 26  
27  
28 34. Lyon E. G., (2011), Beliefs, Practices, and Reflection: Exploring a Science Teacher’s Classroom Assessment Through the  
29 Assessment Triangle Model. *J. Sci. Teacher Educ.*, **22**(5), 417–435.
- 30  
31  
32 35. Mandinach E. B., Honey M., and Light D., (2006), A Theoretical Framework for Data-Driven Decision Making. *Pap.*  
33 *Present. Annu. Meet. AERA*, 1–18.
- 34 670  
35  
36 36. Martone A. and Sireci S. G., (2009), Evaluating Alignment Between Curriculum, Assessment, and Instruction. *Rev. Educ.*  
37 *Res.*, **79**(4), 1332–1361.
- 38  
39  
40 37. Marzano R. J. and Kendall J. S., (2008), *Designing & assessing educational objectives: applying the new taxonomy*, Scott  
41 M. P. and Alpert D. (eds.) Thousand Oaks, California: Corwin Press.
- 42  
43 675 38. Maxwell J. A., (2013), *Qualitative Research Design: An Interactive Approach*, 3rd ed. Knight V. (ed.) Thousand Oaks,  
44 California: SAGE Publications.
- 45  
46  
47 39. McDonald F. J., (1964), Meaningful Learning and Retention: Task and Method Variables. *Rev. Educ. Res.*, **34**, 530–544.
- 48  
49 40. Means B., Chen E., DeBarger A., and Padilla C., (2011), *Teachers’ Ability to Use Data to Inform Instruction: Challenges*  
50 *and Supports*, Washington, DC: Office of Planning, Evaluation and Policy Development, US Department of Education.
- 51  
52  
53 680 41. Merrill M. D., (2007), A Task-Centered Instructional Strategy. *J. Res. Technol. Educ.*, **40**(1), 5–22.
- 54  
55 42. Nakhleh M. B., (1992), Why some students don’t learn chemistry: Chemical misconceptions. *J. Chem. Educ.*, **69**(3), 191.
- 56  
57  
58
-

- 
- 1  
2  
3 43. National Research Council, (2014), *Developing Assessments for the Next Generation Science Standards*, Washington,  
4 DC: The National Academies Press.  
5  
6  
7 44. National Research Council, (2001), *Knowing what students know: The science and design of educational assessment*,  
8  
9 685 Pelligrino J., Chudowsky N., and Glaser R. (eds.) Washington, DC: National Academy Press.  
10  
11 45. National Research Council, (1999), *The Assessment of Science Meets the Science of Assessment: Summary of a*  
12  
13 *Workshop*, Washington, DC: The National Academies Press.  
14  
15 46. Patton M. Q., (2002), *Qualitative Evaluation and Research Methods*, Newbury Park, CA: SAGE Publications, inc.  
16  
17 47. Penuel W. R., Turner M. L., Jacobs J. K., Horne K., and Sumner T., (2019), Developing tasks to assess  
18  
19 690 phenomenon-based science learning: Challenges and lessons learned from building proximal transfer tasks. *Sci. Educ.*,  
20  
21 **103**(6), 1367–1395.  
22  
23 48. Polikoff M. S. and Porter A. C., (2014), Instructional Alignment as a Measure of Teaching Quality. *Educ. Eval. Policy*  
24  
25 *Anal.*, **36**(4), 399–416.  
26  
27 49. Porter A. C. and Smithson J. L., (2001), *Defining, developing, and using curriculum indicators*. CPRE Research Report  
28 695 Series, Philadelphia, PA: Consortium for Policy Research in Education.  
29  
30 50. Praetorius A. K., Pauli C., Reusser K., Rakoczy K., and Klieme E., (2014), One lesson is all you need? Stability of  
31  
32 instructional quality across lessons. *Learn. Instr.*, **31**, 2–12.  
33  
34 51. Remesal A., (2011), Primary and secondary teachers' conceptions of assessment: A qualitative study. *Teach. Teach.*  
35  
36 *Educ.*, **27**(2), 472–482.  
37  
38 700 52. Rothman R., Slattery J. B., Vranek J. L., and Resnick L. B., (2002), *Benchmarking and Alignment of Standards and*  
39  
40 *Testing*. (CSE Technical Report No. CSE-TR-566), Los Angeles, CA.  
41  
42 53. Ruiz-Primo M. A., Li M., Wills K., Giamellaro M., Lan M. C., Mason H., and Sands D., (2012), Developing and  
43  
44 Evaluating Instructionally Sensitive Assessments in Science. *J. Res. Sci. Teach.*, **49**(6), 691–712.  
45  
46 54. Russell J. W., Kozma R. B., Jones T., Wykoff J., Marx N., and Davis J., (1997), Use of Simultaneous-Synchronized  
47 705 Macroscopic, Microscopic, and Symbolic Representations To Enhance the Teaching and Learning of Chemical  
48  
49 Concepts. *J. Chem. Educ.*, **74**(3), 330.  
50  
51 55. Sadler R. D., (1989), Formative assessment and the design of instructional systems. *Instr. Sci.*, **18**, 119–144.  
52  
53 56. Sandlin B., Harshman J., and Yeziarski E., (2015), *Formative Assessment in High School Chemistry Teaching:*  
54  
55 *Investigating the Alignment of Teachers' Goals with Their Items*. *J. Chem. Educ.*, **92**(10), 1619–1625.  
56  
57  
58
-

- 
- 1  
2  
3 710 57. Schafer A. G. L. and Yeziarski E. J., (2020a), Chemistry critical friendships: Investigating chemistry-specific discourse  
4 within a domain-general discussion of best practices for inquiry assessments. *Chem. Educ. Res. Pract.*, **21**(1), 452–468.  
5  
6  
7 58. Schafer A. G. L. and Yeziarski E. J., (2020b), Investigating High School Chemistry Teachers' Assessment Item  
8 Generation Processes for a Solubility Lab. *Chem. Educ. Res. Pract.*, Advance Article.  
9  
10  
11 59. Stiggins R. J., (2001), The Unfulfilled Promise of Classroom Assessment. *Educ. Meas. Issues Pract.*, **20**(3), 5–15.  
12  
13 715 60. Taber K. S., (2013), Revisiting the chemistry triplet: Drawing upon the nature of chemical knowledge and the psychology  
14 of learning to inform chemistry education. *Chem. Educ. Res. Pract.*, **14**(2), 156–168.  
15  
16  
17 61. Tomanek D., Talanquer V., and Novodvorsky I., (2008), What Do Science Teachers Consider When Selecting Formative  
18 Assessment Tasks? *J Res Sci Teach*, **45**(10), 1113–1130.  
19  
20  
21 62. Towndrow P. A., Tan A.-L., Yung B. H. W., and Cohen L., (2010), Science Teachers' Professional Development and  
22 720 Changes in Science Practical Assessment Practices: What are the Issues? *Res. Sci. Educ.*, **40**(2), 117–132.  
23  
24  
25 63. Towns M. H., (2014a), Guide to developing high-quality, reliable, and valid multiple-choice assessments. *J. Chem. Educ.*,  
26 **91**(9), 1426–1431.  
27  
28 64. Towns M. H., (2014b), Guide to developing high-quality, reliable, and valid multiple-choice assessments. *J. Chem.*  
29 *Educ.*, **91**(9), 1426–1431.  
30  
31  
32 725 65. Tyler R., (1949), *Basic Principles of Curriculum and Instruction*, Chicago, IL: University of Chicago Press.  
33  
34 66. Vilardo D. A., MacKenzie A. H., and Yeziarski E. J., (2017), Using Students' Conceptions of Air To Evaluate a Guided-  
35 Inquiry Activity Classifying Matter Using Particulate Models. *J. Chem. Educ.*, **94**(2), 206–210.  
36  
37  
38 67. Wauters K., Desmet P., and Van Den Noortgate W., (2012), Item difficulty estimation: An auspicious collaboration  
39 between data and judgment. *Comput. Educ.*, **58**(4), 1183–1193.  
40  
41  
42 730 68. Webb N. L., (1997), Criteria for alignment of expectations and assessments in mathematics and science education. *Res.*  
43 *Monogr. No.6*, (8), 1–46.  
44  
45 69. Webb N. L., (2006), Identifying Content for Student Achievement Tests. in Downing S. M. and Haladyna T. M. (eds.),  
46 *Handbook of Test Development*. Mahwah, New Jersey: Lawrence Erlbaum Associates, pp. 155–180.  
47  
48  
49 70. Webb N. L., (2007), Issues Related to Judging the Alignment of Curriculum Standards and Assessments. *Appl. Meas.*  
50 *Educ.*, **20**(1), 7–25.  
51 735  
52  
53 71. Webb N. M. and Herman J., (2006), Alignment of Mathematics State-level Standards and Assessments : The Role of  
54 Reviewer Agreement CSE Report 685 Noreen Webb and Joan Herman University of California Norman Webb  
55  
56  
57  
58
-

---

1  
2  
3 University of Wisconsin , Madison June 2006 National Center for Research on Eva, Los Angeles, CA.  
4

5 72. Young K., Lashley S., and Murray S., (2019), Influence of Exam Blueprint Distribution on Student Perceptions and  
6  
7 740 Performance in an Inorganic Chemistry Course. *J. Chem. Educ.*, **96**(10), 2141–2148.  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58