Chemistry Education Research and Practice

Application and testing of a framework for characterizing the quality of scientific reasoning in chemistry students' writing on ocean acidification

Journal:	*Chemistry Education Research and Practice*
Manuscript ID	RP-ART-01-2019-000005.R1
Article Type:	Paper
Date Submitted by the Author:	15-Mar-2019
Complete List of Authors:	Moon, Alena; University of Michigan, Chemistry Moeller, Robert; University of Michigan, Department of Chemistry Gere, Anne; University of Michigan, Sweetland Center for Writing Shultz, Ginger; University of Michigan, Department of Chemistry

Application and testing of a framework for characterizing the quality of scientific reasoning in chemistry students' writing on ocean acidification

Alena Moon, Robert Moeller, Anne Ruggles Gere†, and Ginger V. Shultz

Department of Chemistry, University of Michigan, Ann Arbor, MI, USA
†Sweetland Center for Writing, University of Michigan, Ann Arbor, MI, USA

Corresponding author: Ginger V. Shultz, gshultz@umich.edu, 2521 Chemistry, University of Michigan, Ann Arbor, MI, 48109.

Provide short biographical notes on all contributors here if the journal requires them.

Development and testing of a framework for characterizing the quality of scientific reasoning in students' writing on ocean acidification

Science educators recognize the need to teach scientific ways of knowing and reasoning in addition to scientific knowledge. However, characterizing and assessing scientific ways of knowing and reasoning is challenging. Writing-to-learn offers one way of eliciting and supporting students' reasoning; further, writing serves to externalize and make traceable students' reasoning. For this reason, it is a useful formative assessment of scientific reasoning. The utility hinges on researchers' ability to understand what students can do and think from their writing. Given the challenges in assessing students' writing, this research offers an adapted framework for assessing students' scientific reasoning evident in writing. This work will introduce the adapted framework and show an application to general chemistry students' argumentative writing about ocean acidification. We provide evidence that this framework can be used to validly estimate the quality of students' reasoning. We argue that this framework offers some affordances that overcome challenges reported in the literature. It serves to define scientific reasoning in a domain-general way by breaking it down into its components, but in a way that can produce a composite score that tells us about how students reason using chemistry content. Further, the framework provides a way to characterize the scientific accuracy of students' reasoning that can inform instructors' treatment of alternative conceptions.

Keywords: Writing-to-learn; assessing writing; scientific reasoning

Background

Science educators recognize that it is insufficient to only teach students' scientific knowledge as a collection of concepts and topics. Rather, to enable students to use scientific knowledge, we must support the development of reasoning and thinking skills that scientists use (NRC, 2012; Sevian & Talanquer, 2014; Bulte, Westbroek, De Jong, & Pilot, 2006). Writing-to-learn (WTL) is one way of supporting the development of this skill by activating deep thinking and reasoning in students (Keys, 1999) and, more importantly, making that reasoning visible and traceable (Emig, 1977; Kelly & Takao,

2002; Kelly, Regev, & Prothero, 2007). From an assessment perspective, this evidence of student reasoning is valuable in so far as researchers and practitioners can use it to make an argument about students' abilities to reason scientifically (Laverty et al., 2016; NRC, 2001). However, there are challenges that currently limit the utility of this evidence. There are few widely agreed upon epistemic criteria for characterizing the quality of students' reasoning (i.e., what makes one students' reasoning better than another's). Further, actually applying these criteria to understand and evaluate students' writing is difficult as writing requires the researcher to make choices about grain size, whether to evaluate structure or content or both, and what the presence or absence of a quality criterion actually looks like in students' writing (Kelly & Takao, 2002; Takao & Kelly, 2003a). To address these challenges, we have modified and applied a framework for characterizing and evaluating reasoning in students' argumentative writing. This framework contributes meaningfully to efforts to conceptualize and evaluate scientific reasoning, as well as to efforts to analyse writing, which poses unique challenges.

Writing to Learn

Writing-to-learn refers to the kind of informal writing about science that facilitates learning and ownership of scientific ideas. This informal writing is distinct in that its primary aim is not to communicate or display mastery to an instructor, but to actually facilitate sense-making by activating deep thinking and interaction with the concepts (Keys, 1994). A secondary benefit of writing-to-learn, then, is promoting engagement with disciplinary norms of writing and thinking (Prain & Hand, 2016). There is quite a bit of variation around this primary aim, however; WTL assignments take a variety of forms, lengths, methods of text production, audiences, and genres (Keys, 1994).

A secondary analysis of six writing-to-learn studies revealed some promising gains as a result of writing-to-learn—the treatment condition outperformed comparison

groups on a total test scores and conceptual question scores and this effect was largely due to the treatment (Gunel, Hand, & Prain, 2007). All six studies followed a similar design including a pre-test/post-test design with the test having multiple-choice and conceptual extended response questions. More importantly, all writing interventions were grounded in the same theoretical considerations that have been identified as key for successful learning from writing: 1) opportunities for brainstorming, 2) provision of authentic audiences, 3) drafting and redrafting with feedback, 4) explicit instruction of genre specifications, 5) focus on big ideas, 6) use of rubrics, and 7) diverse opportunities to plan and draft writing (Gere, Limlamai, Wilson, MacDougall Saylor, & Pugh, 2019; Gunel et al., 2007; Klein, 1999, 2015). The theoretical grounding afforded comparisons across domains and writing assignment types and served to reveal the benefits of WTL more broadly (Gunel et al., 2007; Prain & Hand, 2016). However, at the undergraduate STEM level specifically, more work is needed to understand the mechanism of effect for WTL assignments (Reynolds, Thaiss, Katkin, & Thompson, 2012) and we argue that to undertake investigations into the mechanism of effect, we need a reliable and meaningful framework for interpreting and evaluating students' written work.

Characterizing Students' Reasoning in Written Products

Constructed responses reveal rich insight into students' ideas and the coherence of those ideas, but evaluating open responses remains a barrier to implementing such rich assessments (Liu, Rios, Heilman, Gerard, & Linn, 2016). This barrier consists of two distinct but interdependent challenges: characterizing the quality usually in some sort of rubric (Kelly & Bazerman, 2003; Sandoval, 2003; Sandoval & Millwood, 2005) and consistently and reliably applying the quality criteria (Ha, Nehm, Urban-Lurain, & Merrill, 2011; Liu et al., 2016). Additionally, the nature and difficulty of these

challenges vary with the length of constructed response; for example, extended arguments in the form of research reports in oceanography tended to have long and complex chains of reasoning that are difficult to characterize with a single rubric (Kelly, Regev, & Prothero, 2007; Kelly & Takao, 2002). Researchers have sought to overcome these challenges with a variety of approaches for a variety of written products, ranging from short written explanations (Ha et al., 2011; Liu et al., 2016; Moreira, Marzabal, & Talanquer, 2018; Sandoval, 2003; Sandoval & Millwood, 2005; Sandoval & Reiser, 2004) to more extensive writing like laboratory reports or research reports (Grimberg & Hand, 2009; Kelly, Chen, & Prothero, 2000; Kelly et al., 2007; Kelly & Takao, 2002; Takao & Kelly, 2003b). A few of these approaches are distinct in that they break down students' responses into smaller units to then identify patterns in students' reasoning, as opposed to a rubric that considers the quality of the response as a whole (Grimberg & Hand, 2009; Kelly et al., 2007; Kelly & Takao, 2002; Moreira et al., 2018). These approaches will be the focus of this literature review.

Moreira et al. (2018) specifically sought to characterize the causal reasoning of 10th grade chemistry students' explanations of freezing point depression. To do so, the authors modified and applied a discourse analysis framework developed by Russ et al. (2008) to students' written explanations and drawings. The final form of the analysis scheme included four components—entities, properties, activities, and organisation—and the relationships between the components and the students' drawings that were identified in students' responses. Entities are the 'things' in the system that are being considered. Properties are characteristics of those entities and activities are actions of those entities. Organisation refers to the spatial-temporal relationship between the entities and activities or properties of the system. By coding the explanations for these components and relationships, they were able to elucidate patterns in students'

explanations and organize these patterns into four levels according to the quality of causal reasoning. These levels increased in sophistication from descriptive to relational to simple causal, culminating in emerging mechanistic. Unsurprisingly given the authors' previous findings (Sevian & Talanquer, 2014), the majority of students (45%) used relational causal reasoning (Moreira et al., 2018). Explanations at this level could be modelled to show that students generally identified two entities and the properties of one or both entities, and then related either the properties to each other or related entities to properties. Such complex modelling of students' explanations can hopefully equip teachers with more sophisticated approaches to understanding, interpreting, and developing students' reasoning abilities (Moreira et al., 2018).

Grimberg and Hand (2009) similarly identified the presence or absence of dimensions of reasoning and determined *patterns* in students' reasoning evident in their laboratory reports. In this study, Grimberg and Hand (2009) identify cognitive operations used in writing laboratory reports and then construct what they term 'cognitive pathways'—the sequence of cognitive operations used by author(s) of a lab report. The authors argued that because writing a laboratory report was a meaning-making activity, considering the sequence of cognitive operations revealed how students were constructing meaning. Using a list of 11 cognitive operations derived partially from the literature and from the students' data, authors coded students' writing for use of cognitive operations. The cognitive operations included observation, measurement, comparison, analogy, clarifications, claim, cause/effect, induction/generalization, deduction, investigation design, and argumentation (Grimberg & Hand, 2009). When comparing cognitive pathways of low achievers to high achievers, as determined by a standardized skills test, both high and low achievers used the same range of operations, but with a different structure. Though the cognitive structure was partially determined

by the structure of the Science Writing Heuristic (SWH) activity (i.e., clarification questions were posed during the research question portion of SWH activity), high achievers began using complex operations earlier in the text than low achievers. This ultimately demonstrated that SWH scaffolding supported all students in using high-complexity operations, albeit at different rates (Grimberg & Hand, 2009).

Grimberg and Hand's (2009) work demonstrates the capacity of writing to make students' thinking visible and traceable. Emig (1977) argues that writing is unique in its capacity to do this. Kelly and Takao (2002) demonstrate the utility of students' writing for understanding their reasoning by characterizing how undergraduate students use evidence to construct arguments. To make this characterization, they developed a research methodology that models epistemic levels of argument. This framework includes six levels ranging from the lowest, data charts and representations, to the highest, general geological (the specific context for this work) knowledge not specific to the data presented. The levels represented students' ability to abstract from data to make claims. With this framework, the authors analysed a subset of undergraduate oceanography students' assignments by labelling each sentence with an epistemic level. These epistemic criteria were then weighted in order to rank the 24 student arguments (research reports in oceanography) from best to worst. While this framework was very useful for characterizing the quality of students' arguments based on their use of evidence, it did possess limitations. Namely, the assessment of quality determined by the framework did not always align with content experts' evaluation of quality, the framework did not consider inference logic (how the data led to theoretical claims), and the authors had to make inferences in their application of the framework. These limitations are difficult to overcome for everyone aiming to evaluate ill-defined constructs like use of evidence. However, Kelly and Takao (2002) revealed that claims

can be made about students' reasoning from their written work. In order for the insight into students' reasoning provided by writing to be useful for informing students' development, tools for assessing writing must be efficient, systematic, and offer tailored feedback.

Ongoing discussions about writing assessment distinguish between holistic scoring, assigning a single score to a broad variable like writing proficiency, and analytic scoring, breaking down variables like writing proficiency into components that are individually scored (Hamp-Lyons, 2016a, b). High-stakes assessments, such as college entrance examinations, motivated the use of both holistic and analytic scoring approaches to assign students general writing scores, but researchers have begun to identify shortcomings of both (Neill, 2002; Hamp-Lyons, 2016b, Chapman, 2016). With more complex analytical tools (i.e., multivariate analyses), Hamp-Lyons (2016) calls for a movement to multiple trait scoring of writing. In multiple trait scoring, there is no single score given, whether composite or holistic. Rather, a set of scores is assigned with multiple traits each warranting a score, thus lending to a richer description of students' ability (Hamp-Lyons, 2016a, b).

Rationale and Research Objectives

In any effort to measure a student's reasoning, choices must be made about what characterizes quality. Specific to evaluating extensive writing, additional decisions must be made about the grain size, level, and nature of rubric that will be used to determine quality. In order to address these challenges, the cognitive operations used by Grimberg and Hand (2009) were modified and applied to students' writing on ocean acidification. Motivated to leverage the rich insight into student thinking that constructed responses offer, the work presented herein aimed to test an approach for analysing extensive

scientific writing by applying it to a new context. We aimed to answer the following questions regarding this approach:

(1) Can cognitive operations be used to make sense of general chemistry students' argumentative writing? If so, how?

(2) What features of students' argumentative writing do cognitive operations serve to explain?

(3) What is the relationship between framework estimates of quality and conceptual correctness?

Methods

Participants, setting, and data collection

A writing prompt was designed and administered in a first-semester General Chemistry course serving primarily students in the College of Engineering and undeclared students in the College of Literature, Science, and Arts. This course had an enrolment of 1413 students, most of whom were freshman and sophomores. The content of the course covered traditional general chemistry concepts, ranging from dimensional analysis, quantum mechanical atomic models, bonding theories, to reactions, enthalpy, intermolecular forces, chemical equilibrium, and acid-base theories.

This course is structured with three lectures led by an instructor and one discussion session led by a teaching assistant per week. During each discussion section, students complete a quiz. The writing assignment for this study was administered as a substitute for a quiz. The writing assignment was uploaded as a .pdf file to the course management site one week before the due date. This writing assignment directed students to consider a set of concepts in their response and to keep their post between

350 and 500 words. Though the majority of students' responses were within this range, some wrote less and some wrote more.

Students all submitted their writing assignment online the following week at the start time of their specific discussion session. For this reason, some students with discussion sessions later in the week had more time to write than those with earlier recitations. During the discussion, students formed teams of two or three, switched papers and gave feedback to each other. Six hundred seventy-three students gave consent to have their writing analysed. Ethical review board approval was gained in order to collect and analyse written assignments that students consented to have analysed. Students did not receive any feedback on their written work beyond the conversation that took place in their discussion session. We found that students did not make meaningful revisions following the peer review discussion. Additionally, they were not required to submit a revision. However, if they did, that revised draft was used for analysis.

Writing activity development and design

The writing prompt was developed iteratively through correspondence with authors who had expertise in writing to learn and the development of meaningful writing prompts (AG) and the faculty members teaching the course who collectively held more than two decades of experience teaching chemical equilibrium in general chemistry. WTL prompts are generally designed to provide students with an audience, an identity, and an authentic context that require students to engage with a specific concept. This WTL assignment was intentionally designed with elements empirically determined to contribute to meaningful learning through writing (Gere et al., 2019). In this case, the prompt showed a fake social media post, in which 'Ernie Clueless' shares a plot illustrating the trend of concentration of atmospheric carbon dioxide and ocean pH over

time. Ernie claims that these things are unrelated. Students were tasked with explaining the relationship to Ernie, given the relevant equilibria. The prompt targeted the concept of chemical equilibrium, drawing on Le Châtelier's principle. It inherently supported argumentation by requiring students to differentiate their perspective from Ernie's.

Data analysis: Development and application of analytical framework

A list of cognitive operations was modified from a list used by Grimberg and Hand (2009) to analyse reports from a Science Writing Heuristic (SWH) laboratory. In this context, a cognitive operation is a written discursive move that serves some cognitive objective. Cognitive operations, then, determined the grain size for breaking down an essay into smaller analysable units (i.e. the number of sentences that served the objective of a claim, for example, were coded as such). For this reason, a claim could be one sentence in one essay and three sentences in another. The amount of text that was assigned a code was determined by the function of that text. The list of cognitive operations used by Grimberg and Hand (2009) was refined iteratively by testing it against the data. This involved using Grimberg and Hand's original set to code the text, identifying text that could not be coded with this set and operations from this set that did not serve to explain any of the data, and refining the set of operations to a set that were all used to describe virtually all of the text. Multiple initial iterations occurred with the same subset of essays (N=25) and subsequent iterations incorporated more essays on an as needed basis. Table 1 shows the final list of cognitive operations that was used throughout the final analysis. Included in Table 1 is a characterization of the dimensionality of each operation. As will be explained further in the theoretical framework, the complexity of cognitive operations was determined by its dimensionality—number of ideas being drawn upon. Operations with two domains were

more cognitively complex than operations with one dimension. That is, they drew upon and connected more idea units.

Table 1. Finalized list of cognitive operations and descriptors used to analyse all essays, listed in order of increasing cognitive complexity

Cognitive Operation	Description	Dimensionality
1. Definition	Canonical description of a term, concept, idea, or theory	single domain
2. Observation	Qualitative description of change, trend, or transformation of a variable	
3. Measurement	Quantitative description of change, trend, or transformation of a variable	
4. Comparison	Relationship of change, trend, or transformation for two or more variables	
5. Example	Illustration of a class of objects by singling out one object	two domains
6. Claim	Assertion supported with a tentative explanation	
7. Consequences	Cause and effect explanation with either cause or effect falling outside the scope of the writing prompt	
8. Cause and effect	Explanation providing a mechanism with a causal agent and observed effects	
9. Deduction	Application of a theory or principle to a specific system or scenario	multiple domains
10. Argumentation	Explicit differentiation between the author's perspective and the fictional character's perspective*	

*This conceptualization of argumentation was specific to the context of writing prompt used in this study. It is expected that it could be easily translated to other writing contexts that require argumentation.

Once a list of cognitive operations was finalized, the first two authors began coding assignments and built a detailed rubric that included definitions, linguistic markers, and examples for each operation. This rubric was further refined through multiple iterations of analysis by a team of chemistry education researchers in an effort to establish inter-rater reliability (IRR). This stage included a team consisting of the first two authors and another chemistry education researcher trained in qualitative coding of writing. The graduate student was trained on an existing rubric, the whole team coded ten assignments, and an IRR coefficient in the form of Krippendorff's alpha (KA) was calculated. This coefficient was quite low after the first round, so revisions were made to the rubric, another training session was conducted, and subsequent rounds of analysis

were conducted until a Krippendorff's alpha value of 0.69, the minimum acceptable value, was achieved. Training involved discussing the rubric which included examples from many essays and then illustrating the coding process by coding a few essays all together.

The first two authors then coded approximately 200 assignments each, with large overlap. Having two researchers code many of the same assignments lent to the reliability of the coding. Throughout analysis, IRR 'checks' were performed to ensure that the rubric was being applied consistently. This involved selecting overlapping assignments and determining a KA. This stage resulted in a KA of 0.89, which was a desirable value (Krippendorff, 2004).

Estimate of quality

Once all assignments were coded for cognitive operations, estimates of quality were assigned according to the cognitive complexity of the essay. The cognitive operations are ordered in Table 1 according to increasing cognitive complexity; that is, a definition has the lowest cognitive complexity (1) whereas argumentation has the highest cognitive complexity (10). Overall cognitive complexity for the essay was determined by taking a weighted average of operations used. Because the magnitude of text included in an operation varied from one essay to another, the average was weighted by the number of sentences within that operation. An assignment, then, that was 50% argumentation would likely have a higher average complexity than an assignment with only 10% argumentation. This approach resulted in a single number characterizing the quality of students' reasoning in an essay. This process of producing a single number is illustrated with the example essay in Table 2.

Table 2. Example of student essay and determination of cognitive complexity score

Le Chatlier's principle is a way to show how if the equilibrium of a certain reaction is altered by changing certain aspects, then the reaction or equilibrium position will actually shift to fix the change. In Ernie's post there are multiple reactions that follow one after the other, ultimately showing how CO_2 transforms to H_2CO_3, which changes to H^+ and HCO_3, which finally changes to 2H+. These reactions are connected in that as one breaks down, it spurs another reaction to then form. This shows how CO_2 in the atmosphere actually affects PH concentration and ocean acidification (or the number/concentration of H+ ions in the reaction). Because of this correlation, certain components like temperature, pressure, volume, and concentration can affect CO_2 levels and acidification. In our case, we will look at the concentration or amount of CO_2 in correlation to the amount of H^+ that is formed. If the concentration of CO_2 (as a reactant) were to increase, the system would try to decrease it; this would mean that the concentration of the other reactants will increase to react with the CO_2, but then more product would be formed. Thus increasing the concentration of CO_2 (atmos) would cause the system to shift towards the products, producing more CO2(aq). Then because there is more CO_2 (aq), more water would have to be use so more H_2CO_3 would form. The next reaction would then proceed as the others with an increase in the formation of the products H^+ and HCO_3, which in turn would once again increase the next reaction producing more $2H^+$ product. This may seem confusing, but to summarize if CO_2 as a reactant increases, it would need to counteract this change by producing more product of $2H^+$ (acidification). Now, if CO_2 (atmos) reactant were to decrease, the equilibrium will actually try to increase it so as to set the system back at equilibrium. This would mean that a decrease in that reactant would cause more product to be formed so to produce more CO_2; thus $2H^+$ product would increase to make more CO_2. The relationship between CO_2 and pH is that as CO_2 increases, the pH will decrease and make the oceans more acidic. The pH scale runs from 1 to 14, with the acids being numbers 1 to 6. So the lower the pH, the more acidic and the more H^+ ions that are formed (the H^+ are indicators of acidic properties). Le Chatlier's principle shows that increasing one will have to increase the other so that the system is able to be equal again, thus there is correlation between CO_2 (atmos) increase and H^+ ions increase(acidification)."	**Definition (1)** **Observation (2)** **Claim (6)** **Deduction (9)** **Comparison (4)** **Definition (1)** **Comparison (4)** **Cause and Effect (8)**
Cognitive complexity = $\frac{\Sigma(cognitive\ operation\ score * \#\ sentences\ used)}{total\ number\ of\ sentences}$ Definition (1) + 3 * Observation (2) + Claim (6) + 9 * Deduction (9) + Comparison (4) + Definition (1) + Comparison (4) + Cause & Effect (8) = 111 111/18 =6.2 (Average complexity for this essay)	

Faculty ranking of essays

One motivation of this work is to use a framework to systematically characterize students' writing. In order to understand how this framework could accomplish that task, we compared framework estimates of quality with instructors' estimates of quality. The precedent for this approach is offered by Kelly and Takao's testing of the framework for epistemic levels (Kelly & Takao, 2002). Further, evidence from interviews with STEM faculty about writing suggested to us that the knowledge experts use to evaluate writing is tacit (Moon, Gere, & Shultz, 2018). By comparing expert rankings with essays ordered according to a framework, we can identify ways that the framework may capture this tacit knowledge. To compare framework estimates of quality and instructor estimates, we compared essays ranked according to the framework to instructor rankings. To select the essays, we split cognitive complexity into four roughly equivalent ranges (3.1-4.8, 4.8-6.5, 6.5-8.2, 8.2-10), where the highest range included the 'best' essays, according to the framework. Within each range, an essay was randomly selected. The output of this step was four essays ranging from most complex to least complex, as determined by the framework. These essays were then provided to instructors who were tasked with ranking the essays from best to worst according to the quality of scientific reasoning, which they were directed to evaluate as they saw fit. The rationale for not providing faculty with more extensive quality criteria was to target the kind of evaluative work that is inherent to grading this sort of task; that is, we wanted instructors to make the kinds of decisions that are required to define and evaluate scientific reasoning. Five faculty with a range of experience teaching general chemistry from different institutions ranked the writing tasks and one instructor volunteered the reasoning behind their ranking.

Chemistry Content Analysis

Essays were examined to understand how students employed ideas about chemical equilibrium within their argument. Essays were coded for conceptual correctness, which involved flagging all occurrences of inaccurate ideas. Every time an inaccurate idea was identified, the cognitive operation containing that incorrect idea was marked. The pairing of the incorrect idea with the cognitive operation was intentional, based on the theoretical framework explained below, which posits that what students articulate in written forms are representations of the ideas they hold. In this case those ideas related to chemical equilibrium.

Theoretical framework

The primary assumption made in this work is that writing reveals students' cognitive structures or understanding of the meaning of a concept (Emig, 1977; Novak, 2002). Meaning is defined in this case as 'the totality of propositions linked to any given concept,' excluding the emotional association with the concept and the context in which the concept was learned (Novak, 2002). This assumption is grounded in the capacity of writing to 1) connect or relate propositions in the author's mind and 2) make 'evolutionary development of thought graphically visible and available (Emig, 1977).' According to Novak's theory of meaningful learning, the complexity of the meanings can be evaluated, the quantity and quality of which will determine meaningful learning (Novak, 2002).

In this work, units of written text were coded according to their cognitive function as a cognitive operation. Each cognitive operation was constituted by some number of ontological domains, elements within those domains, and relationships between each (Grimberg & Hand, 2009; Halford & McCredden, 1998). Cognitive complexity, then, was defined in terms of dimensionality; in which cognitive operations with higher numbers of domains, elements, and relationships were more complex than

cognitive operations with fewer (Halford & McCredden, 1998). These domains and elements can be conceptualized similarly to the discourse analysis framework used by Moreira et al. (2018) where they refer to 'things' in the system being considered. All cognitive operations were organized on a spectrum from least complex to most complex according to this criterion, illustrated in Table 1. A students' scientific reasoning can then be considered as the progression of operations used in a text.

Therefore, one way of evaluating the quality of meanings was through cognitive complexity. It is possible, however, that a student can employ cognitively complex reasoning without necessarily using correct content (Sandoval & Millwood, 2005; Kelly & Takao, 2002). For this reason, a second way of evaluating the quality of meanings was considering their conceptual correctness; that is, their agreement with scientifically accepted knowledge. A conceptual change perspective suggests that problems of incorrect conceptions arise from Limited or Inappropriate Propositional Hierarchies (LIPHs), the way that concepts are inappropriately organized in the learner's mind, which means that we as instructors must consider both the *content and structure* of incorrect conceptions. Further, this implies a greater instructional effort is needed to remediate stable LIPHs (Novak, 2002). The cognitive operations framework used in this study helps characterize the complexity or stability—as a function of the number of domains, elements, and relationships—of conceptions.

Results

Research Question 1: Can cognitive operations be used to make sense of general chemistry students' argumentative writing? If so, how?

Table 2 illustrates how these cognitive operations were interpreted to analyse the writing with examples from students' essays. The examples are useful for discussing the difficulties of applying this framework. The lower complexity operations (*definition*

through *example*) were relatively easy to identify. The difference between *observation* and *measurement* was essentially a difference between qualitative and quantitative, with *measurements* requiring some numerical component. Because students could sufficiently respond to this writing prompt with qualitative reasoning, *measurement* occurred less frequently (Table 1). A *comparison* was distinct from *observation* and *measurement* in describing more than one variable relative to each other. A *claim* was similar to *comparison* in referencing a relationship between two variables but was distinct in that it required a tentative explanation of the relationship. For the *claim*, then, there were two domains (explanation and observation). *Cause and effect* and *consequences* were similar and easier to identify in text, with common linguistic markers being 'caused,' 'leads to,' or 'drives.' *Consequences* used cause and effect reasoning but relied on causes or effects that fell outside the scope of the prompt. In this case, the student referenced the effect of ocean acidification on coral. A primary marker of *deduction* was the invocation of a principle or theory that was then applied to a specific system. In this context, students frequently invoked Le Chatalier's principle or equilibrium. Finally, *argumentation* was undoubtedly the most difficult to identify as it often drew on multiple operations. So, we used this feature as an identifier. *Argumentation*, then, required indistinguishable use of multiple operations and to explicitly differentiate between the author's perspective and Ernie's (or any opposing position in another context). Table 1, 2, and 3 together can be used to apply this framework to other contexts.

Table 3. Examples of cognitive operations, arranged in order of increasing complexity. Numbers describe cognitive complexity ordering.

Operation	Example from student essay
Definition (1)	Equilibrium is a state where a reaction is occurring forwards and backwards at equal rates with no overall change. When a change occurs to the system, the reaction will shift in a direction to counteract this change.

Observation (2)	The trend in the graphs shows an increase in atmospheric carbon dioxide over time.
Measurement (3)	The ocean pH has dropped from 8.2 to 8.1 since the Industrial Revolution.
Comparison (4)	The plot that you shared illustrates that as the concentration of atmospheric increases over time, the pH of the ocean seawater decreases.
Example (5)	An example is hydrochloric acid, which has hydrogen ions attached and is an acid with a rather low pH level.
Claim (6)	As a matter of fact, Ernie, the correlation between CO_2 levels in the atmosphere and the pH of the oceans makes sense according to chemistry.
Consequences (7)	When this happens, calcifying organisms will become weaker, such as coral. They will be significantly affected as a result and may be unable to live in the current environment in which they live. In addition to ocean acidification wreaking havoc on the environment, other factors such as climate change can do the same thing and increase the amount of damage that is done to it.
Cause and Effect (8)	In each equation, the amount of reactants increases, which drives the reaction forward, meaning the amount of products will increase until the reaction reaches equilibrium.
Deduction (9)	In accordance with Le Châtelier's Principle, increasing the amount (or the concentration) of atmospheric CO_2 will shift this equation towards the dissolved CO_2 in the ocean to make up for the increase in gas on the reactants (left) side. This dissolved CO_2 is indicated as 'CO_2 (*aq*)' in the equation, which stands for 'aqueous CO_2.' Consequently, the dissolved CO_2 relates to the bicarbonate formation equation: CO_2 (*aq*) + H_2O (*l*) ⇄ H_2CO_3 (*aq*) ⇄ H^+ +HCO_3 (*aq*) (Doney et al., 2009 [*reference provided to student*]) Just as increasing the concentration of the atmospheric CO_2 caused the equilibrium between dissolved CO_2 to favor the formation of dissolved CO_2, a similar phenomenon will occur to favor production of bicarbonate (HCO_3) and hydrogen, two byproducts of carbonic acid (H_2CO_3). Increasing dissolved CO2 concentration—a result of increased atmospheric CO_2—will 'push' the equilibrium towards the formation of H_2CO_3. Likewise, a shift towards the formation of H_2CO_3 will also shift the equation forward towards the formation of protons, H^+, and HCO_3. Finally, these products of proton and bicarbonate will be in equilibrium with two protons and carbon trioxide (CO_3^{2-}): H^+ + HCO_3 (*aq*) ⇄ 2H^+ + CO_3^{2-} (Doney et al., 2009)

	Once more, an increase in the H^+ and HCO_3 concentration will push the equilibrium forward towards the formation of two $2H^+$ and CO_3^{2-}.
Argumentation (10)	So using this information, you can now look at the graph you posted and understand the relationship between CO_2 levels and the pH of the water. As CO_2 is absorbed into the water, it produces H+ ions, which then cause the pH of the water to decrease. You can see this trend on the graph. Even though they do not seem like they should be related in any way, a change in one would cause a change in the other. Something that is making the line representing the CO_2 in the atmosphere on the graph to increase so much is the amount of CO_2 humans emit every day. Whenever you drive a car you are releasing CO_2 into the atmosphere. Since there is so much more CO_2 being released, the oceans are absorbing more CO_2. In fact, the oceans have absorbed almost 30% of the CO_2 humans have emitted since the Industrial Revolution. As more CO_2 is absorbed by the oceans because of human activity, the more H+ ions are formed, and the more the pH of the ocean is decreased.

A total of 296 assignments have been coded (average weighted complexity: 6.3; average number of operations per essay: 9). We determined that saturation had been reached when no new codes or patterns were observed in the writing. Table 4 shows the frequency of operations used in the 296 essays. *Observation* was the most frequently used operation by students, with many using more than one observation in a single essay. Students heavily relied on making statements about how a variable was changing to counter 'Ernie's' claim. In this context, this means that students were able to understand how variables were changing from the graph provided. There was very little use of *Example* or *Measurement*. *Argumentation* was also used relatively infrequently. As mentioned above, *argumentation* was distinct as its own operation given indistinguishable use of multiple lower complexity operations. For this reason, *argumentation* required students to combine multiple operations (and domains). This difficulty combined with the infrequent use suggests that *argumentation* was indeed the most complex operation. The high frequency of *claim*, *cause and effect*, and *deduction* is likely tied to this writing context. Students were trying to convince Ernie (*claim* and *cause and effect*) by invoking chemical principles (*deduction*). It is expected that the

distribution of use will vary with different writing contexts, depending upon what the prompt elicits.

Table 4. Descriptive information from application of the framework to the data set presented herein.

Operations	Frequency
Observation (2)	411
Claim (6)	348
Cause and effect (8)	331
Definition (1)	312
Comparison (4)	302
Deduction (9)	235
Consequences (7)	146
Argumentation (10)	63
Example (5)	51
Measurement (3)	34

Research Question 2: What features of students' argumentative writing do cognitive operations serve to explain?

To determine what exactly this framework served to characterize, two comparisons were made. The first comparison was between framework estimates of complexity and instructor estimates of quality. This comparison was intended to demonstrate that this framework was telling us something that faculty would normally have to make a judgment about. Further this comparison was intended to reveal that this framework could make *similar* judgments to an instructor. Table 5 shows how five faculty ranked four assignments, and this is compared to our framework ranking. Instructors were tasked with ranking the assignments according to the quality of the scientific reasoning (as they saw fit to evaluate it). This approach was taken so as to elicit instructors' "gut reaction"—the kind of evaluation they would make if they were grading this sort of task in their class. Instructor rankings reveal a few trends. First and potentially the most

important, there is almost complete consensus on the 'best' essay (J) with the exception of Instructor 3. This finding speaks to the framework's capacity to identify the best. Essay J received a high cognitive complexity score because of the presence of extended argumentation, which aligns with what instructors are valuing when evaluating scientific reasoning. Four of the five instructors ranked Essay G as second best, with the exception of Instructor 3, even though our framework estimates it as second from the worst. Finally, all five instructors consider Essay H and K to be the worst, where as our framework estimates Essay H to be the second best. The difference between instructor rankings for Essays H, G, and K and framework estimates illustrate an important limitation of the framework. Essay H contained a misconception in which the student claimed a relationship between atmospheric temperature and ocean acidification though there was no data regarding heat for any of the chemical reactions provided. However, this student's reasoning was rather sophisticated, with multiple high complexity cognitive operations. Our framework does not account for scientific accuracy of students' essays. We chose to include this essay as it authentically represents what an instructor might encounter with grading writing. One possible explanation for the ranking difference is that for instructors scientific content and reasoning are inextricably linked, which is consistent with feedback from one instructor who explained that content accuracy factored into their ranking. It is for this reason that we also analysed the scientific accuracy of students' essays (see Table 5 below).

Table 5. Instructor ranking of four assignments of varying quality compared to framework estimates of ranking

		Expert rankings (4 worst, 1 best)				
Essay name	Cognitive complexity ranking (4 worst, 1 best)	Instructor 1	Instructor 2	Instructor 3	Instructor 4	Instructor 5
J	1	1	1	2	1	1
H	2	3	3	4	4	4
G	3	2	2	1	2	2

K	4	4	4	3	3	3

The second comparison made was between cognitive complexity—framework estimates of quality—and common student characteristics that are frequently used as measures or predictors of success (Hein & Smerdon, 2013). The purpose of this comparison was to determine if characterizing the quality of students' reasoning in this way was revealing something about students that could have been predicted by a metric that was already collected (e.g., ACT math score). In other words, this framework is useful only in so far as it tells us something interesting about students that other metrics do not. Table 6 shows the correlations between common student characteristics and cognitive complexity. There were no significant correlations between cognitive complexity and any common characteristics, which would not be expected for measures of constructs distinct from that captured by this framework (i.e. math). These findings may mean that the framework captures something distinct from what is measured by other standardized tests. A strong negative correlation exists between the number of operations used and the cognitive complexity. This means that students with higher cognitive complexity essays used fewer moves, which could indicate a synthesis of ideas in order to produce higher complexity operations.

Table 6. Pearson correlations between student characteristics and cognitive complexity (p-values reported for t-tests used for categorical variables: gender and ethnicity [white and non-white students compared])

Variables	Cognitive complexity
Number of operations	-0.649*
Final exam grade	-0.018
Final course grade	-0.025
CHEM placement	-0.081
MATH placement	-0.020
ACT math†	0.003
Current GPA	-0.062
Cumulative GPA	-0.060
Gender	0.401
Ethnicity	0.071

* indicates p (two-tailed) < 0.01

† For students with only SAT math scores, their scores were converted to ACT math scores using contingency tables

Research Question 3: What is the relationship between framework characterizations of complexity and conceptual correctness?

The data above reveal that the cognitive operations framework is characterizing students' reasoning in a way that other measures do not. However, the instructor rankings reveal that there exists a relationship between reasoning and accuracy. The motivation for considering this relationship partially sources from the concern that any information this framework provides is irrelevant if students are largely scientifically inaccurate. In order to explore this relationship, we coded all data that had already been coded according to operations for 'correctness.' That is, when scientifically inaccurate information was identified in an essay, the cognitive operation containing that information was marked as incorrect. In this way, all student writing was coded for both correctness and cognitive function (i.e., content and structure as highlighted in theoretical framework). Table 7 shows the number of incorrect operations per the total number of cognitive operations. Further, there were no correlations between the cognitive complexity and number of incorrect operations or between the number of cognitive operations and the number of incorrect operations. This finding suggests that overall, producing a more complex essay does not make it more likely that a student will use more incorrect ideas, but as Table 7 shows, there may be specific operations that elicit more incorrect ideas. Further, writing more operations, or introducing more separate idea units, does not make a student more likely to put forth incorrect ideas.

Table 7. Number of incorrect cognitive operations relative to total number of operations [def.=definitions, obs.=observation, meas.=measurement, comp.=comparison, ex.=example, claim=claim, cons.=consequences, C&E=cause and effect, Ded.=deduction, Arg.=argumentation]

	Def.	Obs.	Meas.	Comp.	Ex.	Claim	Cons.	C&E	Ded.	Arg.
# incorrect	9	11	0	13	1	8	6	40	18	4
# operations	312	411	34	302	51	348	146	331	235	63

% incorrect per total ops.	3	3	0	4	2	2	4	10	8	6

Evident in Table 7 is a relatively infrequent use of scientifically inaccurate information. That is, given that our unit of analysis is ideas, students are largely generating scientific ideas employing correct scientific information. This further justifies the move beyond simply considering scientific accuracy of students' conceptions towards considering the sophistication of their reasoning about scientific ideas. In this case, only considering the accuracy would have provided a very limited picture of what these students were doing in their writing. Because of the relative infrequence, it became important to consider the nature of the inaccuracies. For this inquiry, categorizing the inaccuracies by operation led to an interesting finding. The highest percentages of inaccuracy, though still relatively small, occurred with *cause and effect*, *deduction*, and *argumentation*. It is possible that higher complexity operations surface alternative conceptions more effectively. Further, the alternative conceptions elicited are potentially more deeply held, keeping in mind the Limited or Inappropriate Propositional Hierarchies (LIPHs). That is, higher complexity operations draw on multiple domains and elements and may have the potential to reveal more of students' mental structures, and thus, expose LIPHs.

Limitations

Though this framework provides a useful way to evaluate students' written work, it has a number of limitations. First, as noted above, this framework does not capture the scientific accuracy of students' written ideas. The utility of this tool, then, is limited to a narrower research goal—characterizing students' reasoning. When combined with an analysis of the content accuracy, however, this framework can provide unique insights about students' understanding. Further, this framework was conceptualized, tested,

refined, and ultimately applied to a corpus of writing in a very specific context—general chemistry argumentative writing about ocean acidification. It is possible that some of the ways that cognitive operations have been conceptualized in this study are specific to this context. For this reason, applications to other contexts are needed to ensure the domain-general nature of this framework. Finally, due to the relatively low occurrence of certain operations in this context, we have a weaker understanding of some of the operations (i.e., measurement). Because of the complete absence of the inductive reasoning operation from Grimberg and Hand's original framework in this set of student writing, it was not included in this application, even though it is likely to be employed in other contexts. Finally, this data was collected at a selective institution and it is likely that different incorrect ideas or reasoning patterns would emerge from other student populations. Again, this can be addressed by applying this framework to student writing in other contexts.

Discussion and Implications

The first research question posed in this work considered how a cognitive operations framework can be used to characterize students' reasoning evident in their argumentative writing. In this article, we show what this framework is like and how it can be applied to students' writing. We refined a list of cognitive operations generated by Grimberg and Hand (2009) and organized them according to complexity, and then used these operations to code general chemistry students' writing on ocean acidification. This framework has some key affordances that make it useful to both research and practice. It is domain general, which means that it can be applied to writing in a variety of contexts. We recommend, then, that others apply this to writing in a variety of contexts across STEM and across levels (introductory to advanced student populations). The domain-general nature of this potentially enables the identification of

differences in students' reasoning across disciplines and levels. For example, do advanced students employ more complex reasoning than introductory students?

Another affordance of this framework is the 'score' that is a product of application—the cognitive complexity. The single score output provides an estimate of construct that is rather difficult to measure —student reasoning. This framework, then, can potentially overcome some of the difficulties with evaluating writing reported in the literature (Hamp-Lyon, 2016; Neill, 2002). This framework provides a novel approach to assigning a holistic score to writing. Further, the use of cognitive operations enables the identification of patterns in students' writing. That is, it can be used to characterize the movement between cognitive operations and the likelihood of moving towards high complexity operations, as shown in Grimberg and Hand's original application (2009). This framework's capacity to capture temporal patterns makes it very useful for understanding how students reason in extensive writing (Grimberg & Hand, 2009; Kelly & Takao, 2002; Moreira et al., 2018).

The second research question aimed to elucidate what features of student thinking were understandable with this framework. That is, what does this framework evaluate the quality of? This was achieved in two ways. The first was to compare framework estimates to instructor estimates of quality. This approach was intended to determine if the framework estimates were similar to the instructor estimates and if both were evaluating a similar construct. This revealed that perhaps for the upper bound of the construct—argumentation—there was agreement between instructors and framework estimates. There was less agreement for the other-than-best essays. Kelly and Takao (2002) identified similar disparities between their framework estimates and expert rankings and explained them as common occurrences when evaluating writing (Wolcott & Legg, 1998). In our case, we argue that the variety was an artefact of the

presence of inaccurate scientific information in one of the essays. Instructors may not separate content and reasoning as this framework does. However, we argue, similar to Kelly and Takao (2002), that this framework may provide a tool for evaluating the validity of instructor's estimates of quality. More research is necessary to establish interrater reliability amongst instructor ratings and identify ways in which the framework can serve as a tool for supporting instructors in systematically assessing students' writing.

To determine if this framework was providing unique information about students' ability, we compared cognitive complexity to other common performance measures. There were no correlations. We posit two potential explanations for this. The first is that this metric of cognitive complexity is indeed measuring something unique from what typical performance metrics measure (National Research Council, 2001). The second is that students who perform well on typical performance metrics do not necessarily perform equally well on more extensive writing tasks (National Research Council, 2001). Both of these explanations warrant further investigation because of the implications for assessment. Specifically, this framework could serve to equip the evaluation of more interesting competencies in students than that measured by typical performance measures *or* assignments of this nature could serve to minimize advantages certain groups bring with them to typical performance measures. However, we also recognize that there may be other performance measures that correlate with the framework estimate. Particularly, we would expect that more generative or authentic assessments might correlate more strongly with cognitive complexity (National Research Council, 2001). Finally, we aimed to characterize the relationship between framework estimates of quality and scientific accuracy. In order to do this, we analysed writing for the presence of scientific inaccuracies and coded the respective operation in

which they appeared. This revealed that scientific inaccuracies occurred relatively infrequently with about 10 percent of *cause and effect* operations including something that did not agree with scientifically accepted knowledge. The percentage among *cause and effect* operations was the highest. However, there appears to be a trend in which higher complexity operations (i.e., cause and effect, deduction, and argumentation) had higher frequencies of incorrect information than low complexity operations. We argue, in light of Novak's work on LIPHs, that higher complexity operations as a representation of students' mental models may better reveal LIPHs (Novak, 2002). That is, employing more complex reasoning may surface more deeply held LIPHs. Students who do not use higher complexity operations may be more limited in both their and their instructors' capacity to address potential alternative conceptions. The relationship between complexity and conceptual correctness warrants further investigation. Understanding this relationship is important for designing formative assessments that better elicit high complexity operations.

This framework also offers some unique implications for instructors who assign similar tasks to their students. Scoring assignments in this way could permit an instructor to draw conclusions about their students' collective access to complex reasoning operations. For example, a low average score of cognitive complexity in their course may motivate instructors to explicitly address and model complex reasoning types for their students. However, we argue that the most important implication of this framework for practice is providing a vocabulary to instructors for giving tailored feedback to students. That is, applying this sort of framework would support an instructor to give specific examples of when a student could have employed more complex reasoning appropriately and instead used a less complex operation.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Emig, J. (1977). Writing as a Mode of Learning. *College Composition and Communication*, *28*(2), 122–128.

Gere, A. R., Limlamai, N., Wilson, E., MacDougall Saylor, K., & Pugh, R. (2019). Writing and Conceptual Learning in Science: An Analysis of Assignments. *Written Communication*, *36*(1), 99–135. https://doi.org/10.1177/0741088318804820

Grimberg, B. I., & Hand, B. (2009). Cognitive Pathways : Analysis of students ' written texts for science understanding. *International Journal of Science Education*, *31*(4), 503–521. https://doi.org/10.1080/09500690701704805

Gunel, M., Hand, B., & Prain, V. (2007). Writing for learning in science: A secondary analysis of six studies. *International Journal of Science and Mathematics Education*, *5*, 615–637. https://doi.org/10.1007/s10763-007-9082-y

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying computerized-scoring models ofwritten biological explanations across courses and colleges: Prospects and limitations. *CBE Life Sciences Education*, *10*(4), 379–393. https://doi.org/10.1187/cbe.11-08-0081

Hein, V., & Smerdon, B. (2013). *Predictors of Postsecondary Success*. *College and Career Readiness and Success Center at American Institutes for Research*.

Kelly, G. J., & Bazerman, C. (2003). How Students Argue Scientic Claims : A Rhetorical-Semantic Analysis. *Applied Linguistics*, *24*(1), 28–55.

Kelly, G. J., Chen, C., & Prothero, W. (2000). The Epistemological Framing of a Discipline : Writing Science in University Oceanography, *37*(7).

Kelly, G. J., Regev, J., & Prothero, W. (2007). Analysis of Lines of Reasoning in Written Argumentation. In *Argumentation in Science Education* (pp. 137–157).

Kelly, G. J., & Takao, A. (2002). Epistemic levels in argument: An analysis of university oceanography students' use of evidence in writing. *Science Education*, *86*(3), 314–342. https://doi.org/10.1002/sce.10024

Keys, C. W. (1994). The development of scientific reasoning skills in conjunction with collaborative writing assignments: An interpretive study of six ninth-grade students. *Journal of Research in Science Teaching*, *31*(9), 1003–1022. https://doi.org/10.1002/tea.3660310912

Klein, P. D. (1999). Reopening Inquiry into Cognitive Processes in Writing-To-Learn. *Educational Psychology Review*, *11*(3), 203–270. https://doi.org/10.1023/A:1021913217147

Klein, P. D. (2015). Mediators and Moderators in Individual and Collaborative Writing to Learn. *Journal of Writing Research*, *7*(1), 201–214.

Krippendorff, K. (2004). Reliability in Content Analysis : Some Common Misconceptions and Recommendations Reliability in Content Analysis : Some Common Misconceptions and. *Human Communication Research*, *30*(3), 411–433.

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, *53*(2), 215–233. https://doi.org/10.1002/tea.21299

Moon, A., Gere, A. R., & Shultz, G. V. (2018). Writing in the STEM classroom: Faculty conceptions of writing and its role in the undergraduate classroom. *Science Education*, *0*(0). https://doi.org/10.1002/sce.21454

Moreira, P., Marzabal, A., & Talanquer, V. (2018). Using a mechanistic framework to characterise chemistry students' reasoning in written explanations. *Chemistry Education Research and Practice*. https://doi.org/10.1039/C8RP00159F

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. *National Academies Press*. Washington DC. https://doi.org/10.17226/10019

Novak, J. D. (2002). Meaningful Learning: The Essential Factor for Conceptual Change in Limited or Inappropriate Propositional Hierarchies Leading to Empowerment of Learners. *Science Education*, *86*(4), 548–571. https://doi.org/10.1002/sce.10032

Prain, V., & Hand, B. (2016). Coming to Know More Through and From Writing. *Educational Researcher*, *45*(7), 430–434.

https://doi.org/10.3102/0013189X16672642
Reynolds, J. A., Thaiss, C., Katkin, W., & Thompson, R. J. (2012). Writing-to-learn in undergraduate science education: A community-based, conceptually driven approach. *CBE-Life Sciences Education*, *11*(1), 17–25.
Sandoval, W. A. (2003). Conceptual and Epistemic Aspects of Students ' Scientific Explanations. *Journal of the Learning Sciences*, *12*(1), 5–51. https://doi.org/10.1207/S15327809JLS1201
Sandoval, W. A., & Millwood, K. A. (2005). The Quality of Students ' Use of Evidence in Written Scientific Explanations. *Cognition and Instruction*, *23*(1), 23–55. https://doi.org/10.1207/s1532690xci2301
Sandoval, W. A., & Reiser, B. J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education*, *88*(3), 345–372. https://doi.org/10.1002/sce.10130
Sevian, H., & Talanquer, V. (2014). Rethinking chemistry: a learning progression on chemical thinking. *Chemistry Education Research and Practice*, *15*(1), 10–23. https://doi.org/10.1039/C3RP00111C
Takao, A. Y., & Kelly, G. J. (2003a). Assessment of Evidence in University Students ' Scientific Writing. *Science & Education*, *12*, 341–363. https://doi.org/10.1023/A
Takao, A. Y., & Kelly, G. J. (2003b). Assessment of Evidence in University Students ' Scientific Writing. *Science & Education*, *12*, 341–363.

Application and testing of a framework for characterizing the quality of scientific reasoning in chemistry students' writing on ocean acidification

Alena Moon, Robert Moeller, Anne Ruggles Gere†, and Ginger V. Shultz

Department of Chemistry, University of Michigan, Ann Arbor, MI, USA
†Sweetland Center for Writing, University of Michigan, Ann Arbor, MI, USA

Corresponding author: Ginger V. Shultz, gshultz@umich.edu, 2521 Chemistry, University of Michigan, Ann Arbor, MI, 48109.

Provide short biographical notes on all contributors here if the journal requires them.

Development and testing of a framework for characterizing the quality of scientific reasoning in students' writing on ocean acidification

Science educators recognize the need to teach scientific ways of knowing and reasoning in addition to scientific knowledge. However, characterizing and assessing scientific ways of knowing and reasoning is challenging. Writing-to-learn offers one way of eliciting and supporting students' reasoning; further, writing serves to externalize and make traceable students' reasoning. For this reason, it is a useful formative assessment of scientific reasoning. The utility hinges on researchers' ability to understand what students can do and think from their writing. Given the challenges in assessing students' writing, this research offers an adapted framework for assessing students' scientific reasoning evident in writing. This work will introduce the adapted framework and show an application to general chemistry students' argumentative writing about ocean acidification. We provide evidence that this framework can be used to validly estimate the quality of students' reasoning. We argue that this framework offers some affordances that overcome challenges reported in the literature. It serves to define scientific reasoning in a domain-general way by breaking it down into its components, but in a way that can produce a composite score that tells us about how students reason using chemistry content. Further, the framework provides a way to characterize the scientific accuracy of students' reasoning that can inform instructors' treatment of alternative conceptions.

Keywords: Writing-to-learn; assessing writing; scientific reasoning

Background

Science educators recognize that it is insufficient to only teach students' scientific knowledge as a collection of concepts and topics. Rather, to enable students to use scientific knowledge, we must support the development of reasoning and thinking skills that scientists use (NRC, 2012; Sevian & Talanquer, 2014; Bulte, Westbroek, De Jong, & Pilot, 2006). Writing-to-learn (WTL) is one way of supporting the development of this skill by activating deep thinking and reasoning in students (Keys, 1999) and, more importantly, making that reasoning visible and traceable (Emig, 1977; Kelly & Takao,

2002; Kelly, Regev, & Prothero, 2007). From an assessment perspective, this evidence of student reasoning is valuable in so far as researchers and practitioners can use it to make an argument about students' abilities to reason scientifically (Laverty et al., 2016; NRC, 2001). However, there are challenges that currently limit the utility of this evidence. There are few widely agreed upon epistemic criteria for characterizing the quality of students' reasoning (i.e., what makes one students' reasoning better than another's). Further, actually applying these criteria to understand and evaluate students' writing is difficult as writing requires the researcher to make choices about grain size, whether to evaluate structure or content or both, and what the presence or absence of a quality criterion actually looks like in students' writing (Kelly & Takao, 2002; Takao & Kelly, 2003a). To address these challenges, we have modified and applied a framework for characterizing and evaluating reasoning in students' argumentative writing. This framework contributes meaningfully to efforts to conceptualize and evaluate scientific reasoning, as well as to efforts to analyse writing, which poses unique challenges.

Writing to Learn

Writing-to-learn refers to the kind of informal writing about science that facilitates learning and ownership of scientific ideas. This informal writing is distinct in that its primary aim is not to communicate or display mastery to an instructor, but to actually facilitate sense-making by activating deep thinking and interaction with the concepts (Keys, 1994). A secondary benefit of writing-to-learn, then, is promoting engagement with disciplinary norms of writing and thinking (Prain & Hand, 2016). There is quite a bit of variation around this primary aim, however; WTL assignments take a variety of forms, lengths, methods of text production, audiences, and genres (Keys, 1994).

A secondary analysis of six writing-to-learn studies revealed some promising gains as a result of writing-to-learn—the treatment condition outperformed comparison

groups on a total test scores and conceptual question scores and this effect was largely due to the treatment (Gunel, Hand, & Prain, 2007). All six studies followed a similar design including a pre-test/post-test design with the test having multiple-choice and conceptual extended response questions. More importantly, all writing interventions were grounded in the same theoretical considerations that have been identified as key for successful learning from writing: 1) opportunities for brainstorming, 2) provision of authentic audiences, 3) drafting and redrafting with feedback, 4) explicit instruction of genre specifications, 5) focus on big ideas, 6) use of rubrics, and 7) diverse opportunities to plan and draft writing (Gere, Limlamai, Wilson, MacDougall Saylor, & Pugh, 2019; Gunel et al., 2007; Klein, 1999, 2015). The theoretical grounding afforded comparisons across domains and writing assignment types and served to reveal the benefits of WTL more broadly (Gunel et al., 2007; Prain & Hand, 2016). However, at the undergraduate STEM level specifically, more work is needed to understand the mechanism of effect for WTL assignments (Reynolds, Thaiss, Katkin, & Thompson, 2012) and we argue that to undertake investigations into the mechanism of effect, we need a reliable and meaningful framework for interpreting and evaluating students' written work.

Characterizing Students' Reasoning in Written Products

Constructed responses reveal rich insight into students' ideas and the coherence of those ideas, but evaluating open responses remains a barrier to implementing such rich assessments (Liu, Rios, Heilman, Gerard, & Linn, 2016). This barrier consists of two distinct but interdependent challenges: characterizing the quality usually in some sort of rubric (Kelly & Bazerman, 2003; Sandoval, 2003; Sandoval & Millwood, 2005) and consistently and reliably applying the quality criteria (Ha, Nehm, Urban-Lurain, & Merrill, 2011; Liu et al., 2016). Additionally, the nature and difficulty of these

challenges vary with the length of constructed response; for example, extended arguments in the form of research reports in oceanography tended to have long and complex chains of reasoning that are difficult to characterize with a single rubric (Kelly, Regev, & Prothero, 2007; Kelly & Takao, 2002). Researchers have sought to overcome these challenges with a variety of approaches for a variety of written products, ranging from short written explanations (Ha et al., 2011; Liu et al., 2016; Moreira, Marzabal, & Talanquer, 2018; Sandoval, 2003; Sandoval & Millwood, 2005; Sandoval & Reiser, 2004) to more extensive writing like laboratory reports or research reports (Grimberg & Hand, 2009; Kelly, Chen, & Prothero, 2000; Kelly et al., 2007; Kelly & Takao, 2002; Takao & Kelly, 2003b). A few of these approaches are distinct in that they break down students' responses into smaller units to then identify patterns in students' reasoning, as opposed to a rubric that considers the quality of the response as a whole (Grimberg & Hand, 2009; Kelly et al., 2007; Kelly & Takao, 2002; Moreira et al., 2018). These approaches will be the focus of this literature review.

Moreira et al. (2018) specifically sought to characterize the causal reasoning of 10th grade chemistry students' explanations of freezing point depression. To do so, the authors modified and applied a discourse analysis framework developed by Russ et al. (2008) to students' written explanations and drawings. The final form of the analysis scheme included four components—entities, properties, activities, and organisation—and the relationships between the components and the students' drawings that were identified in students' responses. Entities are the 'things' in the system that are being considered. Properties are characteristics of those entities and activities are actions of those entities. Organisation refers to the spatial-temporal relationship between the entities and activities or properties of the system. By coding the explanations for these components and relationships, they were able to elucidate patterns in students'

explanations and organize these patterns into four levels according to the quality of causal reasoning. These levels increased in sophistication from descriptive to relational to simple causal, culminating in emerging mechanistic. Unsurprisingly given the authors' previous findings (Sevian & Talanquer, 2014), the majority of students (45%) used relational causal reasoning (Moreira et al., 2018). Explanations at this level could be modelled to show that students generally identified two entities and the properties of one or both entities, and then related either the properties to each other or related entities to properties. Such complex modelling of students' explanations can hopefully equip teachers with more sophisticated approaches to understanding, interpreting, and developing students' reasoning abilities (Moreira et al., 2018).

Grimberg and Hand (2009) similarly identified the presence or absence of dimensions of reasoning and determined *patterns* in students' reasoning evident in their laboratory reports. In this study, Grimberg and Hand (2009) identify cognitive operations used in writing laboratory reports and then construct what they term 'cognitive pathways'—the sequence of cognitive operations used by author(s) of a lab report. The authors argued that because writing a laboratory report was a meaning-making activity, considering the sequence of cognitive operations revealed how students were constructing meaning. Using a list of 11 cognitive operations derived partially from the literature and from the students' data, authors coded students' writing for use of cognitive operations. The cognitive operations included observation, measurement, comparison, analogy, clarifications, claim, cause/effect, induction/generalization, deduction, investigation design, and argumentation (Grimberg & Hand, 2009). When comparing cognitive pathways of low achievers to high achievers, as determined by a standardized skills test, both high and low achievers used the same range of operations, but with a different structure. Though the cognitive structure was partially determined

by the structure of the Science Writing Heuristic (SWH) activity (i.e., clarification questions were posed during the research question portion of SWH activity), high achievers began using complex operations earlier in the text than low achievers. This ultimately demonstrated that SWH scaffolding supported all students in using high-complexity operations, albeit at different rates (Grimberg & Hand, 2009).

Grimberg and Hand's (2009) work demonstrates the capacity of writing to make students' thinking visible and traceable. Emig (1977) argues that writing is unique in its capacity to do this. Kelly and Takao (2002) demonstrate the utility of students' writing for understanding their reasoning by characterizing how undergraduate students use evidence to construct arguments. To make this characterization, they developed a research methodology that models epistemic levels of argument. This framework includes six levels ranging from the lowest, data charts and representations, to the highest, general geological (the specific context for this work) knowledge not specific to the data presented. The levels represented students' ability to abstract from data to make claims. With this framework, the authors analysed a subset of undergraduate oceanography students' assignments by labelling each sentence with an epistemic level. These epistemic criteria were then weighted in order to rank the 24 student arguments (research reports in oceanography) from best to worst. While this framework was very useful for characterizing the quality of students' arguments based on their use of evidence, it did possess limitations. Namely, the assessment of quality determined by the framework did not always align with content experts' evaluation of quality, the framework did not consider inference logic (how the data led to theoretical claims), and the authors had to make inferences in their application of the framework. These limitations are difficult to overcome for everyone aiming to evaluate ill-defined constructs like use of evidence. However, Kelly and Takao (2002) revealed that claims

can be made about students' reasoning from their written work. In order for the insight into students' reasoning provided by writing to be useful for informing students' development, tools for assessing writing must be efficient, systematic, and offer tailored feedback.

Ongoing discussions about writing assessment distinguish between holistic scoring, assigning a single score to a broad variable like writing proficiency, and analytic scoring, breaking down variables like writing proficiency into components that are individually scored (Hamp-Lyons, 2016a, b). High-stakes assessments, such as college entrance examinations, motivated the use of both holistic and analytic scoring approaches to assign students general writing scores, but researchers have begun to identify shortcomings of both (Neill, 2002; Hamp-Lyons, 2016b, Chapman, 2016). With more complex analytical tools (i.e., multivariate analyses), Hamp-Lyons (2016) calls for a movement to multiple trait scoring of writing. In multiple trait scoring, there is no single score given, whether composite or holistic. Rather, a set of scores is assigned with multiple traits each warranting a score, thus lending to a richer description of students' ability (Hamp-Lyons, 2016a, b).

Rationale and Research Objectives

In any effort to measure a student's reasoning, choices must be made about what characterizes quality. Specific to evaluating extensive writing, additional decisions must be made about the grain size, level, and nature of rubric that will be used to determine quality. In order to address these challenges, the cognitive operations used by Grimberg and Hand (2009) were modified and applied to students' writing on ocean acidification. Motivated to leverage the rich insight into student thinking that constructed responses offer, the work presented herein aimed to test an approach for analysing extensive

scientific writing by applying it to a new context. We aimed to answer the following questions regarding this approach:

(1) Can cognitive operations be used to make sense of general chemistry students' argumentative writing? If so, how?

(2) What features of students' argumentative writing do cognitive operations serve to explain?

(3) What is the relationship between framework estimates of quality and conceptual correctness?

Methods

Participants, setting, and data collection

A writing prompt was designed and administered in a first-semester General Chemistry course serving primarily students in the College of Engineering and undeclared students in the College of Literature, Science, and Arts. This course had an enrolment of 1413 students, most of whom were freshman and sophomores. The content of the course covered traditional general chemistry concepts, ranging from dimensional analysis, quantum mechanical atomic models, bonding theories, to reactions, enthalpy, intermolecular forces, chemical equilibrium, and acid-base theories.

This course is structured with three lectures led by an instructor and one discussion session led by a teaching assistant per week. During each discussion section, students complete a quiz. The writing assignment for this study was administered as a substitute for a quiz. The writing assignment was uploaded as a .pdf file to the course management site one week before the due date. This writing assignment directed students to consider a set of concepts in their response and to keep their post between

350 and 500 words. Though the majority of students' responses were within this range, some wrote less and some wrote more.

Students all submitted their writing assignment online the following week at the start time of their specific discussion session. For this reason, some students with discussion sessions later in the week had more time to write than those with earlier recitations. During the discussion, students formed teams of two or three, switched papers and gave feedback to each other. Six hundred seventy-three students gave consent to have their writing analysed. Ethical review board approval was gained in order to collect and analyse written assignments that students consented to have analysed. Students did not receive any feedback on their written work beyond the conversation that took place in their discussion session. We found that students did not make meaningful revisions following the peer review discussion. Additionally, they were not required to submit a revision. However, if they did, that revised draft was used for analysis.

Writing activity development and design

The writing prompt was developed iteratively through correspondence with authors who had expertise in writing to learn and the development of meaningful writing prompts (AG) and the faculty members teaching the course who collectively held more than two decades of experience teaching chemical equilibrium in general chemistry. WTL prompts are generally designed to provide students with an audience, an identity, and an authentic context that require students to engage with a specific concept. This WTL assignment was intentionally designed with elements empirically determined to contribute to meaningful learning through writing (Gere et al., 2019). In this case, the prompt showed a fake social media post, in which 'Ernie Clueless' shares a plot illustrating the trend of concentration of atmospheric carbon dioxide and ocean pH over

time. Ernie claims that these things are unrelated. Students were tasked with explaining the relationship to Ernie, given the relevant equilibria. The prompt targeted the concept of chemical equilibrium, drawing on Le Châtelier's principle. It inherently supported argumentation by requiring students to differentiate their perspective from Ernie's.

Data analysis: Development and application of analytical framework

A list of cognitive operations was modified from a list used by Grimberg and Hand (2009) to analyse reports from a Science Writing Heuristic (SWH) laboratory. In this context, a cognitive operation is a written discursive move that serves some cognitive objective. Cognitive operations, then, determined the grain size for breaking down an essay into smaller analysable units (i.e. the number of sentences that served the objective of a claim, for example, were coded as such). For this reason, a claim could be one sentence in one essay and three sentences in another. The amount of text that was assigned a code was determined by the function of that text. The list of cognitive operations used by Grimberg and Hand (2009) was refined iteratively by testing it against the data. This involved using Grimberg and Hand's original set to code the text, identifying text that could not be coded with this set and operations from this set that did not serve to explain any of the data, and refining the set of operations to a set that were all used to describe virtually all of the text. Multiple initial iterations occurred with the same subset of essays (N=25) and subsequent iterations incorporated more essays on an as needed basis. Table 1 shows the final list of cognitive operations that was used throughout the final analysis. Included in Table 1 is a characterization of the dimensionality of each operation. As will be explained further in the theoretical framework, the complexity of cognitive operations was determined by its dimensionality—number of ideas being drawn upon. Operations with two domains were

more cognitively complex than operations with one dimension. That is, they drew upon and connected more idea units.

Table 1. Finalized list of cognitive operations and descriptors used to analyse all essays, listed in order of increasing cognitive complexity

Cognitive Operation	Description	Dimensionality
1. Definition	Canonical description of a term, concept, idea, or theory	single domain
2. Observation	Qualitative description of change, trend, or transformation of a variable	
3. Measurement	Quantitative description of change, trend, or transformation of a variable	
4. Comparison	Relationship of change, trend, or transformation for two or more variables	
5. Example	Illustration of a class of objects by singling out one object	two domains
6. Claim	Assertion supported with a tentative explanation	
7. Consequences	Cause and effect explanation with either cause or effect falling outside the scope of the writing prompt	
8. Cause and effect	Explanation providing a mechanism with a causal agent and observed effects	
9. Deduction	Application of a theory or principle to a specific system or scenario	multiple domains
10. Argumentation	Explicit differentiation between the author's perspective and the fictional character's perspective*	

*This conceptualization of argumentation was specific to the context of writing prompt used in this study. It is expected that it could be easily translated to other writing contexts that require argumentation.

Once a list of cognitive operations was finalized, the first two authors began coding assignments and built a detailed rubric that included definitions, linguistic markers, and examples for each operation. This rubric was further refined through multiple iterations of analysis by a team of chemistry education researchers in an effort to establish inter-rater reliability (IRR). This stage included a team consisting of the first two authors and another chemistry education researcher trained in qualitative coding of writing. The graduate student was trained on an existing rubric, the whole team coded ten assignments, and an IRR coefficient in the form of Krippendorff's alpha (KA) was calculated. This coefficient was quite low after the first round, so revisions were made to the rubric, another training session was conducted, and subsequent rounds of analysis

were conducted until a Krippendorff's alpha value of 0.69, the minimum acceptable value, was achieved. Training involved discussing the rubric which included examples from many essays and then illustrating the coding process by coding a few essays all together.

The first two authors then coded approximately 200 assignments each, with large overlap. Having two researchers code many of the same assignments lent to the reliability of the coding. Throughout analysis, IRR 'checks' were performed to ensure that the rubric was being applied consistently. This involved selecting overlapping assignments and determining a KA. This stage resulted in a KA of 0.89, which was a desirable value (Krippendorff, 2004).

Estimate of quality

Once all assignments were coded for cognitive operations, estimates of quality were assigned according to the cognitive complexity of the essay. The cognitive operations are ordered in Table 1 according to increasing cognitive complexity; that is, a definition has the lowest cognitive complexity (1) whereas argumentation has the highest cognitive complexity (10). Overall cognitive complexity for the essay was determined by taking a weighted average of operations used. Because the magnitude of text included in an operation varied from one essay to another, the average was weighted by the number of sentences within that operation. An assignment, then, that was 50% argumentation would likely have a higher average complexity than an assignment with only 10% argumentation. This approach resulted in a single number characterizing the quality of students' reasoning in an essay. This process of producing a single number is illustrated with the example essay in Table 2.

Table 2. Example of student essay and determination of cognitive complexity score

Le Chatlier's principle is a way to show how if the equilibrium of a certain reaction is altered by changing certain aspects, then the reaction or equilibrium position will actually shift to fix the change. In Ernie's post there are multiple reactions that follow one after the other, ultimately showing how CO_2 transforms to H_2CO_3, which changes to H^+ and HCO_3, which finally changes to 2H+. These reactions are connected in that as one breaks down, it spurs another reaction to then form. This shows how CO_2 in the atmosphere actually affects PH concentration and ocean acidification (or the number/concentration of H+ ions in the reaction). Because of this correlation, certain components like temperature, pressure, volume, and concentration can affect CO_2 levels and acidification. In our case, we will look at the concentration or amount of CO_2 in correlation to the amount of H^+ that is formed. If the concentration of CO_2 (as a reactant) were to increase, the system would try to decrease it; this would mean that the concentration of the other reactants will increase to react with the CO_2, but then more product would be formed. Thus increasing the concentration of CO_2 (atmos) would cause the system to shift towards the products, producing more CO2(aq). Then because there is more CO_2 (aq), more water would have to be use so more H_2CO_3 would form. The next reaction would then proceed as the others with an increase in the formation of the products H^+ and HCO_3, which in turn would once again increase the next reaction producing more $2H^+$ product. This may seem confusing, but to summarize if CO_2 as a reactant increases, it would need to counteract this change by producing more product of $2H^+$ (acidification). Now, if CO_2 (atmos) reactant were to decrease, the equilibrium will actually try to increase it so as to set the system back at equilibrium. This would mean that a decrease in that reactant would cause more product to be formed so to produce more CO_2; thus $2H^+$ product would increase to make more CO_2. The relationship between CO_2 and pH is that as CO_2 increases, the pH will decrease and make the oceans more acidic. The pH scale runs from 1 to 14, with the acids being numbers 1 to 6. So the lower the pH, the more acidic and the more H^+ ions that are formed (the H^+ are indicators of acidic properties). Le Chatlier's principle shows that increasing one will have to increase the other so that the system is able to be equal again, thus there is correlation between CO_2 (atmos) increase and H^+ ions increase(acidification)."	**Definition (1)** **Observation (2)** **Claim (6)** **Deduction (9)** **Comparison (4)** **Definition (1)** **Comparison (4)** **Cause and Effect (8)**
Cognitive complexity = $\frac{\Sigma(cognitive\ operation\ score * \#\ sentences\ used)}{total\ number\ of\ sentences}$ Definition (1) + 3 * Observation (2) + Claim (6) + 9 * Deduction (9) + Comparison (4) + Definition (1) + Comparison (4) + Cause & Effect (8) = 111 111/18 =6.2 (Average complexity for this essay)	

Faculty ranking of essays

One motivation of this work is to use a framework to systematically characterize students' writing. In order to understand how this framework could accomplish that task, we compared framework estimates of quality with instructors' estimates of quality. The precedent for this approach is offered by Kelly and Takao's testing of the framework for epistemic levels (Kelly & Takao, 2002). Further, evidence from interviews with STEM faculty about writing suggested to us that the knowledge experts use to evaluate writing is tacit (Moon, Gere, & Shultz, 2018). By comparing expert rankings with essays ordered according to a framework, we can identify ways that the framework may capture this tacit knowledge. To compare framework estimates of quality and instructor estimates, we compared essays ranked according to the framework to instructor rankings. To select the essays, we split cognitive complexity into four roughly equivalent ranges (3.1-4.8, 4.8-6.5, 6.5-8.2, 8.2-10), where the highest range included the 'best' essays, according to the framework. Within each range, an essay was randomly selected. The output of this step was four essays ranging from most complex to least complex, as determined by the framework. These essays were then provided to instructors who were tasked with ranking the essays from best to worst according to the quality of scientific reasoning, which they were directed to evaluate as they saw fit. The rationale for not providing faculty with more extensive quality criteria was to target the kind of evaluative work that is inherent to grading this sort of task; that is, we wanted instructors to make the kinds of decisions that are required to define and evaluate scientific reasoning. Five faculty with a range of experience teaching general chemistry from different institutions ranked the writing tasks and one instructor volunteered the reasoning behind their ranking.

Chemistry Content Analysis

Essays were examined to understand how students employed ideas about chemical equilibrium within their argument. Essays were coded for conceptual correctness, which involved flagging all occurrences of inaccurate ideas. Every time an inaccurate idea was identified, the cognitive operation containing that incorrect idea was marked. The pairing of the incorrect idea with the cognitive operation was intentional, based on the theoretical framework explained below, which posits that what students articulate in written forms are representations of the ideas they hold. In this case those ideas related to chemical equilibrium.

Theoretical framework

The primary assumption made in this work is that writing reveals students' cognitive structures or understanding of the meaning of a concept (Emig, 1977; Novak, 2002). Meaning is defined in this case as 'the totality of propositions linked to any given concept,' excluding the emotional association with the concept and the context in which the concept was learned (Novak, 2002). This assumption is grounded in the capacity of writing to 1) connect or relate propositions in the author's mind and 2) make 'evolutionary development of thought graphically visible and available (Emig, 1977).' According to Novak's theory of meaningful learning, the complexity of the meanings can be evaluated, the quantity and quality of which will determine meaningful learning (Novak, 2002).

In this work, units of written text were coded according to their cognitive function as a cognitive operation. Each cognitive operation was constituted by some number of ontological domains, elements within those domains, and relationships between each (Grimberg & Hand, 2009; Halford & McCredden, 1998). Cognitive complexity, then, was defined in terms of dimensionality; in which cognitive operations with higher numbers of domains, elements, and relationships were more complex than

cognitive operations with fewer (Halford & McCredden, 1998). These domains and elements can be conceptualized similarly to the discourse analysis framework used by Moreira et al. (2018) where they refer to 'things' in the system being considered. All cognitive operations were organized on a spectrum from least complex to most complex according to this criterion, illustrated in Table 1. A students' scientific reasoning can then be considered as the progression of operations used in a text.

Therefore, one way of evaluating the quality of meanings was through cognitive complexity. It is possible, however, that a student can employ cognitively complex reasoning without necessarily using correct content (Sandoval & Millwood, 2005; Kelly & Takao, 2002). For this reason, a second way of evaluating the quality of meanings was considering their conceptual correctness; that is, their agreement with scientifically accepted knowledge. A conceptual change perspective suggests that problems of incorrect conceptions arise from Limited or Inappropriate Propositional Hierarchies (LIPHs), the way that concepts are inappropriately organized in the learner's mind, which means that we as instructors must consider both the *content and structure* of incorrect conceptions. Further, this implies a greater instructional effort is needed to remediate stable LIPHs (Novak, 2002). The cognitive operations framework used in this study helps characterize the complexity or stability—as a function of the number of domains, elements, and relationships—of conceptions.

Results

Research Question 1: Can cognitive operations be used to make sense of general chemistry students' argumentative writing? If so, how?

Table 2 illustrates how these cognitive operations were interpreted to analyse the writing with examples from students' essays. The examples are useful for discussing the difficulties of applying this framework. The lower complexity operations (*definition*

through *example*) were relatively easy to identify. The difference between *observation* and *measurement* was essentially a difference between qualitative and quantitative, with *measurements* requiring some numerical component. Because students could sufficiently respond to this writing prompt with qualitative reasoning, *measurement* occurred less frequently (Table 1). A *comparison* was distinct from *observation* and *measurement* in describing more than one variable relative to each other. A *claim* was similar to *comparison* in referencing a relationship between two variables but was distinct in that it required a tentative explanation of the relationship. For the *claim*, then, there were two domains (explanation and observation). *Cause and effect* and *consequences* were similar and easier to identify in text, with common linguistic markers being 'caused,' 'leads to,' or 'drives.' *Consequences* used cause and effect reasoning but relied on causes or effects that fell outside the scope of the prompt. In this case, the student referenced the effect of ocean acidification on coral. A primary marker of *deduction* was the invocation of a principle or theory that was then applied to a specific system. In this context, students frequently invoked Le Chatalier's principle or equilibrium. Finally, *argumentation* was undoubtedly the most difficult to identify as it often drew on multiple operations. So, we used this feature as an identifier. *Argumentation*, then, required indistinguishable use of multiple operations and to explicitly differentiate between the author's perspective and Ernie's (or any opposing position in another context). Table 1, 2, and 3 together can be used to apply this framework to other contexts.

Table 3. Examples of cognitive operations, arranged in order of increasing complexity. Numbers describe cognitive complexity ordering.

Operation	Example from student essay
Definition (1)	Equilibrium is a state where a reaction is occurring forwards and backwards at equal rates with no overall change. When a change occurs to the system, the reaction will shift in a direction to counteract this change.

Observation (2)	The trend in the graphs shows an increase in atmospheric carbon dioxide over time.
Measurement (3)	The ocean pH has dropped from 8.2 to 8.1 since the Industrial Revolution.
Comparison (4)	The plot that you shared illustrates that as the concentration of atmospheric increases over time, the pH of the ocean seawater decreases.
Example (5)	An example is hydrochloric acid, which has hydrogen ions attached and is an acid with a rather low pH level.
Claim (6)	As a matter of fact, Ernie, the correlation between CO_2 levels in the atmosphere and the pH of the oceans makes sense according to chemistry.
Consequences (7)	When this happens, calcifying organisms will become weaker, such as coral. They will be significantly affected as a result and may be unable to live in the current environment in which they live. In addition to ocean acidification wreaking havoc on the environment, other factors such as climate change can do the same thing and increase the amount of damage that is done to it.
Cause and Effect (8)	In each equation, the amount of reactants increases, which drives the reaction forward, meaning the amount of products will increase until the reaction reaches equilibrium.
Deduction (9)	In accordance with Le Châtelier's Principle, increasing the amount (or the concentration) of atmospheric CO_2 will shift this equation towards the dissolved CO_2 in the ocean to make up for the increase in gas on the reactants (left) side. This dissolved CO_2 is indicated as 'CO_2 (*aq*)' in the equation, which stands for 'aqueous CO_2.' Consequently, the dissolved CO_2 relates to the bicarbonate formation equation: CO_2 (*aq*) + H_2O (*l*) ⇄ H_2CO_3 (*aq*) ⇄ H^+ +HCO_3 (*aq*) (Doney et al., 2009 [*reference provided to student*]) Just as increasing the concentration of the atmospheric CO_2 caused the equilibrium between dissolved CO_2 to favor the formation of dissolved CO_2, a similar phenomenon will occur to favor production of bicarbonate (HCO_3) and hydrogen, two byproducts of carbonic acid (H_2CO_3). Increasing dissolved CO2 concentration—a result of increased atmospheric CO_2—will 'push' the equilibrium towards the formation of H_2CO_3. Likewise, a shift towards the formation of H_2CO_3 will also shift the equation forward towards the formation of protons, H^+, and HCO_3. Finally, these products of proton and bicarbonate will be in equilibrium with two protons and carbon trioxide (CO_3^{2-}): H^+ + HCO_3 (*aq*) ⇄ $2H^+$ + CO_3^{2-} (Doney et al., 2009)

	Once more, an increase in the H^+ and HCO_3 concentration will push the equilibrium forward towards the formation of two $2H^+$ and CO_3^{2-}.
Argumentation (10)	So using this information, you can now look at the graph you posted and understand the relationship between CO_2 levels and the pH of the water. As CO_2 is absorbed into the water, it produces H+ ions, which then cause the pH of the water to decrease. You can see this trend on the graph. Even though they do not seem like they should be related in any way, a change in one would cause a change in the other. Something that is making the line representing the CO_2 in the atmosphere on the graph to increase so much is the amount of CO_2 humans emit every day. Whenever you drive a car you are releasing CO_2 into the atmosphere. Since there is so much more CO_2 being released, the oceans are absorbing more CO_2. In fact, the oceans have absorbed almost 30% of the CO_2 humans have emitted since the Industrial Revolution. As more CO_2 is absorbed by the oceans because of human activity, the more H+ ions are formed, and the more the pH of the ocean is decreased.

A total of 296 assignments have been coded (average weighted complexity: 6.3; average number of operations per essay: 9). We determined that saturation had been reached when no new codes or patterns were observed in the writing. Table 4 shows the frequency of operations used in the 296 essays. *Observation* was the most frequently used operation by students, with many using more than one observation in a single essay. Students heavily relied on making statements about how a variable was changing to counter 'Ernie's' claim. In this context, this means that students were able to understand how variables were changing from the graph provided. There was very little use of *Example* or *Measurement*. *Argumentation* was also used relatively infrequently. As mentioned above, *argumentation* was distinct as its own operation given indistinguishable use of multiple lower complexity operations. For this reason, *argumentation* required students to combine multiple operations (and domains). This difficulty combined with the infrequent use suggests that *argumentation* was indeed the most complex operation. The high frequency of *claim*, *cause and effect*, and *deduction* is likely tied to this writing context. Students were trying to convince Ernie (*claim* and *cause and effect*) by invoking chemical principles (*deduction*). It is expected that the

distribution of use will vary with different writing contexts, depending upon what the prompt elicits.

Table 4. Descriptive information from application of the framework to the data set presented herein.

Operations	Frequency
Observation (2)	411
Claim (6)	348
Cause and effect (8)	331
Definition (1)	312
Comparison (4)	302
Deduction (9)	235
Consequences (7)	146
Argumentation (10)	63
Example (5)	51
Measurement (3)	34

Research Question 2: What features of students' argumentative writing do cognitive operations serve to explain?

To determine what exactly this framework served to characterize, two comparisons were made. The first comparison was between framework estimates of complexity and instructor estimates of quality. This comparison was intended to demonstrate that this framework was telling us something that faculty would normally have to make a judgment about. Further this comparison was intended to reveal that this framework could make *similar* judgments to an instructor. Table 5 shows how five faculty ranked four assignments, and this is compared to our framework ranking. Instructors were tasked with ranking the assignments according to the quality of the scientific reasoning (as they saw fit to evaluate it). This approach was taken so as to elicit instructors' "gut reaction"—the kind of evaluation they would make if they were grading this sort of task in their class. Instructor rankings reveal a few trends. First and potentially the most

important, there is almost complete consensus on the 'best' essay (J) with the exception of Instructor 3. This finding speaks to the framework's capacity to identify the best. Essay J received a high cognitive complexity score because of the presence of extended argumentation, which aligns with what instructors are valuing when evaluating scientific reasoning. Four of the five instructors ranked Essay G as second best, with the exception of Instructor 3, even though our framework estimates it as second from the worst. Finally, all five instructors consider Essay H and K to be the worst, where as our framework estimates Essay H to be the second best. The difference between instructor rankings for Essays H, G, and K and framework estimates illustrate an important limitation of the framework. Essay H contained a misconception in which the student claimed a relationship between atmospheric temperature and ocean acidification though there was no data regarding heat for any of the chemical reactions provided. However, this student's reasoning was rather sophisticated, with multiple high complexity cognitive operations. Our framework does not account for scientific accuracy of students' essays. We chose to include this essay as it authentically represents what an instructor might encounter with grading writing. One possible explanation for the ranking difference is that for instructors scientific content and reasoning are inextricably linked, which is consistent with feedback from one instructor who explained that content accuracy factored into their ranking. It is for this reason that we also analysed the scientific accuracy of students' essays (see Table 5 below).

Table 5. Instructor ranking of four assignments of varying quality compared to framework estimates of ranking

		Expert rankings (4 worst, 1 best)				
Essay name	Cognitive complexity ranking (4 worst, 1 best)	Instructor 1	Instructor 2	Instructor 3	Instructor 4	Instructor 5
J	1	1	1	2	1	1
H	2	3	3	4	4	4
G	3	2	2	1	2	2

K	4	4	4	3	3	3

The second comparison made was between cognitive complexity—framework estimates of quality—and common student characteristics that are frequently used as measures or predictors of success (Hein & Smerdon, 2013). The purpose of this comparison was to determine if characterizing the quality of students' reasoning in this way was revealing something about students that could have been predicted by a metric that was already collected (e.g., ACT math score). In other words, this framework is useful only in so far as it tells us something interesting about students that other metrics do not. Table 6 shows the correlations between common student characteristics and cognitive complexity. There were no significant correlations between cognitive complexity and any common characteristics, which would not be expected for measures of constructs distinct from that captured by this framework (i.e. math). These findings may mean that the framework captures something distinct from what is measured by other standardized tests. A strong negative correlation exists between the number of operations used and the cognitive complexity. This means that students with higher cognitive complexity essays used fewer moves, which could indicate a synthesis of ideas in order to produce higher complexity operations.

Table 6. Pearson correlations between student characteristics and cognitive complexity (p-values reported for t-tests used for categorical variables: gender and ethnicity [white and non-white students compared])

Variables	Cognitive complexity
Number of operations	-0.649*
Final exam grade	-0.018
Final course grade	-0.025
CHEM placement	-0.081
MATH placement	-0.020
ACT math[†]	0.003
Current GPA	-0.062
Cumulative GPA	-0.060
Gender	0.401
Ethnicity	0.071

* indicates p (two-tailed) < 0.01

† For students with only SAT math scores, their scores were converted to ACT math scores using contingency tables

Research Question 3: What is the relationship between framework characterizations of complexity and conceptual correctness?

The data above reveal that the cognitive operations framework is characterizing students' reasoning in a way that other measures do not. However, the instructor rankings reveal that there exists a relationship between reasoning and accuracy. The motivation for considering this relationship partially sources from the concern that any information this framework provides is irrelevant if students are largely scientifically inaccurate. In order to explore this relationship, we coded all data that had already been coded according to operations for 'correctness.' That is, when scientifically inaccurate information was identified in an essay, the cognitive operation containing that information was marked as incorrect. In this way, all student writing was coded for both correctness and cognitive function (i.e., content and structure as highlighted in theoretical framework). Table 7 shows the number of incorrect operations per the total number of cognitive operations. Further, there were no correlations between the cognitive complexity and number of incorrect operations or between the number of cognitive operations and the number of incorrect operations. This finding suggests that overall, producing a more complex essay does not make it more likely that a student will use more incorrect ideas, but as Table 7 shows, there may be specific operations that elicit more incorrect ideas. Further, writing more operations, or introducing more separate idea units, does not make a student more likely to put forth incorrect ideas.

Table 7. Number of incorrect cognitive operations relative to total number of operations [def.=definitions, obs.=observation, meas.=measurement, comp.=comparison, ex.=example, claim=claim, cons.=consequences, C&E=cause and effect, Ded.=deduction, Arg.=argumentation]

	Def.	Obs.	Meas.	Comp.	Ex.	Claim	Cons.	C&E	Ded.	Arg.
# incorrect	9	11	0	13	1	8	6	40	18	4
# operations	312	411	34	302	51	348	146	331	235	63

% incorrect per total ops.	3	3	0	4	2	2	4	10	8	6

Evident in Table 7 is a relatively infrequent use of scientifically inaccurate information. That is, given that our unit of analysis is ideas, students are largely generating scientific ideas employing correct scientific information. This further justifies the move beyond simply considering scientific accuracy of students' conceptions towards considering the sophistication of their reasoning about scientific ideas. In this case, only considering the accuracy would have provided a very limited picture of what these students were doing in their writing. Because of the relative infrequence, it became important to consider the nature of the inaccuracies. For this inquiry, categorizing the inaccuracies by operation led to an interesting finding. The highest percentages of inaccuracy, though still relatively small, occurred with *cause and effect*, *deduction*, and *argumentation*. It is possible that higher complexity operations surface alternative conceptions more effectively. Further, the alternative conceptions elicited are potentially more deeply held, keeping in mind the Limited or Inappropriate Propositional Hierarchies (LIPHs). That is, higher complexity operations draw on multiple domains and elements and may have the potential to reveal more of students' mental structures, and thus, expose LIPHs.

Limitations

Though this framework provides a useful way to evaluate students' written work, it has a number of limitations. First, as noted above, this framework does not capture the scientific accuracy of students' written ideas. The utility of this tool, then, is limited to a narrower research goal—characterizing students' reasoning. When combined with an analysis of the content accuracy, however, this framework can provide unique insights about students' understanding. Further, this framework was conceptualized, tested,

refined, and ultimately applied to a corpus of writing in a very specific context—general chemistry argumentative writing about ocean acidification. It is possible that some of the ways that cognitive operations have been conceptualized in this study are specific to this context. For this reason, applications to other contexts are needed to ensure the domain-general nature of this framework. Finally, due to the relatively low occurrence of certain operations in this context, we have a weaker understanding of some of the operations (i.e., measurement). Because of the complete absence of the inductive reasoning operation from Grimberg and Hand's original framework in this set of student writing, it was not included in this application, even though it is likely to be employed in other contexts. Finally, this data was collected at a selective institution and it is likely that different incorrect ideas or reasoning patterns would emerge from other student populations. Again, this can be addressed by applying this framework to student writing in other contexts.

Discussion and Implications

The first research question posed in this work considered how a cognitive operations framework can be used to characterize students' reasoning evident in their argumentative writing. In this article, we show what this framework is like and how it can be applied to students' writing. We refined a list of cognitive operations generated by Grimberg and Hand (2009) and organized them according to complexity, and then used these operations to code general chemistry students' writing on ocean acidification. This framework has some key affordances that make it useful to both research and practice. It is domain general, which means that it can be applied to writing in a variety of contexts. We recommend, then, that others apply this to writing in a variety of contexts across STEM and across levels (introductory to advanced student populations). The domain-general nature of this potentially enables the identification of

differences in students' reasoning across disciplines and levels. For example, do advanced students employ more complex reasoning than introductory students?

Another affordance of this framework is the 'score' that is a product of application—the cognitive complexity. The single score output provides an estimate of construct that is rather difficult to measure —student reasoning. This framework, then, can potentially overcome some of the difficulties with evaluating writing reported in the literature (Hamp-Lyon, 2016; Neill, 2002). This framework provides a novel approach to assigning a holistic score to writing. Further, the use of cognitive operations enables the identification of patterns in students' writing. That is, it can be used to characterize the movement between cognitive operations and the likelihood of moving towards high complexity operations, as shown in Grimberg and Hand's original application (2009). This framework's capacity to capture temporal patterns makes it very useful for understanding how students reason in extensive writing (Grimberg & Hand, 2009; Kelly & Takao, 2002; Moreira et al., 2018).

The second research question aimed to elucidate what features of student thinking were understandable with this framework. That is, what does this framework evaluate the quality of? This was achieved in two ways. The first was to compare framework estimates to instructor estimates of quality. This approach was intended to determine if the framework estimates were similar to the instructor estimates and if both were evaluating a similar construct. This revealed that perhaps for the upper bound of the construct—argumentation—there was agreement between instructors and framework estimates. There was less agreement for the other-than-best essays. Kelly and Takao (2002) identified similar disparities between their framework estimates and expert rankings and explained them as common occurrences when evaluating writing (Wolcott & Legg, 1998). In our case, we argue that the variety was an artefact of the

presence of inaccurate scientific information in one of the essays. Instructors may not separate content and reasoning as this framework does. However, we argue, similar to Kelly and Takao (2002), that this framework may provide a tool for evaluating the validity of instructor's estimates of quality. More research is necessary to establish interrater reliability amongst instructor ratings and identify ways in which the framework can serve as a tool for supporting instructors in systematically assessing students' writing.

To determine if this framework was providing unique information about students' ability, we compared cognitive complexity to other common performance measures. There were no correlations. We posit two potential explanations for this. The first is that this metric of cognitive complexity is indeed measuring something unique from what typical performance metrics measure (National Research Council, 2001). The second is that students who perform well on typical performance metrics do not necessarily perform equally well on more extensive writing tasks (National Research Council, 2001). Both of these explanations warrant further investigation because of the implications for assessment. Specifically, this framework could serve to equip the evaluation of more interesting competencies in students than that measured by typical performance measures *or* assignments of this nature could serve to minimize advantages certain groups bring with them to typical performance measures. However, we also recognize that there may be other performance measures that correlate with the framework estimate. Particularly, we would expect that more generative or authentic assessments might correlate more strongly with cognitive complexity (National Research Council, 2001). Finally, we aimed to characterize the relationship between framework estimates of quality and scientific accuracy. In order to do this, we analysed writing for the presence of scientific inaccuracies and coded the respective operation in

which they appeared. This revealed that scientific inaccuracies occurred relatively infrequently with about 10 percent of *cause and effect* operations including something that did not agree with scientifically accepted knowledge. The percentage among *cause and effect* operations was the highest. However, there appears to be a trend in which higher complexity operations (i.e., cause and effect, deduction, and argumentation) had higher frequencies of incorrect information than low complexity operations. We argue, in light of Novak's work on LIPHs, that higher complexity operations as a representation of students' mental models may better reveal LIPHs (Novak, 2002). That is, employing more complex reasoning may surface more deeply held LIPHs. Students who do not use higher complexity operations may be more limited in both their and their instructors' capacity to address potential alternative conceptions. The relationship between complexity and conceptual correctness warrants further investigation. Understanding this relationship is important for designing formative assessments that better elicit high complexity operations.

This framework also offers some unique implications for instructors who assign similar tasks to their students. Scoring assignments in this way could permit an instructor to draw conclusions about their students' collective access to complex reasoning operations. For example, a low average score of cognitive complexity in their course may motivate instructors to explicitly address and model complex reasoning types for their students. However, we argue that the most important implication of this framework for practice is providing a vocabulary to instructors for giving tailored feedback to students. That is, applying this sort of framework would support an instructor to give specific examples of when a student could have employed more complex reasoning appropriately and instead used a less complex operation.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Emig, J. (1977). Writing as a Mode of Learning. *College Composition and Communication*, *28*(2), 122–128.

Gere, A. R., Limlamai, N., Wilson, E., MacDougall Saylor, K., & Pugh, R. (2019). Writing and Conceptual Learning in Science: An Analysis of Assignments. *Written Communication*, *36*(1), 99–135. https://doi.org/10.1177/0741088318804820

Grimberg, B. I., & Hand, B. (2009). Cognitive Pathways : Analysis of students ' written texts for science understanding. *International Journal of Science Education*, *31*(4), 503–521. https://doi.org/10.1080/09500690701704805

Gunel, M., Hand, B., & Prain, V. (2007). Writing for learning in science: A secondary analysis of six studies. *International Journal of Science and Mathematics Education*, *5*, 615–637. https://doi.org/10.1007/s10763-007-9082-y

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying computerized-scoring models ofwritten biological explanations across courses and colleges: Prospects and limitations. *CBE Life Sciences Education*, *10*(4), 379–393. https://doi.org/10.1187/cbe.11-08-0081

Hein, V., & Smerdon, B. (2013). *Predictors of Postsecondary Success*. *College and Career Readiness and Success Center at American Institutes for Research*.

Kelly, G. J., & Bazerman, C. (2003). How Students Argue Scientic Claims : A Rhetorical-Semantic Analysis. *Applied Linguistics*, *24*(1), 28–55.

Kelly, G. J., Chen, C., & Prothero, W. (2000). The Epistemological Framing of a Discipline : Writing Science in University Oceanography, *37*(7).

Kelly, G. J., Regev, J., & Prothero, W. (2007). Analysis of Lines of Reasoning in Written Argumentation. In *Argumentation in Science Education* (pp. 137–157).

Kelly, G. J., & Takao, A. (2002). Epistemic levels in argument: An analysis of university oceanography students' use of evidence in writing. *Science Education*, *86*(3), 314–342. https://doi.org/10.1002/sce.10024

Keys, C. W. (1994). The development of scientific reasoning skills in conjunction with collaborative writing assignments: An interpretive study of six ninth□grade students. *Journal of Research in Science Teaching*, *31*(9), 1003–1022. https://doi.org/10.1002/tea.3660310912

Klein, P. D. (1999). Reopening Inquiry into Cognitive Processes in Writing-To-Learn. *Educational Psychology Review*, *11*(3), 203–270. https://doi.org/10.1023/A:1021913217147

Klein, P. D. (2015). Mediators and Moderators in Individual and Collaborative Writing to Learn. *Journal of Writing Research*, *7*(1), 201–214.

Krippendorff, K. (2004). Reliability in Content Analysis : Some Common Misconceptions and Recommendations Reliability in Content Analysis : Some Common Misconceptions and. *Human Communication Research*, *30*(3), 411–433.

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, *53*(2), 215–233. https://doi.org/10.1002/tea.21299

Moon, A., Gere, A. R., & Shultz, G. V. (2018). Writing in the STEM classroom: Faculty conceptions of writing and its role in the undergraduate classroom. *Science Education*, *0*(0). https://doi.org/10.1002/sce.21454

Moreira, P., Marzabal, A., & Talanquer, V. (2018). Using a mechanistic framework to characterise chemistry students' reasoning in written explanations. *Chemistry Education Research and Practice*. https://doi.org/10.1039/C8RP00159F

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. *National Academies Press*. Washington DC. https://doi.org/10.17226/10019

Novak, J. D. (2002). Meaningful Learning: The Essential Factor for Conceptual Change in Limited or Inappropriate Propositional Hierarchies Leading to Empowerment of Learners. *Science Education*, *86*(4), 548–571. https://doi.org/10.1002/sce.10032

Prain, V., & Hand, B. (2016). Coming to Know More Through and From Writing. *Educational Researcher*, *45*(7), 430–434.

https://doi.org/10.3102/0013189X16672642
Reynolds, J. A., Thaiss, C., Katkin, W., & Thompson, R. J. (2012). Writing-to-learn in undergraduate science education: A community-based, conceptually driven approach. *CBE-Life Sciences Education*, *11*(1), 17–25.
Sandoval, W. A. (2003). Conceptual and Epistemic Aspects of Students ' Scientific Explanations. *Journal of the Learning Sciences*, *12*(1), 5–51. https://doi.org/10.1207/S15327809JLS1201
Sandoval, W. A., & Millwood, K. A. (2005). The Quality of Students ' Use of Evidence in Written Scientific Explanations. *Cognition and Instruction*, *23*(1), 23–55. https://doi.org/10.1207/s1532690xci2301
Sandoval, W. A., & Reiser, B. J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education*, *88*(3), 345–372. https://doi.org/10.1002/sce.10130
Sevian, H., & Talanquer, V. (2014). Rethinking chemistry: a learning progression on chemical thinking. *Chemistry Education Research and Practice*, *15*(1), 10–23. https://doi.org/10.1039/C3RP00111C
Takao, A. Y., & Kelly, G. J. (2003a). Assessment of Evidence in University Students ' Scientific Writing. *Science & Education*, *12*, 341–363. https://doi.org/10.1023/A
Takao, A. Y., & Kelly, G. J. (2003b). Assessment of Evidence in University Students ' Scientific Writing. *Science & Education*, *12*, 341–363.