Jordan Harshman and Ellen Yezierski

Assessment Data-driven Inquiry: A Review of How to Use Assessment Results to Inform Chemistry Teaching

Abstract

With abundant access to assessments of all kinds, many high school chemistry teachers have the opportunity to gather data from their students on a daily basis. This data can serve multiple purposes, such as informing teachers of students' content difficulties and guiding instruction in a process of data-driven inquiry. In this paper, 83 resources were reviewed to provide a complete description of this process, which has not been previously done. In reviewing the literature, we found that: 1) there is very little research detailing the data-driven inquiry process in a way that can be readily implemented by teachers; 2) the research largely neglects the incorporation of disciplinary content in the data-driven inquiry process; 3) suggestions for teachers' actions provided by the research is general, limiting the impact of these suggestions; and 4) the practical considerations and fidelity of implementation of data-driven inquiry have not been examined. Implications for chemistry teachers are presented along with a call for future research on key areas, thus benefiting researchers of assessment processes. Finally, general data-driven inquiry research is described in the context of chemistry-specific examples that provide useful, practical suggestions for high school chemistry teachers.

Keywords: formative assessment, summative assessment, informing instruction, guiding practice, instructional sensitivity, instructional validity, data use, data driven inquiry, decision-making, assessment as inquiry, reflection, interpretation of assessments, analysis of assessments, and learning objectives in assessment

Introduction

Between homework, quizzes, classroom activities, high stakes summative exams, informal classroom observations, and other inputs, high school chemistry teachers around the globe have access to a wide variety of student data. Through the analysis and interpretation of this data, teachers can uncover great amounts of information, including, but not limited to, their students' conceptions about content and the educational impact of their own instruction. With this information, teachers can tailor instruction to their classroom and even to individual students. The impact of teachers effectively using the results of their many informal and formal, summative and formative assessments on the learning of their students cannot be understated (U.S. Department of Education, 2008; 2011; Institute of Education Sciences, 2009). There is no better source of information for teachers to use to make instructional decisions than data from their own students. Every assignment, homework, quiz, test, activity, lab, in-class question, and discussion yields valuable instructional feedback to high school chemistry teachers. This is free, continuous, and customized-to-your-own-students professional development available every single day to teachers. The conglomeration of literature presented here will not detail the most effective implementation of this process, but it will portray what helpful advice is already available as well as what areas need to be understood better for the use of data to inform teaching.

The United States (U.S. Department of Education, 2011), Caribbeans (Ogunkola & Archer-Bradshaw, 2013; Seecharan, 2001), Hong Kong and Singapore (Towndrow, Tan, Yung, & Cohen, 2010), and Britain (Simon, 1992), just to name a few, have shown increased focus on research of teacher assessment practices. Initially, we sought to investigate how high school chemistry teachers use the results of assessment to make data-driven decisions about their teaching. We searched for research on the use of data resulting from chemistry-specific assessments (e.g., an investigation of how teachers interpret the results from a specific item covering percent yield problems in stoichiometry). After finding no sources that were chemistry-specific and very little that was science-specific, we scoured the general education assessment literature only to find that these resources did not provide a satisfactory level of data-use guidelines for day-to-day instruction. As a result, an in-depth examination of what is available and what is missing in the data-driven inquiry literature is warranted and, therefore, is the goal of this review. It is important to apply the general process of effectively using assessment results to the context of high school chemistry, because the learning goals and modes of assessment tend to vary by discipline and educational setting. Thus, while reviewing the literature, we also illustrate some of the ideas as they apply to the specific context of high school chemistry teaching.

It is important to note that this review derives from bodies of literature that range from informal, formative assessments to high-stakes, national summative assessments. Although the contexts of these assessments vary drastically, all types of assessment generate data that can serve many purposes, one of which is being a guide for instruction. In this light, the process of inquiry that describes how teachers are to inform their instructional practice differs because of differences in design, goals (objectives), and format of items and results, but is the same general process in either formative, summative, diagnostic, proximal, or distal assessments. Stated otherwise, we believe that all assessments produce data that, when analyzed *considering their contexts,* have the potential to inform instruction. Additionally, we use the term "assessment" in a colloquial manner. The term "assessment" as we use it implies two processes: collecting data from students and subjecting these data to criteria that implies evaluation.

As will be reviewed, several sources have described the process that teachers should use in order to guide their practice with the results of assessments. However, there is currently no extensive review of the literature that describes *how* to carry out this process, nor is there any critique of possible limitations of such a process. Since the processes are described as general principles, the task of translating the principles into practice is left entirely up to the instructors, which may create difficulty when translating research into practice (Black & Wiliam, 1998). This review uniquely synthesizes three separate bodies of literature to present an integrated description of the use of data from assessments to guide instruction:

1. Generic descriptions of the process of data use to inform instruction at the classroom level from analysis of high-stakes standardized tests.
2. General suggestions for how this process is carried out in the classroom by teachers using formative assessments.
3. A set of criteria regarding the instructional sensitivity of assessments used for making instructional decisions.

Our hope is that this article will: 1) inspire researchers to investigate this vastly understudied topic of data-driven inquiry; 2) encourage practitioners to consider the potential of the information presented here has to positively impact their instruction; and 3) encourage professional developers to build programs that help teachers enact *the mechanistic, day-to-day details* of how to use the data in their classrooms to inform instruction.

Research Questions

The following research questions guided this review:

1. According to relevant literature, what is the process that teachers should undergo to guide their instruction by assessment results and how can that process be exemplified in a high school chemistry setting?
2. What significant limitations and/or gaps exist in the description of how teachers should guide their instruction by assessment results?

Materials and Methods

This review was conducted via an integrative literature review method described by Torraco (2005). In this approach, common features of a process or concept are integrated towards a comprehensive understanding of that process. The resources for this study were gathered via electronic searches from Web of Knowledge/Science, Google Scholar, and the local libraries at Miami University (in conjunction with OhioLinks). Keywords in electronic searches included various combinations of: *formative assessment, summative assessment, informing instruction, guiding practice, instructional sensitivity, instructional validity, data use, data driven inquiry, decision-making, assessment as inquiry, reflection, interpretation of assessments, analysis of assessments,* and *learning objectives in assessment.* For library searches, the keywords *formative* and *summative assessment* lead the main section about assessment. There were approximately 400-500 titles in this and neighboring sections. To filter through these books as well as the electronic resources, allusions to one or more of the keywords in the chapters (books) or headings, subheadings, or abstracts (articles) had to be present. A large number of articles were also identified by references used in other resources. In total, 83 books and articles were selected for this literature review based on the aforementioned criteria (full list of resources reviewed available in the Online Resource). Data collection occurred primarily from 2011 to 2013, but a few more recent articles have been added to round out the literature review. In describing the scope of this review, it is important to note that not all steps of the assessment process are present. In efforts to answer our research questions in depth, information regarding the purpose of assessing, design of assessments, goal/learning objective setting, and sharing results with stakeholders is not included in this review.

To efficiently present what is documented in the research regarding the process of how teachers are to use the results of their assessments to guide their instruction, we include: 1) an overarching definition and nomenclature for the process; 2) examples of the process both from the literature (in various contexts) and our application to chemistry; and 3) detailed descriptions of what each individual step in the process of informing instruction via assessment results entails as well as major findings of research for each step.

Results

Research Question 1: Data-Driven Inquiry

In response to the first research question, the process by which teachers are to guide their instructional practice is defined by assessment data that drives an inquiry about teaching and learning, or data-driven inquiry. This process goes by many other names: data-informed educational decision-making (U.S. Department of Education, 2008), data-driven decision-making (U.S. Department of Education, 2010; 2011; Brunner, 2005; Ackoff, 1989; Drucker, 1989; Mandinach, 2005), assessment as inquiry (Calfee & Masuda, 1997), cycle of instructional improvement (Datnow, Park, & Wholstet, 2007), formative feedback system (Halverson, Pritchett, & Watson, 2007), action research (Babkie & Provost, 2004), response to intervention [although Deno and Mirkin (1977) did not call it this, they are credited with the central idea], and a review of similar processes from the Institute of Education Sciences (IES) in association with the U.S. Department of Education simply call it the data use cycle (Institute of Education Sciences, 2009). For a graphical example, Figure 1 shows the data-driven decision-making process from the U.S. Department of Education (2010; 2011).

Data-driven inquiry frameworks resembles scientific inquiry in process, namely, defining a problem, collecting data, analyzing and interpreting the data, and then making and assessing a decision. Although the ideas behind the various processes are similar, the name is not, which explains the discrepancy between the labels in Figure 1 and those we use throughout this paper. We describe the process for using data to inform teaching with the terms "data use process" and "data-driven inquiry" throughout this review. As an anecdotal example for how these terms could be used in an educational context, we begin the cycle depicted in Figure 1 with planning. As a note, several authors comment that a teacher can start anywhere on the cycle (Brunner, 2005; U.S. Department of Education, 2010; Institute of Education Sciences, 2009), but this example is presented in the chronological order typically seen in teaching. First

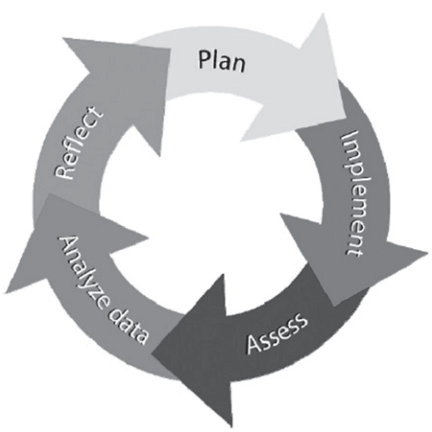

Figure 1. Data-Driven Decision-Making (Department of Education, 2010) This shows a representative data use process, although it does not use the same nomenclature as we do in this review. This cycle includes defining a problem (plan), collecting data (implement and assess), analyzing and interpreting the data (analyze), and making decisions (reflect), similar to scientific inquiry.

a teacher plans a pedagogical strategy based on his/her learning objectives or goals (Plan). As a note, we favor the term "goals" over "objectives" as the latter can hold a connotation (particularly among teachers) as *learning* objectives when data-driven inquiry additionally calls for *instructional* objectives. Thus, "goals" entails both. Then, s/he implements the teaching method (Implement), designs an assessment related to the learning objective, and collects/organizes the assessment results (Assess). The analysis and interpretation (Analyze) of the data can get complicated as s/he can analyze and interpret both in terms of the learning goals and/or problems or questions different from those learning goals (i.e., how effective was the teaching strategy, other compounding factors to performance, impact of educational setting, etc.). Finally, a pedagogical action is hypothesized through reflection on the results and other relevant contextual information (Reflect), and then the assessment process begins anew with the new pedagogical strategy being used.

The Process of Data-Driven Inquiry as Exemplified in Chemistry

To aid the understanding of data-driven inquiry processes, three examples from the reviewed literature were chosen and presented in Tables 1-3. These examples include scenarios in which the data-driven inquiry cycle led the instructors to modify the assessment (Table 1), inform instructional decisions (Table 2), and identify content difficulties in order to refine assessment goals (Table 3). In each table, the first column identifies which step of the data-driven inquiry process is being exemplified and the second column contains an example from the literature. The last column illustrates how these examples might exist in a high school chemistry context, thus addressing the first research question. Table 1 highlights how an original goal is modified after the teacher gets assessment results and also illustrates the importance of considering the alignment of the assessment task and the learning objectives. In the constitution example (Table 2), the use of assessments to directly inform instruction is depicted. Here, the results of the assessment were analyzed with consideration to the original teaching strategy employed. In the final example (Table 3), data-driven inquiry is used to help identify the specific content area in which students are struggling the most. It highlights the importance of the process that can be used in order to isolate the detailed, specific learning objective not met by the students. Information in these tables is then referred to throughout the description of the individual steps.

Steps of Data-Driven Inquiry

Defining a problem – Step 1. There is a semantic difference that identifies the unit for which analysis takes place. Generally when the word "goal" or "problem" is used, it refers to a student outcome, a learning objective, or a problem with students' understandings and is set *prior* to data collection in order to guide the design of assessments. The original goal in Table 2 was to assess students' understandings of atomic structure. Alternatively, when "hypothesizing" or "question posing" appears, it refers to the attempt to explain or address results of the designed assessments and therefore occurs *after* data collection. These are hypotheses about the effect of factors such as educational contexts, individual and class-level history, and even wording of items (as just a few examples) on student performance assessments. In Table 2, the teacher hypothesized that the teaching strategy used may be having a large impact on the outcome. Analysis of both types of questions are important in instructional improvement because in order to know how to adjust instruction, teachers need to know where problems exist in students' understandings, but also need to understand how confounding factors impact the results from which one draws conclusions (Institute of Education Sciences, 2009; U.S. Department of Education, 2011).

In the data-driven decision making model, Knapp (2006), Cuban (1998), and Copland (2003) detail the importance of the ability to reframe potential interpretations of data from multiple

Table 1. Modifying assessment example

Step in Data-Driven Inquiry	4th Grade Vocabulary (Calfee and Masuda, 1997)	Categorizing Reaction Types
Defining a problem	One hypothetical 4th grade boy (Sam) may have a poor vocabulary.	In high school chemistry, a student (older Sam) may not understand chemical reactions.
Designing/Collecting assessment	An assessment is designed and implemented to determine his grade-level in vocabulary.	An assessment that requires Sam to identify reactions as synthesis, decomposition, etc.
Interpretation and analysis	Assessment reveals his vocabulary grade level to be 2.4 (in between 2nd and 3rd grade) so his teacher deems him not very smart.	On the assessment, Sam cannot adequately identify the types of chemical reactions.
Making instructional decisions	Teacher places him in a low ability group in order to give him slower pace.	Figuring Sam doesn't understand chemical reactions, the teacher goes over the definitions and how to identify them multiple times.
Defining an alternative problem/hypothesis to describe results	Teacher thinks that Sam possesses an adequate vocabulary, but does not perform well on the skill of vocabulary recall, which is only one aspect of understanding vocabulary.	Being able to identify types of chemical reactions is not the only aspect of understanding them, and he may understand other aspects of reactions.
Designing/Collecting assessment	The teacher tasks Sam to define words such as "petroleum," use it in an original sentence, and match the word to a definition in different assessments.	The teacher develops an alternative assessment that asks Sam to predict the products including states of matter and balancing of various types of reactions.
Interpretation and analysis	If Sam performs differentially on these tasks and others like it, then he probably understands the word but the type of assessment affects his performance and thus only the pure recall skill required by that assessment type may be what Sam struggles with.	If this assessment yields different results, then Sam probably understands one but not all aspects of chemical reactions.
Making instructional decisions	No instructional decision was provided with this example, however, they do say the following: The teacher needs to consider the consequences that (a) short-term assessment(s) has/have on long-term decisions, such as the decision to place Sam in a low ability group for what could simply be due to the assessment design or context (such as time).	With additional information, the teacher can either target specifically the identification aspect or make curricular change regarding chemical reactions if that teacher decides identification is not as highly valued as other aspects of chemical reactions.

perspectives. These multiple interpretations, formed as questions or hypotheses (Calfee & Masuda, 1997), give teachers the opportunity to access a wide variety of information about their students and their own teaching from one item, one assessment, or a group of assessments. In the U.S. Department of Education large scale study of elementary, middle, and high school teachers (2008), only 38% of teachers reported having professional development that focused on how to formulate questions to answer with data from assessments. To address this, we refer teachers to the IES's guidelines for a good hypothesis: 1) identifies a promising intervention, 2) ensures that outcomes can be measured accurately, and 3) lends itself for comparison study (pre-post or treatment-control designs) (Institute of Education Sciences, 2009). Additionally Suskie (2004) warns against having too many learning goals or inappropriate (too complex or too simple) learning goals as this negatively affects the analysis and interpretation of the resulting data. These suggestions are best illustrated in Table 3 where the teacher modifies a complex goal (understanding of stoichiometry) to a simpler goal (understanding of molar ratios) thereby assessing that goal in a more valid manner.

Designing assessments and collecting data – Step 2. Teachers frequently design, collect, and analyze students' data using formative assessments such as quizzes, homework, in-class activities, and tests. Many of these contain items that are designed (or at least chosen) by the teachers themselves. A constant concern for these items is the extent to which the results can be used to determine the instructional effectiveness. This has been referred to as consequential (Linn & Dunbar, 1991; Messick, 1989), instructional (Yoon & Resnick, 1998), or pedagogical validity (Moran & Malott, 2004), but has recently been called instructional sensitivity (Ruiz-Primo, 2012; Popham, 2007; Polikoff, 2010). In its simplest definition, instructional sensitivity is the extent to which students' performance reflects the quality of instruction received (Koscoff & Klein, 1974).

To demonstrate instructional sensitivity, imagine a chemistry teacher wants to evaluate the effectiveness of a simulation of particles responding to increasing temperature and therefore administers some form of assessment. If the a) content assessed by the assessment aligns with the learning objectives, b) items are not being misinterpreted by the students, and c) format of the response allows the teacher to validly determine students' thought processes, the results can be interpreted to make conclusions about the effect of the simulation on student learning. Factors a-c are aspects of instructional sensitivity, and without them being considered in some way, any conclusion about the student responses would be suspect. For example, an assessment item that simply asks "predict if the volume of

Table 2. Using assessment to inform instruction

Step in Data-Driven Inquiry	U.S. Constitution (Calfee and Masuda, 1997)	Atomic Structure
Defining a problem	A teacher wants students to have a fundamental understanding of the U.S. Constitution.	A teacher wants students to have fundamental understanding of atomic structure.
Designing/Collecting assessment	In order to gain information on what students already know about the Constitution, the teacher asks his students to tell him something about the Constitution. After total silence, he follows up with open questions about the federal government in Washington D.C. After another long silence, he writes keywords on the board such as "President," "Congress," and "Supreme Court" hoping to elicit dialogue, but still no response.	The teacher asks the students to tell them anything they know about the structure of an atom. Some might volunteer something about protons, neutrons, and electrons but provide little follow-up. The teacher could then ask about plum pudding versus solar system models, periodicity, or sub-atomic particles, but gets little response.
Interpretation and analysis	One possible conclusion is that the students actually know nothing about these topics, informing the teacher that he will have to start from scratch.	Considering the lull in discussion, the teacher assumes they know very, very little about atomic structure.
Making instructional decisions	The teacher begins at the absolute basics, tailoring his instruction to include all aspects of the government and the Constitution.	As a result, the teacher begins talking about charges of sub-atomic particles and basic construction of the atoms.
Defining an alternative problem/hypothesis to describe results	A hypothesis is that the students do not normally participate in such open discussions and for lack of knowing how to react, students remain silent.	A hypothesis is that the students do not normally participate in such open discussions and for lack of knowing how to react, students remain silent.
Designing/Collecting assessment	To test this right on the spot, he shifts gears and asks for his students to tell him something about weather because he knows they understand weather.	To test this, the teacher asks more convergent questions, such as what charges do the proton, neutron, and electron have?
Interpretation and analysis	If the response is no longer silence, this may be an indication that it is the pedagogical style (or other contextual aspects) that is yielding the silence, not necessarily a lack of content knowledge regarding the U.S. Constitution.	If the students can answer these questions, then it could have been the open-ended discussion, not the lack in content knowledge that cause students to remain silent.
Making instructional decisions	No instructional decision was provided with this example.	Further, if the teacher's interpretation is that students don't do well with open-ended discussions, that teacher may revise his prompts to elicit more from the students.

gas will increase or decrease at higher temperature," has limited sensitivity to instruction in this case for three reasons: 1) it assesses prediction, whereas the simulation focuses on explanation (related to factor a); 2) "volume of gas" can be misinterpreted by students as "volume of gas particle" (factor b); and 3) the item is susceptible to guessing given its current format (factor c). A more instructionally sensitive item might be "using drawings of gas particles, explain why an increase in temperature will cause an increase in volume" because it minimizes factors a-c, meaning that results can more readily be used to inform instruction.

It is commonly stated that the assessment items must align with the learning objectives of that particular unit (Irons, 2008; Anderson; 2003; Taylor, 2003). In regards to a data-driven inquiry process, this alignment is crucial. Lack of alignment between goals of the assessment and what is assessed by the items leads to misinterpretations and missed opportunities to gather valuable information about teaching and learning. Project 2061 of the American Association for the Advancement of Science provides alignment criteria of this nature that may prove helpful for chemistry teachers (Roseman, Kesidou, & Stern, 1996; Stern & Ahlgren, 2002). Chemistry examples of this alignment (or the lack thereof) can be found in Table 1 and the implications section.

Instructional sensitivity is an important consideration in the data use process because if teachers use items that are sensitive to their instruction, they can use data from those items to adjust instruction (Ruiz-Primo, 2012). Judging both quantitatively and qualitatively the degree to which items are instructionally sensitive has been examined (Popham, 2007; Burstein, 1989; Polikoff, 2010), but no clear standards for evaluation of instructional sensitivity have been published. Outside of "typical" sources of assessment data (i.e., homework, quizzes, etc.), Calfee and Masuda (1997) would argue that, in light of a classroom assessment as applied social science research (Cronbach, 1988), teachers should be open to data collection complete with observations and interviews.

Interpretation and analysis of data – Step 3. Even when an assessment of any kind has been designed so that learning objectives and assessed concepts are aligned, the task of interpreting results in a way that is meaningful to instructors is daunting. Textbooks on classroom assessment (McMillan, 2011; Popham, 2002; Witte, 2012) frequently discuss the importance of:

- understanding validity, reliability, descriptive and (minimal amount of) inferential statistics

Table 3. Identifying content difficulties and refining assessment goals

Step in Data-Driven Inquiry	Perimeter of polygons (Institute of Education Sciences, 2009)	Stoichiometry
Defining a problem	Teachers at an elementary school examine 4th and 5th graders' proficiency rates in language arts and mathematics.	A teacher wants to assess students' knowledge of stoichiometry.
Designing/Collecting assessment	Standardized tests were administered to all students.	The teacher gives assessment with items akin to: If 5.00 g of sodium phosphate react with excess calcium chloride, how much (g) calcium phosphate will precipitate assuming 100% yield?
Interpretation and analysis	Proficiency rates were higher in language arts than in mathematics. In particular, arithmetic was satisfactory, but geometry shapes and measurement skills yielded results that indicated inadequate proficiency for the class. Even more specifically, a teacher noticed that most students struggled on measuring perimeters of polygons, which was surprising as that only required satisfactory performance on arithmetic.	Over half of the students could not answer this question correctly. In looking over the work, she realized that students struggled with many things such as writing formulae and equations, balancing equations, and dimensional analysis.
Making instructional decisions and designing/collecting assessment	Their proposed action in this case was to design and collect more assessment information. They took questions from workbooks regarding perimeters of polygons and gathered more data.	Since the teacher couldn't determine which specific piece needed further instruction, the teacher designed assessment items to only assess mole to mole ratios.
Interpretation and analysis	They began to notice that students performed well on problems where polygons were drawn for them, but did not perform well on real-life application word problems.	She began to notice that about a quarter of the students either didn't include stoichiometric coefficients or reversed them.
Designing/Collecting assessment	As a result, they developed lesson plans focused on the application of perimeter calculations in polygons to word problems, tested again, and found a significant improvement in the performance on these items.	Because of this, she created a few problems that specifically addressed the concept of mole to mole ratios and why and how the coefficients are used.
Making instructional decisions	In the future, the teachers used the same strategies to emphasize application problems to address the problem.	In future years, she will begin with problems that only assess mole-to-mole ratios and move on to problems that assess other components of stoichiometry.

- ethics of assessment
- absence of bias when evaluating students so as to inform an instructor,
- means of summarizing and representing results for oneself as an instructor, other teachers, parents, and other decision-makers.

However, even an understanding of the previously mentioned psychometric aspects does not describe how it applies to a teacher's particular content, context, and pedagogical style and ability. The previously mentioned interpretation strategies generally align with the criterion-referenced era of assessment, meaning the results are interpreted based on predetermined criteria defined by the discipline (Linn, Baker, & Dunbar, 1991). In science classrooms, Bell and Cowie (2002) note that teachers frequently rely on criteria (where the criterion is a scientific understanding of phenomena) because teachers generally want to see that students have learned what they have intended the students learn. This method, often associated with a performance-oriented learning approach, is one of two general approaches. The other interpretation strategy is a growth approach, or student-referenced assessment (Harlen & James, 1996), meaning that teachers reference students' previous assessments in interpreting the current assessments. It is generally accepted that these two methods should be performed in conjunction with each other in interpreting and delivering the results of assessments (Bell & Cowie, 2002). As shown in Table 1, Sam is assessed by a criteria (understanding of chemical reactions), but a growth model could easily be incorporated by repeated measurements of understanding of chemical reactions.

Some years following the implementation of No Child Left Behind Act (NCLB, 2002), the U.S. Department of Education launched a series of large-scale studies to assess teachers' abilities to interpret quantitative assessment data (U.S. Department of Education, 2008; 2010; 2011). First, the U.S. Department of Education found that teachers wanted more opportunities for professional development specific to the interpretation and incorporation of data from standardized tests, and that they were unlikely to do this if they were not confident in their ability to do so (U.S. Department of Education, 2008). A few years later, a different sample revealed that teachers preferred to interact with colleagues on common assessments to interpret data rather than participate in formal professional development (U.S. Department of Education, 2010). With the most recent sample, teachers seemed to express difficulties in fundamental data interpretation skills, such as differentiating bar graphs from histograms, interpreting results of cross-sectional versus longitudinal results, comparing subgroups in more complex data tables, effect of outliers in calculation of the mean or failure to consider distribution when given a mean, and a firm understanding of validity, reliability, and measurement error (U.S. Department of Education, 2011). Over the course of these three studies with participants from the 2004-2008 school years, teachers showed a limited ability to properly understand how to interpret quantitative data from assessments, yet increasingly relied on the support of similar-ability colleagues to assist in this process. An insubstantial number of teachers were shown to possess adequate data use skills in other studies (Brunner, 2005; Cizek, 2001; Herman & Gibbons, 2001; Datnow, Park, & Wohlstetter, 2007, Popham, 1999; Mandinach, 2005) as well. Some authors have tried to give teachers tips on how to present their data visually or organize them into tables for ease of interpretation (Burke & Depke, 2011; Anderson, 2003), which add to a growing list of tools that aid interpretation as opposed to provide a detailed framework for how to use these results.

Quantitative assessment data (from multiple choice, true/false, matching, etc.) is not the only data type available to teachers. Qualitative assessment data (from free response, essay, fill-in-the-blank, etc.) is also widely accessible to teachers. In this type of assessment data, it is more important to interpret results with an absence of bias and alignment with a pre-determined rubric (McMillan, 2011). Without a specific content area and educational context, it is very difficult to describe what is entailed in qualitative data analysis, that rely so heavily on interpretive frameworks.

Making and assessing instructional decisions – Step 4. Outside of comprehending what the data signify, it is also suggested that teachers use the results from assessments to appropriately guide their instruction. Referencing a review of formative assessment literature from Bell and Cowie (2002), actions in response to assessment data can take place at the classroom level, small group, or at the individual student level. Similar to the analysis of qualitative assessment data, it is difficult to comment on how to make instructional decisions in the absence of a specified learning goal, question, and actual student results. Even when this information is present, it is difficult to guide instruction because the awareness of a content deficiency alone does not directly inform or drive pedagogical decisions (Knapp, 2006). However, this awareness does serve two purposes. Firstly, the U.S. Department of Education (2011) and others (Burke & Depka, 2011; Irons, 2008; Institute of Education Sciences, 2009) suggest that teachers use the results of assessment to determine whether they should move forward or recover, reteach, review, or in general allocate more time to the content found to challenge students. Empirical studies suggest that teachers have been doing so for decades (Bennett, 1984; Gipps, 1994; U.S. Department of Education, 2011). Secondly, the Institute of Education Sciences (2009) recommends increasing the students' awareness of their own content deficiencies in order to encourage self-directed assessment.

As Knapp (2006) asserted, both of these recommendations of reteaching and giving students feedback for self-improvement suggest *what to do*, but the assessment results do not elaborate *how to carry it out*. In general, this was fairly common throughout the resources reviewed, as phrases such as "consider what needs to be taught differently" (Irons, 2008), "[attempt] new ways of teaching difficult or complex concepts" (Halverson et al., 2009), "a lesson… can be appropriately modified based on the collected findings," (Witte, 2012) and "[Use] results to revise the unit" (Taylor, 2003) served as main suggestions for teachers. Suskie (2004) elaborates further by claiming that it is not that results simply *cannot dictate* how to adjust instruction, but *should not dictate* how to adjust instruction as only professional judgment in light of results should be used to make such decisions (Bernhardt, 2004). It has also been reported that an intentional plan by teachers for students to self-assess (Yorke, 2003; Institute of Education Sciences, 2009; Irons, 2008) or peer-assess (Falchikov, 1995) can be informed by assessment results, but few details are available. Table 2 presents how a chemistry teacher might teach a concept using a different teaching method while Table 3 shows the teacher did not "go on" with the curriculum when she realized that her students were not understanding the assessed content.

Research Question 2: Conclusions about Gaps and Limitations

In response to the ideas presented by the literature above, we have discovered several limitations and gaps within the existing literature. A "limitation" indicates that a significant amount of research has been conducted, but that literature does not discuss the depth that is required for teachers to adapt research into practice. A "gap" indicates that there is very little to no existing research. Upon reviewing the literature pertaining to data use, we discuss five conclusions:

1. Gap: Data-driven inquiry is only discussed in a general sense and does not address the mechanistic details required to guide day-to-day instruction. While the literature describing assessment as inquiry is valuable, it largely excludes suggestions, instructions, or guidelines that describe *precisely* what teachers should do. That is, the literature would tell a chemistry teacher to conduct the general process of designing, implementing, analyzing, interpreting, and acting on (the results of) an assessment. However, unless a teacher's assessment and results closely mimic the context of a provided example, the process delineated in the research acts as a compass as opposed to a set of directions; it can point you in the right direction, but only with more detailed guidance will you get to your destination. This lack of specificity can be detrimental to the translation of research into practice. In Black and Wiliam's "black box" paper (1998), the authors state:

> "Teachers will not take up ideas that sound attractive, no matter how extensive the research base, if the ideas are presented as general principles that leave the task of translating them

into everyday practice entirely up to the teachers (pg 10)."

2. Gap: The process of guiding instruction by analyzing assessment results is described without reference to educational context or disciplinary content. In accordance with the previous conclusion, we postulate that the data-driven inquiry process does not detail a day-to-day view because, to a large extent, it generalizes across discipline areas and all educational levels of instruction. In most studies reviewed, the process of data-driven inquiry is seemingly identical for elementary, middle school, and secondary level teachers and the students they teach. This is not necessarily incorrect as general social science methods inherent in data-driven inquiry apply to the gamut of student and teacher populations. However, if a researcher were to investigate the very specific, mechanistic details of the process, s/he would need to recognize that the learning goals, assessment types and content, classroom discourse, and all other aspects of assessment evidence are unique at each educational level.

Similarly, the majority of the research does not focus on one particular discipline or another. It can be expected that the assessment goals, format, analysis, and interpretation along with their appropriate pedagogical actions in language arts would differ greatly from that of the physical sciences, for example. Even within an academic discipline, the data-driven inquiry process of chemistry can look entirely different from that of biology, could be different from stoichiometry to gas laws, from conceptual gas law problems to mathematical gas law problems, or even from one conceptual Charles' Law problem to another asked in a different format on the same assessment. This content consideration aligns with the spirit of pedagogical content knowledge (PCK, Shulman, 1987), although few articles mention the role of PCK in the interpretation of assessments (Novak, 1993; Park & Oliver, 2008). Coffey et al. (2011) also claimed that in formative assessment, research widely neglects disciplinary content, which supports the need to consider disciplinary content in assessment interpretation.

3. Limitation: Although the idea that teachers should enact data-driven inquiry (similarly to a social science researcher) to effectively use the results of their assessments is uncontested, the pragmatics and fidelity of implementation of the process have not been studied. In a universal agreement, the resources reviewed point to teachers using an assessment process that includes goal-setting, data collection, interpretation, and analysis in order to inform their instruction. Although the agreement amongst so many authors provides a strong argument for the effectiveness of the process, few short- or long-term studies have examined how well particular teachers implement the *entire process.* Some notable exceptions are recent works in science education (Haug & Ødegaard, 2015; Iczi, 2013; Tomanek, Talanquer, & Novodvorsky, 2008; Ruiz-Primo & Furtak, 2007) and the three Department of Education studies cited earlier. Without this investigation, a characterization of use of data for teachers of a specific discipline is not available. This makes it impossible to determine what, if any, data use training should be developed for current and pre-service teachers. Additionally, since the research lacks a consistent context that weaves together the pedagogy with the consideration of the content, there is little discussion of the pragmatics involved in implementing data-driven inquiry with fidelity: Do teachers value the use of data to inform teaching? What skills do teachers need in order to properly and effectively use data to adjust instruction? How much time will teachers have to dedicate to conduct proper data analysis and interpretation? With other responsibilities and the potential for instructional improvement, is it realistic for teachers to allocate the required time? To what extent does current pre-service teacher training address the skills required for effective data use? These along with many other questions pertaining to fidelity of implementation remain uninvestigated.

4. Limitation: Both "what content to reteach" and "teach it differently" paradigms exist in most of the literature, which limits the true potential of the data to inform practice. As discussed in the literature, the primary reason that teachers analyze and interpret assessment results is to identify the content area(s) on which students perform poorly. Although this is necessary in the data-driven inquiry, the prescribed action is usually to reteach, recover, revisit, or emphasize the suspect content. We again refer to the lack of a context for this finding because without context, one cannot possibly suggest an appropriate action as more information as to what was done previously is required. Instructional strategies and materials used originally help inform how these should be changed in light of assessment results, because a teacher then has evidence to suggest the teaching may have been less effective than desired. This along with the format of the assessment questions, the content being assessed, the wording of the item, and a great many other contextual pieces of information all factor into the interpretation of the results in order to determine the best pedagogical response. As a note, we agree with Knapp (2006) and Suskie (2004) who claim that assessment results by themselves *cannot* inform instruction when considered in isolation. However, we do assert that assessment results along with contextual information *should* guide teachers in their pedagogical decisions.

Implications for High School Chemistry Teachers

The answer to our first research question also addresses the implications of this review to secondary chemistry instruction. Since the literature was not based in the context of chemistry, we are only able to offer recommendations for how teachers should enact data-driven inquiry. However, there are a couple of implications that can be pulled from the general suggestions. First, in defining the goals for assessments (and therefore the focus of the analysis to conduct on resulting data), teachers should ensure that their results can inform a possible intervention. Consider two hypothetical inquiries: 1) Did my students understand movement of gas particles as postulated by kinetic molecular theory? 2) Did my didactic style of instruction best help students understand movement of gas particles as postulated by kinetic molecular theory? The second question will be used to help answer a question about the teacher's performance whereas the first only implies that if the students understand it, then the teacher must have taught it well or vice versa.

As a second implication, emphasis was put on the alignment of learning goals to assessment items. If a teacher wishes to assess students' understandings of molecular polarity, that teacher must realize that asking "Is ammonia polar?" assesses nomenclature (ammonia = NH_3), Lewis structures (and concept of valence electrons), effect of atomic electronegativity on bond polarity, electron and molecular geometry and, finally, molecular polarity as a consideration of individual bond polarities and three dimensional geometries. As a result, this teacher needs to ask the question in a way that will yield results to allow these various factors to be investigated and/or controlled for. Lastly, chemistry teachers will benefit from understanding the limitations of what assessment results can tell them. Interpretation and analysis can identify specific content areas where students struggle, but that needs to be combined with the contextual information only accessible to the teacher of the class. Using the molecular geometry example, if a teacher identifies that 36% of the class labeled the ammonia as trigonal planar as opposed to trigonal pyramidal on account of missing/neglecting the lone pair of electrons on nitrogen (leading to a nonpolar response), that teacher should seek to obtain more information: Who are these 36%? Did they struggle with Lewis structures or molecular geometry? How did I teach this? Have they showed any decreased performance with that instructional strategy previously? Also, instead of "36% of the class labeled ammonia as trigonal planar," what if the results were presented as "36% of the class responded that ammonia was nonpolar?" There are multiple reasons why a student would respond this way. Failure to recognize that the 36% who responded this way specifically struggled with geometries as opposed to any other factor could lead to a misdiagnosis of student difficulties.

Future Directions for Research

Considering the context for which this review was conducted, high school chemistry teachers are the main subjects for which the following suggested research ideas are presented. If research is to inform practice in a significant way, further research must be completed on how the general process of data-driven inquiry is implemented in an everyday context for chemistry teachers. We have already begun data analysis on a study with this goal, but one study cannot possibly capture the variability in the enactment of this process. Studies focused on data-driven inquiry need to incorporate chemistry pedagogical content knowledge, as an appropriate investigation will need to search for *what steps of the process are not present* just as much as (if not more than) *what steps are present*. The latter will describe the current processes in place and inform the state of data-driven inquiry, whereas the former is crucial to identifying areas where chemistry teachers can improve. After an initial, context- and content-oriented data use process is better defined, several inquiries will remain, including:

1. What are the characteristics of high school chemistry teachers' data use process?
2. What are the best practices incorporating data-driven inquiry based on PCK specific to assessment in secondary level chemistry?
3. In what areas can high school chemistry teachers improve their data-driven inquiry skills?
4. What limitations in regards to pragmatic and fidelity of implementation issues exist in proposed interventions targeted at improving high school chemistry teachers' data-driven inquiry?
5. How can professional development of data use skills in either (or both) continuing chemistry teacher training or pre-service training improve high school chemistry teachers' ability to carry out data-driven inquiry?

Research pertaining to the synthesis of best practices is not just a call for the chemistry-specific context, but also a general call for continued research in assessment to incorporate ideas deriving from the pedagogical content knowledge literature. The assessment process cannot be fully articulated speaking only in generalities, but must also be described in consideration of the nature of the content being assessed. Similarly, the content needs to take a significant role in guiding instruction. The effectiveness of general instructional modifications like "reteach" or "change your teaching approach" can only be evaluated fully when the context and nature of the content is given. This is not to say that these suggestions are ineffective, but rather that the use of data to guide instruction is not a general situation and specific actions depend on the context in which the results are generated.

References

Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, *16*, 3-9.

Anderson, L. W. (2003). Classroom assessment: Enhancing the quality of teacher decision making. Mahwah, N.J: L. Erlbaum Associates.

Babkie, A. M., & Provost, M. C. (2004). Teachers as researchers. *Intervention in School and Clinic*, *39*(5), 260-268.

Bell, B., Cowie, B. (2002). Formative assessment and science education. Kluwar Academic Publishers.

Bennett, N., Desforges, C., Cockburn, A., & Wilkinson, B. (1984). The quality of pupil learning experiences. London: Lawrence Erlbaum Associates.

Bernhardt, V. L. (2004). Data analysis for continuous school improvement. Eye on Education, Larchmont, NY.

Black, P., & Wiliam, D. (1998). Inside the black box. *Phi Delta Kappan*, *80*(2), 139.

Brunner, C., Fasca, C., Heinze, J., Honey, M., Light, D., Mandinach, E., et al. (2005). Linking data and learning: The Grow Network study. *Journal of*

Education for Students Placed At Risk, *10*(3), 241–267.

Burke, K., & Depka, E. (2011) Using formative assessment in the RTI framework. Solution Tree Press, Bloomtington, IN.

Burstein, L. (1989). Conceptual considerations in instructionally sensitive assessment. Technical Report 333. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Calfee, R. C., & Masuda, W. V. (1997). Classroom assessment as inquiry. In Phye, G. D. Handbook of classroom assessment. Learning, adjustment, and achievement, San Diego: Academic Press.

Cizek, G. J. (2001). Conjectures on the rise and fall of standards setting: An introduction to context and practice. In G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp. 3-18). Mahwah, NJ: Lawrence Erlbaum & Associates.

Coffey, J. E., Hammer, D., Levin, D. M., & Grant, T. (2011). The missing disciplinary substance of formative assessment. *Journal of Research in Science Teaching*, *48*(10), 1109-1136.

Copland, M. A. (2003). The Bay Area School Collaborative: Building the capacity to lead. In Murphy, J., & Datnow, A. (Eds.), Leadership lessons from comprehensive school reform (pp. 159–184). Thousand Oaks, CA: Corwin Press.

Cronbach, L. J. (1988). Five perspectives on validity argument. Test validity, 3-17.

Cuban, L. (1998). How schools change reforms: Redefining reform success and failure. *Teachers College Record*, *99*, 453–477.

Datnow, A., Park, V., & Wohlstetter, P. (2007). Achieving with data: How high-performing school systems use data to improve instruction for elementary students. Los Angeles, CA: University of Southern California, Center on Educational Governance.

Deno, S. L., & Mirkin, P. K. (1977). Data-Based program modification: A manual.

Drucker, P. F. (1989). The new realities: In government and politics, in economics and business, in society and world view. New York: Harper & Row.

Falchikov, N. (1995). Improving feedback to and from students. In P. Knight (Ed.), Assessment for Learning in Higher Education Birmingham: Kogan Paget.

Gallagher, L., Means, B., & Padilla, C. (2008). Teachers' use of student data systems to improve instruction, 2005 to 2007. U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service.

Gipps, C. (1994) Beyond testing: Towards a theory of educational assessment. London: The Falmer Press.

Halverson, R., Prichett, R. B., & Watson, J. G. (2007). Formative feedback systems and the new instructional leadership. Madison, WI: University of Wisconsin.

Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). Using student achievement data to support instructional decision making (NCEE 2009-4067). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Harlen, W., &James, M. (1996) Creating a positive impact of assessment on learning. Paper presented to the American Educational Research Association Annual Conference, New York.

Haug, B. S., & Ødegaard, M. (2015). Formative assessment and teachers' sensitivity to student responses. *Int. J. Sci. Educ. 37*(4), 629-654.

Herman, J., & Gribbons, B. (2001). Lessons learned in using data to support school inquiry and continuous improvement: Final report to the Stuart Foundation (CSE Technical Report 535). Los Angeles: UCLA Center for the Study of Evaluation.

Izci, K. (2013). Investigating High School Chemistry Teachers' Perceptions, Knowledge and Practices of Classroom Assessment. Dissertation for the University of Missouri – Columbia.

Irons, A. (2007). Enhancing learning through formative assessment and feedback. Routledge.

Knapp, M. S., Swinnerton, J. A., Copland, M. A., & Monpas-Huber, J. (2006). Data-Informed leadership in education. Center for the Study of Teaching and Policy.

Kosecoff, J. B., & Klein, S. P. (1974, April). Instructional sensitivity statistics appropriate for objectives-based test items. Paper presented at the Annual Conference of the National Council on Measurement in Education, Chicago, IL.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational researcher*, *20*(8), 15-21.

Mandinach, E. B., Honey, M., Light, D., Heinze, C., & Rivas, L. (2005, June). Creating an evaluation framework for data-driven decision-making. Paper presented at the National Educational Computing Conference, Philadelphia, PA.

McMillan, J.H. (2011). Classroom assessment: Principles and practice for effective standards-based instruction. 5ed. Pearson.

Means, B., Chen, E., DeBarger, A., & Padilla, C. (2011). Teachers' ability to use data to inform instruction: Challenges and supports. Office of Planning, Evaluation and Policy Development, U.S. Department of Education.

Means, B., Padilla, C., & Gallagher, L. (2010). Use of education data at the local level: From accountability to instructional improvement. U.S. Department of Education.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5-11.

Moran, D. J., & Malott, R. W. (Eds.). (2004). Evidence-based educational methods. Accessed online via Elsevier.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Novak, J. D. (1993). How do we learn our lessons? *The Science Teacher*, *60*(3), 50–55.

Ogunkola, B. J., & Archer-Bradshaw, R. E. (2013). Teacher quality indicators as predictors of instructional assessment practices in science classrooms in secondary schools in Barbados. *Research in Science Education. 43*, 3-31.

Park, S., & Oliver, J. S. (2008). Revisiting the conceptualisation of pedagogical content knowledge (PCK): PCK as a conceptual tool to understand teachers as professionals. *Research in Science Education*, *38*(3), 261-284.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice*, *29*(4), 3-14.

Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership*, *56*, 8-16.

Popham, W. J. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan*, *89*(2), 146-155.

Popham, W.J. (2002). Classroom assessment: What teachers need to know. 3ed. Allyn and Bacon.

Roseman, J., Kesidou, S., & Stern L. (1996, November). Identifying curriculum materials for science literacy: A Project 2061 evaluation tool. Paper distributed at the National Research Council colloquium, "Using Standards to Guide the Evaluation, Selection, and Adaptation of Instructional Materials," Washington, DC.

Ruiz-Primo, M. A. & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *J. Res. Sci. Educ. 44*(1), 57-84.

Ruiz-Primo, M. A., Li, M., Wills, K., Giamellaro, M., Lan, M. C., Mason, H., & Sands, D. (2012). Developing and evaluating instructionally sensitive assessments in science. *Journal of Research in Science Teaching*, *49*(6), 691-712.

Seecharan, H. (2001). Primary school conceptions of assessment and their assessment practices in the Caroni educational division. (Unpublished Master's Thesis). University of the West Indies: Cave Hill.

Stern, L., & Ahlgren, A. (2002). Analysis of students' assessments in middle school curriculum materials: Aiming precisely at benchmarks and standards. *Journal of Research in Science Teaching*. *39*(9), 889-910.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. Harvard educational review, *57*(1), 1-23.

Simon, S. (1992). Curriculum and assessment innovation in science. *Research in Science Education*. *22*, 358-366.

Suskie, L. (2004). What is assessment? Why assess? In Assessing student learning: A common sense guide (pp. 3–17). San Francisco: Jossey-Bass Anker Series.

Taylor, G. R. (2003). Informal classroom assessment strategies for teachers. Scarecrow Press Inc.

Tomanek, D., Talanquer, V., & Novodvorsky, I. (2008). What do science teachers consider when selecting formative assessment tasks? *J. Res. Sci. Educ. 45*(10), 1113-1130.

Torraco, R. J. (2005). Writing integrative literature reviews: Guidelines and examples. *Human Resource Development Review*, *4*(3), 356-367.

Towndrow, P. A., Tan, A., Yung, B. H. W., & Cohen, L. (2010). Science teachers' professional development and changes in science practical assessment practices: What are the issues? *Research in Science Education*. *40*, 117-132.

Witte, R.H. (2012). Classroom assessment for teachers. McGraw-Hill.

Yoon, B., & Resnick, L. B. (1998). Instructional validity, opportunity to learn and equity: New standards examinations for the California mathematics renaissance. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher education*, *45*(4), 477-501.

Jordan Harshman, Graduate Research Assistant, Department of Chemistry and Biochemistry, Miami University, Oxford, OH.

Ellen Yezierski, **PhD,** Professor, Department of Chemistry and Biochemistry, Miami University, Oxford, OH. Correspondence concerning this article should be addressed to: Dr. Ellen Yezierski, Miami University,501 East High Street, Oxford, OH 45056. Email: yeziere@miamioh.edu