Research in Science Education (2020) 50:1885–1916
https://doi.org/10.1007/s11165-018-9757-0

Exploring Effectiveness of Classroom Assessments for Students' Learning in High School Chemistry

Kemal Izci[1] · Nilay Muslu[2] · Shannon M. Burcks[3] · Marcelle A. Siegel[3,4]

Published online: 7 September 2018

Abstract

New notions of science teaching and learning provide challenges for designing and using classroom assessment. Existing assessments are not effective in assessing and supporting desired science learning because they are not designed to capture and aid such learning (NRC 2014; Pellegrino 2013). In addition, it is important to evaluate effectiveness of existing and developed classroom assessments and their usage in supporting desired science learning. Newer notions of validity stress that assessment should have a positive impact on learning and teaching; thus, validity and effectiveness of an assessment should be linked to highlight how an assessment supports learning. Therefore, we suggest that just focusing on the assessment itself, or teachers' understanding and implementing of assessment, to investigate effectiveness of classroom assessment will be incomplete. This qualitative study focuses on the effectiveness of the designed tasks and of the implementation, according to the teacher and aims to (1) provide a new approach for evaluating effectiveness of developed chemistry assessments and (2) use this approach to illustrate the effectiveness of co-developed assessments by five high school chemistry teachers. We utilized multiple sources of data, including teacher-generated assessments, teachers' comments on developed assessments, and students' responses. We designed a rubric to analyze effectiveness and validated it with six expert reviewers. Results showed that the assessments mostly aligned with research-informed principles for effective assessments and helped teachers to achieve their intentions. Our study recommends that teachers develop and utilize various types of classroom assessments that achieve their aims through participation in a collaborative project.

Keywords Effectiveness of assessment tasks · High school chemistry · Assessment development · Chemistry assessment

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11165-018-9757-0) contains supplementary material, which is available to authorized users.

✉ Kemal Izci
kemalizci@gmail.com

Back Affiliation

Introduction

There have been dramatic changes in science education within the last few decades in terms of what it means to teach and learn science. These changes have inspired reforms and new standards in the curriculum of most countries. However, changing standards, as underscored by the National Research Council (NRC 2012), "…will not lead to improvements in K-12 science education unless the other components of the system—curriculum, instruction, professional development, and assessment—change" (NRC 2014, p. 17). In terms of assessment, a reformed view of classroom assessment that draws on cognitive, constructivist, and sociocultural views of learning highlights classroom assessment as a continuous process that provides immediate feedback to both teachers and students to enhance and support learning and teaching, rather than using it at the end of instruction measuring students' aquisition of knowledge (Abell and Siegel 2011; NRC 2014; Pellegrino 2013; Shepard 2000). Therefore, recent definitions and criteria for effective classroom assessments prioritize the "ability to support learning" as an essential criterion. As Elton and Johnston (2002) stated, "Newer notions of validity stress that a 'valid' procedure for assessment must have a positive impact on and consequences for the teaching and learning." (p. 39); thus, validity and effectiveness of an assessment should be linked to highlight how an assessment supports learning. However, there is not an accepted list of standards for classroom assessment to be effective. One of the reasons for this disagreement stems from deliberation of classroom assessment as a task or as a process (Bennett 2011).

Researchers, who consider classroom assessment as a task, consider assessment products to provide and illustrate procedures or frameworks for developing effective classroom assessments in order to elicit and document student learning. Research in this line focuses on experts' or teachers' judgment of developed assessments. According to those researchers, classroom assessment is effective if it (a) addresses intended curriculum or instructional goals (Quellmalz et al. 2012; NRC 2007; Ruiz-Primo et al. 2012), (b) engages students in higher level thinking, including complex scientific reasoning and critical and reflective thinking (Quellmalz et al. 2012; Songer and Gotwals 2012; Liu et al. 2008; Opfer et al. 2012), (c) provides a progressive sequential model to make students' progressively leverage their cognitive skills and differentiate students' level of understanding (Liu et al. 2008; NRC 2001; Ruiz-Primo et al. 2012), and (d) is reliable and valid to yield useful inferences about students' understanding (NRC 2001; Opfer et al. 2012). Alternatively, researchers who mostly have understood and highlighted classroom assessment as a process have provided different standards for effectiveness. In general, classroom assessment should be understood as a process that aims to assess and support learning and instruction (Abell and Siegel 2011; Bell 2007; Bennett 2011; Black and Wiliam 1998, 2009; Coffey et al. 2011; Hattie and Timperley 2007; Shavelson et al. 2008). Researchers in this vein often focus on formative assessment, or assessment for learning, in the literature. They also consider other factors such as the context, teachers' purposes and abilities for implementation, and students' familiarity with an assessment as influencing the quality of instruction (Abell and Siegel 2011, 2013; Black and Wiliam 1998, 2009; Bell and Cowie 2001; Gottheiner and Siegel 2012; Izci 2013; Lyon 2013; NRC 2014). According to these researchers, an assessment is effective if it supports learning and instruction (Bennett 2011; Black and Wiliam 1998, 2009; Bell and Cowie 2001). Toward this aim, researchers have focused on analyzing and supporting teachers' understanding and practices of assessment, which has been alternatively described as assessment understanding (Avargil et al. 2012; Dori and Avargil 2015), assessment literacy (Abell and Siegel 2011; Gottheiner and Siegel 2012; Xu and Brown 2016), and more broadly assessment expertise (Gearhart et al. 2006; Lyon 2013). Having sophisticated assessment literacy is

critical for using assessment processes to support learning; however, research has found that teachers' assessment practice is the main factor that impacts student learning rather than understanding of assessment (Furtak 2012; Herman et al. 2015). This is because sometimes what teachers tell us they know is different than what they do in class (Ateh 2015; Herman et al. 2015; Izci 2013). Therefore, we suggest that just focusing on the assessment itself, or teachers' understanding and implementing of assessment, to investigate effectiveness of classroom assessment will be incomplete. Reasons include that a quality assessment does not warrant improvement in learning, unless it is effectively employed by a teacher to support learning, and conversely, it is difficult for a teacher to elicit, monitor, and aid learning without having a quality assessment (Abell and Siegel 2011; Bennett 2011; Kang and Anderson 2015).

In this study, therefore, we take an integrated perspective of classroom assessment as both a task and process employed by teachers within a classroom to monitor and support learning. Thus, we define classroom assessment as it is a way or a situation for teachers to collect data about students at any moment of instruction for the purposes of assessing students' learning and supporting learning and instruction (Black and Wiliam 2009). However, while our definition focuses on teachers, it also includes assessment tasks, peers, and the individual students as they also form the parts of classroom assessment. Thus, it is important to have quality assessment tasks, effective use of an assessment task by teachers to reveal student learning and take action on the revealed assessment data to aid learning. While it is common to focus on validity and reliability of assessment tasks, this study instead focuses on the effectiveness of the designed tasks and of the implementation, according to the teacher. We propose two simple reasons. One, the scope of the study must be limited. Two, we argue that the areas we focus on are in need of study. We aim to provide an analytical model for evaluating effectiveness of classroom assessments and to employ this model to illustrate evaluation of assessment co-developed by five high school chemistry teachers.

Theoretical Framework

An analytical framework for effectiveness of a practical task (see Fig. 1) was developed by Abrahams and Millar (2008) and then employed (Abrahams and Reiss 2012; Abrahams et al. 2013). Teachers' aims and intentions of practical tasks are related with effectiveness of the

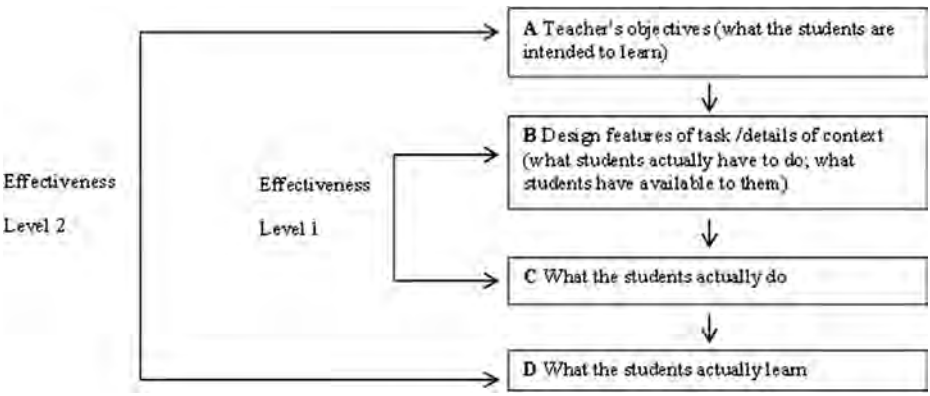

Fig. 1 Model of the process of design and evaluation of a practical task (Abrahams and Reiss, 2012)

tasks within the model. According to the model, teacher's intention to develop and use a task (A) is the starting point. Then, based on the intention, teacher develops a task (B) that has the potential to achieve intended objectives. The third point is related to what students do with the developed task (C) during use, since the teacher's intention might not be achieved. The last point is related to what students actually did with and learned from the task (D). As seen in Fig. 1, effectiveness takes two levels: the first is the alignment between what a teacher intends students to do and what they do (B and C) and the second is the alignment between what the students are intended to learn by the teacher and what the students actually learn (A and D).

Extending this notion of the analytical model of effectiveness of a practical task, we developed a model (Fig. 2) for evaluating the effectiveness of classroom assessments and used it to evaluate classroom assessments co-developed by five high school chemistry teachers. In our model, the starting stage, A, focuses on the purposes for the developers of classroom assessment to illustrate why they want to develop a specific assessment. The second stage, B, centers upon the features of the assessment to see how developers (teachers) chose or developed a task to achieve their aims. The third stage, C, includes how the features of the assessment align with the principles stated in the literature for effective assessment. This shows the opportunities the assessment makes available to students and teachers to monitor and support learning. The final stage, D, is to focus on the real practices of the assessment to show whether the assessment achieved its (teacher's) aim in terms of assessing and supporting learning. As seen from Fig. 2, we also consider effectiveness of an assessment at two levels. Level 1 centers on matching the features of an assessment with fundamentals of effective assessment to assess and support learning as highlighted in the literature and reform documents (B and C). Level 1 is an important aspect of classroom assessment, because an effective assessment should provide rich and formative data for teachers to understand and interpret students' learning and use the results to design instruction to support learning (Abell and Siegel 2011; Siegel and Wissehr 2011; NRC 2007, 2014; Kang and Anderson 2015; Talanquer et al. 2015). Knowledge of assessment tools forms an important part of teacher assessment literacy (Abell and Siegel 2011; Siegel and Wissehr 2011) and requires teachers to know about advantages and disadvantages of different assessment tools and choose/develop appropriate assessments for their own classroom context to assess and support learning. Richness of evidence for students' understanding and achievement leads teachers to make important instructional decisions including judging students'

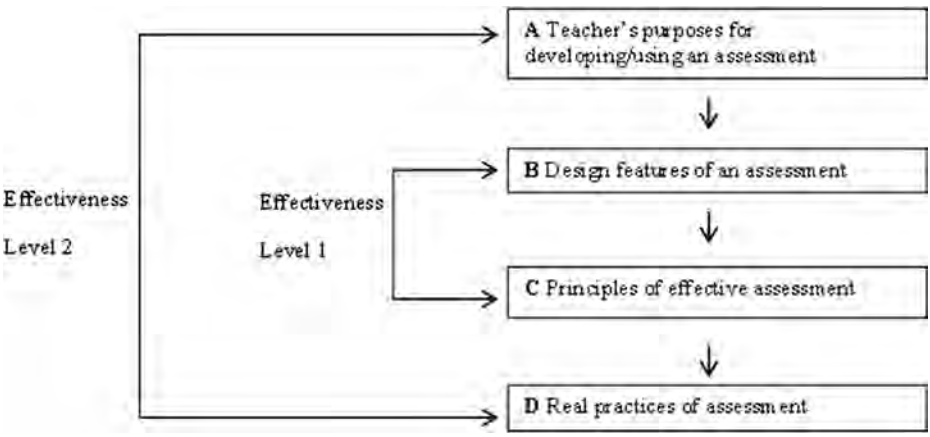

Fig. 2 Our developed model for evaluating the effectiveness of classroom assessments

current situations to decide how to monitor learning, provide feedback, and design and adjust instruction to aid learning (Abell and Siegel 2011; NRC 2007, 2014; Kang et al. 2014). Furthermore, effective assessments motivate students to engage in learning and use their higher level thinking skills, provide equal opportunities for all students to learn and show their learning, and support learning and teaching (see details under the Principles of Effective Classroom Assessment Task). The second level of effectiveness focuses on matching developers' purposes for the assessment with actual practices of the assessment to show if the intended purposes of developers are achieved (A and D). Level 2 is another crucial component of effective assessment, because how teachers employ the assessment influences the degree the assessment impacts learning and teaching (Abell and Siegel 2011; Furtak 2012; Herman et al. 2015; NRC 2007, 2014; Siegel 2012; Siegel and Wissehr 2011; Talanquer et al. 2015). Teachers' beliefs and understanding of assessment is a prerequisite for successful assessment practices (Abell and Siegel 2011; Xu and Brown 2016), while not ensuring effective practice because of other personal and contextual factors (Herman et al. 2015; Izci 2013). Studies have shown that collaborative professional development, including faculty-teacher or teacher-teacher collaborations, is a practical way to overcome challenges limiting teachers' assessment practices and to engage them in effective assessment practices (Avargil et al. 2012; Sato et al. 2008). In summary, to successfully evaluate the *effectiveness* of classroom assessment, we propose that level 1 and level 2 should be considered together to show to what extent assessment meets its ultimate purpose in supporting learning.

Principles for Effective Classroom Assessment

Next, we explain the research-based principles employed in the study (stage C of effectiveness in our model). Different criteria and standards have been suggested by researchers for effective use of assessment to aid learning. For instance, Stiggins (2001) offered five standards that an assessment should satisfy to be effective. They are (1) having a clear purpose to develop and use an assessment, (2) setting explicit learning targets to assess, (3) choosing a suitable assessment to assess the learning target, (4) delivering assessment results to appropriate users in a timely manner, and (5) involving students in the assessment process. Additional researchers provided similar general principles for assessment to be effective (e.g., Siegel 2007; Black and Wiliam 2009; Crooks 1988; Edwards 2013; Gibbs and Simpson 2005; Nicol and Macfarlane-Dick 2006; Ruiz-Primo and Furtak 2007; Stiggins and Chappuis 2005). Siegel (2007) also provides five principles for effective assessment to assess and support diverse students' learning of science. The five principles are (1) matching with learning and instructional targets, (2) be accessible to diverse learners, (3) challenge students to think about big ideas, (4) reveal students' conceptual understanding, and (5) include scaffoldings to support learning. However, with new research and new educational standards, priorities, methods, contents, and function of assessment have evolved. Based on our US context, we emphasized the latest science standards, Next Generation Science Standards (NGSS, NRC 2012). Thus, we extensively reviewed current related assessment and science education literature, which led us to 18 criteria that provide teachers prompts for evaluating their assessments (Table 1) to support learning and teaching. We adopted the five principles of effective assessment suggested by Siegel (2007) and supported each principle with a few teacher-friendly prompts that we developed based on current literature and NGSS emphasize. Table 1 shows the five principles with multiple prompts for each and references that we have benefitted from to develop these prompts.

Table 1 Five principles and related teacher-friendly prompts for effective assessment

Principles	Prompts	Representative references
Cognitively challenge students' thinking	1. The assessment challenges students to think critically to cultivate their scientific habits of mind, develop their capability to engage in scientific inquiry, and lead them how to reason in a STEM context.	(Abell and Siegel 2011; Belland et al. (2016); NRC (2014); Opfer et al. (2012); Siegel 2007; Siegel and Wissehr 2011
	2. The assessment requires a range of thinking and process skills to help students to investigate, evaluate, and develop explanations and solutions via arguing, critiquing, and analyzing data.	Cooper (2015); Coffey et al. (2011); NRC (2014); Liu et al. (2008); Kang et al. (2014); Pellegrino (2013)
	3. The assessment confronts students to develop and use models and engage in argument from evidence to explain their ideas.	Namdar and Shen (2015); NRC (2014); Kang et al. (2014); Pellegrino (2013)
Facilitating student learning	4. The assessment offers scaffolding (e.g., graphics, scenarios, quotes, graphic organizers, analogies) to mediate students' understanding.	Abell and Siegel 2011); NRC (2014); Kang et al. (2014); Shepard (2000); Siegel (2007); Siegel and Wissehr (2011)
	5. The assessment is capable to elicit students' prior knowledge, misconceptions, and conceptual learning to provide formative data for teachers.	Abell and Siegel (2011); Black and Wiliam (1998); NRC (2007, 2014); Siegel and Wissehr (2011)
	6. The assessment produces informative feedback (descriptive and provocative feedback that shows the quality of students' performance toward learning goals and explains how to boost it rather than just providing declarative statement, correct answers, and numerical grades or marks as feedback) that moves learners further.	Black and Wiliam (1998); Hattie and Timperley (2007); NRC (2007); Kang et al. (2014); Ruiz-Primo et al. (2012)
	7. The assessment encourages students to be metacognitive (engage in planning, monitoring, and evaluating their own learning) and reflective to engage in self-regulating learning through planning and carrying out investigations and peer and self-assessment strategies.	Abell and Siegel (2011); NRC (2014); Siegel and Wissehr (2011); Stiggins (2001)
	8. The assessment requires collaborative/group work (e.g., group project), peer review (e.g., peer rubric), and peer assessment to promote peer learning.	Black and Wiliam (1998); Eaton (2009); Gibbs and Simpson (2005); Siegel et al. (2015)
Support teacher's instruction to aid learning	9. The assessment provides written/oral feedback in order to help teachers to modify instruction to aid learning.	Hattie and Timperley (2007); Xu and Brown (2016)
	10. The assessment can elicit higher level thinking (conceptual learning, reasoning, problem solving) to be used into instruction to design teaching activities.	Ateh (2015); Haug and Ødegaard (2015); NRC (2014)
	11. The assessment does not require teachers to spend more time to grade and provide immediate feedback.	Bell and Cowie (2001); Gibbs and Simpson (2005)
Reduce potential bias to aid learning	12. The assessment culturally and linguistically sensitive to provide equal opportunities for all learners regardless to their race, learning styles, language status, gender, and disability.	Abedi et al. (2004); Lyon (2013); Siegel (2007)

Table 1 (continued)

Principles	Prompts	Representative references
	13. The assessment considers students' background (prior daily life experiences) and differences.	Lyon (2013); Siegel, Markey and Swann (2005)
	14. The assessment uses simple and consistent (scaffolded) language to avoid misconcep tions and aid students to understand and engage in assessment task.	Kang et al. (2014); Penfield and Lee (2010); Siegel (2007)
Motivate students to learn and engage in learning process	15. The assessment provides appropriate context (e.g., daily life examples) to engage students in the scientific and engineering practices to understand how scientific knowledge and design principles develop.	Stiggins (2001); Ruiz-Primo et al. (2012); NRC (2014)
	16. The assessment provides a range of opportunities (written, oral) for learners to express their knowledge and skills.	Lyon (2013); Shepard (2000); Siegel (2007)
	17. The assessment provides opportunities for learners to use drawings, diagrams, models, and other formats to motivate learners to involve in learning process and use their creativities to express their ideas.	Lyon (2013); NRC (2014); Siegel (2007)
	18. The assessment motivates students to take responsibility of their own learning (self-motivated).	Nicol and Macfarlane-Dick (2006); Stiggins (2001)

The first principle is that assessment should cognitively challenge students to think critically (NRC 2014). Recent development in science education requires students to engage in complex scientific reasoning because it is linked to conceptual understanding rather than memorization of facts (Liu et al. 2008; NRC 2014; Pellegrino 2013). As assessments involve evidentiary reasoning, an assessment should provide evidence about what kinds of understanding and skills we desire students to gain. The NGSS identifies the ambitious scientific practices for US students. In contrast to previous standards, new standards highlight practice rather than just understanding of learning targets (Pellegrino 2013). Thus, to be effective, an assessment should engage students in scientific practices to conduct, analyze, and interpret data to develop scientific explanation and use models to engage in arguments from evidence to explain their understanding (Belland et al. 2016; Kang et al. 2014 Liu et al. 2008; NRC 2014).

The second principle requires assessments to support students' learning rather than just assess retention of knowledge to provide a grade (Pellegrino 2013). Assessments should provide appropriate forms of material based scaffolding (e.g., graphs, scenarios, quotes, graphic organizers) to mediate students' learning (Abell and Siegel 2011; Shepard 2000; Siegel 2007). Providing scaffolding helps students to organize their thinking and focus on concepts. Mainly, there are two types of scaffolding, material-based and social support that can be used through the entire assessment process to support students to access and engage in the learning process (Puntambekar and Kolodner 2005). While material-based scaffolding is often distributed in the learning environment, across the curriculum materials including assessments and educational software, teachers, peers, and students themselves can act as social scaffolding to facilitate engagement in learning. More open-ended items and fewer closed-ended tests will support students' reasoning, problem solving, and critical thinking ability (Kang et al. 2014;

Liou and Bulut 2017). Various formal and informal assessment strategies should be employed to produce qualitative and quantitative data, which is very important to ensure fairness of measurement and support robust learning (Lyon 2013; Shepard 2000; Siegel 2007). Furthermore, quality assessments should provide written and/or oral feedback in order to enhance learning and instruction. To support students' learning, feedback needs to provide information specific to the task, the learning goals, and the student. Thus, it can help fill the gap between students' current understanding and the learning goal (Hattie and Timperley 2007; Kang et al. 2014). Research has shown that feedback is a crucial influential factor of learning while the types and the ways it is provided determine its real impact on learning (Hattie and Timperley 2007). Another important factor of feedback is timing of feedback. It has been shown that assessments have a large effect size on students' learning when assessment tasks provide immediate feedback (Black and Wiliam 1998; Hattie and Timperley 2007).

Third, assessments should support instruction to aid learning. Classroom assessment can provide critical information for teachers and students in order to support learning and instruction (e.g., Black and Wiliam 1998, 2009; Siegel 2012; Siegel and Wissher 2011). The information can guide students to see what is expected from them to achieve and what they need to do in order to improve their expertise (NRC 2014). Effective assessments elicit students' prior ideas and understanding, provide opportunities for learners to express their thinking, and let teachers use assessment results to monitor and support learning and teaching (Black and Wiliam 1998, 2009; NRC 2007, 2014; Siegel 2007). Eliciting, interpreting, and acting on students' understanding form a vital part of teachers' assessment literacy and expertise (Abell and Siegel 2011; Lyon 2013; Xu and Brown 2016); however, teachers face more difficulties when interpreting and using assessment results to decide how to aid student learning (Ateh 2015; Gottheiner and Siegel 2012; Kang and Anderson 2015). Moreover, using diverse forms of classroom assessment provide a rich data source for teachers to observe, record, and interpret evidence of student learning at multiple levels, which is critical for current science teaching as it requires concurrently focusing on content knowledge, conceptual understanding, and science and engineering practices to prepare students for twenty-first century (NRC 2012, 2014).

Fourth, assessments should reduce potential biases in order to equally and fairly serve all students; assessments should provide equal opportunities for the learners by acknowledging the cultural differences and language abilities (Abedi et al. 2004; Izci 2013). For instance, if an assessment includes an example of rural life while the learners are living in a city, it would be difficult for students to understand the context and answer the related question. It is important to consider differences in students' background and learning style to provide multiple ways for students to express their thinking. Thus, assessment should "…include formats and presentation of tasks and scoring procedures that reflect multiple dimensions of diversity, including culture, language, ethnicity, gender, and disability" (NRC 2014, p. 9). Research has shown that when assessments are culturally and linguistically sensitive and avoid biases, low-level language learner students' learning has been influenced positively (Siegel 2007, 2014; Black and Wiliam 1998; Atkin et al. 2001).

Fifth, assessments should motivate students to learn and engage in the learning process. Assessments need to be provided within an authentic context (such as daily life) that is interesting and enjoyable in order to motivate students to learn and engage in the learning process (NRC 2014; Ruiz-Primo et al. 2012). Recent reform documents highlight the importance of engaging students in science and engineering practices (NRC 2012). Assessments also

need to provide students different forms of written and oral ways to express their thinking and understanding. Providing drawings, diagrams, and models also can motivate students to learn and give the opportunities to express their science process skills (Kang et al. 2014). When assessments put the responsibility of learning on students by providing reflective and metacognitive probes, students are more likely to be engaged in the learning process and become independent thinkers and problem solvers (Atkin, et al. 2001).

Community-Centered Co-Development

To enhance teachers' formative assessment practices, researchers employ various methods. These methods include a focus on enhancing teachers' understanding of assessment (Abell and Siegel 2011; Haug and Ødegaard 2015; Lyon 2011) and other practice-oriented approaches (Ateh 2015; Boud et al. 2018; Kang and Anderson 2015; Sato et al. 2008). Previous research illustrates that while teachers' understanding of assessment improves via professional development (PD) programs, their practices tell a different story (e.g., Herman et al. 2015). In practice, teachers often faced difficulty developing and integrating assessment tasks into their instruction with the aim to assess and support learning (Ateh 2015; Avargil et al. 2012). Therefore, we took a practice-oriented and community-centered approach to PD which involves teachers learning new practices, reflecting on their current practices, collaborating with colleagues and researchers, and changing their performance for instruction using a process of co-development. Teachers can request help from academic collaborators during the PD program regarding how to develop and use a specific assessment task (e.g., Avargil et al. 2012). Thus, a community-centered PD that promotes collaborative work between faculty and K-12 teachers is a productive way to transform teachers' practices of assessment (Furtak et al. 2012). This community-centered PD may encourage teachers and faculty to share ideas, reflect on practices, engage in understanding of challenges faced by colleagues, and transform practices (LePage et al. 2001; Voogt et al. 2015). Infusing this expertise and new knowledge provided learning opportunities for involved partners (Bartholomew and Sandholtz 2009), and the expertise was needed to transform a reform into classroom practice regardless of teaching context (Voogt et al. 2015). Furthermore as Clark (1988) points out, universities represent theoretical spaces and schools represent practical realms of a proposed change. Thus, our collaboration aimed to mutually benefit and facilitate our ability to reform and merge theoretical and practical assessment practices. The study of Sato et al. (2008) also showed how the faculty-teacher collaboration helped teachers to transform reformed assessment view, formative assessment, into classroom practices. Specifically, Fig. 3 illustrates how teachers and researchers in this study engaged in such a community-centered PD program.

Research Questions

In contrast to other studies that focus solely on features of assessment tasks including validity and reliability, and the opportunities they provide to learners to illustrate quality of assessment, this study sought to explore effectiveness of classroom assessment at two levels in order to provide a more comprehensive and practical picture for analyzing effectiveness of classroom assessment practices (see Fig. 2). Specifically, we focused on how classroom assessment was developed and how it was used within classroom context to support learning and teaching. The

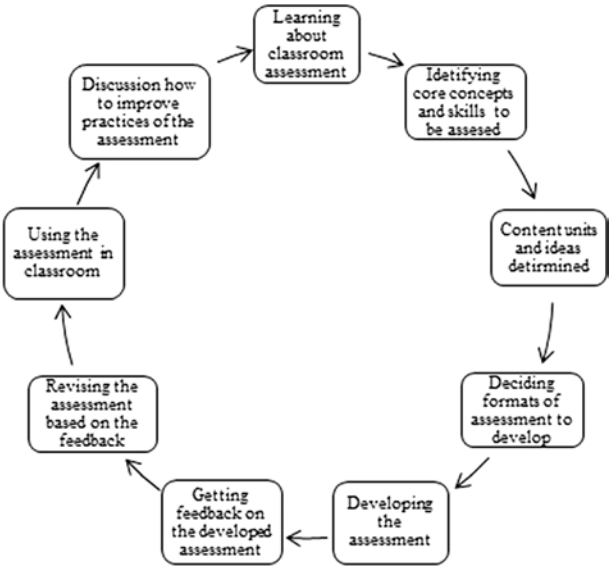

Fig. 3 Teachers' and researchers' engagement in a community-centered PD program

overarching research question posed was: How do we characterize the effectiveness of assessments based on an evaluative model (Fig. 2)? Based on the analytical model (Fig. 2), the following two questions are addressed within the study:

1. To what extent do the developed chemistry assessment tasks meet the principles highlighted in the related literature for effective assessments? (Level 1)
2. How well do the developed chemistry assessment tasks enable teachers to achieve what they intended them to do? (Level 2)

Methodology

Fifty-seven high school chemistry classroom assessments, which go beyond the traditional multiple-choice tests, were co-developed by teachers and researchers during our project. Teachers brought ideas for assessments to meetings where teachers and researchers further developed these assessments by sharing ideas and experiences to improve upon the assessment and meet the state and national science education standards. These assessments represented a variety of assessment practices rather than traditional multiple-choice assessments and align new, innovative forms of instruction with the current understanding of learning. This study is a qualitative study in nature as it uses an analytical model that requires use of multiple data sources to alternatively evaluate effectiveness of assessment tasks. The collaborative assessment development process that teachers and researchers engaged in lets us characterize the effectiveness of assessments based on an evaluative model as other studies employ such alternative approaches to evaluate effectiveness (Abrahams and Reiss 2012; Koh et al. 2018; Ruiz-Primo et al. 2012). The study contains multiple data sources, including developed assessment tasks, teachers' interviews, and reflections and students' responses, which researchers have identified as useful for research design (Yin 2009).

Context

Participants This study included five chemistry teachers co-developing with six researcher classroom assessments. The five teachers were selected from a pool of voluntary applicants who wanted to have more experience with designing and using reform-based assessment to aid their student learning. The selection of participants was based on school types, teaching experiences, and student populations (e.g., urban/rural) to represent a variety of contexts (see Table 2). Thus, we chose to work with two less-experienced teachers (Sophia and Margaret) even if they had less experience in developing and using classroom assessment. However, this representation of teachers from different background and contexts was intended to enrich the findings to a broader context. All of the participants taught chemistry at different high schools located in the Midwestern USA. Participants taught various chemistry courses, including general chemistry, AP chemistry, Honors chemistry, Chemical biology, Medical chemistry, and Physical science courses at the high school level.

Study Context Participants met more than ten times for 3 h during the 2 years of the project. Collaborators, teachers, and researchers met for 3 h three times in the summers and once every 2 months during the fall and spring semesters. During the meetings, (1) researchers presented successful examples of developing and using assessments to improve students' learning, (2) teachers worked together to develop innovative classroom assessments, (3) researchers and peer teachers provided their comments on assessments to improve their quality, and (4) teachers brought students' work on the assessments to discuss results and talk about the challenges they faced during implementation of these assessments. Teachers used the assessments that overlapped with the topics they were teaching rather than using all the developed assessments. In addition to meetings, teachers and researchers used Blackboard to comment on each other's assessments to improve the quality.

Assessment Design Approach Co-developers (teachers and researchers) in this study engaged in an iterative assessment development process. During the assessment development process, one teacher led the development of each assessment, with the cooperative assistance of other teachers and researchers. This co-development process followed a "walking through" approach as described by Horn and Little (2010, p. 207). During the collaboration process, participants' discussions mostly focused on features of effective assessment, chemistry content, students, and instruction. In our case, the iterative process included a nine-step cycle (see Fig. 3) to develop the classroom assessments: (1) teachers were informed about various types of classroom assessments and their advantages and disadvantages, (2) national (NSES and NGSS) and state standards were used to identify core concepts and skills to be assessed, (3) appropriate chemistry units and core ideas were identified by teachers to develop assessments,

Table 2 Participants' personal data

Pseudonym	Level of education	Teaching experience	Gender	School type	Enrolment
John	M.Ed.	11	Male	Urban	1957
Margaret	B.S.	1	Female	Rural	132
Julie	M.Ed.	9	Female	Urban	1790
Sophia	M. Sc.	2	Female	Military	675
Steve	B.S.	7	Male	Rural	632

(4) teachers and researchers engaged in discussion to decide how to assess identified core concepts, (5) teachers developed assessments, completed a "Design" template for each developed assessment, and published them on Blackboard, (6) researchers and peer teachers provided their feedback on Blackboard, (7) assessments were revised by teachers based on received feedback, (8) teachers used assessments in the classrooms and completed a "Use" template for each assessment and posted them on Blackboard, and (9) the difficulties faced by teachers during implementation of these assessments were discussed to improve the successful implementation of developed assessments.

Data Sources We collected a range of data sources during the 2 years of the project. The data sources included the teachers' pre/post-interviews, their online discussions, their reflection on designing and using developed assessments, co-developed assessments, and students' responses to these assessments. In order to illustrate the effectiveness of developed assessments, co-developed classroom assessments, teachers' comments on developed assessments (via two templates, described below), discussions that occurred on discussion boards in the Learning Management System Blackboard, and students' responses to these assessments were used. The qualitative data including teachers' interviews, reflections and comments on developed assessments, online discussions, and students' responses to assessments were employed to illustrate the alignment between teachers' intentions to develop and use an assessment and their real classroom practices (level 2). Classroom assessments were co-developed with teachers and researchers during the iterative process described in Fig. 3. Teachers worked on assessments before, after, and during Face-to-Face meetings. After teachers developed each of the assessments, they were required to fill in the "assessment design template" (ADT) intended to provide opportunities for teachers to reflect on the assessments. The ADT includes metacognitive questions such as, "What was your intended purpose for this assessment? What was the most challenging aspect in creating this assessment?" Furthermore, right after teachers employ the assessments, they completed the "assessment use template" (AUT), which aims to help teachers share their experiences and struggles regarding the enactment of the assessments. The AUT includes questions such as, "How well did the assessment work for your specific population of learners? Did you notice any inhibiting factors for students' learning through the use of this assessment? Do you think the assessment achieved the goal/s you set for? How?" However, some of the developed assessments could not be implemented by these teachers due to the incompatibility of these assessments with the topics they were teaching at the time. Teachers' responses to ADT for all assessments and AUT templates for implemented assessments were also used for effectiveness at level 2.

Teachers continued to improve their assessments by incorporating the advice from other project participants after they filled the ADT for each assessment. Advice was sought as well as provided through discussions in Face-to-Face meetings and through discussions on Blackboard. These comments added to the depth of data for this research study. In addition, data sources included student responses to developed and used assessments to see if the assessments achieved teachers' intentions for their students (level 2).

Rubric of Assessment for Learning One of the purposes of the study was to evaluate to what extent the developed assessment tasks met the principles highlighted in the related literature (level 1). Therefore, we needed to find a way to analyze and compare the assessments with the principles. To achieve this, we became interested in developing a rubric that can provide criteria for us to evaluate the alignment. This rubric is developed to meet the needs of teachers

to scaffold their efforts in the development of assessments that align with recent reform documents (e.g., NRC 2012) and as a tool to evaluate assessments. The need also was identified by others and ultimately National Science Teachers Association (NSTA in the USA) and Achieve (NRC 2014) jointly developed a rubric called "The Educators Evaluating the Quality of Instructional Products (EQuIP)" to support teachers and curriculum developers evaluate instructional materials for alignment with NGSS standards. The EQuIP rubric was designed to evaluate the quality of a lesson or unit with NGSS standards in terms of blending practices, disciplinary core ideas, and crosscutting concepts. It focuses on evaluating a whole lesson and provides evidence for its alignment with NGSS standards, yet the part on "monitoring" for assessment is not very detailed. Therefore, there is a need to have a more detailed rubric that particularly focuses on assessments to guide teachers in developing/choosing effective assessments to achieve their aims. In order to develop such a rubric for assessments, we generated criteria from the assessment literature. We developed 24 different criteria within five sub-dimensions. Each dimension contained detailed items that described how an assessment could meet that dimension. These dimensions are also discussed in the Theoretical Framework section as research-informed principles for effective assessments (Table 1).

After constructing a draft of the rubric that included five dimensions and 24 criteria, we constructed a survey and sent the draft to six different assessment experts whose research interests and studies focused on assessment. These experts work in science education, mathematics education, and educational measurement departments at four different institutions across the USA. In the survey, we provided the draft and asked the expert to provide their reflections for each dimension and for the whole rubric. The experts sent constructive feedback. Based on this feedback, we revised the rubric by adding, excluding, and combining some criteria and developed the final version of the rubric (Table 1) that includes 18 different criteria within five dimensions.

Based on the rubric, an assessment can earn a score of 1, 2, or 3 for each criterion within five dimensions. Score 1 means the assessment is not capable of satisfying the criteria, Score 2 means the assessment partially satisfies the criteria, and Score 3 means the assessment fully satisfies the expectations of a criterion set for effective assessment. This rubric provides many detailed prompts for teachers to consider during development or selection of an assessment and lets teachers score each dimension resulting in a final score. The dimensions are not meant to be obligatory. In other words, an assessment might be strong in one area and not in another and that does not mean the assessment is ineffective overall.

Data Analysis In order to engage in data analysis, first all developed assessments, ADT and AUT, Blackboard discussions, and students' responses to assessments were combined within files. Then, in order to show to what extent each of the developed assessments met the criteria highlighted in the literature for effective assessment, three researchers individually evaluated the assessments using our rubric. A consensus was reached for each sub-dimension for every dimension in the rubric for all assessments between researchers that coded the same assessments. All three researchers coded each assessment and 100% consensus was reached between researchers. This involved a discussion where a consensus was reached so researchers agreed on rubric scores (Creswell 2012). Furthermore, some assessments were independently selected and scored by one outside member who was informed about how to use the rubric to confirm the given scores by the three researchers. Each assessment was scored using the rubric and

assigned a score of 1, 2, or 3. Total scores for each dimension were tallied and used during data analysis and employed to develop figures to present results. In order to present the scores in a more descriptive way, each dimension was sorted into low, medium, and high categories. For each dimension, the low category included assessments that earned a score of 50% and below of the maximum scores of that dimension; the medium category consisted of assessments that received a score between 50 and 75% of the maximum scores; and high category contained assessments that achieved a score more than 75% of the maximum scores. When the related percentage scores calculated for the categories were found decimally (e.g., 4.5), they were completed to the near above integer score if the decimal is .5 and more, and they were completed to the near below integer score if the decimal is below .5.

All the collected qualitative data for this study was deductively analyzed based on the aspects of quality classroom assessment for student learning that was briefly explained in the theoretical framework section. We employed content analysis to see alignment between teachers' goals and practices for implemented assessment. Specifically, teachers' responses to ADT, AUT, discussions on Blackboard, and students' responses to assessments were analyzed in order to see to what extent the assessments in practice accomplished the aims of teachers that they indicated via ADT and AUT. Analysis at this stage was used to show effectiveness of assessments for level 2 while our coding based on the developed rubric was used for level 1. During the analysis of the qualitative data, special attention was given to the goals teachers set to develop an assessment for their classrooms, their self-reported statements about the success of their assessment practices for achieving stated goals for themselves, and our analyzing of students' responses to assessment tasks to see if the stated goals were achieved. Furthermore, teachers' reflection on their implementation revealed difficulties (e.g., providing written feedback, deciding when to move on teaching next concept, asking eliciting questions to students who do not know anything about a concept, reliability of peer feedback) they faced during their implementation of assessments. An example of qualitative coding for level 2 is given in Table 3. The example includes an assessment (see Appendix 1) that Julie, 9-year experienced female teacher, developed for stoichiometry topic to see how her students connect mass and mole concepts to identify elements.

Table 3 Example of coding for level 2

The teacher's goal/s for the assessment	Self-reported achievement of the goal/s by the teacher	Our analysis of the students' work to see if the stated goal/s was achieved
This (the assessment) is designed to have students make the connection between mass and moles and to help them with their experimental design.	Many (students) were confused at first because they thought they would need more materials. I was fine with this, however, because of how I was using this (the assessment) to stretch their thinking. Most students were able to be quite successful so I feel that it (the assessment) was appropriate.	In order to complete the assessment, students used some steps (e.g., measuring, recording, calculating, and interpreting) of experimental design; thus, the assessment achieved Julie's aim of experimental design. The assessment required students to connect mass and moles to calculate molar mass to find the given element, iron. So, the assessment also satisfied Julie's goal for making students to connect mass and mole concepts.

Trustworthiness Data sources and methods were triangulated as teacher developed assessments, teachers' reflections on the assessment development processes, teachers' assessment implementations, and students' responses to assessments provided multiple ways to test the alignment of the developed assessments with the two levels we explained earlier. We also utilized peer debriefing and checks by participating researchers who are not the authors of the study but were members of the larger project (Lincoln and Guba 1985). Finally, this study used theory, logical inferences, and clear reasoning (Brantlinger et al. 2005) during data analysis process to identify the categories we present.

Findings

The aim of this study was to characterize the effectiveness of the assessments, developed in collaborative enquiry between chemistry teachers and assessment researchers, based on a model of effective assessment. The collaborators in this project co-developed 57 different classroom assessments, which are available on our project website (www.dreyfusmu.weebly.com). These assessments focus on 12 different units (e.g., chemical reaction) of high school chemistry. Teachers had summative purposes (e.g., grading), formative purposes (e.g., eliciting students' ideas), or both, to develop and use these assessments. The developed assessments were intended to focus on individual students or groups of students. Various contexts such as laboratory, pre/post-instruction, and embedded within classroom instruction were used to develop these assessments. Beyond content goals, teachers also aimed to assess and enhance laboratory, metacognitive, critical thinking, science process, scientific inquiry, and argumentation skills within their assessments.

By using a model for evaluating effectiveness of assessment and multiple data sources, we next describe to what extent the model can be effective measures of the developed assessments. As a reminder, effectiveness at level 1 focused on the developed assessment itself, and effectiveness at level 2 concentrated on the practice of an assessment.

Effectiveness at Level 1: Alignment Between the Developed Assessments and Principles for Effective Assessments

Dimension 1: Cognitively Challenging Students' Thinking One of the important dimensions of effectiveness for an assessment is to cognitively challenge students' thinking to cultivate their scientific habits of mind. As seen within our rubric (see Table 1), we employed three different prompts to evaluate the effectiveness of an assessment for challenging students' thinking. Therefore, the minimum score an assessment can get on this rubric for the first dimension is 3 and the maximum score is 9. For each dimension of effectiveness, we grouped the assessments as low (50% and below), medium (between 50 and 75%), and high (above 75%) based on the scores received. Thus, the low group for this dimension included assessments that received a score of 5 and below; the medium group consisted of assessments that achieved scores of 6 and 7; and the high group contained assessments that earned a score of 8 and above.

For dimension 1, shown in Fig. 4, within 57 developed assessments, there are only five assessments (two of them had 4 points and three of them 5 points) that were placed within the low group. Twenty-one of the 57 assessments (eight of them received 6 points and 13 of them 7 points) constituted the medium group, while 31 of the 57 assessments formed the high group.

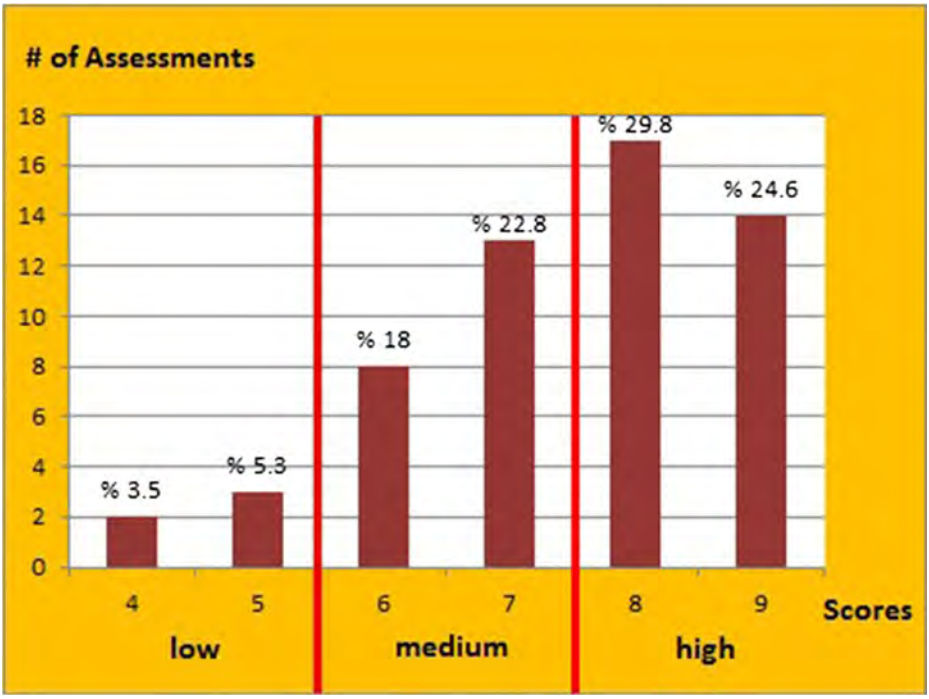

Fig. 4 Dimension 1

Therefore, we found that most of the developed assessments within our project were effective for cognitively challenging students' thinking, as more than 50% of the assessments received the highest score and more than 90% of them got a medium or higher score from our rubric.

Dimension 2: Facilitating Student Learning Facilitating student learning is an important dimension of effective assessments as it provides a context for teachers and students to aid learning. There are five prompts that form this dimension (see Table 1); thus, an assessment can get 5 points at minimum and 15 points at maximum from our rubric. As seen in Fig. 5, for this dimension, low group included scores of 5, 6, and 7; medium group consisted of scores of 8, 9, 10, and 11; and high group contained 12, 13, and 14.

Depending on our scoring rubric, nine assessments (one got 5, one got 6, and seven got 7 points) received a score in the low category for this dimension. Twenty-eight of the 57 assessments were placed into the medium category, as well as 15 of the 57 assessments received a score to rank among the high category. Thus, some of the developed assessments were effective for facilitating student learning, as more than 25% of them received a high score, and most of them were reasonably effective with 60% receiving a medium score.

Dimension 3: Supporting Instruction to Aid Learning Assessment is seen as an important tool for teachers to evaluate effectiveness of their own instruction to aid student learning. Therefore, it is crucial for an effective assessment task to enable teachers to evaluate and modify their instruction to improve learning and teaching. Our rubric has three specific prompts for supporting instruction to aid learning dimension (see Table 1). Hence, an

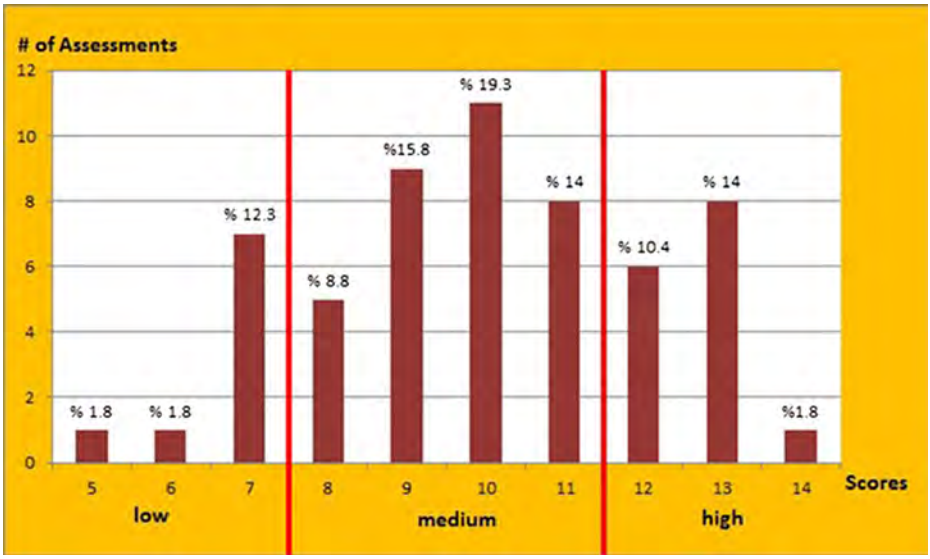

Fig. 5 Dimension 2

assessment can get a minimum of 3 and a maximum of 9 points from the scoring rubric. The low category included 3-, 4-, and 5-point scale, the medium category contained 6- and 7-point scale, while the high category consisted of 8- and 9-point scale.

As it is seen in Fig. 6, only five of the 57 assessments received a low score from our rubric while 23 of them got a medium score and 29 of them received a high score. As a result, most of the assessments were effective for supporting this dimension since more than 50% of the assessments were in the high category and more than 40% of them were practically effective.

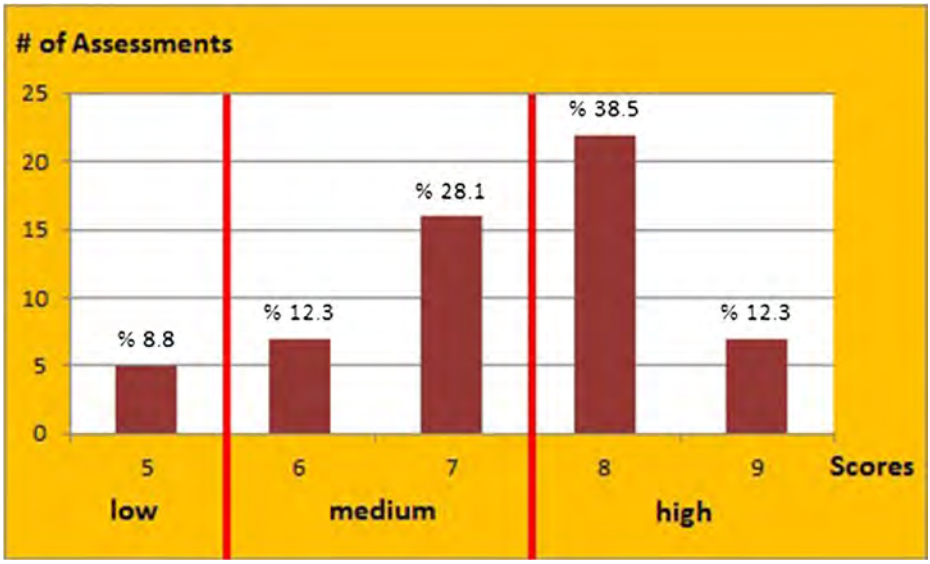

Fig. 6 Dimension 3

Dimension 4: Reducing Potential Biases to Aid Learning One of the crucial dimensions of effective assessment is reducing potential biases in order to fairly and equally assess and support all students' learning. There are three prompts in our rubric that address this dimension (see Table 1); thus, this dimension includes 3 points at minimum and 9 points at maximum from this scoring rubric. The low category included 3-, 4-, and 5-point scale, the medium category contained 6- and 7-point scale, while the high category consisted of 8- and 9-point scale.

As seen from Fig. 7, only one of the assessments received a low score on the scoring rubric, while 13 of them obtained a medium score and 43 a high score. Consequently, most of the assessments were effective for reducing potential biases, as more than 75% of the assessments got a high score while just 1.8% of the assessments got a low score.

Dimension 5: Motivating Students to Learn and Engage in Learning Process Effective assessment should provide motivating probes and context to let students illustrate their knowledge and skills. There are four prompts in our scoring rubric to address this dimension (see Table 1), and therefore, an assessment can receive 4 points at minimum and 12 points at maximum on the rubric. For the dimension, the low category included 3-, 4-, 5-, and 6-point scale, the medium category contained 7-, 8-, and 9-point scales, while the high category consisted of 10-, 11-, and 12-point scale.

As seen in Fig. 8, based on the scoring rubric, only five of the assessments received a low score, 30 of the assessments received a score in medium category, and 22 of the assessments received a score in high category. Therefore, we claim that most of the assessments were effective in the category of motivation of students to engage in learning (38% of the assessments received a high score and 52% a medium score).

In summary, as we quantitatively illustrated, the developed assessments mostly satisfied the principles for effective assessments. The scores each assessment achieved from each of the dimensions of our rubric and the average scores of each dimensions based on all assessments

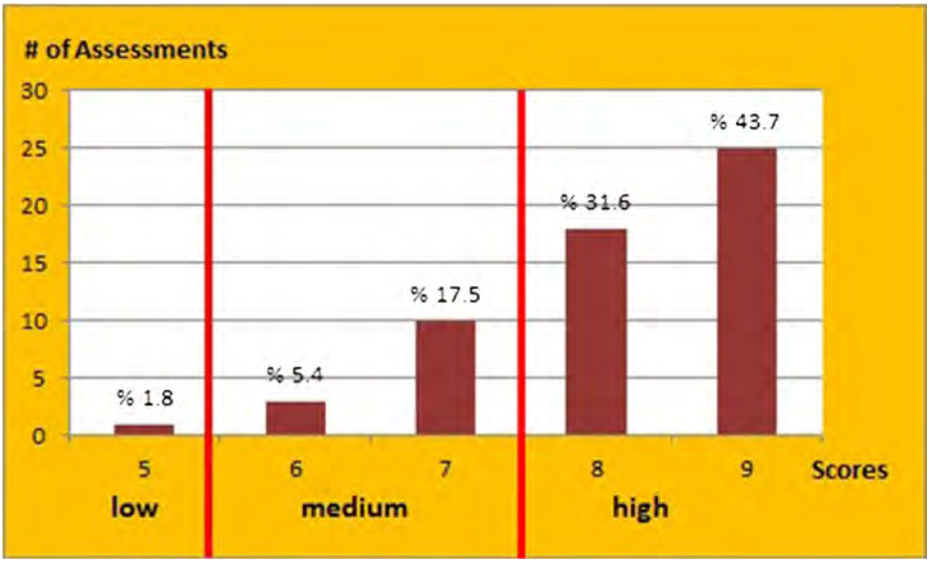

Fig. 7 Dimension 4

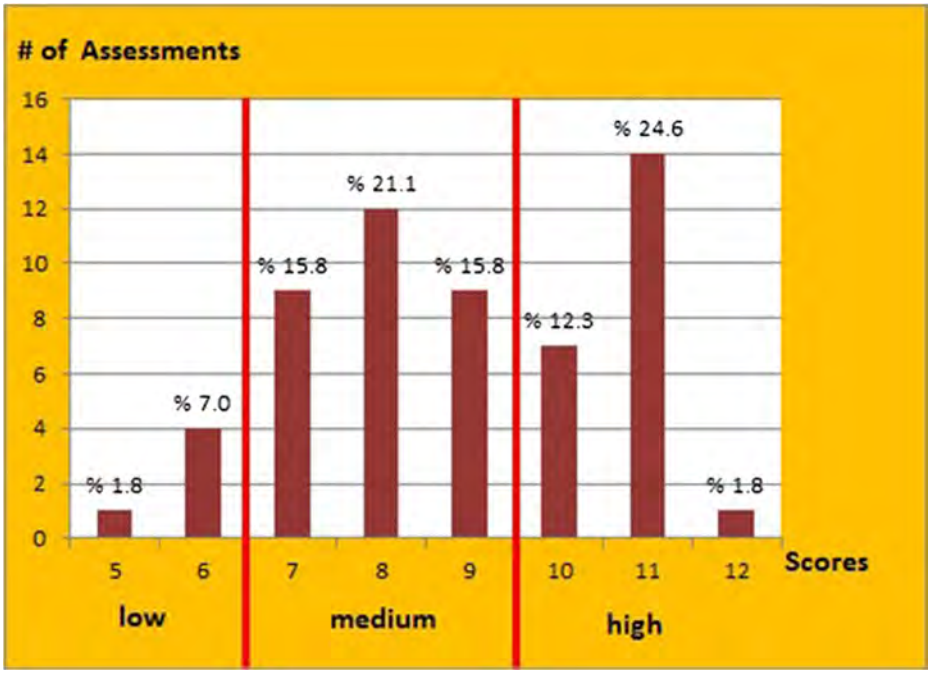

Fig. 8 Dimension 5

are shown in Table 4. When we compare the level of satisfaction, it may be more difficult for the developed assessments to (a) motivate students to engage in learning (73, 6%) and (b) support student learning (66, 5%). However, the developed assessments mostly satisfied the prompts related to the other three dimensions including (a) challenging students' thinking (82, 2%), (b) supporting instruction (81, 7%), and (c) reducing potential biases (89, 9%) to aid learning and instruction. Furthermore, some assessments met the criteria of all dimensions at a high level (assessment 1, 37, 50), while some assessments addressed the criteria well for some dimensions and less well for other dimensions (assessments 5, 27, 40). An assessment is not obliged to be effective at each dimension to be used since teachers may want to focus on a specific aspect, or they may wish to employ multiple assessments to address each dimension during their assessment practices. One of the difficulties in chemistry assessment is to motivate students to engage in assessment and learning process because of the abstract nature of chemistry topics. This also was the case for the teachers in this study, as the assessments that they developed were less successful in motivating and supporting students' learning dimensions. To eliminate the difficulty for developing motivating assessments, we need to think out of our traditional assessment understanding and search for ways that engage today's students. One such assessment (assessment 1) that achieved a higher score based on our rubric for dimension two and five was developed by Sophia in our case. As seen in Appendix 2, Sophia developed the assessment, Make a Movie, to let her students show their understanding of gas concepts by using a list of required vocabulary to make a movie related to a topic of their choices. Sophia developed this because "I (she) would like to see my students be more motivated to successfully complete one of my assessments. So often I will hear my students say things like, 'This is boring/stupid' or 'What is the point of this anyway?" What Sophia's students said is common for most students in terms of classroom assessment. In addition,

Table 4 Scores of each assessments rewarded based on each dimensions of the rubric

Dimension Assessments	D-1 (3*–9**)	D-2 (5–15)	D-3 (3–9)	D-4 (3–9)	D-5 (4–12)
1	9	14	7	9	11
2	8	13	9	9	9
3	7	12	9	8	9
4	8	10	8	9	8
5	6	10	8	8	10
6	7	9	9	8	11
7	5	6	6	6	6
8	9	10	9	8	11
9	6	8	5	8	9
10	9	11	8	7	10
11	9	10	8	8	11
12	4	8	5	6	7
13	6	10	7	8	8
14	8	11	8	9	9
15	8	10	8	7	8
16	8	9	7	7	8
17	7	13	7	9	11
18	6	7	5	6	8
19	5	7	5	7	7
20	8	8	6	8	7
21	8	11	7	7	8
22	8	9	7	7	7
23	8	7	7	7	6
24	6	9	6	8	7
25	4	5	6	8	7
26	7	10	6	7	8
27	8	12	8	8	6
28	7	11	9	9	11
29	7	10	7	8	9
30	7	7	8	8	9
31	9	12	8	8	11
32	6	9	7	8	8
33	7	10	6	9	9
34	7	7	7	9	10
35	8	10	7	9	7
36	7	8	8	9	6
37	7	7	7	9	7
38	8	13	7	9	11
39	8	9	8	8	8
40	8	11	8	8	7
41	7	13	8	9	8
42	8	11	8	9	12
43	9	13	9	9	11
44	7	13	7	5	10
45	6	9	7	9	8
46	7	9	6	7	8
47	8	12	8	9	11
48	9	11	8	8	11
49	9	11	9	9	10
50	9	8	8	8	9
51	9	13	8	9	10
52	5	7	5	7	5
53	8	10	8	9	9
54	9	12	8	9	10
55	9	13	9	9	11
56	6	9	7	9	9
57	9	12	8	9	11
Average scores/%	**7.40 (82.2)**	**9.98 (66.5)**	**7.35 (81.7)**	**8.09 (89.9)**	**8.83 (73.6)**

*The lowest point an assessment earn from related dimension

**The highest point an assessment can earn from related dimension

The bold entries in Column 1 represents the number of assessments that developed within the study

students' unwillingness to engage in assessment processes limits the influence of assessment on their learning since they do not concentrate and stay on assessment tasks longer. Thus, Sophia aimed to "… make an assessment that will show me (her) what they (students) really know about the topic (gases), as well as motivate them (students) to complete a high quality piece of work." Sophia benefitted from technology to develop the assessment because "I (she) feel that by bringing in some technology that they (students) might find entertaining will allow them (students) to stay on task longer and demonstrate their knowledge more effectively than a traditional quiz or book-work assignments." In practice, Sophia used the assessment as an end of unit project and students, in groups of two, prepared and presented their movie to their peers. After students' presentations of their movies, their peers and Sophia asked questions to group members related to their movies. The prepared movies acted as scaffolding for students to apply and illustrate their levels of understanding for selected gas concepts. Plus, as Sophia indicated, "This movie presentation assessment helped reduce some of the anxiety associated with public speaking, particularly for my ELL students." In addition, students' movies and answers of their peers' and Sophia's questions related to the movies produced informative feedback for Sophia to decide how to enhance her students' learning of related concepts. Besides, the peer review process, group work, and preparation of movies helped students to engage in peer learning and self-regulation of their own learning. As Sophia stated, "Students seemed to enjoy." However, Sophia faced with difficulty in "…creating an appropriate scoring guide." She thought that "It is very tough to make something fair and useable at the same time" and had questions in mind such as "How do I put a point value on creativity? What makes a project truly creative? Do you grade for correct grammar?"

In addition, assessments that were categorized in the low category for each dimension (assessments 7, 25, 52) can be easily seen. Because these assessments received low scores for each dimension, this indicated they did not have at least one dimension critical to support learning and teaching, and thus, we do not suggest them for teachers to use. One assessment example from low category, assessment 25, is given as Appendix 5. Furthermore, our special attention to these three assessments showed that (a) their lack of focus on explicit learning objectives (assessments 7 and 52), (b) their low cognitive requirement from students, as students follow a set of steps to make calculation or provide their answers to low-level multiple-choice questions, rather than engaging students using their reasoning to link ideas to show their learning (assessment 25), and (c) their limited use of scaffolding to motivate students to show their learning (assessments 7, 25, 52), as all of them were constructed as plain verbal tasks and require students to provide their responses verbally. However, the teachers developed the low-level assessments did not want to change the assessments because they believed the assessments achieved their aims. For instance, Sophia explained her aim for using assessment 25 (see Appendix 5) as, "I wanted to show students how the calories in their food is related to the calories and joules we have discussed in class and used in calorimetry equations." She believed that the assessment helped her to see how her students connect the calories in food with calories and joules they discussed in class; thus, she did not want to revisze the assessment because it met her goals for using it. The discrepancy in some of the assessment tasks between the teacher's goals and the researcher's critique is explored further in the next section on alignment by focusing on what students achieved on the assessment.

Effectiveness at Level 2: the Alignment Between What Developers Intend Students to Achieve, Through the Use of a Specific Assessment, and What They Actually Achieved

As explained earlier, effectiveness at level 2 compared what teachers intended to what students did on the assessment. In order to represent this effectiveness within the space limitations, we chose three examples of the developed assessments to discuss in detail. However, an example assessment from each of the five teachers was provided in the Appendix (see Appendixes 1, 2, 3, 4, and 6 and see Table 4 for no. 13, no. 1, no. 16, no. 49, and no. 51).

Example 1: Diesel Engine Assessment for Combustion Reactions One of the participating teachers, John, developed and used the assessment in Appendix 3. John, with the feedback of researchers and his colleagues, developed the assessment to assess his students' learning of combustion reactions, which he taught during a chemical reactions unit. He explained his intention to use the assessment in the ADT, "...to evaluate how well the students understand the concept of incomplete combustion reactions. The goal is also to evaluate how well the students can apply the knowledge gained during this unit [chemical reactions] to a context outside of the classroom." By providing a real-life situation, written feedback, and a three-step procedure as seen in Appendix 3, he believed this assessment can challenge students to think critically to evaluate and choose one reasonable cause and predict alternative causes for this phenomenon to show their conceptual understanding of the content and motivate students to apply their knowledge of combustion to outside of the classroom context. Furthermore, while developing the assessment, John predicted that his students would face some difficulty to comprehend the assessment because of their unfamiliarity with diesel engines. He explained this difficulty in the ADT as, "The most challenging aspect of this assessment keeping it simple enough so that students who have no experience with internal combustion engines understand what is being asked."

After using the assessment, John, in AUT, explained that "This assessment was used as a group test question at the end of unit 6 (chemical reactions) in my instructional sequence." In addition, John stated, "The students who had a difficult time with this assessment were not very familiar with how internal combustion engines work." Therefore, his intention of letting students work in groups in order to reduce potential biases such as unfamiliarity of provided context was meaningful but limited since other supporters such as visual representation of working process of the diesel engine in order to reduce biases was lacking (see Appendix 3). Also, during the Blackboard discussion he stated, "Once I informed them [students] that they did not need to know how an engine worked to answer the question, any anxiety they had over the question was removed." Thus, John's explanation showed that the assessment held some bias because a few students thought they need to know how an engine works, and John needed to inform students this was not the case to remove such biases.

Furthermore, one of the intentions of John was to motivate students to engage in learning. John in AUT explained, "I felt that this assessment was very appropriate and motivating for my student population. The scenario provided in the question is one that all of my students have experienced." John also answered the question asking for his students' engagement of the assessment during the Blackboard discussion as, "Judging from the discussions the students seemed engaged in this assessment. I did not notice any inhibiting factors through the use of this assessment." We found that the aim of engaging students in learning was accomplished. The assessment also challenged students to think critically in order to choose and provide a

reasonable explanation for the black smoke released from the engine (see Appendix 3, a). For example, some students had difficulties providing a reasonable cause for the black smoke and as John stated, "Some students believed that the smoke formed during the reaction was simply the un-burnt diesel fuel in solid form. I'm not sure where they came up with this idea." When students' responses were reviewed, it was seen that most of his students had failed to recognize formation of CO at the end of the reaction. Therefore, the assessment achieved its aim for showing students' difficulties and conceptual understandings (see Appendix 3, b). John also aimed to see students' understanding and application of combustion reaction. After his use of the assessment, in the AUT, he stated that, "From the assessment I was able to establish that the majority of my students understand that incomplete combustion reactions occur when there is not enough oxygen present."

Overall, the assessment satisfied most of the teacher's intention for using it. Except for reducing potential biases which was handled verbally, the teacher's reports of use showed effective implementation. This shows that even if an assessment provides some limitations, a teacher's use of the assessment can overcome or reduce its negative influence on students. In summary, the effectiveness was also demonstrated when John responded to a question regarding any changes he would make. His response illustrated the assessment accomplished his aims, "I would not make any changes to this assessment. I felt that the assessment served its purpose and helped me evaluate how well the students understood the concept of combustion reactions." Furthermore, the scores (see no. 16 at Table 4) awarded based on our rubric for effective assessments matched with the successful enactment of the engine assessment.

Example 2: Stoichiometry Recycling Challenge to Make Some Money One of the other assessments seen from Appendix 4 was developed and used by another teacher, Steve, in order to motivate students "…to adapt stoichiometry practices to something a little more real-life" (ADT). Steve chose a real-life context and developed an assessment to scaffold and engage students in applying their knowledge of stoichiometry and scientific practices. Furthermore, as seen from the assessment, Steve also wanted students to engage in "engineering practices by requiring them to make financial decisions based on their calculations" (ADT). Steve stated during his Blackboard discussion, "It (the assessment) requires students to use their mathematics ability to accurately calculate the mathematical processes in order to justify their claims." Therefore, the assessment has potential for supporting students to engage in Science, Technology, Engineering, and Mathematics (STEM) practices. Overall, evidence suggested that Steve had three main aims: (1) make students' use stoichiometry with a real-life example, (2) let students engage in engineering practices, and (3) employ mathematical processes to support their claims.

In practice, Steve used the assessment as a two-person group assignment after they had progressed through mass-mass stoichiometry. As Steve mentioned, most of his students were comfortable and excited during the application of the assessment while "…some students were a little unsure of how to proceed, since the types of questions were unusual to them" (AUT). Steve also explained on Blackboard that "peer learning since they were in groups" helped his students to understand and engage in providing their claims and justifications for the questions within the assessment task. We found the assessment task achieved Steve's aim for making students practice stoichiometry within a real-life context. On the other hand, Steve claimed that some of his students did not fully engage in the assessment process since "some students relied on their partner too much" (AUT). Thus, the context in which the assessment used limited Steve's aim for engaging all his students into practice as some of them did not get

responsibility to answer the task. For Steve's engineering practices aim, as seen from Appendix 4, questions within task (c) asked students to make financial sense by justifying and providing evidence showed the task achieved Steve's second aim, engaging students in the engineering practice of making financial sense of a project. Furthermore, when looking at the Appendix 4, students' responses (a, b) also showed that they used calculations to justify their claims about converting rust into iron. Therefore, the assessment task succeeded in Steve's aim for students using mathematical processes to support their claims.

In summary, the assessment Steve developed and used in his class for stoichiometry concepts mostly achieved the three aims he set for the task. One weakness of the task identified by Steve included using the task as group assessment because this may lead some students to not engage in using their knowledge of stoichiometry concepts but instead relied too much on their partners. The scores of Steve's assessment received from the rubric can be seen at Table 4 (no. 49).

Example 3: What Is This Pile of Stuff? Another assessment developed within our project by Julie was "What is this pile of stuff?" related to mole concepts in high school chemistry. According to Julie, she developed this assessment task "… to have students make the connection between mass and moles and to help them with their experimental design" (ADT). As seen from Appendix 2, in order to accomplish her aim, Julie provided a scenario within the task to have students engage in using experimental procedures. In addition, she stated she aimed "…to see if the students could apply the concept of molar mass" (ADT). Therefore, she used an experimental design scenario and related molar mass within the task; she planned to see how her students understand the concept of molar mass.

In practice, Julie used the task as a quiz after they had discussed moles and molar mass to see if her students could apply the concept, molar mass, to demonstrate their understanding. During the implementation of the assessment task, Julie also required her students to answer the two questions not written in the task as "Why did we do this activity? Explain how this activity relates to what we have done in class?" As seen in Appendix 1, students used the bags to employ experimental design steps to (a) make calculations to find moles, (b) convert moles into molar mass, (c) and to find the name of their elements placed in their bags. Therefore, the task required students to use experimental design steps and achieved Julie's aim for making students utilize experimental design to determine their elements. Furthermore, the assessment task also led students to use data and calculations to support their claims about their elements, which required students to use and show their understanding of the molar mass concept. As Julie, during the Blackboard discussion explained, "Many (students) were confused at first because they thought they would need more materials. I was fine with this, however, because of how I was using this to stretch their thinking." Thus, the task also engaged students in thinking about what they need to have in order to come up with their elements. Furthermore, Julie stated, "They (students) do not realize that using molar mass can be a two-way street. They were very comfortable using it as a conversion factor, but to go 'backwards' to use it to identify something was more difficult." Therefore, the assessment task was successful for Julie's aim for eliciting students' understanding of molar mass concept since the assessment task required students to use their knowledge of molar mass, and the task showed how and where the students had difficulties. On the other hand, some students could not find the elements in their bags because the calculated molar mass did not match with any elements in the periodic table. This was because of the masses of the elements given in the bags to the students. Therefore, Julie stated, "Make sure to measure the masses carefully so the mole amounts on the bags can allow students to get a reasonable molar mass, or identification can be difficult."

Briefly, Julie had three main intentions for the assessment: (a) have students use experimental design, (b) evaluate students' understanding of molar mass concept, and (c) provide context for moles and molar mass concepts. As seen from Appendix 1 and above explanations, we found that the assessment mostly achieved her aims for designing the assessment. The awarded scores based on the rubric can be seen in Table 4 (no. 13).

Conclusion and Discussion

Our aim for this study was to characterize the effectiveness of the assessments based on a model of effective assessment by focusing on their alignment with research-informed principles of effective assessments and their achievement for transforming teachers' intentions into classroom practices. Classroom assessments co-developed by five teachers were used as unit of analysis to illustrate the utilization of this model. The results showed that teachers' developed assessments had different levels of effectiveness from high to low for each of the five research-informed dimensions. Furthermore, based on teachers' self-reports and our analysis on students' works, the results showed that the assessments the teachers practiced mostly enabled them to achieve their intentions for design and use.

Assessment of science learning and reasoning is crucial for effective science instruction and the quality of science instruction will not be complete without using quality classroom assessments to enhance learning (Liu et al. 2008; NRC 2014; Kang et al. 2014; Pellegrino 2013). Existing assessments may not be effective in assessing and driving the new forms of science learning since they are not designed to capture and aid such learning (NRC 2014; Pellegrino 2013; Siegel and Wissehr 2011). Therefore, the design and/or selection and use of classroom assessments to effectively aid such forms of learning are desired, although it is hard for many teachers (Gottheiner and Siegel 2012; Lyon 2013; Kang and Anderson 2015; Pellegrino 2013). It is important to evaluate the effectiveness of existing and developed classroom assessments to be sure that they meet the expectations for using them in supporting such a desired science learning, which research is needed for (NRC 2014; NSTA 2015; Pellegrino 2013). Earlier efforts to evaluate effectiveness of assessments have mostly focused on providing a checklist for teachers to choose an assessment for their purposes (e.g., Brown et al. 2003), evaluating assessment items for ability to assess taxonomy such as the SOLO and Bloom (e.g., Opfer et al. 2012), and evaluating effectiveness of teachers' use and feedback for supporting learning (e.g., Haug and Ødegaard 2015). However, the current study provided a more comprehensive model that guides researchers and teachers to evaluate assessments based on their alignment with effectiveness principles *and* teachers' goals.

The newer notion of effectiveness for assessment values the ability of assessments for positively influencing learning and teaching rather than other psychometric factors such as reliability (Elton and Johnston 2002). Earlier notions of effectiveness for classroom assessment conceptualize assessment as a task or process and separately try to develop effective ways to enhance the quality of assessment tasks or teachers' use of assessment process (e.g., Wu et al. 2014). While developing and making quality classroom assessments available for use is reasonable (Kang et al. 2014), it is difficult to ensure they are used in a way to support learning and teaching as well as their effectiveness. Plus, focusing on teachers' use of assessment to evaluate effectiveness is sound, it is difficult for a successful teacher to elicit, monitor, and aid learning and teaching without having quality assessment tasks (Abell and Siegel 2011; Bennett 2011; Kang et al. 2014). However, this study, in contrast to others, goes

beyond the task only or process only studies to examine teachers' aims and executions and provides a more complete picture for effectiveness of assessments and considers both the theoretical and practical realm. One of the advantages of our approach is it focused both quantitative and qualitative approaches to elicit the quality of classroom assessments. This model lets researchers and teachers to evaluate assessments based on research-informed principles for their potential to support learning and teaching. The potential quality of an assessment is important, and using low quality assessments does not help teachers and students to aid learning-wasting instructional time (Abell and Siegel 2011; Gottheiner and Siegel 2012; NRC 2014; Kang et al. 2014; Pellegrino 2013; Sandlin, Harshman and Yezierski 2015). Furthermore, the alignment between teachers' intentions and assessments has been highlighted as an integral criterion for instructional sensitivity and is crucial for valid inferences that teachers make to support learning (e.g., Popham 2007; Ruiz-Primo et al. 2012; Sandlin et al. 2015). This alignment also forms an important component of data-driven inquiry, an assessment process (Sandlin et al. 2015) and is considered important for eliciting and supporting disciplinary content knowledge that mostly is neglected during assessment processes (Coffey et al. 2011; Harshman and Yezierski 2015). Thus, evaluating effectiveness of assessments is complex and requires consideration on both potential of an assessment and its success for transforming teachers' pedagogical and content-related aims in practice. However, there is little research conducted in this area in the forms of data-driven inquiry (Sandlin et al. 2015; Harshman and Yezierski 2015) and instructional sensitivity (Popham, 2007; Ruiz-Primo et al. 2012) and more research is needed in order to provide a more clear picture of effectiveness of assessments because of the importance of powerful assessments in driving new forms of learning. As a contribution to this area, the current study takes our attention to the complex but important concept effectiveness of assessment and provides examples of effectiveness at two levels based on a model. Level 1 is useful for teachers thinking about ways to improve their design of new or refinement of developed assessments. Level 2 is useful for teachers in showing particular ways others have implemented the assessments and how teachers have conceptualized the purpose and intent of an assessment and how it went within real classroom context. Teachers can use level 2 to share the successes and difficulties they faced during usage of a specific assessment task and get colleagues ideas for overcoming the challenges to improve their practice. This study also provides evidence for effectiveness of assessments and a rubric for teachers' use.

In the current study, science educators, chemistry specialist, and teachers collaboratively engaged in the assessment development process. As shown in Fig. 3, it was an iterative development process. A successful assessment design requires collaboration among different peoples such as educators, content experts, and teachers (DeBarger et al. 2013). Thus, this study engaged the teachers in a co-development process within a community-centered PD program. This process is also known as community of practice and provides benefits for both teachers and researchers (Wenger 1998; Wenger et al. 2002). Community of practice provides an effective model for collaboration among individuals at different institutions to change and enhance a desired practice, in this case, classroom assessment (Kislov et al. 2011). There are three essential features of a community of practice, domain, community, and practice, and when these three features are merged, community of practice is most effective (Wenger 1998). In our case, the community of practices helped teachers to reflect on their current practices, share their experiences and struggles with a community formed by colleagues and researchers, and gain theoretical and practical understanding about assessment process. Providing a supportive community is important and the contributions of each participant should be valued

in order to continue the partnership to change teachers' practices (Bartholomew and Sandholtz 2009). Otherwise, even if PD programs help to enhance teachers' understanding of assessment, their practices will not be changed as shown by researchers (DeBarger et al. 2013; Herman et al. 2015). Thus, we need to be careful when trying to enhance teachers' assessment literacy because just designing PDs to inform teachers about different strategies, purposes, and ways for use of assessment to improve teachers' theoretical understanding of and orientation to assessment, which is necessary but not enough to reform teachers' assessment practices (Abell and Siegel 2011; Gottheiner and Siegel 2012; Xu and Brown 2016). On the other hand, the co-development process can also assist researchers to share their knowledge and experiences with teachers, learn what work or what does not work in practice in terms of classroom assessment, reflect on and change their efforts for transforming teachers' assessment practices, and refine and regulate their theoretical ideas in light of practice (Bartholomew and Sandholtz 2009). However, we need to be aware that community of practice provides new problems because of having different goals, different understanding of the roles of teachers and researchers, and different context the partners came from (Bartholomew and Sandholtz 2009; LePage et al. 2001). Thus having knowledgeable, experienced, and well-trained teacher educators is important to support teachers' assessment practices (Abell and Siegel 2011; Bell 2007). Luckily, we did not experience the same struggles within the collaborative program in this study because the goals were the same: all participants' knowledge and skills were valued, and participants had the freedom to choose, develop, and use assessment tasks during the collaboration.

An essential goal of assessment design is to understand what students know and can do. The articulation of claims about what students should know and be able to do need to be guided by educators and content experts. Developing appropriate assessments and scoring guidelines requires both content knowledge and experience with students and should involve teachers. Plus, assessment development should be an iterative process that provides revisiting and revising goals, contents, and items (DeBarger et al. 2013).

The results provided examples of assessments that are not effective at all dimensions but are still worthy of use in classrooms. It is suggested teachers need to build an assessment system that should include a variety of assessments that can be used for different purposes to aid integrated science learning (NRC 2014; Pellegrino 2013). No one assessment can provide adequate information for measuring NGSS learning (NRC 2012). Thus, the five-dimensional rubric used in this study can guide teachers to design and choose a variety of assessments for various purposes to assess and support learning. The results of the study showed that the guidance and supportive community helped teachers to construct assessments that mostly transfer their intentions into classroom practices as exemplified by John and Steve's assessment practices. Thus, even if all developed assessment within the collaborative program did not satisfy all five dimensions of our rubric, they mostly achieved a teacher's intention for use and were effective at level 2 of our model. Furthermore, as John and Steve's practices showed that even if the assessment tasks they used provided limitations for their practices, they were aware of these and took actions (such as using the task as group activity, warning students they did need to know how an engine works to response the task) to overcome these issues to achieve their goals. This is important because a main obstacle has been that most novelties, even if they are high quality, are not adopted by most of teachers because of the lack of researchers' support and teachers' involvement in development process (Penuel and Yarnall 2005). Thus, as Penuel and Yarnall (2005) reported the benefits of co-development process for instructional software design, the co-development process in our study also supported teachers' adoption and practices of assessment or assessment literacies and made them to

adjust their implementation in a way to accomplish their goals for use. One of the other ways to support teachers effectively design and use assessment tasks to aid learning is design-based research, which is a process requiring use of design principles and scientific methods to let collaborators develop appropriate products and solutions for educational context (Easterday, Lewis, and Gerber 2014). Design-based research requires an iterative design process since education is a complex process and the outcomes of an educational intervention are not easily predictable (Easterday et al. 2014). Thus, the iterative assessment design process (see Fig. 3) used in the study is important since it has the potential to support researchers and teachers to engage in iteration to develop effective assessment tasks (McKenney 2018). This process also can support teachers in aligning assessment with their learning goals and finding different ways to elicit and assess their student learning to avoid development of ineffective assessment tasks. Just three assessments the teachers developed within the program did not satisfy any dimension of our rubric. Details of these assessments showed that main problems with these assessments were (a) they lacked a clear objective and (b) provided limited scaffolding for students to engage and show learning. Thus, we, as researchers, need to provide ways to for teachers to start assessment development process by choosing a clear learning goal. Teachers also faced difficulties to provide scaffolding; thus, teachers should be informed about types and effectiveness of different scaffolding and be engaged in using various scaffoldings to achieve their aims of assessing and supporting learning (Kang et al. 2014).

Limitations and Implications

There are a few limitations that the authors recognized for this study. Firstly, the study used a rubric to show alignment of assessments with five important dimensions for an assessment to be effective. However, it is possible to set other criteria beyond the scope of the rubric to define effectiveness and use different methods to illustrate effectiveness. Therefore, evaluating effectiveness of assessments for supporting learning and teaching is a hard task and still needs to be investigated. Also we emphasize that an assessment does not need to succeed at each of the dimensions of the rubric to be an effective assessment for a teacher to use. Secondly, the study, in order to illustrate alignment between teachers' goals and practiced assessments, used teachers' self-reports and researchers' review of students' responses on three assessments. However, more classroom observations could have been done by researchers for teachers' use of assessments to provide a clearer picture of alignment. Furthermore, student interviews could be conducted to see how teachers' assessment practices are successful and perceived by students. The study also focused on a small sample of teachers and a specific content area, chemistry. Thus, while the implications seem important for other content areas, the results are not intended to generalize for other teachers and content areas. Furthermore, conducting more observations of teachers' assessment practices can provide robust pictures of effectiveness for the alignment between teachers' intentions and real practices of assessments.

The results of the study have some important implications for teachers and science educators. Firstly, the results of the study showed that engaging in an iterative and collaborative assessment development process including reflections, revisions, and implementations supported teachers for selecting and developing their own assessments to reach their own aims. Thus, it is important for science educators to collaborate with science teachers to help them engage in developing their own assessments to meet their own aims by taking ownership that can make teachers to effectively use those assessments in own classrooms since other assessments developed by outsiders may not satisfy their aims and can be useless to

implement. Secondly, the results suggested that considering assessment both as a process and task is more complete, thus providing and exemplifying a complete picture of assessment during the teacher training program will be helpful for teachers to design and engage in effective assessment practices. Just providing quality assessment tasks does not aid using assessment to support learning since the interactions between a teacher, the instructional practice, and the school context impact decision-making process of the teacher to adapt the practice (Shaharabani and Tal 2017). Furthermore, the rubric developed and used in the study to evaluate effectiveness of the assessment tasks will be helpful both for teachers to select and design their own assessments and for science educators to highlight the principles of effective assessments during training teachers and providing PD for practicing teachers. This rubric also can be an example for science educators and researchers who may aim to develop their own rubric for evaluating effectiveness of instructional materials including assessments tasks too.

References

Abedi, J., Hofstetter, C., & Lord, C. (2004). Assessment accommodations for English language learners: implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1–28.

Abell, S. K., & Siegel, M. A. (2011). Assessment literacy: What science teachers need to know and be able to do? In D. Corrigan, J. Dillon, & R. Gunstone (Eds.), *The professional knowledge base of science teaching* (pp. 205–221). The Netherlands: Springer.

Abrahams, I., & Millar, R. (2008). Does practical work really work? A study of the effectiveness of practical work as a teaching and learning method in school science. *International Journal of Science Education, 30*(14), 1945–1969.

Abrahams, I., & Reiss, M. J. (2012). Practical work: its effectiveness in primary and secondary schools in England. *Journal of Research in Science Teaching, 49*(8), 1035–1055.

Abrahams, I., Reiss, M. J., & Sharpe, R. M. (2013). The assessment of practical work in school science. *Studies in Science Education, 49*(2), 209–251.

Ateh, C. M. (2015). Science teachers' elicitation practices: insights for formative assessment. *Educational Assessment, 20*(2), 112–131.

Atkin, J. M., Black, P., Coffey, J., & National Research Council. (2001). *Classroom assessment and the national science education standards*. Washington, DC: National Academies Press.

Avargil, S., Herscovitz, O., & Dori, Y. J. J. (2012). Teaching thinking skills in context-based learning: teachers' challenges and assessment knowledge. *Journal of Science Education and Technology, 21*, 207–225.

Bartholomew, S. S., & Sandholtz, J. H. (2009). Competing views of teaching in a school-university partnership. *Teaching and Teacher Education, 25*(1), 155–165.

Bell, B. (2007). Classroom assessment of science learning. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 1105–1149). Mahwah: Lawrence Erlbaum.

Bell, B., & Cowie, B. (2001). *Formative assessment and science education*. Dordrecht: Kluwer Academic.

Belland, B. R., Walker, A. E., Kim, N. J., & Lefler, M. (2016). Synthesizing results from empirical research on computer-based scaffolding in stem education: a meta-analysis. *Review of Educational Research, 87*(2), 309–344.

Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–74.

Black, P., & Wiliam, D. (2009). Developing a theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31.

Boud, D., Dawson, P., Bearman, M., Bennett, S., Joughin, G., & Molloy, E. (2018). Reframing assessment research: through a practice perspective. *Studies in Higher Education, 43*(7), 1107–1118.

Brantlinger, E., Jimenez, R., Klingner, J., Pugach, M., & Richardson, V. (2005). Qualitative studies in special education. *Exceptional Children, 71*, 195–207.

Brown, E., Gibbs, G., & Glover, C. (2003). Evaluation tools for investigating the impact of assessment regimes on student learning. *Bioscience Education, 2*(1), 1–7.

Clark, R. (1988). School–university relationships: an interpretive review. In K. Sirotnik & J. Goodlad (Eds.), *School–university partnerships in action: concepts, cases, and concerns* (pp. 32–65). New York: Teachers College Press.

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.

Cooper, M. M. (2015). Why ask why? *Journal of Chemical Education, 92*(8), 1273–1279.

Creswell, J. W. (2012). *Qualitative inquiry and research design: choosing among five approaches*. Sage Publications.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438–481.

DeBarger, A. H., Penuel, W. R., & Harris, C. J. (2013). *Designing NGSS assessments to evaluate the efficacy of curriculum interventions*. Austin: Educational Testing Service.

Dori, Y.J. & Avargil, S. (2015). Teachers' understanding of assessment. In R. Gunstone (Ed.) Encyclopedia of science education. Springer Reference (www.springerreference.com) Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/SpringerReference3032482013-08-21.

Easterday, M. W., Lewis, D. R. & Gerber E. M. (2014). Design-based research process: problems, phases, and applications. *ICLS 2014 Proceedings, 317-324.*

Eaton, T. T. (2009). Engaging students and evaluating learning progress using collaborative exams in introductory courses. *Journal of Geoscience Education, 57*(2), 113–120.

Edwards, F. (2013). Quality assessment by science teachers: five focus areas. *Science Education International, 24*(2), 212–226.

Elton, L., & Johnston, B. (2002). *Assessment in universities: a critical review of research*. York: Learning and Teaching Support Network (LTSN) Generic Centre.

Furtak, E. M. (2012). Linking a learning progression for natural selection to teachers' enactment of formative assessment. *Journal of Research in Science Teaching, 49*(9), 1181–1210.

Furtak, E. M., Heredia, S., Morrison, D., & Renga, I. (2012). *Teacher development in the collaborative design of common formative assessment.* Paper presented at the American Educational Research Association. Vancouver, BC, Canada.

Gearhart, M., Nagashima, S., Pfotenhauer, J., Clark, S., Schwab, C., Vendlinski, T., Osmundson, E., Herman, J., & Bernbaum, D. J. (2006). Developing expertise with classroom assessment in K-12 science: learning to interpret student work. Interim findings from a 2-year study. *Educational Assessment, 11*(3&4), 237–263.

Gibbs, G., & Simpson, C. (2005). Conditions under which assessment supports students' learning. *Learning and teaching in higher education, 1*, 3–31.

Gottheiner, D. G., & Siegel, M. A. (2012). Experienced middle school science teachers' assessment literacy: Investigating knowledge of students' conceptions in genetics and ways to shape instruction. *Journal of Science Teacher Education, 23*, 531–557.

Harshman, J., & Yezierski, E. (2015). Guiding teaching with assessments: high school chemistry teachers' use of data-driven inquiry. *Chemistry Education Research and Practice, 16*(1), 93–103.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Haug, B. S., & Ødegaard, M. (2015). Formative assessment and teachers' sensitivity to student responses. *International Journal of Science Education, 37*(4), 629–654.

Herman, J., Osmundson, E., Dai, Y., Ringstaff, C., & Timms, M. (2015). Investigating the dynamics of formative assessment: relationships between teacher knowledge, assessment practice and learning. *Assessment in Education: Principles, Policy & Practice, 22*(3), 344–367. https://doi.org/10.1080/0969594X.2015.1006521.

Horn, I. S., & Little, J. W. (2010). Attending to problems of practice: routines and resources for professional learning in teachers' workplace interactions. *American Educational Research Journal, 47*(1), 181–217.

Izci, K. (2013). *Investigating high school chemistry teachers' perceptions, knowledge and practices of classroom assessment*. (Unpublished PhD Thesis), Columbia: University of Missouri-Columbia

Kang, H., & Anderson, C. W. (2015). Supporting preservice science teachers' ability to attend and respond to student thinking by design. *Science Education, 99*(5), 863–895.

Kang, H., Thompson, J., & Windschitl, M. (2014). Creating opportunities for students to show what they know: the role of scaffolding in assessment tasks. *Science Education, 98*(4), 674–704. https://doi.org/10.1002/sce.21123.

Kislov, R., Harvey, G., & Walshe, K. (2011). Collaborations for leadership in applied health research and care: lessons from the theory of communities of practice. *Implementation Science, 6*(64), 1–10.

Koh, K., Burke, L. E. C., Luke, A., Gong, W., & Tan, C. (2018). Developing the assessment literacy of teachers in Chinese language classrooms: a focus on assessment task design. *Language teaching research., 22*(3), 264–288.

LePage, P., Boudreau, S., Maier, S., Robinson, J., & Cox, H. (2001). Exploring the complexities of the relationship between K-12 and college faculty in a nontraditional professional development program. *Teaching and Teacher Education, 17*(2), 195–211.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills: Sage.

Liou, P. Y., & Bulut, O. (2017). The effects of item format and cognitive domain on students' science performance in TIMSS 2011. *Research in Science Education*. https://doi.org/10.1007/s11165-017-9682-7.

Liu, O., Lee, H., Hofstetter, C., & Linn, M. C. (2008). Assessing knowledge integration in science: construct, measures, and evidence. *Educational Assessment, 13*(1), 23.

Lyon, E. G. (2011). Beliefs, practices, and reflection: exploring a science teacher's classroom assessment through the assessment triangle model. *Journal of Science Teacher Education, 22*(5), 417–435.

Lyon, E. G. (2013). Learning to assess science in linguistically diverse classrooms: tracking growth in secondary science preservice teachers' assessment expertise. *Science Education, 97*(3), 442–467. https://doi.org/10.1002/21059.

McKenney, S. (2018). How can the learning sciences (better) impact policy and practice? *Journal of the Learning Sciences, 27*(1), 1–7.

Namdar, B., & Shen, J. (2015). Modeling-oriented assessment in K-12 science education: a synthesis of research from 1980 to 2013 and new directions. *International Journal of Science Education, 37*(7), 993–1023.

National Research Council. (2001). Knowing what students know: The science and design of educational assessment. Committee on the Foundations of Assessment. In J. W. Pellegrino, N. Chudowsky, & R. Glaser (Eds.), *Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education*. Washington, DC: The National Academy Press.

National Research Council. (2007). *Taking science to school: learning and teaching science in grades K-8*. Washington, D.C.: The National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: practices, crosscutting concepts, and core ideas*. Washington, D.C.: The National Academies Press.

National Research Council. (2014). *Developing assessments for the next generation science standards*. Washington, D.C.: The National Academies Press.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218.

NSTA (2015). *Educators evaluating the quality of instructional products (EQuIP) rubric for lessons and units: Science (Version 2.0)*. Retrieved from http://www.nextgenscience.org/sites/ngss/files/EQuIP%20Rubric%20for%20Science%20052714.pdf.

Opfer, J. E., Nehm, R. H., & Ha, M. (2012). Cognitive foundations for science assessment design: knowing what students know about evolution. *Journal of Research in Science Teaching, 49*(6), 744–777.

Pellegrino, J. W. (2013). Proficiency in science: assessment challenges and opportunities. *Science, 340*, 320–340. https://doi.org/10.1126/science.1232065.

Penfield, R. D., & Lee, O. (2010). Test-based accountability: potential benefits and pitfalls of science assessment with student diversity. *Journal of Research in Science Teaching, 47*(1), 6–24.

Penuel, W. R., & Yarnall, L. (2005). Designing handheld software to support classroom assessment: analysis of conditions for teacher adoption. *The Journal of Technology, Learning and Assessment, 3*(5), 50–70.

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

Puntambekar, S., & Kolodner, J. L. (2005). Toward implementing distributed scaffolding: helping students learn from design. *Journal of Research in Science Teaching, 42*, 185–217.

Quellmalz, E. S., Timms, M. J., Silberglitt, M. D., & Buckley, B. C. (2012). Science assessments for all: integrating science simulations into balanced state science assessment systems. *Journal of Research in Science Teaching, 49*(3), 363–393.

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 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.

Sandlin, B., Harshman, J., & Yezierski, E. (2015). Formative assessment in high school chemistry teaching: investigating the alignment of teachers' goals with their items. *Journal of Chemical Education, 92*(10), 1619–1625.

Sato, M., Wei, R. C., & Darling-Hammond, L. (2008). Improving teachers' assessment practices through professional development: the case of National Board Certification. *American Educational Research Journal, 45*(3), 669–700.

Shaharabani, Y. F., & Tal, T. (2017). Teachers' practice a decade after an extensive professional development program in science education. *Research in Science Education, 47*(5), 1031–1053.

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., et al. (2008). On the impact of curriculum-embedded formative assessment on learning: a collaboration between curriculum and assessment developers. *Applied Measurement in Education, 21*(4), 295–314.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*, 4–14.

Siegel, M. A. (2007). Striving for equitable classroom assessments for linguistic minorities: Strategies for and effects of revising life science items. *Journal of Research in Science Teaching, 44*(6), 864–881.

Siegel, M. A. (2012). Filling in the Distance Between Us: Group Metacognition During Problem Solving in a Secondary Education Course. *Journal of Science Education and Technology, 21*(3), 325–341.

Siegel, M. A. (2014). Developing preservice teachers' expertise in equitable assessment for English learners. *Journal of Science Teacher Education, 25*(3), 289–308. https://doi.org/10.1007/s10972013-9365-9.

Siegel, M. A., & Wissehr, C. (2011). Preparing for the plunge: Preservice teachers assessment literacy. (2011). *Journal of Science Teacher Education, 22*, 371–391. https://doi.org/10.1007/s10972-011-9231-6.

Siegel, M. A., Markey, D., & Swann, S. (2005). *Life science assessments for English learners: A teacher's resource*. Berkeley: University of California.

Siegel, M. A., Roberts, T. M., Freyermuth, S. K., Witzig, S. B., & Izci, K. (2015). Aligning assessment to instruction: Collaborative group testing in large enrollment science classes. *Journal of College Science Teaching, 44*(3), 74–82.

Songer, N. B., & Gotwals, A. (2012). Guiding explanation construction by children at the entry points of learning progressions. *Journal of Research in Science Teaching, 49*(2), 141–165.

Stiggins. (2001). *Student-involved classroom assessment* (3rd ed.). Upper Saddle River: Prentice-Hall.

Stiggins, R., & Chappuis, J. (2005). Using student-involved classroom assessment to close achievement gaps. *Theory Into Practice, 44*(1), 11–18.

Talanquer, V., Bolger, M., & Tomanek, D. (2015). Exploring prospective teachers' assessment practices: noticing and interpreting student understanding in the assessment of written work. *Journal of Research in Science Teaching, 52*(5), 585–609.

Voogt, J., Laferrière, T., Breuleux, A., Itow, R. C., Hickey, D. T., & McKenney, S. (2015). Collaborative design as a form of professional development. *Instructional Science, 43*(2), 259–282.

Wenger, E. (1998). *Communities of practice: learning, meaning and identity*. Cambridge: University of Cambridge.

Wenger, E., McDermott, R. A., & Snyder, W. (2002). *Cultivating communities of practice: a guide to managing knowledge*. Brighton: Harvard Business Press.

Wu, P. H., Wu, H. K., & Hsu, Y. S. (2014). Establishing the criterion-related, construct, and content validities of a simulation-based assessment of inquiry abilities. *International Journal of Science Education, 36*(10), 1630–1650.

Xu, Y., & Brown, G. T. (2016). Teacher assessment literacy in practice: a reconceptualization. *Teaching and Teacher Education, 58*, 149–162.

Yin, R. K. (2009). *Case study research: Design and methods* (3rd ed.). Thousand Oaks: Sage Publications, Inc..

Affiliations

Kemal Izci[1] · Nilay Muslu[2] · Shannon M. Burcks[3] · Marcelle A. Siegel[3,4]

[1] Department of Educational Sciences, Ereğli College of Education, Necmettin Erbakan University, Konya, Turkey

[2] Department of Secondary Science and Mathematics Education, Mugla Sıtkı Koçman University, Mugla, Turkey

[3] Department of Learning, Teaching, & Curriculum: MU Science Education Center, University of Missouri, Columbia, MO, USA

[4] Department of Biochemistry, University of Missouri, Columbia, MO, USA