International Journal of Gender, Science and Technology
http://genderandset.open.ac.uk

Selected papers presented at the 5th Network Gender & STEM Conference, 29–30 July 2021, in Sydney, Australia

In association with
NETWORK GENDER & STEM educational and occupational pathways and participation

Building Elementary School Teachers' Capacity in the Design and Implementation of Authentic STEM Assessments for Girls

Kim Koh, Olive Chapman, & Shimeng Liu
University of Calgary, Canada

ABSTRACT
Improving science, technology, engineering and mathematics (STEM) curricula and promoting the participation of historically underrepresented groups, including girls in STEM-related majors and careers, have become important policy initiatives in many countries. Teachers' use of innovative pedagogical and assessment approaches in the classroom learning environment has the greatest impact on student learning and choices of STEM-related fields of study and careers. Therefore, there is an urgent need to build teacher capacity in the design and development of instructionally sensitive, authentic assessments with the potential to promote engagement of elementary school girls in STEM learning. This article is part of a larger design-based research study on building capacity for Canadian elementary school teachers in the design and implementation of authentic STEM assessments to promote girls' STEM self-efficacy and interest in STEM. Participants included three Grades 5 and 6 teachers and their students. Included data sources are teacher-designed assessment tasks, teacher interviews, classroom observations and student self-reflections. Findings indicate that despite their capacity to design assessment tasks with a real-world problem, teachers tended to focus on the solicitation of students' factual and procedural knowledge, lacked capacity to promote students' integration of mathematics and sciences, and encountered some implementation challenges. Although there were positive effects of the authentic STEM assessment on girls' development of a growth mindset, interest in mathematics and investigative skills, teachers should be more intentional in increasing the intellectual engagement of girls in STEM through the creation of authentic assessment tasks that are more gender-responsive and that focus on eliciting higher-order cognitive skills. These findings underscore the importance of providing elementary school teachers with sustained professional development to build their capacity to design and implement high-quality authentic STEM assessments (i.e., high cognitive demand tasks) for girls.

KEYWORDS
elementary school teachers; authentic STEM assessments; girls; building teacher capacity; research partnership; STEM integration; design and implementation challenges; professional development

Building Elementary School Teachers' Capacity in the Design and Implementation of Authentic STEM Assessments for Girls

INTRODUCTION

To have the greatest impact on girls' STEM learning, intervention should be implemented as early as the elementary school level because many girls begin to develop STEM conceptions and identities before and during their elementary education (e.g., Koh et al., 2021; Murphy & Mancini-Samuelson, 2012). Archer et al. (2012, 2013), Ceci and Williams (2010), DeWitt et al. (2011), Heaverlo et al. (2013) and Maltese and Tai (2010) found that gender disparity in STEM self-efficacy and interest begins in elementary school. Such a disparity is typically attributed to gender stereotypes arising from girls' socialization. A substantial body of literature has also pointed out that girls tend to fall behind boys in standardized tests of science and mathematics achievement due to test anxiety (Devine et al., 2012; Putwain & Daly, 2014), false beliefs about their mathematics and science abilities (Perez-Felkner et al., 2017) and the nature of the assessment (Jovanovic et al., 1994; Reardon et al., 2018). For instance, Jovanovic et al. (1994, p. 354) note that when the assessment is focused on problem solving, reasoning and critical thinking (e.g., authentic assessment), the 'female disadvantage' in science disappears.

An authentic assessment is composed of open-ended, performance-based tasks (authentic tasks) that can create realistic learning experiences for girls as they are actively engaged in solving real-world problems through reasoning, critical thinking, collaboration, communication and self-directed learning (Koh, 2017; Wiggins, 1989). For a girl, her successful performance on authentic tasks in STEM will likely be viewed as having better 'utilitarian, aesthetic or personal value' (Newmann et al., 1996, p. 284) compared to multiple-choice or short-answer tests that tend to reinforce rote learning and memorization. Hence, we posit that an intervention approach focusing on authentic STEM assessment in a formal learning environment may help elementary school girls improve their STEM self-efficacy and interest in STEM subjects and careers.

Based on our previous research, we learned that elementary school teachers need continuing professional learning and support to build their capacity in the design and implementation of authentic assessments. If we envision effecting change in girls' STEM self-efficacy (growth mindset included) and interest in STEM, there is a need to explore how to best support teachers in authentic STEM assessment, particularly in designing and using authentic assessment tasks to facilitate girls' STEM learning, with an eye toward closing the gender gap in STEM. Our previous research has shown that ongoing, sustained professional development in authentic assessment is effective for inservice teachers to improve their assessment literacy, especially in the design and implementation of authentic tasks to promote the learning of elementary school students in

single subjects such as English, mathematics, science and Mandarin Chinese (Koh, 2011a; 2014; Koh et al., 2018).

In the current study, instead of providing a series of traditional workshops for teachers, we pursued a research partnership in which the elementary school teachers collaborated with us (i.e., university researchers) to design and implement authentic STEM assessments. This collaboration serves as a means to enhance the teachers' professional learning and reflective practice (Schön, 1983). Research-practice partnerships are defined by Coburn and Penuel (2016) as 'long-term collaboration[s] between practitioners and researchers that are organized for investigating problems of practice and for developing solutions for improving school practice and even school districts' (p. 48). Many scholars have pointed out that teachers must be treated as active agents in research partnerships and that teachers' active involvement in designing, implementing and reflecting on teaching units helps improve their instructional and assessment practices (Garet et al., 2001; Henrick et al., 2017). The teachers in our study worked collaboratively to plan, design, enact and reflect upon an authentic STEM assessment prototype.

This paper is part of a larger design-based research (DBR) study investigating the design of authentic STEM assessment and its effects on elementary school girls' STEM self-efficacy and interest in STEM subjects and careers. We analysed and reported the data collected from the first two phases of the study, namely teacher-designed authentic STEM assessment tasks, teacher interviews, classroom observations and student self-reflections. This paper presents our exploration of the intellectual quality of the teacher-designed authentic STEM assessments to understand the teachers' design and implementation challenges, particularly when engaging girls in STEM learning with boys in mixed classrooms. Specifically, we examine what aspects of the authentic assessment task design and implementation can be more gender-responsive, enhancing girls' STEM learning experience in a formal classroom environment.

The focus of our work corresponds to a call by the United Nations Children's Fund (2020) underscoring the importance of designing and implementing gender-responsive pedagogical approaches to improve formal learning opportunities for girls in STEM. We posit that gender-responsive assessments for learning and equity can lead to improved cognitive and affective learning outcomes for girls. An intentional focus on assessment for learning and equity (or formative assessment for equity) would enable STEM teachers to address gender stereotypes; design classroom assessment tasks that are conducive to girls' growth mindsets; and encourage boys and girls to participate, collaborate and support each other while completing the tasks.

The intellectual quality of the authentic assessment tasks designed by the teachers in our study was measured by the eight criteria for authentic intellectual quality: *depth of knowledge, knowledge criticism, knowledge manipulation, extended communication, clarity and organization, making connections to the real world beyond the classroom, student control and explicit performance standards or success criteria* (Koh, 2011a, 2011b; Koh et al., 2020). The following research questions were answered:

(1) What is the intellectual quality of teacher-designed authentic assessment tasks in STEM?
(2) What are the teachers' perceptions of their experiences when
    (a) designing authentic STEM assessments for addressing gender disparity in STEM learning, and
    (b) implementing authentic STEM assessments in mixed-gender classrooms?

The findings will help the researchers determine which aspects of the authentic assessment task design and implementation can be made more gender-responsive and conducive to STEM learning for girls.

LITERATURE REVIEW

Gender disparity in STEM and standardized tests

To date, both policy documents and extant research have consistently urged for finding effective ways to attract and retain women in STEM to ensure a nation's competitiveness. However, a common phenomenon encountered by many countries is that women comprise a relatively lower proportion of STEM graduates and careers, including countries with greater gender equality such as Canada, the UK, Finland and Sweden (Stoet & Geary, 2018). In Canada, women aged 25 to 34 were underrepresented at 39% of total STEM graduates (Statistics Canada, 2011) and women aged 25 to 64 made up only 23% of the STEM workforce (Statistics Canada, 2016). In the United States, only 23% of female high school students major in computer science and 29% in physics (National Science Foundation, 2018). According to Perez-Felkner et al. (2017), 'women are particularly underrepresented in physical, engineering, mathematics, and computer (PMEC) sciences' (p. 1) globally. In their latest report, the United Nations Children's Fund, ITU (2020) notes that 'female workers make up an estimated 26% of workers in Data and Artificial Intelligence roles, 15% of workers in Engineering roles and 12% of workers in Cloud Computing roles' (p. 13). This evidence suggests that gender disparity in STEM persists even though we are in the modern era of information technology.

The gender gap in STEM begins as early as the elementary school level (Ceci & Williams, 2010). Perez-Felkner et al. (2017) have pointed out that research shows girls tend to have lower perceptions of and beliefs about their mathematics and science abilities when compared to boys. For example, Beghetto (2007), Cooper (2006) and Master et al. (2017) found that compared with boys, girls reported less interest in and confidence in science and technology. Such a phenomenon is due to gender stereotypes girls encounter in their socialization (e.g., boys are better in mathematics and sciences; boys are better than girls at robotics and programming, science and technology are not for women; men are more suitable for careers in engineering and physical and computer sciences). Gender stereotypes are found to be common across different cultures, including developed countries (Perez-Felkner et al., 2017).

Research has consistently found that boys outperformed girls on standardized tests that rely more heavily on multiple-choice items, which is a commonly used item format to measure mathematics and science knowledge. In contrast, girls outperform boys on assessments that rely more heavily on open-ended or constructed-responses items (e.g., writing an essay) or performance-based tasks (Jovanovic et al., 1994). Reardon et al. (2018) analysed eight million fourth- and eighth-grade standardized test scores drawn from a representative

sample of students in 42 states across the United States. The authors found that gender gaps and multiple-choice items were correlated and that the assessment item format explained approximately 25% of the variation in achievement gaps between male and female students. In addition, girls tended to experience a higher level of test anxiety than boys on standardized paper-and-pencil tests (Devine et al., 2012; Putwain & Daly; 2014). These gender achievement gaps on standardized tests persist until high school and college. Typically, 'women and underrepresented minorities are found to score significantly lower than men on standardized tests designed to predict achievement in undergraduate and graduate physics and [mathematics] courses' (Ripin, 1996, p.3).

The need for authentic assessment in STEM

The literature on STEM has shown that elementary school teachers lack preparation to teach and assess in STEM fields (e.g., Epstein & Miller, 2011; Kurup et al., 2019; Murphy & Mancini-Samuelson, 2012; Nadelson et al., 2013). For example, Lamberg and Trzynadlowski (2015) have pointed out that STEM academies or schools in the United States focus primarily on high school students. STEM curriculum at the elementary school level, which is critical for fostering students' interest in STEM, is not well-developed. Many elementary school teachers have limited access to high-quality instructional materials that enable them to engage students in rigorous learning about STEM. Moreover, there is a lack of assessment tools for measuring students' learning and performance in STEM (Gao et al., 2020; Harwell et al., 2015; Margot & Kettler, 2019). Even though some researchers attempt to develop assessment tools for STEM, the format of the assessment (i.e., multiple choice) is often contrived in nature and tends to measure decontextualized factual and procedural knowledge in separate STEM subjects (e.g., Harwell et al., 2015). Such a traditional assessment approach does not require students to engage in deep understanding and applications of conceptual knowledge, which is essential for their success in STEM (Saxton et al., 2014). Worst, it perpetuates gender disparity and reinforces girls' fixed mindsets about their ability in mathematics and sciences.

The Common Core State Standards for Mathematics and the Next Generation Science Standards in the United States, 'call for the engagement of students in authentic tasks that require integration across the STEM disciplines and support for the development and application of conceptual knowledge and reasoning' (*National Academy* of *Engineering* & *National Research Council*, 2014, p. 108). Scholars such as Roehrig et al. (2021) also point out an urgent need to create tools and rubrics assessing the quality of written integrated STEM curricula. Authentic assessment is deemed to be a viable approach due to its alignment with inquiry- and competency-based pedagogical approaches, which have been recommended to Canadian teachers to help achieve the vision of the *Canada 2067 STEM Learning Framework*. Such an approach allows for the design of a series of open-ended, performance-based tasks that support an integrated STEM curriculum. Students' engagement is also promoted toward a multidisciplinary or transdisciplinary approach to solve complex problems typically faced by scientists, mathematicians, engineers and other professionals in the real world. Therefore, teachers who can design and implement authentic STEM assessments would create exciting, hands-on learning experiences for girls. Additionally, girls

can be encouraged to act out the roles of male-oriented STEM professionals such as engineers and mathematicians.

In essence, authentic STEM assessment plays a pivotal role in realizing high-quality STEM education. The goal is to promote an equitable opportunity for every student, regardless of their gender and other sociodemographic backgrounds to develop not only STEM content knowledge, but also 21st-century competencies such as critical thinking, creativity and innovation, communication, collaboration and informational technology skills. These cross-cutting competencies also referred to as 'readiness skills' for STEM college and careers (Saxton et al., 2014), prepare girls and young women to thrive in a competitive world.

Research has shown that authentic and formative assessments are more effective than standardized tests and summative evaluations to capture students' demonstrations of conceptual understanding or disciplinary knowledge and competencies such as critical thinking, creative problem-solving, innovative design, collaboration, communication, inquiry-based, analytic habits of mind and growth mindsets (Koh, 2011a; Koh & Chapman, 2019). These competencies are the essential learning outcomes of STEM education (*National Academy of Engineering & National Research Council*, 2014). Therefore, we posit that girls' early exposure to authentic STEM assessment tasks that embed formative assessment (e.g., setting of learning goals, descriptive feedback rather than evaluative feedback, and self-assessment using high-quality rubrics) and instructional scaffolding during the learning process of STEM may help them develop a growth mindset (Koh et al., 2015; Koh et al., 2022).

The term 'growth mindset' was first coined by Stanford psychologist Carol Dweck (2016) to refer to the belief that a person's ability and talent can be improved through effort and persistence over time. When girls have a growth mindset, they believe that their success in STEM requires hard work rather than innate ability (Cheryan et al., 2017). Many girls engage less often with mathematical and scientific tasks during adolescence; however, girls who have developed a growth mindset are more likely to persevere on challenging mathematics tasks. As such, they are more willing to take on challenges and learn from trial and error, thereby increasing their confidence, self-efficacy, interest and achievement. There is also recognition of the value of embedding formative assessment or assessment for learning seamlessly into the design and implementation of authentic assessment tasks that provide insight into girls' mindsets and habits of learning associated with STEM. Formative assessment strategies such as self and peer assessments could help girls develop their metacognitive skills, communication, collaboration, confidence and self-efficacy. Taken together, these findings suggest that it is vital to build elementary school teachers' capacity in the design and implementation of authentic STEM assessments, especially for girls as early as the elementary school level.

Elementary school teachers' assessment literacy in STEM

Research has consistently pointed out that teachers generally lack assessment literacy, particularly in the design and use of authentic tasks to support students' learning and assess specific learning outcomes in the day-to-day classroom (DeLuca, 2016; Koh, 2011a; Koh et al., 2018). More specifically, many teachers reported that although they valued STEM education, they encountered

challenges such as making a shift from teacher-led to student-led pedagogies (i.e., inquiry-based/project-based/problem-based learning); integrating the STEM curriculum into their existing curricula (Qablan, 2021); struggling to design STEM lessons that integrate multiple disciplines; a lack of quality assessment tools, materials and resources; insufficient planning time; inadequate knowledge of STEM disciplines; and tension arising from a misalignment of STEM curriculum and standardized tests (Capraro et al., 2016; Margot & Kettler, 2019). In Falloon et al.'s (2021) case study, they found that teachers struggled to understand the experiential interdisciplinary STEM pedagogy, as they believed 'we just play…there is no assessments, there's no curriculum' (p.117) and the delivery to students was spontaneous.

Several problems associated with inservice teachers' inadequate assessment literacy in STEM have emerged in initial teacher preparation programs. Preservice teachers in mathematics and science are found to have a solid foundation in applying standards-based instruction but a weak development of their confidence in and efficacy for teaching STEM content (Moon et al., 2021; Nadelson et al., 2013). Among the inservice teachers, an increase in teaching experience does not automatically generate a higher efficacy for teaching STEM (Nadelson et al., 2013). In contrast to assessing student work using grading scales, the multiplicity of solutions to the design problem and the creativity and innovation emerging in the iterative engineering process can give rise to unprecedented difficulties (Bartholomew et al., 2018) for teachers' authentic assessment tasks design and implementation. Teachers' lack of assessment literacy has also been questioned by parents who misconstrued student-led, project-based STEM assessment as equivalent to loosening up the academic rigour mandated by traditional pedagogies (Breiner et al., 2012). Some studies (e.g., Lesseig et al., 2016) underscore the importance of creating a culture of collaboration between teachers and other teachers and between teachers and university researchers in the design and delivery of STEM programs/initiatives. Such professional support and access to expertise are critical for the successful design, planning and implementation of the interdisciplinary or transdisciplinary teaching and assessment required for STEM lessons. We also believe that elementary school teachers need support to know how to use authentic STEM assessment tasks to create an enabling learning environment in the day-to-day classrooms to promote girls' STEM self-efficacy and interest.

THEORETICAL FRAMEWORK

The following frameworks guided our work with the two elementary school teachers in the design of authentic STEM assessment.

Authentic STEM assessment

Authentic assessment includes authentic tasks that are rich and contextualized within real-world problems and that replicate the genuine intellectual challenges and performance standards that are typically faced by professionals in the field (Koh, 2017; Wiggins, 1989). As such, an authentic STEM assessment should create opportunities for elementary school girls to engage in thinking and acting like they were scientists, engineers, mathematicians and designers. In real life, these professionals are involved in collaborations to solve complex, multidisciplinary or transdisciplinary problems over an extended period. Therefore, the design of an authentic STEM assessment should consider performance tasks that enable an integration of STEM disciplines to solve real-

world problems. This emphasis will allow girls to apply STEM concepts and skills to solve the problem in an integrated manner through collaboration and communication with the other gender. Such an authentic learning experience could increase girls' understanding of how things work and improve gender-equitable opportunities in playing some professional roles, thereby improving girls' STEM self-efficacy and interest in STEM subjects and careers.

To promote girls' interest and intrinsic motivation, authentic tasks should be intellectually engaging (i.e., high cognitive demand tasks) and perceived as having 'utilitarian, aesthetic or personal value' (Newmann et al., 1996, p. 284). To engage all students' interests in STEM, Howes et al. (2013) state that 'the core disciplines in STEM will therefore need to be strengthened – adding opportunities to motivate STEM learning through its connections to enjoyment, aesthetics, societal needs, and real problem solving' (p. 15). This suggestion supports Newmann et al.'s (1996) emphasis on students' perceived relevance and value (i.e., value beyond school) of authentic assessment. These theoretical underpinnings led us to the use of a patchwork text approach, the structure of the observed learning outcomes (SOLO) taxonomy and the authentic intellectual quality (AIQ) criteria in informing the design of an authentic STEM assessment for girls. We posit that an authentic STEM assessment involves patches of performance tasks at different levels of intellectual challenge or cognitive demand, which align with the identified learning outcomes in a STEM unit of work, culminating in a final piece (e.g., a design project).

Criteria for authentic intellectual quality (AIQ)

The levels of intellectual challenge in the authentic STEM assessment tasks were informed by the AIQ criteria (Koh, 2011a, 2011b; Koh & Chapman, 2019; Koh & Luke, 2009; Koh et al., 2020), which were adapted from Newmann et al. (1996) authentic intellectual work and the revised Bloom's taxonomy of knowledge (Anderson & Krathwohl, 2001). The AIQ consists of eight criteria and their respective elements: depth of knowledge (factual knowledge, procedural knowledge, advanced concepts), knowledge criticism (presentation of knowledge as a given, comparing and contrasting knowledge, critiquing knowledge), knowledge manipulation (reproduction; organization, interpretation, analysis, synthesis and/or evaluation of information; application or problem solving; generation or construction of new knowledge), extended communication, clarity and organization, making connections to the real world beyond the classroom, student control and explicit performance standards or success criteria. A high-quality authentic assessment should place greater demands on higher-order cognitive outcomes such as advanced concepts; comparing and contrasting knowledge; critiquing knowledge; organization, interpretation, analysis, synthesis and/or evaluation of information; application or problem solving and generation or construction of new knowledge. Due to word limits, please see Koh (2011a, 2011b) and Koh & Luke (2009) for detailed descriptions of AIQ criteria.

The structure of the observed learning outcome taxonomy (SOLO)

The structure of the observed learning outcome (SOLO) taxonomy (Biggs & Collis, 1982) was used to guide the teachers in our research to identify STEM learning outcomes when they co-wrote a STEM unit of work. The taxonomy describes levels of increasing intellectual challenge or cognitive demand in a student's understanding of a subject (i.e., learning progression; Hattie, 2012), through five stages: pre-structural, uni-structural, multi-structural, relational

and extended abstract. At the pre-structural level, assessment tasks and learning activities are used to gauge and develop students' fundamental knowledge and skills. In progressing from uni-structural to multi-structural, students acquire relevant knowledge and skills. At the relational level, students' higher-order thinking manifests in their ability to make and explain connections between different ideas around a related topic or theme. This is when they are required to synthesize and construct knowledge (Koh & Burke, 2018). The performance tasks in the authentic STEM assessment (See Figure 1) were mapped to the identified learning outcomes in the SOLO taxonomy, which were clearly laid out in the teachers' instructional and assessment plan.

Patchwork text approach

Integrated STEM programs and their assessments should identify the knowledge and skills to be monitored during learning activities and tests at the culmination of a project (Crismond, 2001). A patchwork text approach to assessment is an integrated overall design made up of small segments or tasks, each of which is complete and related to a common theme (Winter, 2003). These small tasks are carried out at regular intervals throughout a unit of work. The tasks are stitched together or culminated into the final integrative task (e.g., a design project). Each of the small tasks assesses a range of learning outcomes, with increasing cognitive demands or intellectual challenges. The final integrative task or project enables students to apply the knowledge and skills that they have learned from the previous tasks. Hence, its design principles align well with the SOLO taxonomy and authentic assessment. The patchwork approach enables teachers to provide instructional scaffolding and formative feedback to help students move forward in their learning and improve the quality of their work. As such, students become more motivated to tackle difficult tasks, resulting in increased confidence and interest in STEM subjects and careers.

Table 1

Design Template for Authentic STEM Assessment

SOLO taxonomy	Identified learning outcomes	Example of patchwork task	Assessment
Pre-structural: Pre-knowledge or surface understanding and knowledge Uni-structural: Deepening surface understanding and more Multi-structural: Performing the task based on knowledge and understanding of extracting relevant information to solve real-world problems Relational: Appreciating the significance of the parts in relation to the whole (integration and connection of knowledge—ideas are understood in connection with other ideas) Extended Abstract: Making connections in STEM and beyond; extending/ transferring learning to new, real-world contexts			

Table 1 illustrates the design template for mapping the identified learning outcomes and assessment tasks across different levels of intellectual challenges or cognitive demands using the SOLO taxonomy and patchwork text approach.

METHOD

We employed a DBR methodology (Design-Based Research Collective, 2003) to guide our collaboration with the teachers in our study to design, develop and implement an authentic STEM assessment prototype. DBR 'requires more than simply showing a particular design works but demands that the researcher move beyond a particular design exemplar to generate evidence-based claims about learning that address contemporary theoretical issues and further the theoretical knowledge of the field' (Barab & Squire, 2004, pp. 5–6). Additionally, experimentation and testing of new classroom interventions must be done as a collaborative undertaking between teachers, researchers and school administrators (Brown, 1992). Hence, our DBR is defined by being situated in a real school context. We focus on the design and experimentation of authentic STEM assessments as a significant intervention and collaborative partnership between researchers and the three teachers. Moreover, design principles are evolved for authentic STEM assessments using mixed-method data sources to inform the refinement of the design and the generation of evidence-based design principles. The goal is to practically impact teacher assessment capacity in task design and implementation, which in turn, promotes the engagement of elementary school girls in STEM learning.

Participants

The participants included three female elementary school teachers (Dawn, Sharon and Lauren, all pseudonyms) who taught Grades 5 and 6 mathematics and science at a Canadian urban public school and their students (10–11 years of age). The third teacher, Lauren, only took part in the implementation phase. Dawn and Sharon both completed a master's course in educational assessment and had experience designing inquiry-focused and project-based STEM activities. They also decided to collaborate with the university researchers as partners in design. Hence, both gravitated toward an authentic assessment (Koh et al., 2021) as a new classroom assessment practice. Our research was approved by the research ethics board of the university and the school district. Written consent was obtained from the participating teachers and parental consent and assent were gathered from the student participants.

Setting

We worked with the two teachers (Dawn and Sharon), providing them with relevant resources and support in the design of authentic STEM assessments. During the design phase (July to August 2019), the teachers were briefed on the design principles of authentic assessment including the AIQ criteria, SOLO taxonomy and the patchwork text approach before co-designing the authentic STEM assessment prototype. During the implementation phase (November 2019 to February 2020), the first and third authors visited the school to conduct observations. The implementation of the authentic assessment took place over eight sessions in the school's learning commons. For each session, students collaborated in small groups of four to five. Most groups had a mixture of boys and girls. There were about 75 students, including both participating and non-participating students. Non-participating students refer to students who participated in the authentic STEM assessment, which was implemented during

regular class time; but did not give parental consent and assent for their data to be used in the study. Due to ethical considerations, we could not exclude boys and non-participating girls from the classroom intervention. In the first ten minutes, teachers gave instructions about authentic STEM assessment tasks (i.e., research safety regulations, create a list of materials and begin to develop a budget) and provided materials (i.e., Chromebooks, paperboard, hot glue guns, etc.) that might be needed for completing the tasks. Working in their small groups, students were engaged in completing authentic STEM assessment tasks, which culminated in a final design project (i.e., the building of a multiple-purpose events centre). A final exhibit was held at the school for the students to showcase their design prototypes to students and teachers who were from other grade levels.

Data sources and analysis

Teacher-Designed Assessment Tasks. The teachers' designed authentic STEM assessment tasks centre around the problem scenario of inviting students as professionals to develop a blueprint proposal and a prototype of a multi-purpose events centre on a twenty-acre vacant lot in the local downtown area. The final proposal needs to be presented to teachers representing the 'city council' for approval before making a 3D prototype. Students were distributed to fifteen groups of 4 to 5 (mixed-gender groups), each performing a self-selected role among the choices of 'project manager', 'researcher', 'recorder', 'architect' or 'engineer'.

Figure 1

The Patchwork Structure of The Teacher-Designed Authentic STEM Assessment

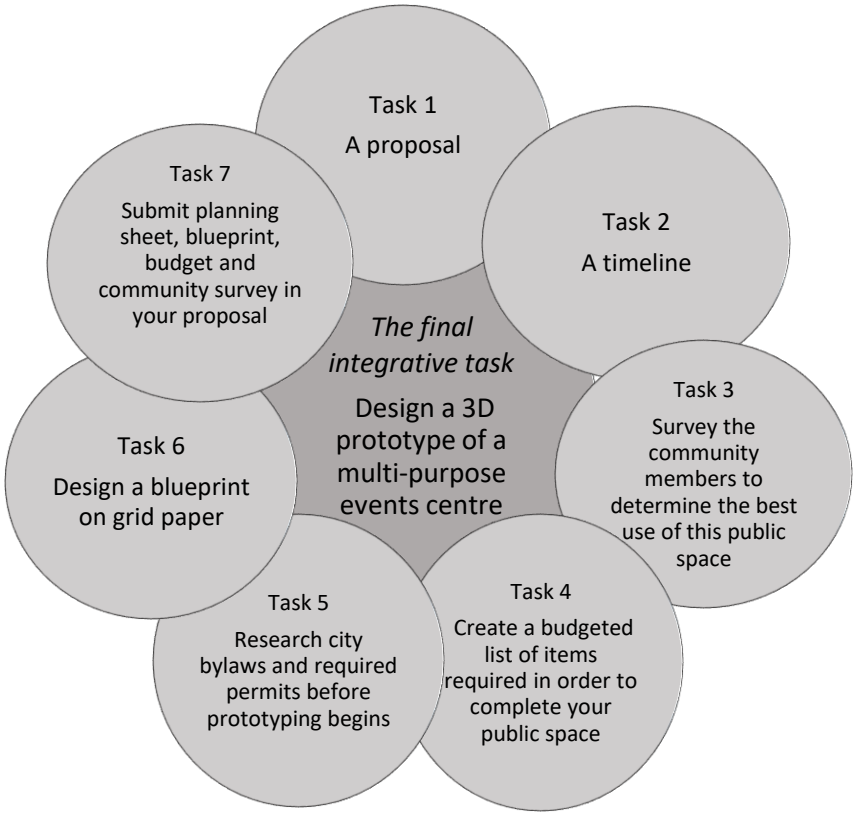


Figure 1 depicts teacher-designed assessment tasks to facilitate students' completion of the culminating project (i.e., design of a 3D prototype of a multi-purpose events centre). In initializing the assessment, teachers laid out all tasks

to students up front. Students were given the agency to approach Tasks 1–7 one by one or concurrently.

The teacher-designed assessment tasks and associated rubrics were analysed by the researchers using the AIQ criteria for integrated STEM (Koh et al., 2020). The findings shed light on the intellectual quality of the tasks, that is, the extent to which higher-order cognitive outcomes are assessed. Furthermore, the results aid the researcher and teachers in identifying issues in the design and implementation of authentic STEM assessment in classrooms.

Interviews. The teachers' one-on-one, semi-structured interviews were conducted after the study and were transcribed verbatim. Using NVivo 12, we first coded the interview transcripts using a priori codes informed by the research question (i.e., the challenges in design and implementation) and inductive codes that emerged from the data to identify descriptive patterns. In the second coding cycle, we grouped the codes into higher-level categories based on the patterns and used our memos to generate emerging themes. We reflected individually on the themes and followed up with discussions in which we reached an agreement about our themes (Miles et al., 2020). Analysis of the interview data serves to tap into the teachers' perceptions of their experiences in designing assessment tasks and in their implementation with students in classrooms. Table 2 shows the coding schemes for analysing the interview data.

Table 2
The Coding Schemes

First coding cycle		Second coding cycle	
Priori codes	Inductive codes	Categories	Themes
Design of authentic STEM assessment tasks	Students' resistance	Confusion about assessment types	Lack of assessment literacy in STEM
	Learning outcomes	Struggles to integrate STEM curriculum into existing curricula	
	Timeline management		
Implementation of authentic STEM assessment tasks	Time and space constraints	Measurement of students' engineering/design process but not science, mathematics and technology	Lack of district-wide guidance on what is STEM curriculum
	Student collaboration		
	Formative assessment		

Observations. Classroom observations took place while the authentic STEM assessment was implemented by the teachers. A coding protocol was written to conduct the observations at multiple intervals. The observational data were used to augment the other three data sources.

Student Self-Reflections. To gain a better understanding of students' experiences in the authentic STEM assessment, each student was asked to complete a self-reflection sheet by the end of the STEM showcase. The self-reflection sheet consisted of 12 open-ended questions about students' reflections on their STEM knowledge, challenges, problem-solving, group work dynamics and an overall evaluation of their experiences. Only 20 student participants (9 boys and 11 girls) completed and submitted self-reflections. Note that this article reports only key findings highlighting gender differences in boys' and girls' experiences in the teacher-designed authentic STEM assessment.

RESULTS

We present the results organized by each research question.

What is the intellectual quality of teacher-designed authentic assessment tasks in STEM?

The overall AIQ scores of the teachers' designed authentic STEM assessment tasks are presented in Table 3. As can be seen, in terms of *depth of knowledge*, the scores in each of the elements (Rows 2–4) indicate that the assessment tasks focused on testing students' factual and procedural knowledge, that is, recognizing basic concepts, facts or principles and performing a set of routine steps. There was little opportunity for students to demonstrate their understanding of advanced concepts, such as how mathematical concepts related to scientific concepts and technological ideas in the design of a multi-purpose events centre. The scores of *knowledge criticism* (Rows 6–8) show that the assessment tasks asked students to merely follow a set of procedures, such as designing a survey based on clearly defined steps. In collecting and analysing survey data, the tasks required students to organize and classify the response data to their surveys. The teachers did not prompt students to justify their choices of survey questions, analytic methods or the overall survey design. The criterion of *knowledge manipulation* in the assessment intends to assess students' higher-order thinking and reasoning skills, such as making mathematical conjectures and scientific hypotheses to solve the problem, reach a conclusion and make discoveries. The scores (Rows 10–13) imply that the teacher-designed tasks focused more on students' reproduction of facts, concepts and procedures rather than the creative production of new knowledge. Despite that, the tasks required students to analyse and interpret information while designing and modifying their blueprints and prototypes. For the remaining AIQ criteria (Rows 15–19), scores were high on *extended communication* and *making connections to the real world beyond the classroom*, indicating that the tasks provided students with the opportunity to elaborate their reasoning, thinking, explanation or conclusion using multimodal methods such as graphs, drawings, physical constructions and symbolic representations.

Students were also asked to make connections to the real world because the central problem was derived from a realistic scenario on the news. The score on *clarity and organization* means the whole set of tasks in the authentic STEM assessment had a clear structure of composition, yet the instructions for some of the tasks were vague. In designing the blueprint and building the 3D prototype, students had agency (*student control*) to determine sources of information and the materials they intended to use. A generic rubric was written by the two teachers but detailed qualitative descriptions of performance criteria were missing that were needed to assess student performance across a variety of

tasks. This absence means the assessment lacked *explicit sharing of performance standards and success criteria* with the students.

Table 3
The Authentic Intellectual Quality (AIQ) of Teacher-Designed Authentic STEM Assessment

Row	AIQ criteria	Scores
1	A. Depth of knowledge:	
2	a. Factual knowledge	4
3	b. Procedural knowledge	4
4	c. Advanced concepts	2
5	B. Knowledge criticism:	
6	a. Presentation of knowledge as a given	3
7	b. Comparing and contrasting knowledge	3
8	c. Critiquing knowledge	2
9	C. Knowledge manipulation:	
10	a. Reproduction	4
11	b. Organization, interpretation, analysis, synthesis and/or evaluation of information	3
12	c. Application/problem solving	2
13	d. Generation of new knowledge	1
14	Other:	
15	D. Extended communication	4
16	E. Clarity and organization of the tasks	3
17	F. Making connections to the real world beyond the classroom	4
18	G. Student control	3
19	H. Explicit performance standards/success criteria	1

Note. All criteria were scored using 4-point rating scales (ranging from 1 = no requirement to 4 = high requirement).

Regarding the integration of STEM subjects, the teachers tended to focus on the solicitation of students' factual and procedural STEM knowledge despite their attempt to make the assessment authentic with a real-world problem, that is, the design of a multi-purpose event centre and to promote students' extended communication. The design of a prototype of the building seemed to increase boys' and girls' interest in mathematics (e.g., measurement, survey). However, there is a lack of evidence on how mathematical and scientific concepts can be applied to the engineering design challenge.

Table 4 delves into the seven tasks in the authentic STEM assessment, clarifying the STEM learning goals, assessments and the corresponding intellectual quality indicators characterising each of the tasks. Tasks 1 and 2 were parts of the students' preparation, serving as prerequisites for other tasks and Task 7 was the submission of required documents to teachers. Thus, the tasks were not intended for assessing STEM. As seen in the column of STEM learning targets,

Tasks 3 and 4 focused on assessing students' knowledge of mathematics—statistics, probability and number sense—via a group-based performance task (i.e., creating and analysing a survey) and a group-based arithmetic calculation (i.e., creating a budget for their building). In the assessment of mathematical knowledge, teacher-designed tasks relied more on students' reproductive, procedural and organizational cognitive capabilities. Task 5 set the application of technology as the priority, assessing students' abilities to use Chromebooks to search for information. The assessment of technology required low-level thinking (i.e., browsing the internet) and did not challenge students to go beyond basic ICT skills.

Task 6 and the final integrative task assessed students' science and engineering capabilities using group presentations (i.e., designing, building and modifying a 2D blueprint as well as prototyping a 3D construction). In assessing the engineering component, teacher-designed tasks required students to experience the process of asking using survey questions, imagining through group discussions, planning via the 2D blueprint and creating a physical model. However, there was no mention of testing and improving their designs. Although the teachers identified the prototype as a science outcome and implied the construction of simple machines as a part of the design, only two groups of students included their designed electrical devices in the final exhibition of their prototypes. This absence suggests the ambiguity of instructions in the assessment tasks, which might have compromised the validity of the designed tasks.

Table 4
STEM Learning Goals, Assessments and AIQ Alignment

Authentic Assessments	STEM Learning Targets	AIQ Alignment
Task 1: A group-based performance task asking students to write a proposal	N/A	N/A
Task 2: A group-based performance task asking students to write a timeline recording weekly accomplishments	N/A	N/A
Task 3: A group-based performance task of surveying community members and analysing the result	Mathematics—Statistics & Probability	Ab, Cb, F, G
Task 4: A group-based performance task asking students to make a budget for construction items	Mathematics—Number sense	Ba, Ca, F

Task 5: A group-based performance task of searching for online information about city bylaws and required permits	Technology—Gathering information	Aa, Ba, Ca, F, G
Task 6: A group presentation of a visual illustration of the blueprint	Engineering—Designing	Ab, Ba, Ca, D, E, G
Task 7: A group presentation of delivering planning sheet, blueprint, budget and community survey to teachers	N/A	N/A
The Final Integrative Task: A group presentation of a 3D prototype of students' design (A STEM showcase)	Science—Prototype; Engineering—Building & Modifying	Ab, Cc, D, F, G

Note. The alphabets used to code AIQ indicators are identical to those used in Table 3, for example, Aa refers to 'factual knowledge' in Depth of Knowledge.

When student participants were asked to reflect on their interest in the authentic STEM assessment (i.e., Did you enjoy this STEM project?), 78% of boys and 64% of girls confirmed their interest and enjoyment, indicating that boys enjoyed doing this STEM project more than girls did. The reasons for enjoyment provided by boys tended to focus on the application of skills, whereas girls thought that they could demonstrate their creativity with the design of the prototype. For example, a boy stated, "I enjoyed this project because I get to use my skills and put it into the project." As several girls commented, "Yes, I liked it because we could use our creativity…," "It really got our minds thinking and so we could use creativity," and "We all could work together and have a good time putting our creativity in this project." This finding is in line with Newmann et al.'s (1996) aesthetic value of authentic tasks and Howes et al. (2013), who argue for adding opportunities to motivate STEM learning for all students (girls included) through the connections to enjoyment, aesthetics, societal needs and real-world problem solving. The teacher-designed authentic assessment did provide students with the opportunity to make connections to the real world (i.e., using STEM knowledge and skills to design and build the prototype of a multi-event centre). As Teacher Dawn commented during the one-on-one interview, she thought that the authentic STEM assessment provided students with a great opportunity "to work together and use skills that maybe they don't get an opportunity to use when they're sitting down in the classroom at their desk," also "to tap into their creative side and use their critical thinking skills and put that all together with collaborating with their classmates."

Our analysis of student participants' responses to self-refection questions about the STEM concepts that they had used in completing the authentic STEM assessment tasks reveals that boys tended to consider materials and electronic design for the final building prototypes, as evidenced by the 33% of boys who listed the science concept of materials, such as "slime and chemicals," "climate-friendly materials" or "materials." Twenty-two per cent of boys mentioned the concept of building, for instance "a supportive base and solar panels for energy." Eleven per cent of boys referred to the concept of a lever, as a boy commented,

"I used science with building and levers to help make my project better." None of the girls, however, mentioned materials, buildings or levers. Instead, they reported that they tried to apply their investigative skills (e.g., estimating, predicting, measuring, surveying, collecting data and using variables). Although the authentic STEM assessment designed by the teachers had some limitations (i.e., a lack of STEM integration), it did enable girls to engage in scientific investigation while solving real-world problems embedded within several tasks. Investigative skills are critical in engaging students in STEM (Bybee, 2010; Kennedy & Odell, 2014). Regarding the use of technological elements, more girls (64%) than boys (11%) reported that they had used online documents as a source of reliable information to inform their design, budget and selection of construction materials.

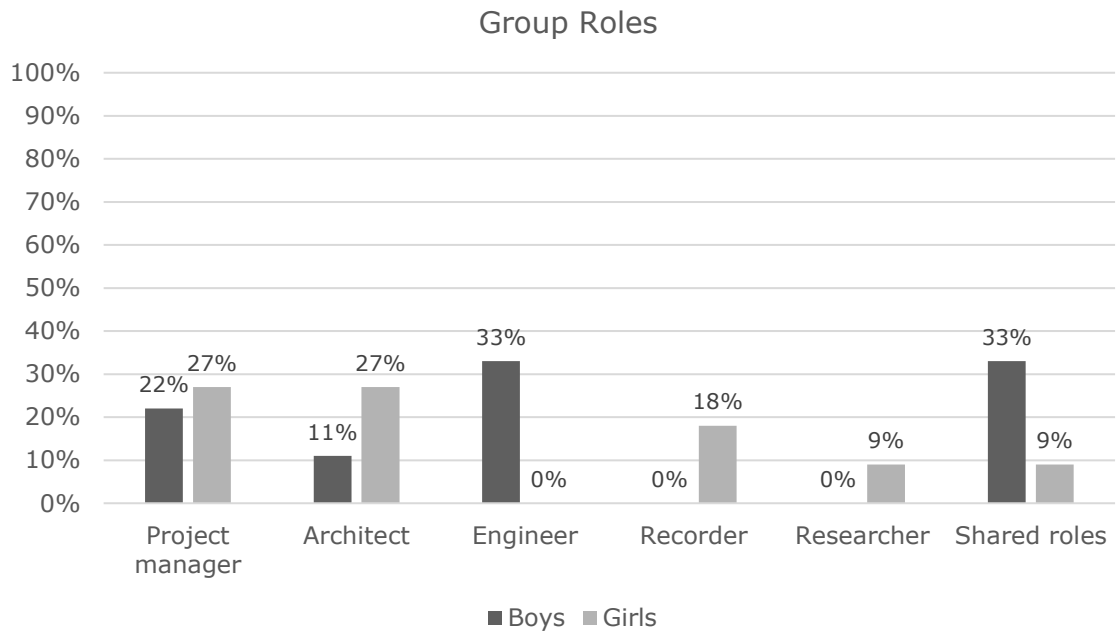
Girls also mentioned the various mathematical concepts (e.g., addition, multiplication, measurement and geometry) they had learned and put more weight on the cost and budget of the building; boys were proud of the survey and graphs they had created (i.e., measurement). The result indicates that the authentic STEM assessment provided girls with the opportunity to be more confident in mathematics, as evidenced by a slight increase in one of the self-efficacy domains—growth mindsets and a significant improvement in interest in mathematics (self-reported questionnaires, See Koh et al., 2021). Nineteen girls completed the pre- and post-questionnaires. The mean scores on 'belief that STEM ability grows with effort' before and after girls' experience in the authentic STEM assessment were 3.88 (SD = .719) and 4.22 (SD = .752), t = 1.58 (p > .01); Cohen's d was found to be 0.4, a medium effect size. According to Cohen (1988), d values of 0.20, 0.50 and 0.80 represent 'small', 'medium' and 'large' effect sizes, respectively. The mean score difference in girls' interest in mathematics (M before authentic STEM assessment = 4.60, SD = 1.27 and M after authentic STEM assessment = 5.66, SD = 1.29) was statistically significant (t = 3.202, p < .01) and had a large effect size (Cohen's d = 0.89). These statistics indicate a considerable improvement in girls' interest in mathematics after completing the authentic STEM assessment. Additionally, girls' interest in STEM careers had slightly improved (See Koh et al., 2021 for details).

The statistical findings are corroborated by the teachers' comments during the one-on-one interviews. For example, one of the teachers noted that students, especially girls, persevered with their design project in the authentic STEM assessment as they kept generating, reflecting and modifying their ideas. Unlike one-shot traditional assessments, an authentic assessment approach affords girls the opportunity to realise they can develop their mathematics and science abilities over time if they are persistent in their tasks.

> "I think giving them some time to work on it and then giving them a break so that they could keep thinking about it in their head. Plus, they had the shared Google document that they could always go back to if they found other ideas or whatever. We left bringing in materials up to them. So, if they were at home and they found something, like a recycled item, that might work as a basketball hoop or something, I think it was just always kind of on their brain, this project. And so they were bringing more materials in. They were talking about other things they could do. And then when they got together again, they would share those ideas again and make modifications as they needed to." (Teacher Dawn)

Interestingly, girls' selection of their roles in the STEM project was diverse (e.g., project manager, architect, recorder and researcher), but none chose the role of engineer (Figure 2). Additionally, none of the boys reported that they played the role of either the recorder or the researcher in their group project. To gain more experience, boys tended to share roles (e.g., project manager, architect and engineer) with others.

Figure 2
Comparison of Boys' and Girls' Selection of Roles in the STEM Project



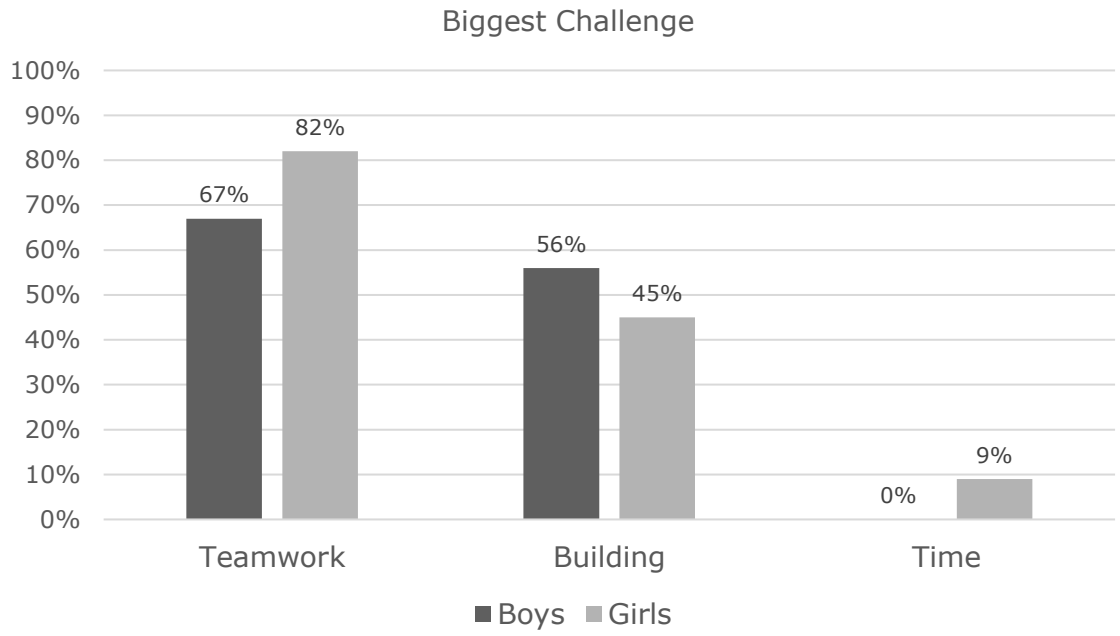
While reflecting on the most significant challenges in completing their final designs (the question was 'What challenges did your design team face while completing this project?'), students' responses were grouped into three broad categories—teamwork, building the prototype and time constraints (See Figure 3). Eight-two per cent of girls and 67% of boys claimed that the greatest challenge was in the process of working with other group members. More girls than boys seemed to face challenges in collaboration. A boy described, "At first we got excited and all worked on different things without teamwork, so we had to restart." A girl noted, "We did not always cooperate and we talked a lot." Fifty-six per cent of boys and 45 per cent of girls mentioned issues in building their prototypes as the biggest obstacle. Some participants listed both the troubles in teamwork and building the design prototype as the greatest challenges they encountered. For example, a boy expressed, "We did have a little trouble with deciding on what to build because someone made a bit hard for us to make a final decision. We also had trouble with materials." As a girl stated, "A lot of us faced the challenge of talking and being distracted. Also having the problem of running out of hot glue since everyone wanted walls/decorations being glued by hot glue." Finally, one girl (9%) thought that time constraints were the biggest challenge.

Despite some challenges, teachers Sharon and Lauren thought that the authentic STEM assessment facilitated growing girls' confidence by challenging

stereotypes about gender and assessment. Sharon said, "It does help to eliminate any kind of stereotype because we can come up with projects and activities that are appealing and interesting to both genders, and it is the way to instill confidence or motivation, increased confidence in both genders, but particularly girls."

Figure 3

Comparison of Boys' and Girls' Biggest Challenges Faced While Completing the STEM Project



Teachers can benefit from authentic STEM assessment by designing tasks to combat the gender stereotype in authentic contexts, which is more difficult with traditional assessments as they tend to be based on separate subjects and decontextualized knowledge. This view was confirmed by Lauren who normally had mathematics assessments with her students of "a quick checklist of who's getting it, who's on track, just by what they've shown me on their whiteboards." Her science assessments varied depending on the unit, which could be "lots of rubrics, lots of group work, lots of peer evaluations, self-evaluations," or "small little quizzes or even exit slips, just things like that along the way rather than one big unit test at the end." After implementing the authentic STEM assessment, Lauren believed it could change girls' fixed mindset of "I'm not good at it and I'll never be good at it, so why should I even try" because it focuses on the process of "working it out", as she commented, "I enjoyed watching the kids work towards that final design. I enjoyed them working it out and hashing out their differences, and, in the end, they created some amazing event centres. Some of them were quite detailed and specific and creative, and then it was interesting to see the process along the way, starting from a blueprint and then coming to the actual what they created."

By affording teachers the agency to design tasks that challenge gender stereotypes and the fixed mindsets of girls, the authentic STEM assessment empowered and engaged girls in creative designs for solutions instead of simply searching for the one correct answer. Girls' development of a growth mindset

was further reinforced after overcoming all the challenges (i.e., collaboration and time constraints) in completing the authentic STEM assessment. Teacher Sharon said, "once they get through the rough patches of figuring things out…by the time we've gone through the process and they get to the part where they're presenting it…you see an amazing growth, you see an amazing amount of excitement." In essence, it is the authentic STEM assessment, which affords students time and resources to develop abilities and affords teachers agency to combat gender and assessment stereotypes, that worked to facilitate the 'amazing growth' of students, in particular creating a formal learning environment to foster girls' growth mindsets in STEM.

What are the teachers' perceptions of their experiences when designing authentic STEM assessments for addressing gender disparity in STEM learning?

Introducing Authentic STEM Assessment to Students. All three teachers in the study perceived the benefits of introducing authentic STEM assessment to the mixed-gender classroom. Sharon noted that she is "a strong advocate for this type of learning," although most students were unaware of the application of integrated STEM knowledge in completing the design project, they could see the value in it. Dawn's statement emphasised creativity and critical thinking as cross-cutting competencies that students can utilise, as she commented, "[Authentic STEM assessment] is a great opportunity for students (boys and girls) to work together and use skills that maybe they don't get an opportunity to use when they're sitting down in the classroom at their desk, just completing assignments that are given to them by teachers."

Lauren shared the same sentiment with Dawn, adding that creativity is a competency generated from collaboration with peers. An authentic assessment represents an alternative form of assessment unlike a traditional test or a quiz. Teachers as assessors "were watching students in how they interacted with each other, how they problem solved, how they planned, how they execute the whole presentation at the end, all of [those activities]" (Dawn). The newly designed authentic STEM assessment affords students—especially girls—more time and resources to develop problem-solving and critical thinking skills, as well as the competence of persistency despite facing challenges in their group work.

The teachers held different views concerning how to assess students' learning of STEM in classrooms. Lauren did not take part in the design phase and expressed her unacquaintance with the assessment of individual students' STEM learning: "Because you have a big group of people working together." However, she suggested giving students a rubric "so they know what their focus is on and what the end result should be." Dawn believed STEM assessment is principally about "how they're using those 21st-century skills, how they're putting them all together and being able to effectively use them when given a task," including the creative and critical thinking she mentioned earlier. Sharon understood that the assessment needs to focus more on the interdisciplinary knowledge of STEM subjects. Previously, Sharon was confused about whether a STEM assessment should consist of standalone mini-tasks or an integrated project. She considered the authentic STEM assessment in this research partnership project "a good framework" to assist teachers in assessing a STEM project as a whole. According to her, "It also gives us direction on how to move forward, in addition to seeing how the students are performing in those different areas."

The teachers' comments in the interviews suggest that their prior conceptions of STEM assessment helped them easily make connections with project-based learning, group work and 21st-century competencies. However, their students might not have been able to understand the learning behind a hands-on STEM project, indicating that the complexity of the STEM curriculum did not translate into clearly defined assessment practices in the STEM classroom. Despite this caveat, the teachers' experiences in the design of the authentic STEM assessment prototype seem to be beneficial for guiding them to start thinking about how an authentic assessment allows for integrated STEM. Additionally, their participation in the research partnership project gave them some assessment design guidelines to follow.

Co-Designing Authentic STEM Assessment. While reflecting on how to come up with their designed authentic STEM assessment tasks, Sharon and Dawn shared their co-design experiences. Lauren did not contribute to the design. It was suggested from Sharon's and Dawn's interviews that both developed the assessment tasks by referring to the design principles shared by the university researchers. However, they admitted that they placed greater emphasis on the engineering design process. As Dawn commented:

> "So first we present them with the challenge. Then we give them time to brainstorm ideas and think about how they could tackle that challenge or that project. We give them time to plan and to research with their group members. And then we give them time to do a blueprint. They need to get that checked by one of us, and then they have some time for building their prototype and sharing their ideas and figuring out what works and what doesn't work and modifying. And then they work on presenting and they would present it to the rest of the students, present their project to the rest of the students as well as teachers and school personnel who would be in the building and any other invited guests that we have invited to see."

Dawn's comments described a straightforward linear progression of engineering design: proposing a challenge, brainstorming ideas, researching as a group, planning a blueprint, building a prototype and presenting to the public. The only assessment mentioned was that they would need to check students' blueprints and the final product on the paper regarding the engineering design process. Sharon shared the same approach to design: "We designed a package where they have all the information that they need with a timeline from start to end." In the end, there was time for reflection. Both Dawn and Sharon neglected to embed an assessment of students' conceptual understandings of mathematics, science and technology in their designed engineering process.

In describing their authentic assessment design, Dawn extended her comments to 'authenticity,' which distinguished their tasks from other conventional STEM projects. As she stated: "I think the authentic piece comes from you're making it a real-life learning experience, and that's where you need to really pay attention to what the challenge is. Because building the event center…is current…They know it's happening…So that was relevant to them. That makes it authentic…If we were just to give them a challenge or we could give them the same challenge every year, and it doesn't really have a purpose other than it's a group project…". This statement means that Dawn thoroughly investigated finding an

authentic and real-life topic that would provide her students with a challenge of great relevance and meaning. This approach represents how Dawn understood the essence and purpose of an authentic assessment in STEM. When prompted during the interview why they approached the design of an authentic STEM assessment in this way, Sharon and Dawn believed that their past experiences had a great influence on their decision. According to Sharon, their design is consistent with their past experiences of developing STEM activities, each time with "some slight alterations." "Our very first one …" said Sharon, "was just to get an activity to engage the students." For the current ones, they put more thought into "what actually are we covering in terms of the curriculum" (Sharon). When it came to the value of using the design principles of authentic assessment, including the frameworks provided by the university researchers in the design, Sharon thought that the principles and frameworks gave them a structure that led to the great ideas behind each design of the authentic STEM assessment. Dawn insisted that they kept "the same sort of format" and changed the intellectual challenge for each of the STEM tasks. The benefits of using an authentic assessment approach in their task design deepened their thinking about how to align their lessons with authentic tasks in the STEM unit of work.

Due to the lack of assessment components in their designed tasks, Dawn considered student resistance as one of the barriers. She explained that many of her students felt confused about the term 'assessment', as they tended to equate assessment with a mark on their report cards from "a test or some sort of quiz." Although Dawn introduced the designed tasks to students as a form of assessment, she reassured students that "it's not going on your report card," and "you're not going to be graded on it." In addition, Sharon said her students did not take it as an assessment, "they considered it a project."

Other barriers included logistics issues and alignment with existing curriculum outcomes. Viewing the design and implementation of the authentic STEM assessment as an extra duty, Sharon commented, "figuring out timelines, structuring it around our schedule and a school calendar, and the timelines of guest speakers is … tricky to implement." Moreover, "when we look at curriculum outcomes, it can be difficult to make sure that you have enough time to introduce those outcomes to the students before we start the project." In this sense, without finding a way to intentionally integrate the authentic STEM assessment into their daily instruction and gaining support from the school administration, the teachers found it challenging to invest considerable time into designing an authentic assessment.

Based on Dawn's and Sharon's narratives, it is evident that the teachers' design of authentic STEM assessment tasks followed the same format they used in the past. This format is a linear progression of an engineering design process that moves students from ideating around a real-life challenge to the final presentation of their solution. Although the teachers had completed a master's course in educational assessment and were introduced by the researchers to the design principles of authentic assessment, the teachers' assessments were still heavily influenced by their pre-existing conceptions. Teachers were introduced to the frameworks of the SOLO taxonomy, the patchwork text approach and the AIQ criteria. Nevertheless, the teachers' designs of the authentic STEM assessment were shaped by their prior experiences. Applying some of the

principles of authentic assessment to the teachers' design helped improve the structure of their tasks and made the teachers more aware of the integrated STEM outcomes and the tasks they were intended to cover. They also appreciated the value of using authentic assessments to address gender and assessment stereotypes. However, the designed tasks of the teachers still showed many gaps in assessing mathematics, science and technology as a whole. Three barriers were identified in the teachers' design: (1) students' resistance to the assessment that would be graded, (2) the incorporation of the authentic STEM assessment into existing schedules, and (3) the struggle to integrate STEM curricula into existing school curricula.

What are the teachers' perceptions of their experiences when implementing authentic STEM assessments in mixed-gender classrooms?

Barriers to the design of authentic STEM assessment also created challenges in its implementation at school. First, failure to integrate STEM curriculum into existing school curricula prevented students to have a full understanding of the assessment tasks and the required interdisciplinary knowledge and competencies. As Dawn found the challenge in her implementation was to make sure "that they all understood what the task was and the terminology that we were using." And students did not notice the connections between authentic tasks and STEM subjects due to the teacher's inadequate instructions before the implementation of the tasks. Thus, it was unclear to students that the curriculum outcomes were covered by the teacher-designed assessment tasks. After posing questions such as 'What math are you using right now?' 'What science is incorporated?' and 'What technology is incorporated?' to students in the self-reflection sheet, Dawn concluded, "I think they just think of it as a big project that they're working on, but they don't draw those connections."

Dawn's observation was supported by Sharon, who added that "we were going to have all those mini-lessons" before the implementation if there would be a second authentic STEM assessment. Some students in Lauren's class had even lost their motivation to complete the authentic tasks because of their insufficient understanding of STEM-related knowledge. Lauren clarified that what, "they loved the best was actually making the model and bringing in things from home…a lot of mine struggled with the research part of it, like finding prices…so many kids do give up, and don't want to do anymore."

Second, conflicts between the teacher-designed assessment tasks and the existing curricula led to issues of coordination. Sharon called their designed tasks "an extra project," because "it wasn't taking the place of what we were covering in school, but I can see the value that STEM projects could take the place of that kind of more traditional instruction because you are covering those outcomes." Despite that, she was able to envision the value of the STEM curriculum. The most significant challenge of implementing their designed tasks came from time and space constraints. As Sharon continued, "that was the hardest part because they would just have all of their materials out and so eager to start and we'd have to clean up…the scheduling is a bit of a nightmare sometimes."

Third, student resistance to the idea of assessment combined with their insufficient understanding of STEM-related knowledge included in the tasks

offered teachers scant opportunity to evaluate the depth of student understanding and their application of STEM knowledge. When teachers were asked whether they used any form of formative assessment in their implementation of the designed authentic STEM assessment tasks, Dawn replied:

> "Just constantly monitoring group work and offering, providing feedback. If they were stuck on something, giving them some advice or questioning them as to how they could fix their problem or modify their plan so that they can achieve what they're trying to do. But also then we always had checkpoints for them. So, before they could begin to build their prototype, they needed to create their blueprint, and they needed us to check that that was complete before they went a step further. So just kind of checking in with them and providing any clarification and feedback that we could along the way so that they were constantly kind of focused on the right path."

Dawn's comment suggests that she did use a formative assessment (i.e., group work facilitation, checkpoints at different phases) to monitor students' progress during the engineering design process. This informal assessment enabled her to ensure that each group was 'on the right path.' However, no questions or prompts were given to the students concerning their understanding and application of mathematics, science and technology throughout the overall design process. It was also confirmed by Sharon, as she stated:

> "So, while the students were creating the project or going through the assessment, we would be taking anecdotal notes on what they were doing. For us, because the time constraints and everything else, it was more about how they were working together in a group and the kinds of the social interaction, the problem solving and all of those other competencies or skills. Those were a little bit easier for us to look at and comment on and make anecdotal notes on, rather than all of the curricular outcomes that are tied with this project, because there were so many and there were so many students."

Sharon's formative assessment was documented in her anecdotal notes. She observed group collaboration and problem-solving as manifested in students' overt behaviours that could be easily observed rather than making inferences about students' mental activities. Again, her formative assessment was not about task-related STEM curricular outcomes. Hence, students were unable to see the value of integrating STEM into solving real-world problems.

Our thematic analysis of the observation data shows: (1) Each group of students approached the tasks differently. Some mixed-gender groups spent considerable time discussing and assigning team-member roles, while others quickly moved to drawing the blueprint; (2) Time became a constraint for completing the planned task as some groups were able to complete it within the session and some had to pause in the middle; (3) Some groups struggled to find a direction due to the lack of specific instructions for sub-tasks and descriptive rubrics; (4) STEM innovations were embodied in occasions such as when students had the agency to choose materials for making their prototypes, decide the way to design surveys and collect data and negotiate the procedures of engineering

design (student control) instead of being instructed by teachers step by step; (5) Students were unaware of the mathematics and science knowledge they used to complete the tasks until the last day of completing the tasks. Some groups were able to identify a few general curricular outcomes such as multiplication, measurement, geometry, proportion, pattern and electricity; (6) Students felt immensely excited about handcrafting their design prototypes; and (7) Formative assessment and instructional scaffolding by the teachers were rare. Most interactions between teachers and students concerned the design of the final building prototype. Themes (2), (3), (5), (6) and (7) corroborate the teachers' narratives of their experiences in the design and implementation of the authentic STEM assessment tasks.

In short, several themes emerging from the interview data were related to the design and implementation challenges encountered by teachers. Design challenges included themes such as struggling to integrate the STEM curriculum into existing school curricula, a heavy focus on the design challenge rather than mathematical and scientific concepts as well as integration of STEM, helping students understand task demands and instructions and managing timelines (structuring authentic STEM assessments around class schedules and the school calendar). Implementation challenges included students' perceptions of STEM learning, student collaboration (boys and girls) and time constraints.

DISCUSSION AND CONCLUSION

Drawing from the results of a larger DBR study, this paper reported our research partnership with three elementary school teachers in the design and implementation of authentic STEM assessments that aimed to promote girls' self-efficacy and interest in STEM subjects and careers. This type of research partnership also serves as the teachers' continuing professional learning in authentic STEM assessment. At the beginning of the project, the researchers introduced the design principles to the two teachers involved in the design phase of the research partnership. These principles included developing an authentic STEM assessment based on the theoretical frameworks of authentic assessment, AIQ criteria, SOLO taxonomy and a patchwork text approach. The teachers attempted to comply with the principles in the design phase; however, they relied heavily on the engineering design process with which they were familiar through their prior experiences.

The teachers' main takeaway from the theoretical frameworks mentioned above is the authenticity emphasised by the theory of authentic assessments (Koh, 2017; Koh et al., 2020). However, as we have seen in the teacher-designed authentic STEM assessment, 'authenticity' only manifests in creating a problem scenario that reflects the real world to students (as mentioned in the Introduction and Theoretical Framework, besides 'making connections to the real world beyond the classroom', there are seven criteria defining the intellectual quality of authentic assessment tasks). The teachers lacked a full understanding of the design principles of authentic STEM assessment despite their earlier completion of master's degree coursework in educational assessment and their participation in the research partnership project. Another limitation might be due to the interdisciplinary STEM concepts and competencies required for students to complete the design prototype.

Despite seeing some benefits of authentic assessment for girls, the teachers' partial understanding of what comprises a valid authentic assessment in STEM prevented them from developing an effective design. The teachers' lack of assessment literacy in STEM was also shown in the dilemma of assessing students' 21st-century competencies and STEM interdisciplinary knowledge, as well as questions considering the authentic assessments as a project-based inquiry or a formative or summative assessment during the implementation stage. With these confusions, teachers' experiences of designing the authentic STEM assessment indicate consistent struggles for teachers integrating the STEM curriculum into the existing school curricula.

Our findings are consistent with the research in the United States and Australia demonstrating that elementary school teachers lack preparation to teach and assess in the STEM fields (e.g., Epstein & Miller, 2011; Kurup et al., 2019; Murphy & Mancini-Samuelson, 2012; Nadelson et al., 2013; Timms et al., 2018). Therefore, school districts should develop a quality integrated STEM curriculum and make it available for teachers so they are not left alone to consider ways of integrating STEM subjects into their pedagogy and assessment. As Margot and Kettler (2019) aptly pointed out, STEM content knowledge should be one of the areas in focus for inservice teachers' professional learning. Additionally, schools need to give more time for teachers to work together with their colleagues to plan lessons and assessments. Such collaboration will facilitate teachers' design and implementation of authentic STEM assessments focusing on STEM integration, higher-order cognitive outcomes and gender-responsiveness. According to Garet et al. (2001), two of the key features of effective professional learning and development for teachers are: (1) a focus on content knowledge and (2) the collective participation of teachers from the same school, grade or subject.

We analysed the intellectual quality of the teacher-designed authentic STEM assessment using the AIQ criteria. The results showed that the tasks predominantly assessed students' factual and procedural knowledge that resort to learners' abilities to reproduce knowledge. There was a lack of focus on eliciting the evidence of students' higher-order thinking and understanding of advanced STEM concepts including how integrated STEM can be used to solve a real-world problem, that is, the design of a multi-purpose events centre. As indicated in the student self-reflection data, some girls developed an interest in a variety of mathematics concepts and reported their use of investigative skills. They also appreciated the nature of the authentic STEM assessment, enabling them to develop their creativity and a growth mindset in STEM through the design prototypes. This suggests that teachers should be more intentional in increasing the intellectual engagement of girls in STEM through creating authentic tasks that elicit higher-order cognitive skills (high cognitive demand tasks). We believe the AIQ criteria can serve as guideposts for this purpose.

Similar to studies conducted by Koh and Luke (2009), Koh (2011a, 2014), and Koh et al. (2018) in other subjects, inservice teachers need more ongoing, sustained professional development in the area of authentic assessment task design. As shown in the current study, teachers were also unable to seamlessly integrate formative assessment into both the design and implementation phases of the authentic STEM assessment. Formative assessment enables teachers to use information gathered during the implementation of authentic STEM

assessment to support students' STEM learning through the provision of quality feedback and differentiated instruction in a diverse classroom (Tomlinson & Moon, 2013). Formative assessment is also deemed to help develop a growth mindset in students (Yan et al., 2021). This assessment approach is crucial for girls who may lack confidence in their STEM abilities. Our research corroborated Margot and Kettler (2019), who call for research into teachers' effective use of formative assessment in STEM education.

Findings from the teacher's interviews and classroom observations indicate that the teachers faced challenges in their design and implementation of the authentic STEM assessment due to time and space constraints. The teachers considered their involvement in co-designing an authentic STEM assessment to be an added strain on their daily instructional duties. Structuring the authentic assessment tasks using the patchwork text approach introduced by the researcher and the engineering design process they were more familiar with also oriented the teachers to focusing on students' final performance on designing and building rather than on students' application of STEM knowledge to solve a real-world problem. Thus, in the process of completing the authentic tasks, students seldom made connections between what they were doing and the related STEM knowledge. As a result, teachers failed to assess students' understanding of science, mathematics and technology in the implementation. Nevertheless, we believe there is potential in the teachers' designed tasks to become a more powerful authentic STEM assessment should they have greater integration in STEM subjects and the incorporation of AIQ criteria and SOLO taxonomy into the content and format.

On the day when students showcased their design prototypes during our final classroom observation, an event alerted to us the strength and possibility of teachable moments for STEM. A Grade 5 girl had embarked on tackling a real-world problem, introducing a parking lot her group had built to the researchers. The lot had two levels with a pole in the centre supporting the building. Recalling some STEM concepts she learned from third grade, the student then asked whether the lot should have one pole in the centre or four poles standing in the four corners. The researchers did not provide a definite answer for her but wondered how her teacher would have helped the girl relate the parking lot building to the girl's prior knowledge. The researchers further wondered how the girl's teacher would help her apply old learning to a new situation or even prompt her to generate new concepts.

In sum, the first two phases of this DBR study focused on a partnership with three elementary school teachers to design and implement an authentic STEM assessment, which aimed to promote girls' STEM self-efficacy and interest in STEM subjects and careers. We analysed the teachers' designed authentic STEM assessment and interview data, which were complemented by classroom observation and student self-reflection findings. Results show that the teachers faced some design and implementation challenges even though they were able to see the benefits of the research partnership project in informing their thinking about authentic STEM assessment and gender disparity.

LIMITATIONS AND FUTURE DIRECTIONS

The DBR methodology that we adopted requires an iterative cycle of testing and refinement of authentic STEM assessments in practice. Our analysis of the

teacher-designed authentic STEM assessment using the AIQ criteria provides further insight into the refinement of the authentic STEM assessment design in the next phase of the study. Opportunities for ongoing, sustained professional development in authentic STEM assessment design should be provided to elementary school teachers. This will enable them to build their capacity to design assessment tasks that place greater emphasis on eliciting girls' higher-order cognitive skills and that are more gender-responsive by considering the characteristics of girls (e.g., that girls need more time to complete a project, encouragement to choose the roles of male-oriented STEM professionals, and that girls are more persistent than boys in a challenging working environment). To promote girls' growth mindset, patchwork tasks should incorporate more frequent formative assessments (e.g., feedback, self-assessment and peer assessment). In an ideal way of implementing authentic STEM assessment, teachers can facilitate student discussion on the roles of Project Manager (recording and reporting), Researcher, Accountant, Architect and Engineer. This approach allows boys and girls to explicitly address gender stereotypes and gender disparities in STEM subjects and careers. It may also help improve their collaboration.

Due to the disruption of the COVID-19 pandemic, we were unable to implement our second authentic STEM assessment (i.e., an upgraded version of the authentic STEM assessment) with the same group of student participants. Another limitation is the small sample size of teachers as our research partnership involved only one elementary school. Only three Grades 5 and 6 classes were at the school. Future studies should include a wide range of teacher participants from different types of schools and a longer duration of the implementation of the authentic STEM assessment.

AUTHOR NOTE

The authors have no conflict of interest to disclose. The study 'Designing Authentic STEM Assessments for Girls' was supported by Alberta Education Research Partnerships grants 2019-0024 (2019-2022) to Koh and Chapman, which supported Liu's postdoctoral fellowship and our co-development of this manuscript. The authors would like to thank the two teachers who participated in the research partnership. Please direct correspondence to Kim Koh: Werklund School of Education, 2500 University Drive NW, University of Calgary, Alberta T2N 1N4, Canada. E: khkoh@ucalgary.ca; https://orcid.org/0000-0001-8315-4299.

REFERENCES

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Archer, L., DeWitt, J., Osborne, J., Dillon, J., Willis, B., & Wong, B. (2012). Science aspirations, capital, and family habitus: How families shape children's engagement and identification with science. *American Educational Research Journal*, *49*(5), 881–908. https://doi.org/10.3102/0002831211433290

Archer, L., DeWitt, J., Osborne, J., Dillon, J., Willis, B., & Wong, B. (2013). 'Not girly, not sexy, not glamorous': Primary school girls' and parents' constructions of science aspirations. *Pedagogy, Culture & Society, 21*(1), 171–194. https://doi.org/10.1080/14681366.2012.748676

Barab, S., & Squire, K. (2004). Design based research: Putting a stake in the ground. *The Journal of Learning Sciences*, *13*(1), 1–14. https://doi.org/10.1207/s15327809jls1301_1

Bartholomew, S. R., Strimel, G. J., Zhang, L., & Homan, J. (2018). Examining the potential of adaptive comparative judgment for elementary STEM design assessment. *The Journal of Technology Studies*, *44*(2), 58–75. https://www.jstor.org/stable/26730731

Beghetto, R. A. (2007). Factors associated with middle and secondary students' perceived science competence. *Journal of Research in Science Teaching, 44*(6), 800–814. https://psycnet.apa.org/doi/10.1002/tea.20166

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. Academic Press.

Breiner, J. M., Harkness, S. S., Johnson, C. C., & Koehler, C. M. (2012). What is STEM? A discussion about conceptions of STEM in education and partnerships. *School Science and Mathematics*, *112*(1), 3–11. https://doi.org/10.1111/j.1949-8594.2011.00109.x

Brown, A. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences, 2*(2), 141–178. https://doi.org/10.1207/s15327809jls0202_2

Bybee, R. W. (2010). Advancing STEM education: A 2020 vision. *Technology and Engineering Teacher*, *70*(1), 30–35.

Capraro, R. M., Capraro, M. M., Scheurich, J. J., Jones, M., Morgan, J., Huggins, K. S., Corlu, M. S., Younes, R., & Han, S. (2016). Impact of sustained professional development in STEM on outcome measures in a diverse urban district. *The Journal of Educational Research, 109*(2), 181–196. https://doi.org/10.1080/00220671.2014.936997

Ceci, S. J., & Williams, W. M. (2010). *The mathematics of sex: How biology and society conspire to limit talented women and girls*. Oxford University Press.

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2017). Why are some STEM fields more gender balanced than others? *Psychological Bulletin, 143*(1), 1–35. http://dx.doi.org/10.1037/bul0000052

Coburn, C. E., & Penuel, W. R. (2016). Research-practice partnerships: Outcomes, dynamics, and open questions. *Educational Researcher, 45*(1), 48–54. https://doi.org/10.3102/0013189X16631750

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cooper, J. (2006). The digital divide: The special case of gender. *Journal of Computer Assisted Learning, 22*(5), 320–334. https://doi.org/10.1111/j.1365-2729.2006.00185.x

Crismond, D. (2001). Learning and using science ideas when doing investigate-and-redesign tasks: A study of naïve, novice and expert designers doing constrained and scaffolded design work. *Journal of Research in Science Teaching, 38*(7), 791–820. https://doi.org/10.1002/tea.1032

DeLuca, C. (2016). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability, 28*(3), 251–272. https://doi.org/10.1007/s11092-015-9233-6

Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher*, *32*(1), 5–8. https://doi.org/10.3102/0013189X032001005

Devine, A., Fawcett, K., Szűcs, D., & Dowker, A. (2012). Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. *Behavioral and Brain Functions*, *8*(1), 1–9. https://doi.org/10.1186/1744-9081-8-33

DeWitt, J., Archer, L., Osborne, J., Dillon, J., Willis, B., & Wong, B. (2011). High aspirations but low progression: The science aspirations-careers paradox among minority ethnic students. *International Journal of Science and Mathematics Education, 9*(2), 243–271. https://doi.org/10.1007/s10763-010-9245-0

Dweck, C. S. (2006). *Mindset: The new psychology of success*. Ballantine Books.

Epstein, D., & Miller, R. T. (2011). Elementary school teacher and the crisis in STEM education. *The Education Digest, 77*(1), 4–10.

Falloon, G., Stevenson, M., Beswick, K., Fraser, S., & Geiger, V. (2021). Building STEM in Schools: An Austraian cross-case analysis. *Educational Technology & Society*, *24*(4), 110–122. https://www.jstor.org/stable/48629249

Gao, X., Li, P., & Shen, L. (2020). Reviewing assessment of student learning in interdisciplinary STEM education*. International Journal of STEM Education, 7*(24). https://doi.org/10.1186/s40594-020-00225-4

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915–945. https://doi.org/10.3102/00028312038004915

Harwell, M., Moreno, M., Phillips, A., Guzey, S. S., Moore, T. J., & Roehrig, G. H. (2015). A study of STEM assessments in engineering, science, and mathematics for elementary and middle school students. *School Science and Mathematics, 115*(2), 66–74. https://doi.org/10.1111/ssm.12105

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge.

Heaverlo, C., A., Cooper, R., & Lannan, F. S. (2013). STEM development: Predictors for 6^{th}-12^{th} grade girls' interest and confidence in science and math. *Journal of Women and Minorities in Science and Engineering, 19*(2), 121–142.

Henrick, E. C., Cobb, P., Penuel, W. R., Jackson, K., & Clark, T. (2017). *Assessing research-practice partnerships: Five dimensions of effectiveness*. William T. Grant Foundation. https://wtgrantfoundation.org/library/uploads/2017/10/Assessing-Research-Practice-Partnerships.pdf

Howes, A., Kaneva, D., Swanson, D., & Williams, J. (2013). Re-envisioning STEM education: Curriculum, assessment and integrated /interdisciplinary studies. *Vision for C&A Report*. The University of Manchester. https://royalsociety.org/~/media/education/policy/vision/reports/ev-2-vision-research-report-20140624.pdf

Jovanovic, J., Solano-Flores. G., & Shavelson, R. J. (1994). Performance-based assessments: Will gender differences in science achievement be eliminated? *Education and Urban Society, 26*(4), 352–366. https://doi.org/10.1177/0013124594026004004

Kennedy, T. J., & Odell, M. R. L. (2014). Engaging students in STEM education. *Science Education International, 25*(3), 246–258. https://files.eric.ed.gov/fulltext/EJ1044508.pdf

Koh, K. (2011a). Improving teachers' assessment literacy through professional development. *Teaching Education, 22*(3), 255–276. https://doi.org/10.1080/10476210.2011.593164

Koh, K. (2011b). *Improving teachers' assessment literacy.* Pearson.
Koh, K. (2014). Teachers' assessment literacy and student learning in Singapore mathematics classrooms. In B. Sriraman, J. Cai, K. Lee, L. Fan, Y. Shimizu, C. Lim, & K. Subramaniam (Eds.), *The first sourcebook on Asian research in mathematics education: China, Korea, Singapore, Japan, Malaysia and India* (pp. 981–1010). Information Age Publishing.
Koh, K. (2017). Authentic assessment. In G. W. Noblit (Ed.), *Oxford Research Encyclopedia of Education*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264093.013.22
Koh, K., & Burke, LECA. (2018, November 20). *Design of authentic assessments to enrich mathematics learning experiences for girls: An integrated patchwork approach* [Paper presentation]. STEM in Education Conference 2018, Brisbane, Queensland, Australia.
Koh, K., Burke, LECA., Luke, A., Gong, W. G., & 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. https://doi.org/10.1177/1362168816684366
Koh, K., & Chapman, O. (2019). Building teachers' capacity in mathematics authentic assessment. In D. Potari & O. Chapman (Eds.), *International handbook of mathematics teacher education: Knowledge, beliefs, and identity in mathematics teaching and teaching development* (2nd ed., 43–76). Brill│ Sense.
Koh, K., Chapman, O., & Lam, L. (2022). An integration of virtual reality into the design of authentic assessment for STEM learning. In J. Keengwe (Ed.), *Handbook of Research on Transformative and Innovative Pedagogies in Education* (pp. 18–35). IGI Global.
Koh, K., Chapman, O., & Liu, S. M. (2020). *Manual for the design principles of authentic STEM assessment* [Unpublished manuscript]. Werklund School of Education, University of Calgary.
Koh, K., Chapman, O., & Liu, S. M. (2021, April 9–12). *Designing authentic assessments to promote girls' self-efficacy and interest in STEM subjects and careers* [Paper presentation]. American Educational Research Association Annual Meeting, Virtual.
Koh, K., Hadden, J., Parks, C., Monaghan, L., Sanden, L., Gallant, A., & LaFrance, M. (2015). Building teachers' capacity in authentic assessment and assessment for learning: A critical inquiry approach. In P. Preciado Babb, M. Takeuchi, & J. Lock (Eds.), *Proceedings of the IDEAS Designing Responsive Pedagogy* (pp. 43–52), University of Calgary.
Koh, K., & Luke, A. (2009). Authentic and conventional assessment in Singapore schools: An empirical study of teacher assignments and student work. *Assessment in Education: Principles, Policy & Practice, 16*(3), 291–318. https://doi.org/10.1080/09695940903319703
Kurup, P. M., Li, X., Powell, G., & Brown, M. (2019). Building future primary teachers' capacity in STEM: Based on a platform of beliefs, understandings and intentions. *International Journal of STEM Education, 6*(10). https://doi.org/10.1186/s40594-019-0164-5
Lamberg, T., & Trzynadlowski, N. (2015). How STEM academy teachers conceptualize and implement STEM education. *Journal of Research in STEM Education, 1*(1), 45–58. https://doi.org/10.51355/jstem.2015.8
Lesseig, L., Nelson, T. H., Slavit, D., & Seidel, R. A. (2016). Supporting middle school teachers' implementation of STEM design challenges. *School Science and Mathematics, 116*(4), 177–188. https://doi.org/10.1111/ssm.12172

Maltese, A. V., & Tai, R. H. (2010). Eyeballs in the fridge: Sources of early interest in science. *International Journal of Science Education, 32*(5), 669–685. https://doi.org/10.1080/09500690902792385

Margot, K. C., & Kettler, T. (2019). Teachers' perception of STEM integration and education: A systematic literature review. *International Journal of STEM Education, 6*(2), 1–16. https://doi.org/10.1186/s40594-018-0151-2

Master, A., Cheryan, S., Moscatelli, A., & Meltzoff, A. N. (2017). Programming experience promotes high STEM motivation among first-grade girls. *Journal of Experimental Child Psychology, 160*, 92–106. https://psycnet.apa.org/doi/10.1016/j.jecp.2017.03.013

Miles, M. B., Huberman, A. M., & Saldaña, J. (2020). *Qualitative data analysis: A methods sourcebook* (4th Ed.). Sage.

Moon, S., Carpenter, S. L., Hansen, A. K., Bushong, L., & Bianchini, J. A. (2021). Examining the effects of undergraduate STEM teacher recruitment and teacher education programs on preservice secondary science and mathematics teacher readiness and teacher performance assessment (edTPA) scores. *School Science and Mathematics, 121*(8), 452–465. https://doi.org/10.1111/ssm.12498

Murphy, T. P., & Mancini-Samuelson, G. J. (2012). Graduating STEM competent and confident teachers: The creation of a STEM certificate for elementary education majors. *Journal of College Science Teaching, 42*(2), 18–23.

Nadelson, L. S., Callahan, J., Pyke, P., Hay, A., Dance, M., & Pfiester, J. (2013). Teacher STEM perception and preparation: Inquiry-based STEM professional development for elementary teachers. *The Journal of Educational Research*, *106*(2), 157–168. https://doi.org/10.1080/00220671.2012.667014

National Academy of Engineering and National Research Council. (2014). *STEM integration in K–12 education: Status, prospects, and an agenda for research*. The Academic Press.

National Science Foundation. (2018). Science and engineering indicators 2018. National Science Foundation.

Newmann, F, M., Marks, H. M., & Gamoran, A. (1996). Authentic pedagogy and student performance. *American Journal of Education*, *104*, 280–312. https://www.jstor.org/stable/1085433

Perez-Felkner, L., Nix, S., & Thomas, K. (2017). Gendered pathways: How mathematics ability beliefs shape secondary and postsecondary course and degree field choices. *Frontiers in Psychology, 8*, 1–11. https://doi.org/10.3389/fpsyg.2017.00386

Putwain, D., & Daly, A. L. (2014). Test anxiety prevalence and gender differences in a sample of English secondary school students. *Educational Studies*, *40*(5), 554–570. https://doi.org/10.1080/03055698.2014.953914

Qablan, A. (2021). Assessing teachers education and professional development needs to implement STEM after participating in an intensive summer professional development program. *Journal of STEM Education*, *22*(2), 1–6. https://www.jstem.org/jstem/index.php/JSTEM/article/view/2495/2214

Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zárate, R. C. (2018). The relationship between test item format and gender achievement gaps on math and ELA tests in fourth and eight grades. *Educational Researcher, 47*(5), 284–294. https://psycnet.apa.org/doi/10.3102/0013189X18762105

Ripin, B. H. (1996, July). Fighting the gender gap: Standardized tests are poor indicators of ability in physics. *APS News, 5*(7), 3. https://www.aps.org/publications/apsnews/199607/gender.cfm

Roehrig, G. H., Dare, E. A., Ring-Whalen, E., & Wieselmann, J. R. (2021). Understanding coherence and integration in integrated STEM curriculum. *International Journal of STEM Education, 8*(2). https://doi.org/10.1186/s40594-020-00259-8

Saxton, E., Burns, R., Holveck, S., Kelley, S., Prince, D., Rigelman, N., & Skinner, E. A. (2014). A common measurement system for K–12 STEM education: Adopting an educational evaluation methodology that elevates theoretical foundations and systems thinking. *Studies in Educational Evaluation, 40*, 18–35. https://doi.org/10.1016/J.STUEDUC.2013.11.005

Schön, D. A. (1983) *The reflective practitioner: How professionals think in action*. Basic Books.

Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science, 29*(4), 581–593. https://doi.org/10.1177/0956797617741719

Statistics Canada. (2011). *Education in Canada: Attainment, field of study and location of study.* https://www12.statcan.gc.ca/nhs-enm/2011/as-sa/99-012-x/99-012-x2011001-eng.cfm#a4

Statistics Canada. (2016). *Insights on Canadian society women in scientific occupations in Canada*. https://www150.statcan.gc.ca/n1/pub/75-006-x/2016001/article/14643-eng.htm

Timms, M., Moyle, K., Weldon, P., & Mitchell, P. (2018). Challenges in STEM learning in Australian schools. *Report*. Australian Council for Educational Research. https://research.acer.edu.au/policy_analysis_misc/28

Tomlinson, C. A., & Moon, T. R. (2013). *Assessment and student success in a differentiated classroom*. ASCD. https://files.ascd.org/staticfiles/ascd/pdf/siteASCD/publications/assessment-and-di-whitepaper.pdf

United Nations Children's Fund, ITU (2020). *Towards an equal future: Reimagining girls' education through STEM*. https://www.unicef.org/media/84046/file/Reimagining-girls-education-through-stem-2020.pdf

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*(9), 703–713. https://doi.org/10.1177/003172171109200721

Winter, R. (2003). Contextualizing the patchwork text: Addressing problems of coursework assessment in higher education. *Innovations in Education and Teaching International, 40*(2), 112–122. https://doi.org/10.1080/1470329031000088978

Yan, Z., King, R. B., & Haw, J. Y. (2021). Formative assessment, growth mindset, and achievement: Examining their relations in the East and the West. *Assessment in Education: Principles, Policy & Practice, 28*(5–6), 676–702. https://doi.org/10.1080/0969594X.2021.1988510