Education Tech Research Dev (2021) 69:1729–1752
https://doi.org/10.1007/s11423-021-10007-6

DEVELOPMENT ARTICLE

Development and validation of a measure of authentic online work

Jennifer Darling-Aduana[1]

Accepted: 21 May 2021 / Published online: 1 June 2021
© Association for Educational Communications and Technology 2021

Abstract
Researchers tout digital learning as a tool that can increase the authenticity of student learning and assessment tasks but lack a psychometrically valid instrument to test this hypothesis. Further, there are several complementary definitions of authentic work, versus a single agreed upon definition, presented in academic literature. I synthesized this literature to develop the Authentic Online Work Rubric that measures two primary components of authentic online work – opportunities for higher-order thinking and real-world relevance. Data were collected from online courses developed by one of the largest online course providers in the United States. I validated the scale using principal component analysis before generating a lesson-level standardized coefficient using item response theory for both the higher-order thinking and real-world relevance subscales. The use of this rubric facilitates the measurement of authentic work and is targeted for use in the evaluation of learning tasks and assessments in online instructional settings to support researcher, developer, and school-based continuous improvement processes.

Keywords Online learning · Authentic work · Pedagogical issues · Secondary education · Teaching/learning strategies · Scale development

Introduction

Online course-taking is an increasingly widespread learning context, with 14 percent of secondary school students in the United States enrolling in at least one online course each year pre-COVID-19 (Gemin et al., 2015). Further, the recent almost universal transition to virtual learning due to COVID-19 is likely to have lasting consequences for the education sector, with more students predicted to enroll in virtual schools and supplemental online courses in the future (McDonald, 2020). Research also highlights the potential for online education to improve student learning and engagement by increasing access to authentic learning tasks and assessments (Hwang et al., 2014; Hwang et al., 2018; Msonde & Van Aalst, 2017; Means et al., 2009; Pérez-Sanagustín et al., 2012). Examples include

✉ Jennifer Darling-Aduana
jdarlingaduana@gsu.edu

[1] College of Education & Human Development, Georgia State University, 30 Pryor Street SW, Atlanta, GA 30303, USA

providing opportunities for interactivity, problem-based learning, and reflection to facilitate higher-order thinking and provide real-world relevance (Marks, 2000; Means et al., 2009; Newmann et al., 1996; Reeves et al., 2002).

Despite this potential, asynchronous, online course-taking - one of the most common forms of K-12 online learning (Gemin & Pape, 2017) - is often associated with lower learning outcomes, particularly for students belonging to marginalized groups (Ahn & McEachin, 2017; Heinrich et al., 2019; Heppen et al., 2017). With the quality of students' educational experiences at stake, developing a measurement tool to document the prevalence of authentic work and support its expansion in online instructional settings is essential. Further research in this area is particularly important given the increased reliance on online course-taking nationally, the lack of oversight on the predominately for-profit companies developing the most popular K-12 online course systems, and the disproportionate enrollment in online courses of students at risk of course failure and from marginalized backgrounds (Burch & Good, 2014; Clements et al., 2015; Heinrich et al., 2019; Molnar, 2013).

Researchers have established complementary definitions and bullet-pointed lists describing authentic work (Bidwell et al., 1997; Bloom, 1984; Hiebert et al., 2005; Newmann, 1992; Newmann et al., 1996; Reeves et al., 2002; Siddiq et al., 2016; Stein et al., 1996). Yet, despite the importance of authentic work (and its potential in online learning contexts), there exists to my knowledge no psychometrically valid scale of authentic work for use in either online or in-person instructional settings. For instance, most related, existing instruments were developed to support qualitative or descriptive analysis in traditional, face-to-face classrooms (Bloom, 1984; Hiebert et al., 2005; Newmann, 1992; Newmann et al., 1996; Siddiq et al., 2016; Stein et al., 1996). The one psychometrically valid scale with similar items measures the conceptually distinct construct of teacher orientations to learning (i.e., Bidwell et al., 1997) versus rating the instructional activities. Furthermore, the only researchers to examine authentic work specifically in an online setting provided only a bullet-pointed definition of components of authentic online work (Reeves et al., 2002). As a learning environment with unique strengths, challenges, and features, there is a need for a measure designed to evaluate one of the most common and increasingly widespread forms of technology-based learning – asynchronous, online course-taking (Gemin & Pape, 2017).

The development of a scale for authentic online work has the potential to clarify what researchers and practitioners mean when they discuss authenticity in online learning contexts. The validated rubric can also be used as a tool to aid in more consistent measurement, evaluation, and comparisons of authentic online work across learning platforms and contexts. For these reasons, this study included an examination of the following research questions. (1) What items should be included in an instrument to evaluate the extent to which authentic work is present in an asynchronous, K-12 online instructional environment? And (2) to what extent is the Authentic Online Work Rubric developed using those items a valid and reliable measure? By examining these questions, I aimed to establish a framework to inform future study, with the subsequently developed measure of authentic online work representing the first step toward a valid and reliable evaluative tool for researchers and practitioners. For instance, instructional designers can use the rubric when developing or updating online course material to identify lessons requiring redesign to improve authenticity and highlight successful strategies used in highly rated lessons that could be integrated elsewhere.

Specifically, the online course system used for pilot testing was fully standardized, without opportunities for students to communicate with the prerecorded instructor or peers.

Due to the scalability and profitability of this course format, this is the most prevalent form of fully online learning at the K-12 level, and thus merits documentation (Christensen et al., 2013). The research process included consolidating definitions of authentic work, developing items in alignment with that definition, refining the rubric based on expert feedback and pilot coding, and collecting data using the rubric to code lessons developed by one of the largest online course providers in the United States. I then used the data collected to establish construct validity and test reliability. The use of this rubric facilitates the measurement of authentic work and is targeted for use in the evaluation of learning tasks and assessments in asynchronous, K-12 online instructional settings.

Defining authentic work

Several researchers have defined authentic work. Marks (2000) conceptualized authentic work as (a) asking students to solve new and interesting questions, (b) prioritizing deep dives into a single topic, (c) applying content to situations outside of school, and (d) communicating ideas with others. The related Framework for Authentic Intellectual Work emphasizes the importance of student construction of knowledge through higher-order thinking, disciplined inquiry, and value beyond school (Newmann et al., 1996). Disciplined inquiry requires students to demonstrate and communicate deep understanding by building upon prior knowledge, while value beyond school necessitates an application outside of the school context (Newmann et al., 1996). Similarly, Reeves and colleagues (2002) defined authentic work in an online context as consisting of complex tasks that were open to multiple interpretations to allow for competing solutions and a diversity of outcomes. Development of these complex tasks is often accomplished by integrating real-world examples that draw on students' existing *funds of knowledge*, the knowledge students gain through participation in daily familial and community life (Lebow & Wager, 1994; Moll & González, 2004).

I also turned to prior literature to inform the development of rubric items (See Table 1 for a list of these items.) Some items were pulled word for word from rubrics on related topics (including problem-based learning, higher-order thinking, dialogic instruction, and critical mathematics), since a single authentic work instrument was not available (Hiebert et al, 2005; Hunsader et al., 2014; Land et al., 2018; Munter et al., 2015; Newmann, 1992; Osler, 2007). Other times, I adapted measurement strategies or summarized key concepts, such as those identified from bulleted lists or features of authentic work in online instructional settings (Hill & Hannafin, 2001; Reeves et al., 2002). Through this work, I generated a list of instructional strategies used to facilitate authentic work, including open-ended tasks, multiple perspectives, collaboration, reasoned communication, deep (versus surface-level) examinations, reflection opportunities, real-world connections, life skill development, meaningful work product, critical lens, and student choice. Although these strategies can be used to facilitate authentic work, they do not guarantee it. For instance, collaboration with peers on a group project can be designed to facilitate interactive, student-directed knowledge generation, but depending on the assignment it may not (Chi, 2009). Put another way, while these strategies (i.e., reasoned communication) can be used to make instructional activities authentic, not all authentic activities require the integration of each of these strategies.

Thus, to achieve the ultimate goal of creating a single, measurable definition, I returned to the three primary definitions of authentic work presented above (Marks, 2000;

Table 1 Authentic work related items, questions, and criteria

Area	Related items, questions, and criteria (and source)
Higher-order thinking	How many minutes would it take to complete each instructional task? (Hiebert et al., 2005)
	Students were encouraged to examine content from different perspectives. (Hill & Hannafin, 2001)
	Content was connected to other subject areas or domains. (Hunsader et al., 2014)
	How many minutes were devoted to student generation of knowledge (versus direct instruction)? (Land et al., 2018)
	How many minutes were dedicated to each of the following tasks? (a) introducing new material, (b) reviewing old material, (c) practicing new material, and (d) students using procedures or solving problems? (Munter et al., 2015)
	(1) There was sustained examination of a few topics rather than superficial coverage of many. (2) Lesson displayed substantive coherence and continuity. (3) Students were given an appropriate amount of time to think (i.e., prepare responses to questions). (4) Students were asked challenging questions and/or to perform challenging tasks. (5) The instructor was a model of thoughtfulness. (6) Students were asked to offer explanations and reasons for their conclusions. (Newmann, 1992)
	(1) What proportion of assessment questions, practice problems, and other instructional tasks were (a) multiple-choice and (b) allowed for various correct responses (versus a single correct answer)? (2) Activities required the student to define the tasks and subtasks needed to complete an assignment. (Reeves et al., 2002)
Real-world relevance	Assessment questions, practice problems, and other instructional tasks were embedded in a specific and meaningful context. (Hunsader et al., 2014)
	(1) Was content connected to political, economic, and social issues (i.e., racial profiling, poverty, minimum wage, gentrification, military spending, public health, health insurance, educational funding and equity, pollution)? (2) Was content connected to financial education (i.e., managing debt, high-cost loans, paying for college)? (Osler, 2007)
	(a) Did the student create a product? (2) Did the product have value in its own right? (Reeves et al., 2002)

Newmann et al., 1996; Reeves et al., 2002) to organize and distill the items identified in Table 1 around the essential components and purpose of authentic work. I accomplished this by mapping the complementary definitions, items, and instructional strategies onto Newmann and colleagues' (1996) seminal definition of authentic work as requiring higher-order thinking, disciplined inquiry, and value beyond school. Of the instructional strategies Marks (2000) identified as facilitating authentic work, the purpose of asking students to solve new and interesting questions and prioritizing deep dives into a single topic was to facilitate higher-order thinking. Applying content to situations outside of school - the third means of facilitating authentic work identified by Means (2000) - maps directly onto the goal of providing value beyond school, although, the emphasis on interesting questions could also support this goal. Lastly, communicating ideas to others is one of the primary (but not the only) means to accomplish disciplinary inquiry (Newmann et al., 1996).

Reeves and colleagues' (2002) characteristics of authentic activities also map easily onto the components of authentic work identified by Newmann and colleagues (1996). The purpose of ill-defined, complex tasks with a diversity of outcomes that require students to reflect and examine a task from multiple, interdisciplinary perspectives are all to trigger higher-order thinking. The authentic work component *supporting value outside of school* can be accomplished through authentic activities, such as ensuring real-world relevance and that work products have value outside a school setting. At the same time, providing the opportunity to collaborate is one means to accomplish disciplined inquiry.

Similarly, I could assign each instructional strategy (and the associated items) I identified through my review of prior literature to either higher-order thinking, disciplined inquiry, or value outside of school. Higher-order thinking can be facilitated through open-ended tasks, multiple perspectives, deep (versus surface-level) examinations, and reflection opportunities. Value outside of school can be facilitated through real-world connections, life skill development, meaningful work product, critical lens, and student choice. Disciplined inquiry builds on many of the same strategies identified above but is also often associated with opportunities for reasoned communication and collaboration.

The next step was to adapt these components and complementary definitions of authentic work for use in an online instructional setting. As technology-facilitated courses can be as varied in classroom environment, expectations, and instructional activities as face-to-face courses, I chose to focus on one of the most common K-12 online course structures, which requires students to log in to a highly-structured, third-party developed, asynchronous course with anytime, anywhere access (Gemin & Pape, 2017). The largest differences between conceptualizing authentic work in traditional, face-to-face settings and this type of online learning setting include the inability of students to interact directly with peers or the instructor delivering instruction as well as limited capabilities to facilitate or provide substantive feedback on open-format assignments within the self-contained online course structure (Rosé & Ferschke, 2016). It is important to note that some asynchronous models of online course-taking (primarily at the postsecondary level) have developed processes for encouraging peer support, feedback, and other forms of collaboration (i.e., Baikadi et al., 2018; Cade et al., 2014; Demmans Epp et al., 2020; Dowell et al., 2019; Rosé & Ferschke, 2016; Vassileva et al., 2016). However, these advances are not standard features within the most prevalent K-12 online course-taking platforms.

As such, an examination of authentic work online must recognize the distinct forms of instructional activities feasible within the type of asynchronous, online instructional setting studied, which is one of the most common K-12 models for offering online course-taking within a traditional school setting. The primary modification I made to address this unique learning context was to remove items that assumed strategies for facilitating authentic work (i.e., student-teacher interactions and other forms of collaboration), which by the nature of the course structure, could not be facilitated. However, I aimed to incorporate the purpose behind these tasks within the appropriate subscale. For instance, while it was not possible within the online environment studied for "students to share their knowledge with others," I included the overlapping item "students were asked to create work product that had value in its own right outside of the school setting" under the real-world relevance subscale. In this way, the purposes (if not the specific strategies) were integrated in the proposed rubric. All other items and topics (see Table 1) appeared applicable to an asynchronous, anytime, anywhere online course structure.

Relatedly, it is important to note that the definition of disciplined inquiry was problematic when attempting to classify instructional strategies, as the focus on demonstrating (and communicating) deep understanding by building upon prior knowledge speaks to the combination of strategies designed to facilitate higher-order thinking (i.e., requiring deep versus surface-level understanding) and building upon prior knowledge (one of the primary means through which real-world relevance can be facilitated). In fact, prior research indicates that the goals of peer feedback and interactions, where feasible, are also to facilitate higher-order thinking (Comer et al., 2014; Usher & Barak, 2018). For these reasons, as organized in Table 1, Icollapsed the disciplined inquiry subcategory, assigning items measuring instructional strategies designed to facilitate higher-order thinking versus real-world relevance in the appropriate subcategory. Thus, for this study, I conceptualized authentic online work as technology-supported instructional activities that facilitate student-directed learning by encouraging higher-order thinking and real-world application. Below, I define higher-order thinking and real-world relevance in greater detail and summarize how they have been operationalized in prior research.

Newmann (1992) conceptualized higher-order thinking in the classroom as requiring the posing of challenging questions or tasks, sustained examination of a few related topics, appropriate time to think, and expectations for reasoned communication (also Hiebert et al., 2005; Stein et al., 1996). Munter, Stein, and Smith (2015) extended upon this work through the definition of an instructional model that emphasized student generation of knowledge (versus direct instruction) using dialogue, collaborative work, real-time feedback, and student ownership. The facilitation of higher-order thinking, and student-generated knowledge more specifically, therefore requires open-ended (rather than closed-ended) assessment questions, practice problems, and other instructional tasks that allow for multiple correct answers based on how students choose to define the various tasks required to complete an assignment (Gamoran & Nystrand, 1992; Land et al., 2018; Lebow & Wager, 1994; Reeves et al., 2002; Stein et al., 1996). This type of instructional activity also supports a deeper understanding of underlying processes by allowing students to examine content from multiple perspectives (Chi, 2009; Hill & Hannafin, 2001; Reeves et al., 2002; Reeves & Reeves, 1997).

Real-world relevance likewise takes a variety of forms. At the most basic level, real-world relevance involves embedding instructional tasks in a meaningful context (Hiebert et al., 2015; Hunsader et al., 2014; Lebow & Wager, 1994; Newmann, 1992). For instance, opportunities for students to address a social problem encourage critical thought while providing information and skills essential for civil discourse and action in a democratic society (Au, 2012; Griner & Stewart, 2013). Providing the opportunity for students to create meaningful work products – assignments with value beyond an academic context - can also give classroom work more intrinsic meaning (Brown et al., 1989; Reeves et al., 2002).

Method

Based on the components and definitions described above, I developed and validated a reliable measure of authentic online work that was used in the coding of approximately 200 hours of online course content. First, I developed a rubric based on the review of literature described above. I then added, dropped, and refined items based on feedback from experts in online learning and authentic work. Next, I conducted a round of pilot coding. As a result of this process, I identified and remedied item wording that required additional precision. Four raters (including myself) then completed the rubric for each lesson in the courses studied. Throughout the lesson coding process, my three research assistants and I discussed item clarity and overlap. We refined (and recoded) items as needed and ultimately decided to drop unclear or repetitive items. I then conducted principal component analysis using Stata to better understand the relationship between items and to affirm the appropriate items to consolidate into each scale. Through this process, I generated two standardized scales (with a mean of zero and standard deviation of one) using item response theory, one for higher-order thinking and one for real-world relevance. Each of these steps is described in greater detail below.

Item selection

First, I developed an original rubric to measure authentic work in online contexts (see Table 2 for the scale items, Appendix A for the full rubric). I relied on a review of prior literature and preexisting instruments related to authentic work, which tended to be more theoretical than psychometrically validated, to define and operationalize the constructs of interest (i.e., Bidwell et al., 1997; Bloom, 1984; Hiebert et al., 2005; Newmann, 1992; Newmann et al., 1996; Reeves et al., 2002; Siddiq et al., 2016; Stein et al., 1996). The higher-order thinking subscale was designed to measure the extent to which the lesson lecture, assignments, practice problems, or assessments asked students to think deeply and critically about course content, often requiring students to generate new knowledge. The real-world relevance scale was created to identify the extent to which the instructor or instructional material placed lesson content in an applied context.

At this stage, I also generated items to create a communication and collaboration subscale (see Appendix B for a list of items). However, due to the sample and learning context studied, I was unable to validate this subscale and test whether communication and collaboration might, in fact, represent a third component of authentic online work (instead of strategies that could facilitate the other subscales as hypothesized). Thus, discussion of this subscale is excluded from the rest of the rubric development process.

I used a four-point Likert-type scale for each item: never, rarely, sometimes, and often. *Never* was only selected if the item was never observed. *Rarely* indicated that the item occurred once or twice during the approximately 40-minute lesson, and *often* indicated that the item occurred in all cases or all but one or two components of the lesson. We rated an item *sometimes* if that item occurred three or more times (more than the cutoff for Rarely) but did not meet the criteria to be classified as often. Deciding on the appropriate rating was accomplished by first determining the level of analysis (i.e., each new topic introduced, each assessment question) depending on the item. For instance, in an algebra 1

Table 2 Authentic online work subscale items

Higher-order thinking items

1. Students spent instructional time generating knowledge (versus direct instruction).
2. Assessment questions, practice problems, and other instructional tasks were delivered in an open-response format (i.e., NOT multiple-choice or true/false).
3. Assessment questions, practice problems, and other instructional tasks allow for various correct responses (i.e., open response questions that allow students to apply concepts to a topic of their choosing).
4. There was more than one method for generating an acceptable response.
5. Assignments required students to gather information on their own.
6. Students were asked challenging questions and/or to perform challenging tasks (such as those requiring extensive prior content knowledge, multiple steps, or the application of multiple concepts.)
7. Students were asked to offer reasoning to support responses.

Real-world relevance items

1. Assessment or instructional tasks were embedded in a specific, meaningful context.
2. Assessment or instructional tasks asked students to synthesize, interpret, explain, or evaluate complex information in addressing a concept, problem, or issue.
3. Students were asked to create work product that had value in its own right outside of the school setting.
4. Assessment or instructional tasks asked students to elaborate on their understanding, explanations, or conclusions through extended writing.

lesson, if five new mathematical techniques were introduced in the lesson and each time it was embedded in a specific, meaningful context this example would be rated *often*. If three out of five mathematical techniques were embedded in a specific, meaningful context, this example would be rated *sometimes*. Conversely, if out of 20 assessment questions, only two allowed for various correct responses, this item would be rated *rarely*, while if the same lesson contained four out of 20 assessment questions that allowed for various correct responses, the item would be rated *sometimes*.

The rubric was then refined based on feedback from six content experts and pilot coding. I recruited content experts through my professional network, asking for introductions where necessary. I targeted scholars familiar with the online platform studied along with scholars who had published peer-reviewed articles on online learning or authentic work. The expert review consisted of sharing a copy of the proposed instrument for feedback. Most expert review suggestions consisted of rephrasing items to be clearer, which I implemented as suggested. Next, I engaged in a pilot coding process, whereby I used the rubric to rate several lessons across different courses. As a result, I consolidated items that captured the same online lesson characteristics and provided additional definition for terms with multiple possible interpretations.

Online course setting

This section details features of the district and online course system used to pilot the rubric. All courses were developed by a for-profit company that contracts with over 16,000 schools in the United States. Despite the focus on a single online course platform, the focus on standardization and scalability within the more prevalent online course vendors contributes to the enactment of similar systems across developers (Cottom, 2017; Molnar, 2013). Across the large, urban school district studied for the pilot, 44 different online courses had enrollments of at least 50, and up to 700, high school students a year. Enrollment was distributed across 46 high schools and allowed students to earn the course credits required for high school graduation. Despite this, the online course provider classified some of the courses as being designed for a middle school versus high school (reading) level.

Students accessed the courses through a hybrid blended model (Christensen et al., 2013), where they could log in to the system from a school-based computer lab with a lab monitor available for assistance during an assigned class period. Students could also log in and receive credit for completing course content outside the school day from any internet-enabled device. This is one of the most common models for online course-taking within brick and mortar schools due to often limited infrastructure for more transformative technology-supported instruction (Christensen et al., 2013).

Each course was broken into 29 to 58 lessons of approximately 40 minutes in length. These lessons each included a pre-recorded, teacher-directed video lecture. Students had to watch the lecture, complete assignments, and earn a minimum score on the end-of-lesson assessment to earn credit for lesson completion. The most common assignments included responding to practice problems that required students to remember and recite lesson content. Less often, assignments might require students to write an essay, complete a worksheet (i.e., to develop a family budget), or research a topic (i.e., potential careers). In addition, most end-of-lesson assessments consisted entirely of closed-response questions. Most of these questions required students to remember and recite lesson content. Some questions required surface-level application to a new context or situation. Open-ended questions – as well as questions requiring extended application, evaluation, or synthesis - were relatively rare.

Data collection

My research team and I collected data by reviewing and coding 412 lessons according to the newly developed Online Authentic Work Rubric using access to the online course system provided by a district partner. I limited the data validation process to the 10 courses in which the most students were enrolled. The ten courses included in the analysis were algebra 1, career planning and development, citizenship, ninth-grade English/language arts (ELA), healthy living, geometry, personal finance, physical science, United States history, and world history. In addition to sampling courses from across subject areas, these courses varied in grade level, with the online course vendor identifying the courses sampled as targeting seventh through twelfth-grade students.

Where more than one semester of a course fell into the top 10, I selected the semester in which more students enrolled for analysis after spot-checking course content to establish similar levels of authenticity between semesters. For instance, the algebra 1 course was a yearlong course that required the completion of both algebra 1 semester one and algebra 1 semester two. Since slightly more students enrolled in the first semester of the course, I only watched and coded lessons in the first-semester algebra 1 course, so algebra 1 was not represented twice in the pool for analysis. Sampling only one semester of yearlong courses allowed me to code courses across a wider range of subject areas and grade levels instead of, for instance, including two algebra 1 courses. The courses excluded for this reason demonstrated comparable levels of authentic work across semesters.

In total, the courses examined represented 60 percent of all online course enrollments in the district during the study period. Coding the top 10 courses was selected to provide variability in terms of grade level and subject while still representing the typical experience of a student completing an online course. The top 10 courses were also selected due to the time-intensive nature of data collection to maximize the impact of limited resources. On average, 440 students in the district studied enrolled in each of the top 10 courses during the study period compared to an average of only 206 students enrolled in the next 10 most frequently enrolled in courses. Further, courses not in the top 10 were each enrolled in by less than two percent of online course-takers in the district.

After course selection, I trained three additional raters. Two raters were graduate students. One was an advanced undergraduate student. All were pursuing degrees in the education field. Training consisted of discussing the rubric and walking through a sample coding process in-person. Then, each researcher rated a lesson that I also coded. We discussed any discrepancies, repeating the coding and discussion process until consistent before the rater proceeded to code an entire course on their own. I continued to code additional courses with anyone who did not code satisfactorily and compared interrater reliability monthly, retraining as necessary to re-calibrate.

Each online lesson was evaluated on the extent to which higher-order thinking and real-world relevance were present. All responses were entered in Qualtrics for analysis. There was a primary rater assigned to each course who rated every lesson in the course based on their content expertise. For instance, I assigned a former math educator to rate the algebra 1 course. Others coded a few lessons from that course to establish interrater reliability and ensure continued consistency in rubric interpretation. Raters assigned a rating within one point of each other on the four-point Likert-type scale in 93 percent of cases.

Throughout the coding process and after all courses were rated, the other raters and I discussed any discrepancies or confusion regarding the interpretation of items. I revised or dropped these items and culled items whose meanings overlapped substantially with other items. For instance, the original rubric asked raters to evaluate the extent to which lessons "asked students to communicate responses verbally or in written form" and "asked students to offer reasoning to support responses." In this instance, the first item was removed, because it provided no additional information after accounting for the second item. Additionally, a few items related to applicability to social issues and applying a critical lens to content originally thought to represent real-world relevance were removed due to a factor analysis conducted after all courses were rated that identified that these items loaded on a separate factor. Similarly, two questions related to learning life (and career-relevant) skills loaded on a separate factor, and thus were not included in the real-world relevance scale. Table 3 provides a list of all items removed from the final rubric and the reason(s) for exclusion.

Results

Factor analysis

I first conducted principal component factor analysis on items collected at the lesson-level from both the higher-order thinking and real-world relevance subscales. The second factor had an eigenvalue of 1.355 versus 1.028 for a third factor. Convention dictates that eigenvalues above one should exist (Cattell, 1978). Since the eigenvalue for the third factor was within rounding error and the items that loaded on the third factor did not align with prior theory, I chose to proceed with two versus three factors. Together, the two factors accounted for 57.5 percent of the variance in the underlying data. Factor loadings on the higher-order thinking subscale ranged from 0.61 to 0.78, while factor loadings on the real-world relevance subscale ranged from 0.62 to 0.85, as seen in Fig. 1. Each item contributed a level of uniqueness to its respective subscale, ranging from 0.270 to 0.635.

Scale development

After coding the 412 online lessons using the Authentic Online Work Rubric, I used IRT graded response models (Samejima, 2016) to place the extent to which higher-order thinking and real-world relevance were facilitated in each lesson on standardized, continuous scales with a mean of zero and standard deviation of one. The grading response model used information from each ordinal response category on the rubric while also allowing each item to vary in terms of its difficulty and discrimination. In this way, the analysis mapped the interrelationship between each subscale items (after accounting for the varied difficulty level of each subscale) while generating a single scale that captured both the quality (the presence of each item) and quantity (the Likert-type scale frequency responses) into a single standardized variable. Generating a single number for each subscale was used to identify general trends within and across courses. For this reason, a similar strategy would likely be helpful for researchers or practitioners interested in identifying more macro

Table 3 Removed items with reason for exclusion

Removed higher-order thinking item	*Reason(s) for exclusion*
Activities allowed students to decide the process they wanted to undertake to complete an assignment.	There was substantial overlap with the retained item "There was more than one method for generating an acceptable response."
Students were encouraged to think about course content or procedures from more than perspective.	There was substantial overlap with retained items #1-4 on the higher-order thinking subscale. As the item with the least specificity, this item was excluded.
Course content encouraged students to reflect upon their own values	This item loaded on a separate culturally relevant factor.
Content was connected to other subject areas.	This item loaded on a separate factor. Further, the qualitative analysis confirmed that being connected to other subject areas was not associated with increased higher-order thinking.
Content was connected to other domains within the same subject.	There was consistency in ratings across lessons within courses. Further, the qualitative analysis confirmed that being connected to domains within the same subject was not associated with increased higher-order thinking.
There was sustained examination of topics.	Attempts to further define the meaning of sustained examination led to substantial overlap with retained items. For instance, sustained examination meant that sufficient time needed to devoted to the topic that students could engage with the content in a deep (versus surface-level) manner, which could be accomplished by time devoted to student-generation of knowledge, performing challenging tasks etc. Further, the word "sustained" alluded to time devoted to a particular topic, which we found to not necessarily be correlated with the facilitation of higher-order thinking this subscale was designed to measure.
Lesson displayed substantive coherence and continuity.	This item loaded on a separate factor. Further, the qualitative analysis confirmed that displaying coherence and continuity was not associated with increased higher-order thinking.
Students were asked to apply content (i.e., answer practice problems) following the introduction of new content or skills.	There was consistency in ratings across lessons within courses on this item, indicating that this was more of a structural versus instructional decision. Further, the qualitative analysis indicated that most practice problems in the course studied did not facilitate higher-order thinking.
Removed real-world thinking item	*Reason for exclusion*
Content was explicitly connected to political, economic, or social issues.	There was substantial overlap with the retained item, "Assessment or instructional tasks were embedded in a specific, meaningful context."
Students were asked to connect course content to their daily lives.	There was substantial overlap with the retained item, "Assessment or instructional tasks were embedded in a specific, meaningful context."
Instruction aimed to developed skills relevant for adult life (i.e., finances, nutrition).	This item loaded on a separate life skills factor.
Content was explicitly linked to a potential career.	This item loaded on a separate life skills factor.

Table 3 (continued)

Students were asked to create a work product.	There was substantial overlap with the retained item, "Students were asked to create work product that had value in its own right outside of the school setting."
Multiple forms of assessment, such as self-assessment and portfolios, were used to evaluate student learning.	There was substantial overlap with the retained item, "Students were asked to create work product that had value in its own right outside of the school setting."
Content highlighted the experiences of populations that have been historically marginalized.	This item loaded on a separate culturally relevant factor.
Content highlighted the contributions of non-White and/or female individuals.	This item loaded on a separate culturally relevant factor.
Content provided information essential for civil discourse or action in a democratic country.	This item loaded on a separate culturally relevant factor.
Students were provided opportunities for extended learning.	There was substantial overlap with the retained item, "Students were asked to create work product that had value in its own right outside of the school setting."

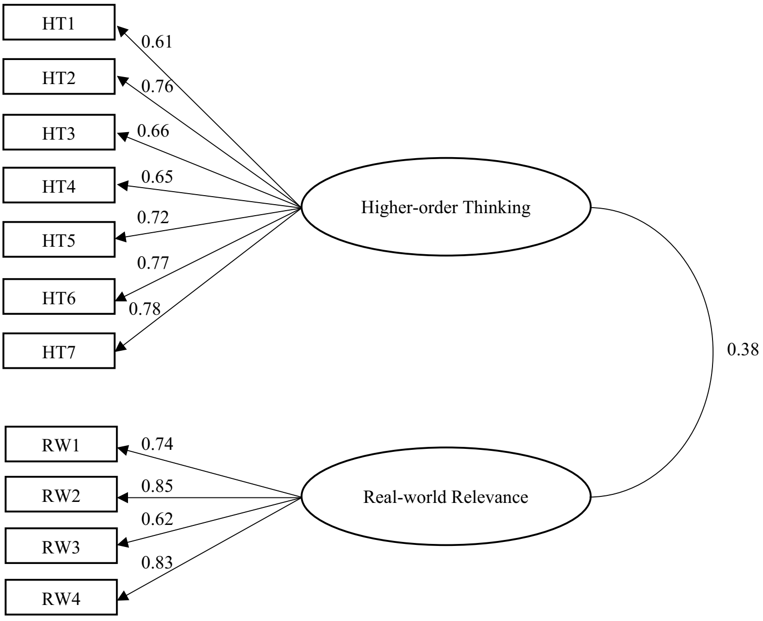
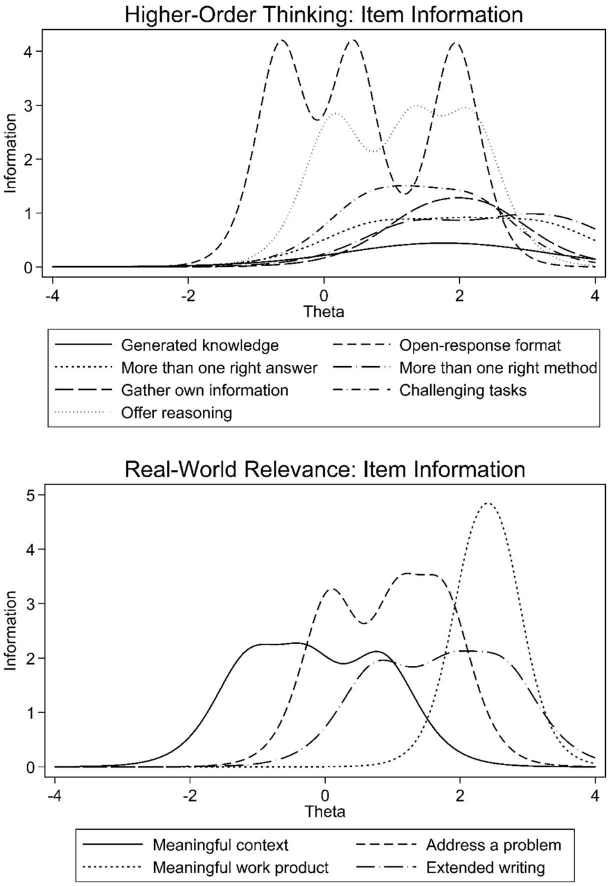


Fig. 1 Factor structure of the authentic online work rubric

course or system-level trends, while instructional designers and educators evaluating a single lesson or course would likely find individual item ratings more informative.

The use of IRT graded response models also assisted in defining authentic work in alignment with the first aim of this study. Specifically, resulting item information functions placed each item in context with other items on the subscale in terms of difficulty. The

Fig. 2 Higher-Order thinking and real-world relevance item information functions



resulting figures (see Fig. 2) provide a layered approach to understanding the facilitation of each higher-order thinking and real-world relevance, with tasks shown on the left side of the x-axis representing potentially more easily implementable tasks that may not contribute as much to the overall level of authentic work in a given lesson. In contrast, tasks shown on the right side of the x-axis are tasks that while potentially more challenging to implement will likely also contribute more to the overall level of authentic work.

The resulting Authentic Online Work rubric is an 11-item instrument consisting of two subscales. The higher-order thinking subscale includes seven items, while the real-world relevance subscale includes four items. The Cronbach's alpha for the higher-order thinking scale is 0.82, and the Cronbach's alpha for the real-world relevance scale is 0.74. Conventions in the social sciences identified the internal consistency of the real-world relevance scale as acceptable and the internal consistency of the higher-order thinking scale as good (DeVellis, 2016). The two scales represent two distinct but correlated constructs, r = 0.384, $p < 0.001$.

As summarized in Table 2, seven items loaded onto the higher-order thinking scale. Of those, the extent to which the lesson asked students to respond in an open-response format and offer reasoning to support their assertions provided the most influential item information for scale development (see Fig. 2). Lessons that asked students challenging questions and/or to perform challenging tasks (such as those requiring extensive prior content knowledge, multiple steps, or the application of multiple concepts), also scored more highly on higher-order thinking. Whether students were expected to gather their own information and answer questions with more than one right answer was often necessary to achieve the highest scores in higher-order thinking, as demonstrated by the higher relative item information provided by these items at higher values on the x-axis of the higher-order thinking portion of Fig. 2. In contrast, whether students were allowed opportunities to generate (versus being given) knowledge provided relatively less information, although it is unclear whether this was because the item lacked valuable information or due to low levels of prevalence in the lessons rated.

Of the four items that loaded onto the real-world relevance scale, whether the instructor provided a meaningful context for lesson content was most influential on the low end of the scale. This indicated that providing a meaningful context often distinguished between lessons rated low versus very low on the real-world relevance scale. Whether students were asked to evaluate, apply, or synthesize complex information to solve a problem or issue provided the most information in the middle range of the scale, while whether students were asked to create work product with meaning outside of a school context distinguished the lessons with the highest level of real-world relevance from lessons with moderate levels of real-world relevance. Work product in this context refers to any output created by completing instructional activities, including but not limited to an essay, multimedia presentation, business plan, or family budget. In contrast, solving a purely symbolic algebra problem that differs from problems in the lecture only through the replacement of number values in the equation would not be considered work product, because there is no material generated distinguishable from the course content created by the online course vendor. Expectations of writing as a means for students to elaborate on their understanding, explanations, or conclusions often accompanied other elements of real-world relevance.

Construct validity

To evaluate construct validity, I examined correlations between each scale and the types of tasks raters identified as present within each lesson. These correlations also help establish face validity, demonstrating that the types of tasks and course components known to contribute to higher-order thinking and real-world relevance were correlated with the appropriate subscales. As shown in Table 4, lessons that required more higher-order thinking were also more likely to include student-directed tasks that required writing and students to actively interact with the online system. Lessons that demonstrated more real-world relevance were more likely to require the evaluation and synthesis of ideas, instructional tasks in the online lessons which were often presented within an applied context. Both higher-order thinking and real-world relevance were also likely to be present in lessons requiring students to create work product (i.e., an essay, presentation, or science lab report). High correlations between real-world relevance and critical thinking, application, and evaluation tasks reinforced observational findings that integrating real-world examples were one of the most common means used in the online courses to facilitate these processes. However, the higher-order thinking scale was better at distinguishing between the inclusion of recitation tasks (which were correlated with real-world relevance but not high-order thinking) and more complex tasks (which were correlated with higher-order thinking).

Nonetheless, lower correlations between the higher-order thinking scale and tasks requiring critical thinking, application, and evaluation indicate an important distinction

Table 4 Correlations between subscales, rubric ratings, and course components

	Higher-order thinking	Real-world relevance
Higher-order thinking	1.000	
real-world relevance	0.384***	1.000
rubric ratings		
Proportion skill introduction	−0.069	0.014
Interactive task(s)	0.369***	0.284***
Reading task(s)	−0.107**	0.053
Writing task(s)	0.480***	0.182***
Recite task(s)	−0.074	0.158***
Demonstrate task(s)	0.206***	0.238***
Critical thinking task(s)	0.244***	0.442***
Application task(s)	0.184***	0.424***
Evaluation task(s)	0.210***	0.498***
Synthesis task(s)	0.235***	0.609***
Creation task(s)	0.479***	0.425***
Vendor−provided course components		
Assignment	−0.066	0.097**
Lab	0.073*	0.189***
Material title	0.053	0.129***
Online resource	−0.116***	−0.026
Summary	−0.073*	0.109***
Vocabulary	−0.167***	−0.151***
Warm-up	−0.030	0.161***

*$p < 0.10$ **$p < 0.05$ ***$p < 0.01$

between more surface-level measures of higher-order thinking and this scale, in that this scale prioritized processes that require students to take ownership of learning processes and generate their own knowledge. For example, a math problem that required students to solve an equation almost identical to one introduced in the lecture might require critical thinking or the application of recently introduced skills to a new context. But solving this math problem would not meet the higher bar for this higher-order thinking scale, since students were expected to replicate a process to determine a solution that had only one correct answer. However, an in-depth worksheet on budgeting that asked students to research trends in household expenses in the United States and apply that knowledge along with their mathematical skills to develop current and future personal budgets was rated highly on higher-order thinking (as well as real-world relevance).

There was comparatively less association between vendor-provided information on course components and the higher-order thinking and real-world relevance scales. Notably, the inclusion of additional activities (i.e., assignments, labs, material titles) in addition to direct instruction by the vendor when designing lessons was generally associated with more real-world relevance. This makes sense because these additional activities often provided an in-depth example, with warm-up and summary components often focusing specifically on framing the content the lecture will introduce in terms of real-world applicability. In contrast, the inclusion of additional vendor-developed activities such as assignments, labs, or material titles did not appear to be associated with higher-order thinking. Lessons that included more technology-directed, non-interactive features (i.e., vocabulary, online resources) were often rated lower in higher-order thinking. This makes sense, as technology-directed features often left less time for more in-depth activities such as research and writing that facilitated higher-order thinking.

Scale ratings

When using the Authentic Online Work Rubric to rate the 412 lessons within the top 10 most frequently enrolled in courses, the mean rating was 0.009 (SE = 0.910) and −0.001 (SE = 0.897) for the higher-order thinking and real-world relevance subscales respectively. The mean and distribution reflect that the scale generation process employed creates standardized coefficients. I identified variability in the level of higher-order thinking and real-world relevance both within and across courses (see Table 5) that aligned with qualitative observations of course content. Mean ratings of higher-order thinking ranged from −0.691 in the healthy living course, a health course, to 0.907 for the personal finance course. Mean ratings of real-world relevance ranged from −0.862 for the algebra 1 course to 1.148 for the personal finance course. Standard errors by course for each subscale ranged from 0.060 to 0.211. Statistical outliers (i.e., lessons with ratings more than two standard deviations away from the mean) were consistent with trends identified qualitatively when observing lesson

Table 5 Lesson ratings of higher-order thinking and real-world relevance by course

	N	Mean	SE	Min	Max
Higher-order thinking					
Algebra 1	33	0.166	0.065	−0.390	0.922
Career planning and development	45	0.464	0.066	−1.215	1.307
Citizenship	37	0.619	0.113	−1.215	1.560
ELA 9	32	−0.004	0.096	−1.215	1.295
Geometry	46	−0.093	0.109	−1.215	2.128
Healthy living	58	−0.691	0.102	−1.215	2.141
Personal finance	32	0.907	0.211	−1.215	3.237
Physical science	29	0.490	0.146	−1.215	2.404
Survey of U.S. history	49	−0.907	0.079	−0.670	1.494
Survey of world history	51	−0.002	0.125	−1.215	3.237
Real-world relevance					
Algebra 1	33	−0.862	0.077	−1.328	−0.050
Career planning and development	45	−0.835	0.075	−1.328	0.606
Citizenship	37	−0.127	0.060	−0.670	0.606
ELA 9	32	−0.869	0.092	−1.328	0.191
Geometry	46	0.253	0.093	−0.670	1.973
Healthy living	58	−0.173	0.118	−1.328	1.267
Personal finance	32	1.148	0.164	−0.670	2.806
Physical Science	29	0.653	0.146	−0.670	2.353
Survey of U.S. History	49	0.345	0.080	−0.670	1.494
Survey of World History	51	0.468	0.090	−0.670	2.806

lectures, assignments, and assessments. As this triangulation between sources indicated that outliers represented valid data points, outliers were included in all analyses.

When plotting higher-order thinking and real-world relevance scores by lesson, I also identified an increasingly strong relationship between the two subscales as ratings increased. In the bottom-left quadrant of Fig. 3, I identified a weak relationship between the two subscales. However, in the top-right quadrant, there emerged a strong, positive relationship. This varied association indicated that at high levels, it was more likely that higher-order thinking facilitated real-world relevance and vice versa. Thus, lessons plotted on the top-right quadrant of Fig. 3 where the two components studied occurred in tandem appeared to represent a fuller realization of authentic online work.

Limitations and future research

The present study examined only the 10 most frequently enrolled in online courses within a single large, urban district. All 10 courses were developed by the same, prominent vendor. This means that I was not able to evaluate the rubric on unobserved levels or types of authenticity that may be present in other courses developed by the same vendor or online courses developed by another vendor. As such, I encourage researchers to apply (and adapt) this rubric to other types of online learning systems and instructional environments. Continued refinement could focus on improving measures of reliability and validity as well as a more explicit focus on the quality of authentic work.

More specifically, in upcoming work, I intend to use this rubric to measure authentic work within an online learning environment that incorporates both asynchronous and synchronous online learning environments. This study will allow me to validate the communication and collaboration subscale. I will also use this study to test whether this component represents a distinct, third subscale versus strategies that should more appropriately be integrated into the higher-order thinking or real-world relevance subscales. Extending rubric use to this new context will also help establish generalizability and reliability to a unique online learning context. Promising topics for future research also include examining preconditions (i.e., scaffolding, student-teacher dynamics, orientations to learning) to the successful enactment of the authentic tasks measured with this rubric.

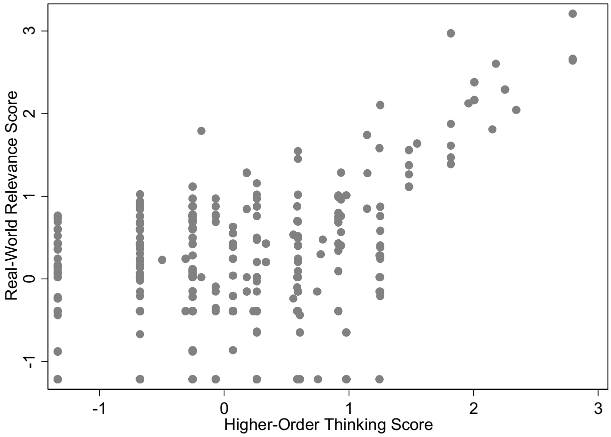


Fig. 3 Associations between higher-order thinking and real-world relevance scales

There is also room for more K-12 online courses to integrate instructional strategies that support student-directed and community-grounded learning (Scardamalia & Bereiter, 2010). Without observing these instructional strategies, it is not possible to develop psychometrically-validated instruments to measure them. For example, despite some definitions of authentic work including elements of collaboration or communication (Hiebert et al., 2005; Newmann, 1992; Stein et al., 1996), I was unable to validate related items, since the online system used for validation (and similar, commonly used asynchronous, online course systems) did not facilitate these types of activities. At the postsecondary level, innovations such as peer assessment within Massive Open Online Courses (MOOC) and similar structures provide a road map for the potential integration and measurement of interaction-based strategies for facilitating authentic work (Baikadi et al., 2018; Cade et al., 2014; Demmans Epp et al., 2020; Dowell et al., 2019; Rosé & Ferschke, 2016; Vassileva et al., 2016). As the K-12 online course structure continues to evolve and learn from similar educational products and services, there will likely be the potential to add – and validate – related items.

Discussion

Increasing access to authentic work, which is associated with improved student engagement and learning outcomes (Marks, 2000), represents one of the greatest potential benefits to the use of online learning systems (Reeves et al., 2002). Yet that promise is difficult to achieve and rarely observed in practice (Darling-Aduana, 2021; Heinrich et al., 2019; Hohlfeld et al., 2017; Reeves et al., 2002). The goal of improving authentic work in online settings requires the development of a shared definition and tools for measuring its presence. Through the development and validation of such a measure, I aimed to establish a framework for the future study, identification, and implementation of authentic online work.

The resulting rubric represents a first step to the definition and measurement of authentic online work, which can be used by researchers and educators to inform evaluation and continuous improvement processes of online course systems. Notably, the rubric consistently measures two important components of authentic work – higher-order thinking and real-world relevance. The higher-order thinking scale was designed to measure the extent to which students were asked to think deeply and critically about course content. The real-world relevance scale was created to identify the extent to which course content was embedded in a meaningful context. Correlations between these subscales and related instructional tasks demonstrated alignment with these definitions.

In particular, the higher-order thinking subscale was driven by the extent to which students were asked to engage with tasks that could be completed using multiple correct methods, accomplished through more open-ended tasks, and required students to explain their reasoning. At the highest level, higher-order thinking was often facilitated by complex tasks, such as those requiring extensive prior content knowledge, multiple steps, or the application of multiple concepts. The real-world relevance subscale required, at minimum, the identification of a context for course content. Higher-levels required the use of course content to address a problem. Integrating opportunities for students to develop meaningful work product – something of value outside of an academic context – generated the highest ratings of real-world relevance. Beyond using the rubric as an evaluative tool, the identification of and relationships between each item and the larger constructs of higher-order thinking and real-world highlighted here provides insight into strategies for the integration of authentic work into online courses. For instance, educators at the school district who provided data for rubric validation are currently using the results to target and reimagine lessons rated low on authentic work by integrating active and inquiry-based learning strategies identified from lessons with higher ratings on higher-order thinking and real-world relevance rubric items.

Correlations between the higher-order thinking and real-world relevance subscales, rubric ratings, and course components established face and concurrent validity, while the Cronbach alphas for each subscale demonstrated a minimum level of reliability. Further, the rubric possessed enough nuance to detect varying levels of authentic work across and within courses, which is essential for practical use, with the personal finance and physical science courses facilitating the most consistently high levels of higher-order thinking and real-world relevance. I also identified that when observed at high levels, a strong, positive relationship between each subscale emerged, potentially indicating that high levels of higher-order thinking were facilitating real-world relevance and vice versa.

Conclusion

The Authentic Online Work Rubric provides a consistent definition and measurement tools for researchers and practitioners interested in evaluating authentic online work. The rubric was validated across over 200 hours of lessons from 10 courses within a single, widely used online learning interface. Subscales for the extent to which course content and instructional tasks facilitated higher-order thinking and real-world relevance were developed and validated. The rubric can be used to compare various forms of online instruction and provide information essential to inform online course redesigns and other continuous improvement processes. Additional research is needed to refine subscales and determine what adaptions are necessary for use across varied online learning contexts.

Appendix A: authentic online work rubric

Lesson information

Assign lesson id and instructor id using a 00 format, where the first lesson and instructor are assigned a 01 id and the second lesson and instructor are assigned a 02.

Observer Name:
Course Name:
Lesson Name:
Lesson Id:
Instructor Id:
Which of the following components are included in the lesson?

Warm-up
Lecture
Practice
Assessment

Writing
Interactive (i.e., lab, performance)
Other (please describe)

Total time required to watch any lecture component to the lesson (rounded to the nearest minute): ___

The number of minutes spent related to each of the following instructional expectations (You may allocate the same minute to more than one instructional expectation. The total number of minutes will likely exceed the total lecture length):

___ Skill introduction
___ Drilling/practice
___ Review
___ Assessment
___ Games
___ Enrichment/accelerated instruction
___ Other (please describe)

Which of the following orders of thinking are required to complete instructional tasks? (Refer to Bloom's Taxonomy (Bloom, 1984) for definitions and examples of each component.

Listen
Recite/remember
Demonstrate
Think critically
Apply
Synthesize
Evaluate
Create

Higher-order thinking

*Rate each item, where **rarely** indicates the item occurred once or twice during the lesson and **often** indicates that the item occurred all but once or twice during the lesson.*

	Never	Rarely	Sometimes	Often
Students spent instructional time generating knowledge (versus direct instruction).	1	2	3	4
Assessment questions, practice problems, and other instructional tasks were delivered in an open-response format (i.e., NOT multiple-choice or true/false).	1	2	3	4
Assessment questions, practice problems, and other instructional tasks allow for various correct responses (i.e., open response questions that allow students to apply concepts to a topic of their choosing).	1	2	3	4
There was more than one method for generating an acceptable response.	1	2	3	4
Assignments required students to gather information on their own.	1	2	3	4

	Never	Rarely	Sometimes	Often
Students were asked challenging questions and/or to perform challenging tasks (such as those requiring extensive prior content knowledge, multiple steps, or the application of multiple concepts.)	1	2	3	4
Students were asked to offer reasoning to support responses.	1	2	3	4

Real-world relevance

*Rate each item, where **rarely** indicates the item occurred once or twice during the lesson and **often** indicates that the item occurred all but once or twice during the lesson.*

	Never	Rarely	Sometimes	Often
Assessment or instructional tasks were embedded in a specific, meaningful context.	1	2	3	4
Assessment or instructional tasks asked students to synthesize, interpret, explain, or evaluate complex information in addressing a concept, problem, or issue.	1	2	3	4
Students were asked to create work product that had value in its own right outside of the school setting.	1	2	3	4
Assessment or instructional tasks asked students to elaborate on their understanding, explanations, or conclusions through extended writing.	1	2	3	4

Describe and include personal reflections on the content, skill focus, and instructional tasks included in this lesson. Also describe any implicit (or explicit) values, expectations, norms, or beliefs expressed by the instructor or course content.

Appendix B: unvalidated communication and collaboration subscale items

The following items were generated based on a review of literature and related scales to create a communication and collaboration subscale for the Authentic Online Work Rubric. This subscale could not be validated due to the sample and learning context studied. However, future research will include validating this subscale and examining use of the entire scale in alternative online instructional settings.

1. Students worked together to accomplish learning tasks.
2. Students shared their knowledge with others.
3. Students were provided prompt feedback (from their teacher and/or peers).
4. Students were provided opportunities to respond to feedback (i.e., by improving work).
5. Instructional activities provided opportunities for multiple forms of communication (i.e., public presentation, discussion, debate, writing).

Acknowledgements I thank the American Educational Research Association, National Science Foundation, and William T. Grant Foundation for generous funding of this research; Mason Shero, Hillary Bendert, and

Jacqueline Aboulafia for their excellent research assistance; Carolyn Heinrich, Gary Henry, Jason Grissom, and Mark Warschauer for their invaluable feedback; Vendor researchers for their support in accessing data from the online instructional system; District administrators and educators for their assistance in providing student record data and support of the project data analysis and fieldwork; and Vanderbilt University and the Wisconsin Evaluation Collaborative at the Wisconsin Center for Education Research, University of Wisconsin-Madison, for ongoing support of this initiative.

Funding This research was supported by a grant from the American Educational Research Association which receives funds for its "AERA Grants Program" from the National Science Foundation under NSF award NSF-DRL #1749275. Opinions reflect those of the author and do not necessarily reflect those AERA or NSF.

Declarations

Conflict of interest The author declares that they have no conflict of interest.

Ethical approval This study was reviewed and approved by the University of Wisconsin-Madison Education and Social/Behavioral Science IRB, protocol #2014-1239-CP005.

Consent to participate This study included the collection of data from archival documentation. As data were not collected from individuals, informed consent was not required.

References

Ahn, J., & McEachin, A. (2017). Student enrollment patterns and achievement in Ohio's online charter schools. *Educational Researcher, 46*(1), 44–57. https://doi.org/10.3102/0013189X17692999

Au, W. (2012). *Critical curriculum studies: Education, consciousness, and the politics of knowing*. Routledge.

Baikadi, A., Demmans Epp, C., & Schunn, C. D. (2018). Participating by activity or by week in MOOCs. *Information and Learning Science*. https://doi.org/10.1108/ILS-04-2018-0033

Bidwell, C. E., Frank, K. A., & Quiroz, P. A. (1997). Teacher types, workplace controls, and the organization of schools. *Sociology of Education, 70*(4), 285–307.

Bloom, B. S. (1984). *Taxonomy of educational objectives, Handbook 1: Cognitive domain* (2nd ed.). . Longman.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32–42.

Burch, P., & Good, A. G. (2014). *Equal scrutiny: Privatization and accountability in digital education*. Harvard Education Press.

Cade, W., Dowell, N., Graesser, A., Tausczik, Y., & Pennebaker, J. (2014). Modeling Student Socioaffective Responses to Group Interactions in a Collaborative Online Chat Environment. *Educational Data Mining (EDM),* 399-400.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavior and life sciences*. Plenum.

Chi, M. T. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*(1), 73–105.

Christensen, C. M., Horn, M. B., & Staker, H. (2013). Is K-12 blended learning disruptive? An introduction to the theory of hybrids. *Clayton Christensen Institute for Disruptive Innovation.*

Clements, M., Stafford, E., Pazzaglia, A. M., & Jacobs, P. (2015). *Online course use in Iowa and Wisconsin public high schools: The results of two statewide surveys*. Regional Educational Laboratory Midwest.

Comer, D. K., Clark, C. R., & Canelas, D. A. (2014). Writing to learn and learning to write across the disciplines: Peer-to-peer writing in introductory-level MOOCs. *International Review of Research in Open and Distributed Learning, 15*(5), 26–82.

Darling-Aduana, J. (2021). Authenticity, engagement, and performance in online high school courses: Insights from micro-interactional data. *Computers & Education., 167*, 104175.

Demmans Epp, C., Phirangee, K., Hewitt, J., & Perfetti, C. A. (2020). Learning management system and course influences on student actions and learning experiences. *Educational Technology Research and Development, 68*(6), 3263–3297. https://doi.org/10.1007/s11423-020-09821-1

DeVellis, R. F. (2016). *Scale development: Theory and applications*. Sage Publications.

Dowell, N., Lin, Y., Godfrey, A., & Brooks, C. (2019). Promoting inclusivity through time-dynamic discourse analysis in digitally-mediated collaborative learning. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Artificial intelligence in education* (pp. 207–219). Springer International Publishing.

Gamoran, A., & Nystrand, M. (1992). Taking students seriously. In F. Newmann (Ed.), *Student engagement and achievement in American secondary schools* (pp. 40–61). Teachers College Press.

Gemin, B., Pape, L., Vashaw, L., & Watson, J. (2015). *Keeping pace with k-12 digital learning: An annual review of policy and practice*. Evergreen Education Group.

Gemin, B., & Pape, L. (2017). *Keeping pace with K-12 online learning, 2016*. Evergreen Education Group.

Griner, A. C., & Stewart, M. L. (2013). Addressing the achievement gap and disproportionality through the use of culturally responsive teaching practices. *Urban Education, 48*(4), 585–621. https://doi.org/10.1177/0042085912456847

Heinrich, C. J., Darling-Aduana, J., Good, A. G., & Cheng, H. (2019). A look inside online educational settings in high school: Promise and pitfalls for improving educational opportunities and outcomes. *American Educational Research Journal, 56*(6), 2147–2188. https://doi.org/10.3102/0002831219838776

Heppen, J. B., Sorensen, N., Allensworth, E., Walters, K., Rickles, J., Taylor, S. S., & Michelman, V. (2017). The struggle to pass algebra: Online vs face-to-face credit recovery for at-risk urban students. *Journal of Research on Educational Effectiveness, 10*(2), 272–296. https://doi.org/10.1080/19345747.2016.1168500

Hiebert, J., Stigler, J. W., Jacobs, J. K., Givvin, K. B., Garnier, H., Smith, M., Hollingsworth, H., Manaster, A., Wearne, D., & Gallimore, R. (2005). Mathematics teaching in the United States today (and tomorrow): Results from the TIMSS 1999 video study. *Educational Evaluation and Policy Analysis, 27*(2), 111–132. https://doi.org/10.3102/01623737027002111

Hill, J. R., & Hannafin, M. J. (2001). Teaching and learning in digital environments: The resurgence of resource-based learning. *Educational Technology Research and Development, 49*(3), 37–52.

Hohlfeld, T. N., Ritzhaupt, A. D., Dawson, K., & Wilson, M. L. (2017). An examination of seven years of technology integration in Florida schools: Through the lens of the Levels of Digital Divide in Schools. *Computers & Education, 113*, 135–161. https://doi.org/10.1016/j.compedu.2017.05.017

Hunsader, P. D., Thompson, D. R., Zorin, B., Mohn, A. L., Zakrzewski, J., Karadeniz, I., Fisher, E. C., & MacDonald, G. (2014). Assessments accompanying published textbooks: the extent to which mathematical processes are evident. *ZDM, 46*(5), 797–813. https://doi.org/10.1007/s11858-014-0570-6

Hwang, G. J., Hung, P. H., Chen, N. S., & Liu, G. Z. (2014). Mindtool-assisted in-field learning (MAIL): An advanced ubiquitous learning project in Taiwan. *Educational Technology & Society, 17*(2), 4–16.

Hwang, G. J., Lai, C. L., Liang, J. C., Chu, H. C., & Tsai, C. C. (2018). A long-term experiment to investigate the relationships between high school students' perceptions of mobile learning and peer interaction and higher-order thinking tendencies. *Educational Technology Research and Development, 66*(1), 75–93.

Land, T. J., Bartell, T. G., Drake, C., Foote, M. Q., Roth McDuffie, A., Turner, E. E., & Aguirre, J. M. (2018). Curriculum spaces for connecting to children's multiple mathematical knowledge bases. *Journal of Curriculum Studies*. https://doi.org/10.1007/978-3-319-92907-1_15

Lebow, D. G., & Wager, W. W. (1994). Authentic activity as a model for appropriate learning activity: Implications for emerging instructional technologies. *Canadian Journal of Educational Communication, 23*, 231–231.

Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American Educational Research Journal, 37*(1), 153–184. https://doi.org/10.3102/00028312037001153

McDonald. K. (2020, May 11). Four K-12 education models that may gain popularity during COVID-19. Forbes. https://www.forbes.com/sites/kerrymcdonald/2020/05/11/four-k-12-education-models-that-may-gain-popularity-during-covid-19/#27b4e6936b77

Cottom, T. M. (2017). *Lower ed: The troubling rise of for-profit colleges in the new economy*. The New Press.

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2009). *Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies*. U.S. Department of Education, Office of Planning, Evaluation, and Policy Development.

Moll, L. C., & González, N. (2004). Engaging life: A funds-of-knowledge approach to multicultural education. In J. A. Banks (Ed.), *Handbook of research on multicultural education* (pp. 699–715). Jossey-Bass.

Molnar, A. (2013). *School commercialism: From democratic ideal to market commodity*. Routledge.

Msonde, S. E., & Van Aalst, J. (2017). Designing for interaction, thinking and academic achievement in a Tanzanian undergraduate chemistry course. *Educational Technology Research and Development, 65*(5), 1389–1413.

Munter, C., Stein, M. K., & Smith, M. A. (2015). Dialogic and direct instruction: Two distinct models of mathematics instruction and the debate(s) surrounding them. *Teachers College Record, 117*(11), 1–32.

Newmann, F. M. (1992). Higher-order thinking and prospects for classroom thoughtfulness. In F. Newmann (Ed.), *Student engagement and achievement in American secondary schools* (pp. 62–91). Teachers College Press.

Newmann, F. M., Marks, H. M., & Gamoran, A. (1996). Authentic pedagogy and student performance. *American Journal of Education, 104*(4), 280–312. https://doi.org/10.1086/444136

Osler, J. (2007). A guide for integrating issues of Social and economic justice into mathematics curriculum. RadicalMath. Retrieved from http://www.radicalmath.org/docs/SJMathGuide.pdf

Pérez-Sanagustín, M., Santos, P., Hernández-Leo, D., & Blat, J. (2012). 4SPPIces: A case study of factors in a scripted collaborative-learning blended course across spatial locations. *International Journal of Computer-Supported Collaborative Learning, 7*(3), 443–465.

Reeves, T. C., Herrington, J., & Oliver, R. (2002). Authentic activities and online learning. In A. Goody, J. Herrington, & M. Northcote (Eds.), *Quality conversations: Research and Development in Higher Education* (Vol. 25, pp. 562–567). HERDSA.

Reeves, T. C., & Reeves, P. M. (1997). Effective dimensions of interactive learning on the World Wide Web. In B. H. Khan (Ed.), *Web-based instruction* (pp. 59-66).

Rosé, C. P., & Ferschke, O. (2016). Technology support for discussion based learning: From computer supported collaborative learning to the future of massive open online courses. *International Journal of Artificial Intelligence in Education, 26*(2), 660–678.

Samejima, F. (2016). Graded response models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of item response theory* (Vol. 1, pp. 85–100). Springer.

Scardamalia, M., & Bereiter, C. (2010). A brief history of knowledge building. *Canadian Journal of Learning and Technology*. https://doi.org/10.21432/T2859M

Siddiq, F., Hatlevik, O. E., Olsen, R. V., Throndsen, I., & Scherer, R. (2016). Taking a future perspective by learning from the past–A systematic review of assessment instruments that aim to measure primary and secondary school students' ICT literacy. *Educational Research Review, 19*, 58–84. https://doi.org/10.1016/j.edurev.2016.05.002

Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal, 33*(2), 455–488. https://doi.org/10.3102/00028312033002455

Usher, M., & Barak, M. (2018). Peer assessment in a project-based engineering course: comparing between on-campus and online learning environments. *Assessment & Evaluation in Higher Education, 43*(5), 745–759.

Vassileva, J., McCalla, G. I., & Greer, J. E. (2016). From small seeds grow fruitful trees: How the PHelpS peer help system stimulated a diverse and innovative research agenda over 15 years. *International Journal of Artificial Intelligence in Education, 26*(1), 431–447. https://doi.org/10.1007/s40593-015-0073-9

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Jennifer Darling-Aduana is an Assistant Professor in the Department of Learning Sciences at Georgia State University. Her research focuses on the equity implications of K-12 educational policies and practices, such as the widespread expansion of digital learning, as well as the more micro student-teacher and student-curriculum interactions that inadvertently contribute to social reproduction in the classroom.