Authentic Online Work Scale Development and Rubric

I quantified the extent to which curricular content and instructional activities in online course videos and assignments facilitated authentic work through the development and use of the Authentic Online Work Rubric. The rubric asked raters to evaluate the extent to which courses provided opportunities for higher-order thinking and real-world relevance, two primary components of authentic work identified in prior research (Marks, 2000; Newmann, Marks, & Gamoran, 1996; Reeves, Herrington, & Oliver, 2002).

Scale Development

First, I developed an original rubric to measure authentic work in online contexts (see Appendix A for the final rubric). I relied on a review of prior literature and preexisting instruments on authentic work, which tended to be more theoretical than psychometrically validated, to define and operationalize the constructs of interest (i.e., Au, 2012; Bidwell, Frank, & Quiroz, 1997; Hiebert et al., 2005; Newmann, 1992; Newmann et al., 1996; Reeves et al., 2002; Siddiq, Hatlevik, Olsen, Throndsen, & Scherer, 2016; Stein, Grover, & Henningsen, 1996). The higher-order thinking scale was designed to measure the extent to which students were asked 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 course content resonated with or were applicable to students' lives, interests, and/or aspirations. The rubric was then refined based on feedback from content experts and pilot coding. I trained three additional raters using the rubric, establishing interrater reliability at the beginning of and throughout the coding process. Each online lesson was then 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. Others coded a few lessons from each course to establish interrater reliability and ensure consistent rubric interpretation. Before reconciliation, raters assigned a rating within one point of each other on the four-point Likert-type scale items in 93 percent of cases. To minimize concerns regarding variability in ratings based on the rater, I used only ratings from the primary rater of each course to ensure each lesson within a given course was rated by the same person.

Throughout the coding process and after all courses were rated, the research team and I discussed any discrepancies in or confusion regarding the interpretation of items. We revised or dropped these items and culled items whose meanings overlapped substantially with other items.[1] Additionally, I removed several items about applying a critical lens to social issues that I originally categorized under real-world relevance based on a review of prior research. Exploratory factor analysis conducted after all courses were rated indicated that these items represented a unique construct, but there was insufficient information to create a third, psychometrically valid subscale. The same factor analysis identified that two questions related to learning life (and career-relevant) skills also represented a separate construct and thus were excluded from the real-world relevance scale.

After coding, I use item response theory (IRT) rating scale models to place the extent to which higher-order thinking and real-world relevance were present in each lesson on standardized, continuous scales. The Cronbach's alpha for the higher-order thinking scale was 0.82, while the Cronbach's alpha for the real-world relevance scale was 0.74. Conventions in the social sciences identified the internal consistency of the real-world relevance scale as acceptable

[1] 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 to offer reasoning to support responses." The first item was removed, because it provided no additional information after accounting for the second item.

and the internal consistency of the higher-order thinking scale as good (DeVellis, 2016). The two scales represented two distinct but correlated constructs, $r=0.384$, $p<0.001$.

As shown in Figure 1, 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 information for scale development. Of the four items that loaded onto the real-world relevance scale, not providing meaningful context for lesson content was most influential on the low end of the 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. The resulting scales had close to a normal distribution, as shown in Figure 2, apart from a second peak on the higher-order thinking scale at the extreme low end of the distribution.

[Insert Figure 1]

[Insert Figure 2]

To evaluate convergent validity, I examined correlations between each scale and the type of tasks raters identified as present within each lesson. As shown in Table 1, lessons that required more higher-order thinking were also more likely to include student-directed tasks that required interactivity and writing, while lessons that demonstrated more real-world relevance were more likely to require the evaluation and synthesis of ideas. Both higher-order thinking and real-world relevance were likely to be present in lessons requiring students to create work product. 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.

High correlations between real-world relevance and critical thinking, application, and evaluation tasks reinforce 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 demonstration tasks (which higher-order thinking was correlated with).

[Insert Table 1]

Nonetheless, lower correlations between the higher-order thinking scale and tasks requiring critical thinking, application, and evaluation indicates an important distinction between some measures of higher-order thinking and this scale, in that this scale prioritizes 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 might require critical thinking or the application of recently introduced skills to a new context, but would not meet the higher bar for this higher-order thinking scale, since students were expected to replicate a process to determine the solution, which 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 the additional activities often provide a more 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 guarantee higher-order thinking. However, lessons that included more technology-directed, non-interactive features (i.e., vocabulary, online resources) were often rated lower in higher-order thinking.

References

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

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.

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

Hiebert, J., Stigler, J. W., Jacobs, J. K., Givvin, K. B., Garnier, H., Smith, M., ... & 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.

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.

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). New York, NY: 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.

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*, Volume 25 (pp. 562-567). Hammondville, NSW: HERDSA.

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.

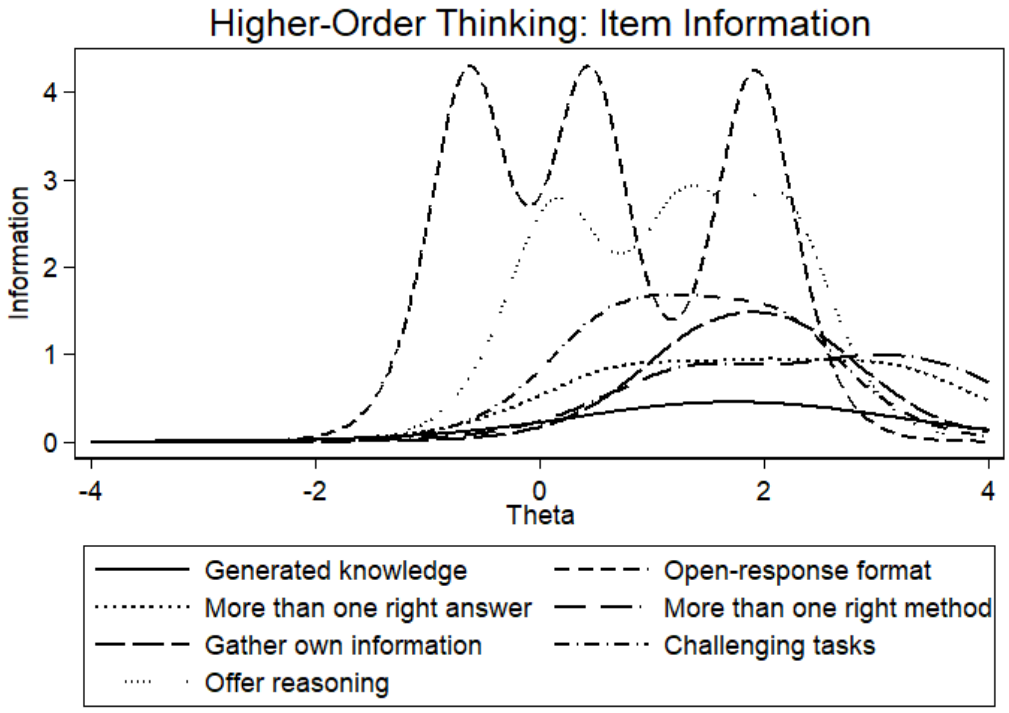
Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal, 33*(2), 455-488.

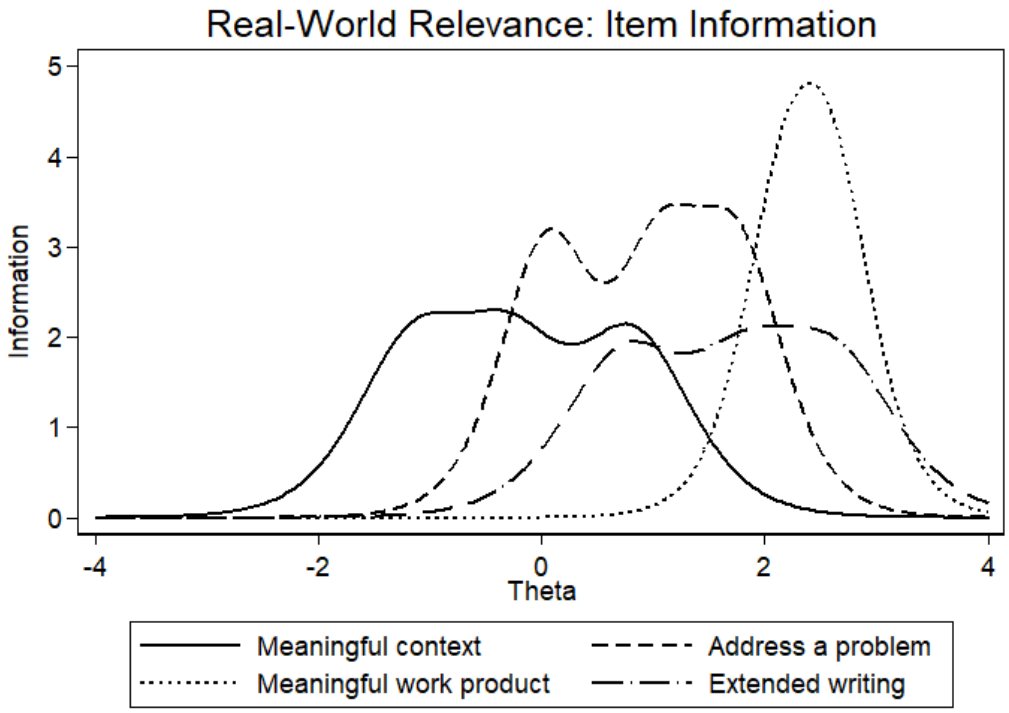
Table 1. Correlations between Subscales, Rubric Ratings, and Course Components

	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***

Excerpt from: Darling-Aduana, J. (2019). Authenticity, engagement, and performance in online high school courses: Insights from micro-interactional data.

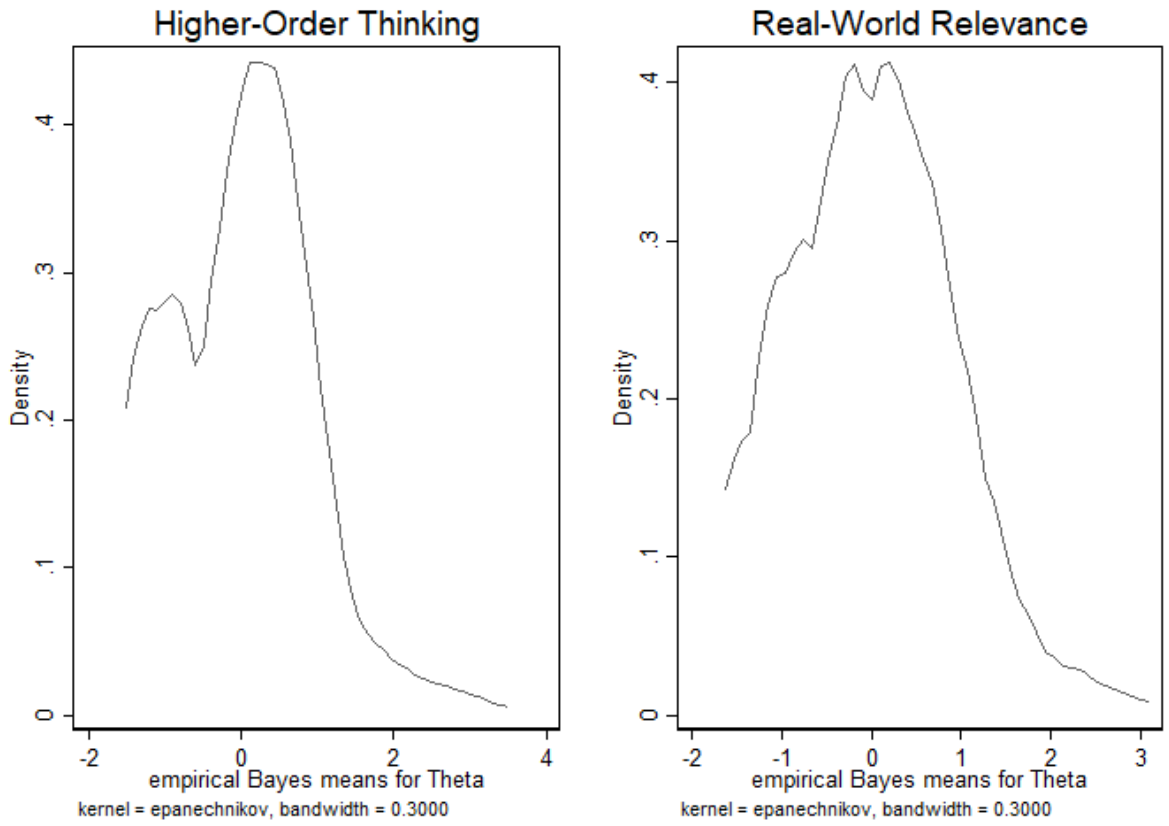
Figure 1. Higher-Order Thinking and Real-World Relevance Item Information Functions





Excerpt from: Darling-Aduana, J. (2019). Authenticity, engagement, and performance in online high school courses: Insights from micro-interactional data.

Figure 2. Distribution of Higher-Order Thinking and Real-World Relevance Scales



Appendix A: Authentic Online Work Rubric

The following rubric was developed based on a review of relevant literature and preexisting instruments on authentic work and were refined based on feedback from content experts and pilot coding. Training consisted of discussing the rubric and walking-through a sample coding process in-person. Then, each observer rated a lesson that I also coded. We discussed any discrepancies, repeating the coding and discussion process until consistent before the observer proceeds to coding 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.

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 number of minutes required to watch the lecture videos (round to the nearest minute): ___

Number of minutes spent related to particular 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?

- ☐ Listen
- ☐ Recite/remember
- ☐ Demonstrate
- ☐ Think critically
- ☐ Apply
- ☐ Synthesize
- ☐ Evaluate
- ☐ Create

Excerpt from: Darling-Aduana, J. (2019). Authenticity, engagement, and performance in online high school courses: Insights from micro-interactional data.

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	NA
Students spent instructional time generating knowledge (versus direct instruction).	1	2	3	4	NA
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	NA
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	NA
There was more than one method for generating an acceptable response.	1	2	3	4	NA
Assignments required students to gather information on their own.	1	2	3	4	NA
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	NA
Students were asked to offer reasoning to support responses.	1	2	3	4	NA

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	NA
Assessment or instructional tasks were embedding in a specific, meaningful context.	1	2	3	4	NA
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	NA
Students were asked to create work product that had value in its own right outside of the school setting.	1	2	3	4	NA
Assessment or instructional tasks asked students to elaborate their understanding, explanations, or conclusions through extended writing.	1	2	3	4	NA

Excerpt from: Darling-Aduana, J. (2019). Authenticity, engagement, and performance in online high school courses: Insights from micro-interactional data.

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.

INSTRUCTOR CHARACTERISTICS

Which of the following best describes the lesson instructor's presenting gender?

- Female
- Male
- Other (please describe)

Provide comments on any information relevant to your classification as well as any concerns you have regarding the accuracy of your assessment.

Which of the following best describes the lesson instructor's presenting race or ethnicity?

- White or Caucasian
- Black or African American
- Hispanic or Latino
- Asian, Pacific Islander, Hispanic Native, or Alaskan Native
- Other (please describe)

Provide comments on any information relevant to your classification as well as any concerns you have regarding the accuracy of your assessment.

Excerpt from: Darling-Aduana, J. (2019). Authenticity, engagement, and performance in online high school courses: Insights from micro-interactional data.