Educational Technology Research and Development

Does Audience Matter? Comparing Teachers' and Non-Teachers' Application and Perception of Quality Rubrics for Evaluating Open Educational Resources

--Manuscript Draft--

Manuscript Number:	ETRD-D-16-00356R5	
Full Title:	Does Audience Matter? Comparing Teachers' and Non-Teachers' Application and Perception of Quality Rubrics for Evaluating Open Educational Resources	
Article Type:	Research Section	
Keywords:	Open Educational Resources; Quality Rubrics; Audience; Rubric Perception; Rubric Application	
Corresponding Author:	Min Yuan University of Utah Salt Lake City, UT UNITED STATES	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:	University of Utah	
Corresponding Author's Secondary Institution:		
First Author:	Min Yuan	
First Author Secondary Information:		
Order of Authors:	Min Yuan	
	Mimi Recker	
Order of Authors Secondary Information:		
Funding Information:	Utah State University	Dr. Min Yuan
Abstract:	<p>While many rubrics have been developed to guide people in evaluating the quality of Open Educational Resources (OER), few studies have empirically investigated how different people apply and perceive such rubrics. This study examines how participants (22 teachers and 22 non-teachers) applied three quality rubrics (comprised of a total of 17 quality indicators) to evaluate 20 OER, and how they perceived the utility of these rubrics. Results showed that both teachers and non-teachers found some indicators more difficult to apply, and displayed different response styles on different indicators. In addition, teachers gave higher overall ratings to OER, but non-teachers' ratings had generally higher agreement values. Regarding rubric perception, both groups perceived these rubrics as useful in helping them find high-quality OER, but differed in their preferences for quality rubrics and indicators.</p>	
Response to Reviewers:	N/A	

Does Audience Matter? Comparing Teachers' and Non-Teachers' Application and Perception of Quality Rubrics for Evaluating Open Educational Resources

Min Yuan
min.yuan@eccles.utah.edu
University of Utah

Mimi Recker
mimi.recker@usu.edu
Department of Instructional Technology and Learning Sciences, Utah State University

Compliance with Ethical Standards:

Funding: Portions of this research were supported by Utah State University.

Conflict of Interest: The authors declare that they have no conflict of interest.

Acknowledgements:

This research is partially supported by Utah State University. Portions of this research were previously presented at the American Educational Research Association Annual Meeting (AERA 2016) in Washington, DC. We thank Drs. Anne Diekema and Andy Walker for their valuable input.

Abstract

While many rubrics have been developed to guide people in evaluating the quality of Open Educational Resources (OER), few studies have empirically investigated how different people apply and perceive such rubrics. This study examines how participants (22 teachers and 22 non-teachers) applied three quality rubrics (comprised of a total of 17 quality indicators) to evaluate 20 OER, and how they perceived the utility of these rubrics. Results showed that both teachers and non-teachers found some indicators more difficult to apply, and displayed different response styles on different indicators. In addition, teachers gave higher overall ratings to OER, but non-teachers' ratings had generally higher agreement values. Regarding rubric perception, both groups perceived these rubrics as useful in helping them find high-quality OER, but differed in their preferences for quality rubrics and indicators.

Keywords

Open Educational Resources; Quality Rubrics; Audience; Rubric Perception; Rubric Application

Biography

Min Yuan is an Instructional Designer at University of Utah. She completed her Ph.D. in the Department of Instructional Technology and Learning Sciences at Utah State University in 2015. Her research interests include technology integration in higher education and evaluating the quality of online resources available for educational purposes.

Mimi Recker is Professor in the department of Instructional Technology and Learning Sciences at Utah State University. She received degrees from the University of Pennsylvania and the University of California, Berkeley. Her research interests include design and evaluation of interactive learning environments; educational uses of new media, information, and communication technologies; human-computer interaction; models of multimedia information access; social context of dynamic information ecologies.

Does Audience Matter? Comparing Teachers' and Non-Teachers' Application and Perception of Quality Rubrics for Evaluating Open Educational Resources

Introduction

Open Educational Resources (OER) are online teaching, learning, and research materials that reside in the public domain or under open licenses permitting reuse and repurposing (UNESCO, 2016). The free and open nature of OER offers great educational opportunities. OER offer access to large numbers of resources in many disciplines and in a wide variety of instructional types and media formats. For example, the site [OERCommons.org](https://oercommons.org) provides access to OER developed by over 100 partner organizations, ranging from universities and government agencies to school districts and non-profit organizations. These OER take different forms, including full courses, interactives and simulations, digital textbooks, and instructional artifacts such as lessons plans, worksheets, and assessments. In addition, OER can be reused unchanged, remixed with existing instructional materials, adapted to other learning contexts, and redistributed to other users (Atkins, Brown, & Hammond, 2007; Richter & McPherson, 2012).

With these great possibilities, OER are continuing to garner increased interest in all levels and sectors of education, especially for online and distance education (UNESCO, 2016). In the U.S., for example, recent U.S. Secretaries of Education stated:

“OER can not only accelerate and enrich learning, they can also substantially reduce costs... We want everyone to benefit from the universal access to **high quality** educational opportunities that OER have placed within our hands” (Duncan, 2012, emphasis ours).

“Openly licensed educational resources can increase equity by providing all students, regardless of zip code, access to **high quality** learning materials that have the most up-to-date and relevant content” (King, 2016, emphasis ours).

1
2
3
4 However, many issues and challenges must be addressed to fully meet the broad vision
5
6 put forth in these quotes (Atkins et al., 2007; Murphy, 2013). One important aspect (as
7
8 highlighted in the bolded text) is the notion of “quality.” In particular, evaluating the quality of
9
10 OER has taken on new importance, as some OER have been found to have low quality, which
11
12 may negatively impact instructional use (Bundsgaard & Hansen, 2011; Fitzgerald, Lovin, &
13
14 Branch, 2003).
15
16
17

18
19 Many approaches have been proposed to examine this problem. For example, OER can
20
21 be vetted by expert panels and stored in curated repositories, an expensive and laborious process.
22
23 Alternatively, users may be asked to provide ratings of OER in order to *crowdsource* their
24
25 quality ratings (Abramovich & Schunn, 2012; Porcello & Hsi, 2013). Recently, machine-
26
27 learning algorithms have been developed to perform the OER evaluation task (Wetzler et al.,
28
29 2013). Similarly, analyses of usage logs as a result of user activities can help reveal which OER
30
31 users find more useful (Abramovich, Shunn, & Correnti, 2013).
32
33
34
35

36 Finally, a common approach involves the use of quality rubrics (Achieve, 2011;
37
38 UNESCO, 2016). In previous work, we published a comprehensive review of rubrics developed
39
40 for evaluating the quality of OER (Yuan & Recker, 2015). As described below, we found that
41
42 most rubrics are composed of multiple dimensions that serve to decompose the overall quality
43
44 construct into more concrete terms (e.g., Achieve, 2011; Nesbit, Belfer, & Leacock, 2007). We
45
46 also found that rubrics often use rating scales for each dimension, and sometimes include scoring
47
48 guides to help the user apply them.
49
50
51
52

53 Our review, however, revealed little empirical literature examining if the use of quality
54
55 rubrics was effective in guiding people’s evaluation of OER. In particular, little is known about
56
57 how reliably people apply these rubrics, and how they perceive the utility of these rubrics.
58
59
60
61
62
63
64
65

Further, as quality rubrics have different audience types, for example teachers and non-teachers, it is also important to examine how different audiences may vary in the way they approach evaluation tasks. Therefore, the purpose of this study is to empirically examine three quality rubrics, with a focus of exploring people's application characteristics, their interrater reliability, their perceptions of the rubrics' utility, and whether teachers and non-teachers differed in their approaches. To align with these purposes, the study addressed the following research questions:

1. How do people apply quality rubrics (and their indicators) when evaluating OER? Are there differences between teachers and non-teachers?

2. How do people perceive the utility of quality rubrics (and their indicators) when evaluating OER? Are there differences between teachers and non-teachers?

Literature Review

In our investigations of the research literature related to rubrics for evaluating OER quality, we draw on three areas to guide our work. These are: 1) how rubrics define quality and constituent dimensions, and how these dimensions are measured, 2) how rubric designers address reliability and utility issues, and 3) user factors affecting rubric use. As a key user group is teachers, the latter is positioned in terms of a framework for characterizing teacher knowledge.

Characteristics of Quality Rubrics

Defining quality in OER. OER quality is a complex construct, comprised of multiple *dimensions*, such as content quality, pedagogical value, and technological utility (Custard & Sumner, 2005; Haughey & Muirhead, 2005; Sumner, Khoo, Recker, & Marlino, 2003; Wetzler et al., 2013). Each quality dimension can in turn be comprised of one or more *quality indicators*. For example, content quality is often considered the most important dimension of OER quality;

1
2
3
4 and can consist of multiple quality indicators, such as accuracy, completeness, and clarity
5
6 (Fitzgerald et al., 2003; Nesbit et al., 2007). Thus, the complexity inherent in the quality
7
8 construct requires people to consider multiple quality dimensions (as well as multiple indicators)
9
10 when evaluating OER.
11
12
13

14 Rubrics designed to evaluate OER quality thus often consist of several dimensions or
15
16 indicators, with each dimension or indicator measuring one aspect of OER quality. However,
17
18 rubrics may choose to emphasize different dimensions and quality indicators (Yuan & Recker,
19
20 2015). For example, the OER rubric from Achieve (2011) focuses on the pedagogical value of
21
22 OER, while the rubric from Kurilovas, Bireniene, & Serikoviene (2011) highlights the
23
24 reusability of OER. Further, rubrics often contain unique indicators. For example, the rubric
25
26 from Achieve (2011) uniquely includes “opportunities for deeper learning” as an important
27
28 quality indicator, while the rubric from Wetzler et al. (2013) considers “sponsor” of OER as an
29
30 important factor influencing OER quality. These differences make it necessary to explore which
31
32 quality rubrics and indicators lead to more effective application and user acceptance.
33
34
35
36
37

38 **Scoring mechanism.** Different quality rubrics provide different rating scales. For
39
40 example, the rubric from Nesbit et al. (2007) uses a 5-point rating scale, the rubric from Wetzler
41
42 et al. (2013) uses a binary (yes—no) rating scale, and the rubric from Fitzgerald et al. (2003)
43
44 does not provide a rating scale. In addition, some rubrics may choose to provide scoring guides.
45
46 For example, the rubric from Achieve (2011) lists different requirements for different points on
47
48 the rating scale, while the rubric from Nesbit et al. (2007) provides examples to help facilitate the
49
50 evaluation task. This study examines how participants’ ratings are distributed over the response
51
52 scale as well as whether participants tend to pick midpoint and/or extreme values.
53
54
55
56
57

58 **Reliability and Utility of Rubrics**

59
60
61
62
63
64
65

1
2
3
4 **Reliability.** One benefit of using rubrics in evaluation tasks is that they are assumed to
5
6 make judgments more reliable (Jonsson & Svingby, 2007). For example, the rubric developed by
7
8 McMartin, McKenna, and Youssefi (2000) to assess students' knowledge of engineering practice
9
10 had high reliability, and the rubric developed by Harris, Grandgenett, and Hofer (2010) to
11
12 measure the quality of the integration of technology in teaching proved to be reliable in two
13
14 rounds of testing. Finally, after reviewing 75 studies, Jonsson and Svingby (2007) concluded that
15
16 rubrics generally lead to more reliable evaluation.
17
18
19

20
21 However, in some cases, rubric application does not result in a satisfactory level of
22
23 reliability (Rezaei & Lovorn, 2010; Stellmack, Konheim-Kalkstein, Manor, Massey, & Schmitz,
24
25 2009). For example, Stellmack et al. (2009) developed a rubric to evaluate students' APA-style
26
27 writing, but found the resulting reliability lower than their expectation. Rezaei and Lovorn
28
29 (2010) argued that rubrics would not produce good reliability without proper training of users.
30
31 Thus, in order to understand the effectiveness of a quality rubric, it is important to first examine
32
33 its reliability.
34
35
36

37
38 There are several different methods to measure reliability. This study focuses on
39
40 interrater reliability, which measures the agreement of ratings obtained with a rubric when used
41
42 by different raters (Moskal & Leydens, 2000). A higher interrater reliability indicates that users
43
44 have better agreement when applying the rubric in evaluation tasks (Moskal & Leydens, 2000).
45
46 Low reliability, on the other hand, can provide useful data to rubric designers by indicating areas
47
48 needing improvement.
49
50
51

52
53 **Utility.** Perceived utility greatly influences users' adoption of a tool (Legris, Ingham, &
54
55 Collette, 2003). For example, Davis (1989) found that perceived utility was positively
56
57 correlated with users' current usage and potential future usage. Chismar and Wiley-Patton (2003)
58
59
60
61
62
63
64
65

1
2
3
4 considered the perceived utility as an important factor impacting users' intention to use a tool.
5
6 Perceived utility consists of multiple dimensions, including whether the tool can save users'
7
8 time, simplify users' work, increase users' productivity, etc. (Davis, 1989). Thus, examining
9
10 participants' perception of the utility of rubrics forms an important component of this study.
11
12
13

14 **Studies of rubric reliability and utility.** As noted above, few studies have empirically
15
16 examined the reliability and utility of rubrics designed to evaluate OER quality (Gligora
17
18 Marković, Kliček, & Plantak Vukovac, 2014; Vargo, Nesbit, Belfer, & Archambault, 2003;
19
20 Yuan & Recker, 2015). However, a related area is the growing use of assessment rubrics in
21
22 instruction – as a means to communicate instructor expectations as well as help students learn. In
23
24 this area, a few studies have examined rubric reliability or perceived utility (Reddy & Andrade,
25
26 2010).
27
28
29

30
31 For example, East's (2009) rubric for assessing the quality of foreign language writing,
32
33 and Newell, Dahm, and Newell's (2002) rubric for assessing students' learning outcomes both
34
35 achieved good inter-rater reliability. Other studies found that rubric reliability can vary when
36
37 used by different user groups. For example, Chong and Romkey (2017) developed a rubric to
38
39 assess the quality of undergraduate thesis reports, and found good inter-rater reliability among
40
41 generalist experts but only fair agreement between supervisors and generalist experts. Similarly,
42
43 Oakleaf (2009) found varying levels of agreement between different user groups (e.g., librarians,
44
45 faculty, students) when using a rubric for assessing information literacy.
46
47
48
49

50
51 Regarding the perceived utility of rubrics, Bolton (2006) and Andrade and Du (2005)
52
53 found that students had a positive attitude towards rubrics, believing that rubrics can help them
54
55 understand assignments instructions, check their own work, and improve their performances. For
56
57 teacher users, some believed rubrics made their assessment more efficient and objective
58
59
60
61
62
63
64
65

(Campbell, 2005), while some were resistant to rubric use, as they did not have enough time or they thought they did not need a rubric to help with grading (Parkes, 2006).

User Characteristics Associated with Rubric Use

Evaluation tasks involve complex individual judgments and decision-making processes, which are affected by a set of personal characteristics, such as beliefs, expertise, and experience (Kammerer, Bråten, Gerjets, & Strømsø, 2013; Venkatesh, Morris, & Ackerman, 2000). For example, Kammerer et al. (2013) found that people with strong beliefs about the accuracy and reliability of online resources were less likely to reflect on the credibility of resources and exhibited more certainty in their decisions regarding resource evaluation and selection. Brand-Gruwel, Wopereis, and Vermetten (2005) found that when solving information tasks, experts spent more time than novices scanning and judging the quality, relevance, and usefulness of information, and were more likely to activate their prior expertise. Van Vaerenbergh and Thomas (2013), after reviewing the literature, concluded that people's education and length of employment can influence the way they apply rubrics, and thus may lead to different response styles.

Teacher knowledge and expertise. In investigating differences between teachers and non-teachers, we posit that through professional experience and training, teachers develop a unique set of beliefs and knowledge that makes them experts in this task domain compared to non-teachers (Colton & Sparks-Langer, 1993; Livingston & Borko, 1989; Maloch et al., 2003; Shavelson & Stem, 1981).

In a seminal article, Shulman (1986) proposed that effective teachers draw upon a broad range of expertise. This resulted in the distillation of a theoretical framework characterizing what teachers know. This includes disciplinary or content knowledge (CK), teaching or pedagogical

1
2
3
4 knowledge (PK), and their intersection, pedagogical content knowledge (PCK). The latter goes
5
6 beyond content knowledge to include knowledge about how to teach particular subject areas.
7
8 Together, these forms of knowledge are identified as influential factors in how teachers use and
9
10 interpret curricular materials (Remillard, 2005) and how they interpret and perceive design
11
12 elements of rubrics.
13
14

15
16 **Expertise reversal effect.** Research within the cognitive load framework has identified
17
18 the “expertise reversal effect,” showing that instructional tools that benefit novices may not
19
20 necessarily benefit or can even be detrimental to experts (Kalyuga, Ayres, Chandler, & Sweller,
21
22 2003; Kalyuga & Renkl, 2010). For example, Reisslein, Atkinson, Seeling, and Reisslein (2006)
23
24 exposed students to a series of instructional procedures in a computer-based environment and
25
26 found that detailed instructions with examples were beneficial for inexperienced learners, but
27
28 less so for experienced learners. Similarly, Oksa, Kalyuga, and Chandler (2010) found that when
29
30 using the same explanatory scaffolds, people with low prior knowledge outperformed the control
31
32 group in understanding target content, while people with high prior knowledge performed worse
33
34 than the control group.
35
36

37
38 In sum, teachers possess pedagogical knowledge, experience, and expertise that differs
39
40 from non-teachers. We anticipate that these differences will influence how teachers perceive,
41
42 apply, and interpret rubrics, indicators, and their supporting materials when evaluating OER.
43
44
45
46
47
48
49

50 51 **Methods**

52
53 Table 1 summarizes the research questions, data sources, and analyses guiding this study.

54
55 Table 1

56
57 *Research Questions, Data Sources, and Data Analysis Techniques*
58
59
60
61
62
63
64
65

Research question	Data sources	Data analysis
1. How do people apply quality rubrics when evaluating OER? Are there differences between teachers and non-teachers?	<ul style="list-style-type: none"> • Main Survey: participants' ratings of 20 OER using the three rubrics 	<ul style="list-style-type: none"> • Descriptive statistics • Pearson correlation • Chi-squared test • Intra-class correlation coefficient (ICC)
2. How do people perceive the utility of quality rubrics when evaluating OER? Are there differences between teachers and non-teachers?	<ul style="list-style-type: none"> • Reflection Survey: participants' ratings of the utility of the three rubrics • Interviews: participants' perceptions of the utility of the three quality rubrics 	<ul style="list-style-type: none"> • Descriptive statistics • Mann-Whitney U test • Code (<i>In vivo</i> coding, open coding, pattern coding)

Participants and Study Procedures

Twenty-two teachers and 22 non-teachers were recruited for this study using Amazon Mechanical Turk (AMT), a web service for coordinating and paying people to perform online tasks (Paolacci, Chandler, & Ipeirotis, 2010). The 22 teachers were all K-12 teachers; among them, three were math teachers, three were science teachers, two taught all subjects, and the remaining taught non-STEM subjects, including English, history, and dance. Table 2 summarizes teachers' and non-teachers' demographics. Results from a Mann-Whitney test showed a significant difference between teachers and non-teachers on their self-rated familiarity with the target domain for OER (mathematics fractions), confirming that teachers self-reported more relevant prior content knowledge than non-teachers. There were no significant differences on other variables.

To address RQ1, participants used the "main" survey to evaluate 20 OER (described below) using the three selected rubrics and their indicators (also described below). To address RQ2, participants completed the "reflection" survey to rate the utility of the three rubrics and their indicators.

Table 2

Participant Demographics

Variable	Mean	Median	SD	Min	Max
Teacher Participants					
Years of teaching experience	11.67	11.00	7.36	1	25
Comfort level with technology ^a	8.32	9.00	1.91	4	10
Ability to search for online educational resources ^a	8.50	8.50	1.41	6	10
Ability to use online educational resources in teaching, learning, and working ^a	8.32	9.00	1.64	5	10
Familiarity with fractions ^{a b}	7.95	8.00	1.86	4	10
Non-teacher Participants					
Comfort level with technology ^a	8.68	9.00	1.21	6	10
Ability to search for online educational resources ^{a c}	8.62	9.00	1.36	5	10
Ability to use online educational resources in teaching, learning, and working ^{a c}	8.14	9.00	1.65	5	10
Familiarity with fractions ^{a b}	6.50	6.50	2.43	0	10

Note. ^a Possible scores range from 1 to 10, with 1 denoting poor, and 10 denoting excellent.

^b Difference between teachers and non-teachers is significant (Mann-Whitney test; $p < .05$)

^c One non-teacher did not specify his / her ability to search and use online educational resources.

Further, five teachers and five non-teachers were invited to participate in a semi-structured interview examining their perceptions of the three rubrics. These interviewees were selected following the maximum variation principle based on their ratings of the utility of the three rubrics. This enabled an examination of both positive and negative perspectives regarding indicators and rubrics, as well as an identification of common themes (Creswell, 2009).

The teachers (assigned identifiers from T1 to T5) had substantial teaching experience (17, 21, 21, 10, 5 years of experience respectively) and taught a variety of subjects and grade levels (8th grade math and reading intervention, special education, 8th grade math, high school English,

1
2
3
4 and high school science, respectively). The non-teachers (assigned identifiers from N1 to N5)
5
6 reported a variety of professional experiences (e.g., server, middle management, project
7
8 management, self-employed).
9

10 11 **Rubric Selection** 12

13
14 This study examined three rubrics: the Achieve OER rubric (Achieve, 2011), the
15
16 Gateway to Educational Materials rubric (GEM; Fitzgerald et al., 2003), and the Learning Object
17
18 Review Instrument (LORI; Nesbit et al., 2007). These rubrics were selected based on the
19
20 following criteria: they 1) evaluate OER quality instead of relevance or other features, 2) are
21
22 composed of different quality indicators, which 3) are described in a way that can be easily
23
24 understood by users; 4) are suitable for human (and not machine) evaluation; and 5) are
25
26 applicable to multiple domains.
27
28
29

30
31 The Achieve OER and LORI rubrics were developed to help educators identify useful
32
33 educational resources. Both rubrics used Likert scales for users to rate resources. GEM was
34
35 developed to help educators determine if an online resource was of sufficient quality to be
36
37 included into an educational website. The rubric did not include a rating scale, simply an
38
39 indication if the criterion had been met or not. As a result, for this study, we adopted the 5-point
40
41 Likert scale from LORI.
42
43
44

45
46 In addition, several indicators were excluded. From Achieve OER, we excluded two
47
48 indicators (Degree of alignment to standards, Assurance of accessibility) and from LORI, we
49
50 excluded four indicators (Standards compliance, Accessibility, Feedback and adaptation, and
51
52 Interaction usability). The two standards-related items were excluded as many existing OER did
53
54 not specify which standards they were aligned to; the two accessibility-related items were
55
56 excluded, as many existing OER were not designed to accommodate students with disabilities.
57
58
59
60
61
62
63
64
65

The final two indicators were excluded as few OER contained adaptation and navigation elements. Table 3 shows the indicators used and scoring mechanisms for each of the three rubrics.

Table 3

Selected Quality Rubrics

Source	Quality Indicators	Scoring Mechanisms
Achieve OER (Achieve, 2011)	Six quality indicators used in this study: <ol style="list-style-type: none"> 1. Quality of explanation of the subject matter 2. Utility of materials designed to support teaching 3. Quality of assessment 4. Quality of technological interactivity 5. Quality of instructional and practice exercises 6. Opportunities for deeper learning 	<ul style="list-style-type: none"> • Rating scale: 0-3 & N/A (3: superior, 2: strong, 1: limited, 0: very weak/none, N/A: item not applicable) • Scoring guide: detailed scoring guide for all five points on the scale
GEM (Fitzgerald, Lovin, & Branch, 2003)	Six quality indicators used in this study: <ol style="list-style-type: none"> 7. Accuracy 8. Appropriateness 9. Clarity 10. Completeness 11. Motivation 12. Organization 	<ul style="list-style-type: none"> • Rating scale: None provided • Scoring guide: None provided
LORI (Nesbit, Belfer, & Leacock, 2007)	Five quality indicators used in this study: <ol style="list-style-type: none"> 13. Content quality 14. Learning goal alignment 15. Motivation 16. Presentation design 17. Reusability 	<ul style="list-style-type: none"> • Rating scale: 1-5 & N/A (1: poor, 5: excellent, N/A: item not applicable) • Scoring guide: detailed scoring guide for points 1 and 5, and an example for point 3

Materials: OER

The [TeachersPayTeachers.com](https://www.teacherspayteachers.com) (TPT) website served as the source for the OER that participants evaluated. TPT is a popular platform for educators to exchange educational resources, with more than 3.8 million registered users and more than 2 million resources (About

1
2
3
4 TPT, 2016). The TPT website provides access to a variety of OER and can be searched
5
6 according to grade levels, subjects, and resource types. In addition, users of this site can rate
7
8 OER quality on a 0-4 scale. The great variety of OER as well as the available rating information
9
10 facilitated the selection of OER for this study.
11
12
13

14 From this site, OER were selected that met the following criteria: the OER 1) was free,
15
16 and thus aligned with the definition of OER; 2) had at least 3 user ratings; 3) was in the form of
17
18 lesson plans or activities, thus having appropriate granularity; and 4) was relevant to the
19
20 following teaching scenario involving a lesson about fractions:
21
22

23 A teacher plans to teach a lesson about fractions using OER. He/she has collected some
24
25 OER from the Internet, but he/she is not sure which resources have appropriate quality
26
27 and should be used in his/her teaching. Could you please help him/her rate the quality of
28
29 these OER?
30

31 These criteria led to a potential pool of approximately 400 OER. As previous studies
32
33 showed that both rating levels and the number of ratings affect users' selection of resources
34
35 (Abramovich et al., 2013), a stratified random sampling strategy was used based on the ratings
36
37 and the number of ratings to select the final pool of 20 OER.
38
39
40

41 **Data Collection**

42
43 The two surveys ("main" and "reflection") completed by 44 participants and the
44
45 interviews of 10 participants served as the data sources for this study (see Table 1). The first
46
47 "main" survey prompted participants to rate 20 OER using each indicator from the three rubrics
48
49 while using the scale associated with that indicator (see Table 3). All "N/A" responses were
50
51 coded as "-1". The second "reflection" survey prompted participants to rate the utility of each
52
53 indicator on a 1-5 scale, with "1" indicating participants "strongly disagree" that this indicator
54
55
56
57
58
59
60
61
62
63
64
65

was useful in helping them select resources, and “5” indicating participants “strongly agree” that this indicator was useful in helping them select resources.

Note that to prevent an order effect, the order of the three rubrics and their indicators was randomized in the “main” survey. Finally, the survey also included the option “prefer not to respond” for each multi-choice question.

Semi-structured interviews were used to examine individual participants’ perceptions of the quality rubrics. In particular, during the interviews, interviewees were asked about their experiences using rubrics, how they perceived the utility of different rubrics, reasons for their perceptions, and how they would improve these rubrics.

Data Analysis

The 44 participants’ evaluation of 20 OER using 17 quality indicators resulted in 14,960 data points. Among them, 15 responses (0.1%) were “prefer to not respond,” coded as missing data, and thus were excluded using the listwise deletion method.

To address the first question, descriptive statistics provided an overview of participants’ ratings. In addition, Pearson correlation was used to investigate the relationship between teachers’ and non-teachers’ ratings, and chi-squared tests were used to compare the differences between teachers’ and non-teachers’ ratings. A significance level of .05 was used in both analyses.

Then the intra-class correlation coefficient (ICC) was used to examine agreement between users. The ICC statistic was chosen since it is versatile, can calculate the agreement among three or more raters, and can work with both ordinal and interval data (Hallgren, 2012). In particular, a two-way mixed ICC was used, since the 44 participants rated all 20 OER, and these participants were not random representatives of the whole population. The ICC values were

1
2
3
4 examined to see whether different indicators and/or different user groups showed different
5
6 reliability levels.
7

8
9 To address the second research question, teachers' and non-teachers' ratings in the
10
11 reflection survey were used to calculate descriptive statistics (Mean, SD) for each rubric and
12
13 quality indicator. The responses from the two user categories were then compared using the
14
15 Mann-Whitney test (due to non-normal distributions).
16
17

18
19 The analysis of interview data followed six-phases suggested by Braun and Clarke
20
21 (2006). Table 4 lists the key activities involved in each phase with an example showing the
22
23 coding process. Phase 1 begins with an immersion into the interview data. In Phase 2, initial
24
25 codes were generated via *in vivo* coding, in which codes were directly identified using the words
26
27 and phrases from the interview data, and open coding, in which codes were created to describe
28
29 segments of data in the interviews (Saldaña, 2012). In phase 3, pattern coding was used to group
30
31 codes into a smaller number of candidate themes (Saldaña, 2012). Phases 4 and 5 featured a
32
33 cyclical process of continuous refinement of themes. In Phase 6, after all themes were refined
34
35 and finalized, teachers' and non-teachers' data were compared to examine similarities and
36
37 differences on each theme.
38
39
40
41

42
43 To establish the credibility and trustworthiness of the qualitative results, we reconciled
44
45 the quantitative data from surveys and the qualitative data from interviews to yield more accurate
46
47 and valid findings. We also asked an external researcher to audit the analysis, interpretations, and
48
49 conclusions of the study.
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Table 4

The Six Phases for Analyzing Interview Data (Braun & Clarke, 2006)

Phase	Key Activities	Examples
1. Familiarizing with the data	<ul style="list-style-type: none"> • Transcribe interviews • Read through and familiarize with transcript 	<p>Data extracts:</p> <p>“I think the rubrics really helped when I was looking at stuff because it forced me to look at different qualities... It made me think about each aspect of it [OER] separately as opposed to just overall good or bad.”</p> <p>“[Rubrics] definitely can help because it structures your thinking, you know, rather than just saying what do you think? Or being totally subjective, it forces you to consider one aspect of whatever it is you’re ranking, and then structure that, you know, is it excellent, good, fair, very good, and so forth... [Rubrics] help different people or more than one person that’s doing the ranking. It structures the thinking so that they’re all doing it the same way, and you have some consistency between people.”</p>
2. Generating initial codes	<ul style="list-style-type: none"> • Identify interesting words and phrases from transcripts • Create initial codes 	<p>Initial codes:</p> <ul style="list-style-type: none"> • Looking at different aspects of OER quality • Avoiding subjectiveness • Distinguishing good resources from inferior ones • Achieving reliability
3. Searching for themes	<ul style="list-style-type: none"> • Review initial codes • Group related codes to generate candidate themes 	<p>Candidate themes:</p> <ul style="list-style-type: none"> • Rubrics focus users’ evaluation • Rubrics improve objectivity and reliability
4. Reviewing themes	<ul style="list-style-type: none"> • Review candidate themes in relation to the coded data extracted • Refine themes 	
5. Defining and naming themes	<ul style="list-style-type: none"> • Identify themes most related to the research questions • Define and rename relevant themes 	<p>Renamed theme:</p> <ul style="list-style-type: none"> • Benefits of rubrics
6. Producing the report	<ul style="list-style-type: none"> • Examine similarities and differences between teachers and non-teachers • Write up findings 	

Results

Research Question 1: Application of Rubrics

Overview and comparisons of participants’ ratings

Participants' ratings. Figure 1 depicts three bar charts, with each one displaying the frequency distribution of participants' ratings of OER for each of the three rubrics. For each chart, the legends show the different rating scales, the x-axis represents the different indicators included in the rubric, and the y-axis represents the frequency of each rating.

As shown in Figure 1, participants more frequently picked “-1”, which denotes “N/A”, in the Achieve OER rubric, as compared to the other two rubrics. This was especially evident for indicators 1 (explanation), 3 (assessment), 4 (technological interactivity), and 6 (opportunities for deeper learning), suggesting that participants find these indicators less applicable.

Figure 1 also shows different response tendencies for different indicators. For example, participants tended to pick more middle ratings for indicators 1 (explanation), 2 (utility of materials designed to support teaching), 5 (instructional and practice exercises), and 6 (opportunities for deeper learning), tended to pick more maximum ratings on indicators 7 (accuracy) and 8 (appropriateness), and tended to equally pick all five scores on indicators 10 (completeness) and 14 (learning goal alignment).

Figure 1. Frequency distributions of participants' ratings for the three rubrics

Correlations between teachers' and non-teachers' average ratings. Pearson correlations showed that teachers' average ratings of 20 OER were positively and significantly correlated with non-teachers' on all 17 indicators, with magnitudes from .51 to .97 (all $p < .05$; see Figure 2). This suggests similarities in how teachers and non-teachers apply rubrics, despite the fact that teachers reported greater education-related expertise and more content knowledge as shown in Table 2. Indicator 4 (technological interactivity) had the lowest correlation, the indicator that also had the highest number of "N/A" selections.

Figure 2. Correlations between teachers' and non-teachers' average ratings (* indicates $p < .01$)

Comparing teachers' and non-teachers' ratings. To compare the frequency distribution of ratings between teachers and non-teachers, three chi-squared tests (one for each rubric) were conducted, with the values on the rating scale as the column and user category (teachers, non-teachers) as the row. Tables 5-7 present the results from the three chi-squared tests, which all revealed a statistically significant association between participants' ratings and their user category; Achieve OER: $\chi^2(4, N = 5273) = 47.23, p < .01$; GEM: $\chi^2(5, N = 5275) = 32.12, p$

$< .01$; LORI: $\chi^2 (5, N = 4397) = 41.77, p < .01$, respectively.

Table 5

Cross Tabulation of Ratings and User Categories for the Achieve OER Rubric

User category	Rating					χ^2
	-1	0	1	2	3	
Teachers	539 (5.6)	195 (1.5)	593 (-4.3)	809 (-2.4)	501 (1.3)	47.23**
Non-teachers	384 (-5.6)	168 (-1.5)	729 (4.3)	890 (2.4)	465 (-1.3)	

Note. Adjusted standardized residuals appear in parentheses below group frequencies.

** $p < .01$.

Table 6

Cross Tabulation of Ratings and User Categories for the GEM Rubric

User category	Rating						χ^2
	-1	1	2	3	4	5	
Teachers	39 (-1.0)	271 (1.1)	413 (-.8)	540 (-3.9)	598 (-1.0)	778 (4.8)	32.12**
Non-teachers	48 (1.0)	247 (-1.1)	435 (.8)	657 (3.9)	627 (1.0)	622 (-4.8)	

Note. Adjusted standardized residuals appear in parentheses below group frequencies.

** $p < .01$.

Table 7

Cross Tabulation of Ratings and User Categories for the LORI Rubric

User category	Rating						χ^2
	-1	1	2	3	4	5	
Teachers	39 (2.6)	246 (2.8)	377 (-2.9)	460 (-3.0)	484 (-1.2)	593 (4.1)	41.77**
Non-teachers	19 (-2.6)	191 (-2.8)	453 (2.9)	543 (3.0)	517 (1.2)	475 (-4.1)	

Note. Adjusted standardized residuals appear in parentheses below group frequencies.

** $p < .01$.

The post-hoc tests used pairwise comparisons with Bonferroni corrections of the p -values (Gardner, 2001). The post-hoc tests showed that teachers gave significantly more maximum ratings, while non-teachers gave significantly more neutral ratings.

Inter-rater reliability of rubrics

Figure 3 depicts ICC values (a measure of interrater agreement) for teachers and non-teachers for the 17 quality indicators. Note that the ICC ranges from 0 to 1 with “1” indicating perfect agreement and “0” no agreement.

Figure 3. Teachers' and non-teachers' interrater reliability for indicators

Overall, teachers' and non-teachers' ratings showed good reliability, with 16 of the 17 indicators having ICC values higher than .6, a commonly accepted threshold indicating satisfactory reliability (Shoukri, Asyali, & Donner, 2004), and only indicator 17 (reusability)

1
2
3
4 resulted in ICC values less than .6. At the rubric level, indicator 5 (quality of instructional and
5
6 practice exercises) produced the highest ICC among the Achieve OER indicators, indicator 11
7
8 (motivation) produced the highest ICC among the GEM indicators, and indicator 15 (motivation)
9
10 produced the highest ICC among the LORI indicators. Interestingly, all of these three indicators
11
12 were related to pedagogical value.
13
14

15
16 Figure 3 also shows in general non-teachers' agreement levels are higher than teachers'.
17
18 In particular, among the 17 quality indicators, non-teachers' agreement levels are higher than
19
20 teachers for 16 indicators. The only exception is indicator 16 (presentation design) from LORI.
21
22

23
24 In summary, in terms of the rubric application, both teachers and non-teachers picked the
25
26 "N/A" choice many times for indicators 1 (explanation of the subject matter), 3 (assessment), 4
27
28 (technological interactivity), and 6 (opportunities for deeper learning), and both groups tended to
29
30 rate OER more highly when using indicator 7 (accuracy) and indicator 8 (appropriateness). In
31
32 addition, ICC values for most indicators were above the widely accepted threshold value for
33
34 satisfactory reliability.
35
36

37
38 Their application differed in that teachers tended to pick more maximum ratings while
39
40 non-teachers tended to pick more neutral ratings. This phenomenon has been observed in prior
41
42 studies of teachers' evaluations of OER (Abramovich et al., 2013; Leary et al., 2011). Moreover,
43
44 despite these generally higher ratings by teachers, non-teachers' ratings had higher ICC values
45
46 than teachers on all but one indicator.
47
48
49

50 51 **Research Question 2: Perceptions of Rubrics**

52 53 *Ratings of the utility of rubrics*

54
55 Table 8 reports participants' ratings (Mean, *SD*) for the utility of each indicator from the
56
57 three rubrics on a 1-5 rating scale. Overall, participants gave similar ratings for the utility of the
58
59
60
61
62
63
64
65

three rubrics, slightly favoring the GEM rubric the most and the Achieve OER rubric the least.

Table 8

Perceived Utility of Indicators from the Three Rubrics

	Teachers		Non- teachers		All participants	
	Mean	SD	Mean	SD	Mean	SD
Achieve OER rubric						
Indicator 1: Quality of explanation of the subject matter	4.36	.49	4.45	.67	4.41	.58
Indicator 2: Utility of materials designed to support teaching	4.36	.58	4.18	.85	4.27	.73
Indicator 3: Quality of assessment	3.82	.91	3.82	.80	3.82	.84
Indicator 4: Quality of technological interactivity	3.86	1.04	3.64	1.00	3.75	1.01
Indicator 5: Quality of instructional and practice exercises	4.14	.71	4.36	.79	4.25	.75
Indicator 6: Opportunities for deeper learning	4.27	.88	4.32	.78	4.30	.82
Overall	4.14	.35	4.13	.53	4.13	.44
GEM Rubric						
Indicator 7: Accuracy	4.18	.73	4.32	.57	4.25	.65
Indicator 8: Appropriateness	4.41	.73	4.45	.74	4.43	.73
Indicator 9: Clarity	4.50	.67	4.55	.74	4.52	.70
Indicator 10: Completeness	4.27	.63	4.23	.81	4.25	.72
Indicator 11: Motivation	4.32	.95	4.64	.73	4.48	.85
Indicator 12: Organization	4.05	.72	4.27	.98	4.16	.86
Overall	4.29	.36	4.41	.50	4.35	.44
LORI Rubric						
Indicator 13: Content quality	4.23	.75	4.45	.60	4.34	.68
Indicator 14: Learning goal alignment	4.55	.74	4.36	.66	4.45	.70
Indicator 15: Motivation	4.23	1.02	4.55	.74	4.39	.89
Indicator 16: Presentation design	3.77	1.11	4.09	.97	3.93	1.04
Indicator 17: Reusability	3.91	1.34	3.50	1.01	3.70	1.19
Overall	4.14	.65	4.19	.51	4.16	.59

Note. Participants' perceived utility of each indicator was on a five-point Likert scale (1 – “strongly disagree that the indicator is useful” to 5 – “strongly agree that the indicator is useful”)

At the indicator level, participants reported positive ratings for the utility of all 17 quality indicators. Only four indicators averaged lower than 4 points on the 5-point rating scale of

indicator utility: indicator 3 (assessment), indicator 4 (technological interactivity), indicator 16 (presentation design), and indicator 17 (reusability). Results from Mann-Whitney tests showed no significant differences between teachers' and non-teachers' ratings of the utility of these indicators.

Participant Interviews

The analysis of interview data, following the six-phase process described above, distilled three themes: benefits of rubrics, features of a good rubric, and opinions about indicators (see Table 9). The following paragraphs describe the three themes and compare teachers' and non-teachers' perceptions in terms of each theme.

Table 9

Themes Identified and Example Codes

Themes Identified	Example Codes
Benefits of rubrics	<ul style="list-style-type: none"> • Distinguishing good resources from inferior ones • Simplifying users' work • Achieving reliability
Features of a good rubric	<ul style="list-style-type: none"> • Indicators selected • Descriptions/explanations provided • Flexibility allowed
Opinions about indicators	<ul style="list-style-type: none"> • Relevance • Applicability • Importance

Benefits of rubrics. All 10 interviewees affirmed the benefits of the rubrics. In particular, T1 and N1 believed rubrics could help users distinguish good resources from inferior ones. T4 and N5 stated that the use of rubrics could lead to a more comprehensive evaluation of OER, as the rubrics provided a set of quality indicators that allowed users to focus on the different aspects of OER. T2, T5, N2, and N5 thought that the use of rubrics could help simplify users' work, as

1
2
3
4 they provided a set of criteria to evaluate OER. Finally, N3 emphasized that rubrics could help
5
6 users provide more objective evaluation, which can make ratings more consistent:
7
8

9 So yeah, it's definitely useful because it structures your thinking when you're evaluating
10 something. Very useful, and also to help different people or more than one person that's
11 doing the ranking it structures the thinking so that they're all doing it the same way and
12 you have some consistency between people.
13
14

15
16 *Features of a good rubric.* Teachers and non-teachers stated that a good rubric is one that
17
18 contains useful indicators. For example, a teacher (T4) expressed a preference for the LORI
19
20 rubric, as its indicators focused on student learning and were comprehensive:
21
22

23 Oh, you know I think that I liked that one [LORI rubric] was the fact that it really looked
24 at how students learned. It looked at motivation, how the information is visually
25 designed to enhance learning, and different backgrounds. That one rubric focused more
26 on reusability in the classroom, and different learners.
27
28

29 Teachers and non-teachers articulated different views regarding the rubrics'
30
31 accompanying guidelines and scoring guides. Three non-teachers (N2, N3, and N4) indicated
32
33 they wanted rubrics with detailed guidelines and explanations. For example, N4 stated that a
34
35 rubric would be useless when the guidelines for rubric application were not well explained:
36
37
38

39 Maybe if it's not a very well explained rubric, that could probably hinder you. Like if
40 you're supposed to go based off of this but the guideline is kind of bland or it doesn't
41 explain something very well, that would probably hinder you a little bit.
42
43

44 In contrast, a teacher, T3, stated that if a rubric used too specific language and a too strict
45
46 rating scale, it would deprive users of the freedom to make their own decisions or to adjust their
47
48 ratings:
49
50

51 The hindering thing, I think sometimes we get so specific in a rubric stating these are the
52 things that qualify for superior, these are the things that qualify as strong, that sometime
53 we'll use language of the rubric to describe a lesson and maybe it wouldn't be language
54 we would use. For example, there were lessons in the tasks that I evaluated according to
55 the language of the rubric and the words strong, for example, would come up, whereas
56 overall I didn't think it was strong at all. I didn't like the activity, or didn't think it was
57 worthwhile. But according to the framework I had to go on that's what it would have to
58 be evaluated as.
59
60
61
62
63
64
65

1
2
3
4 Similarly, three teachers (T1, T2, and T5) claimed they wanted more flexible rubrics that
5
6 gave them more freedom to revise or adjust based on their students' needs or school context. In
7
8 particular, T1 specified that a good rubric should be modifiable to satisfy the needs of different
9
10 schools and different students:
11
12

13
14 Well I think that the expectation for a good rubric is that it's sort of customized...
15 Schools are in all types of different areas and locations. I think a rubric needs to make
16 sure that you're checking to see if differences in your population are being met.
17
18

19
20 *Opinions about indicators.* Teachers and non-teachers consistently highlighted the
21
22 importance of indicators related to students' learning. For example, six interviewees (T1, T3, T4,
23
24 T5, N2, and N3) spoke highly of the "accuracy" indicator, as they believed that errors, typos, and
25
26 inaccuracies in OER could interfere with learning.
27
28

29
30 Interviewees expressed concern about indicators they found less applicable. For example,
31
32 T3 and N5 considered the "quality of technological interactivity" indicator less applicable, as
33
34 they noted that many OER did not have interactive components. N5 thought it would be hard to
35
36 apply the "quality of assessment" indicator unless teachers actually used the OER in classroom.
37
38 This statement helped explain why many participants picked "none applicable" when using these
39
40 two indicators to evaluate the OER.
41
42

43
44 Additionally, interviewees indicated that they found the "reusability" indicator difficult to
45
46 apply, as they were confused about its definition. For example, T2 thought that "reusability"
47
48 pertained to whether users could freely use other people's OER, while T3 thought that
49
50 "reusability" judged whether OER could be used with students with varying backgrounds. These
51
52 varying interpretations offer an explanation for why this indicator had the lowest ICC value.
53
54

55
56 When talking about their favorite indicators, non-teachers (N2, N3, N4, and N5)
57
58 emphasized the importance of the "motivation" indicator. For example, N3 believed that students
59
60
61
62
63
64
65

1
2
3
4 need to be motivated to learn:
5

6
7 Well, like I said before, motivation would be one of the things that I would have in there
8 because it's important for the students to be motivated to actually do the lesson. So that
9 would be important.
10

11 In contrast, four out of the five teachers (T1, T2, T3, and T4) emphasized the importance
12 of the "alignment with standards" indicator. As stated by T3:
13
14

15
16 I would want a category to rate how aligned the material or the book was to the standards
17 that I needed to be using. Such as eighth grade common core or whatever state standards.
18

19 In summary, analyses of interviews indicated some similarities between teachers and
20 non-teachers regarding the rubrics. In particular, both teachers and non-teachers acknowledged
21 the benefits of rubrics, identified some indicators as more useful, and recognized some indicators
22 as less applicable. On the other hand, the interviews also revealed differences between teachers'
23 and non-teachers' perceptions of the rubrics. In particular, the teachers asked for more flexible
24 rubrics that can be revised or customized to meet their own needs, while the non-teachers asked
25 for rubrics with detailed guidelines and explanations. At the indicator level, teachers emphasized
26 the importance of "alignment with standards," while non-teachers emphasized the importance of
27 "motivation."
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43

44 Discussion

45
46 Findings from this study address the call for studies examining the effectiveness and
47 usefulness of quality rubrics and their use by different audiences. In particular, this study
48 examined how 44 participants applied and perceived three OER quality rubrics. In terms of
49 rubric application, participants found some quality indicators difficult to use, and displayed
50 different response styles on different indicators. The reliability analysis showed that 16 out of the
51 17 indicators led to good interrater reliability. When comparing teachers' and non-teachers'
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 application, teachers' overall ratings of OER were higher than non-teachers', but non-teachers'
5
6 ratings had higher ICC values than teachers.
7

8
9 In terms of rubric perception, participants gave positive ratings to the utility of all 17
10
11 quality indicators, and there were no significant differences between teachers' and non-teachers'
12
13 ratings. However, the interviews revealed that teachers preferred more flexible rubrics, while
14
15 non-teachers preferred rubrics with detailed guidelines. In addition, teachers stressed the
16
17 importance of the "alignment with standards" indicator, which is consistent with findings from a
18
19 recent study indicating that teachers are very standards-driven when searching for online
20
21 instructional materials (Opfer, Kaufman, & Thompson, 2017). In contrast, non-teachers
22
23 highlighted the importance of the "motivation" indicator.
24
25
26
27

28
29 A closer look at the findings revealed agreement between how participants applied
30
31 quality rubrics to evaluate OER and their subsequent stated perception of their utility. For
32
33 example, participants frequently chose "N/A" for indicators 3 (assessment) and 4 (technological
34
35 interactivity) in OER evaluation, while also reporting that these two indicators were less useful.
36
37 Similarly, participants achieved the lowest ICC when using indicator 17 (reusability), while also
38
39 giving it a low utility rating.
40
41
42

43
44 Second, the findings revealed similarities between teachers' and non-teachers'
45
46 application and perception of the rubrics. For example, teachers' average ratings of 20 OER
47
48 when using the 17 indicators were significantly correlated with those of non-teachers, and
49
50 teachers' ratings of the utility of quality indicators were similar to those of non-teachers. These
51
52 similarities suggest that the rubrics were applicable by both those with pedagogical knowledge
53
54 (teachers) as well as those without (non-teachers).
55
56
57

58 Third, the findings revealed differences between teachers and non-teachers in their
59
60
61
62
63
64
65

1
2
3
4 perception and application of rubrics. In terms of rubric perception, teachers' preference for
5
6 flexible rubrics and the "alignment with standards" indicator can be understood in the context of
7
8 teachers' expertise. Through experience and education, teachers have developed higher levels of
9
10 PK and PCK (Shulman, 1986). This allowed them to better understand which OER better
11
12 facilitate learning, suit certain students, and integrate into the curriculum. Thus, teachers perhaps
13
14 wanted more flexible rubrics so that they could adjust the ratings of an OER based on students'
15
16 different backgrounds or educational contexts (Davis & Krajcik, 2005). Similarly, teachers also
17
18 understood the importance of educational standards in the modern classroom and the importance
19
20 of aligning teaching materials.
21
22
23

24
25
26 Regarding rubric application, non-teachers' ratings generally achieved higher reliability
27
28 (ICC values) than teachers. One possible explanation for this is the expertise reversal effect.
29
30 Specifically, teachers (in their role as experts) possess both PK and PCK regarding what
31
32 constitutes a useful resource for their teaching context. They may also have developed their own
33
34 techniques for evaluation (Parkes, 2006). Then, when applying the rubrics, teachers might have
35
36 encountered redundant, extraneous, or even conflicting information within the rubric. This, in
37
38 turn, may have caused extensive cognitive load and thus did not benefit or even hindered rubric
39
40 application (Kalyuga et al., 2003; Kalyuga & Renkl, 2010; Reisslein et al., 2006). In contrast,
41
42 these rubrics may have provided important scaffolding for non-teachers, for example essential
43
44 information about OER use and application, which supported non-teachers in completing the
45
46 evaluation tasks (Kalyuga et al., 2003). As further evidence for the expertise reversal effect, non-
47
48 teacher interviewees expressed a preference for more detailed rubrics, while teachers, with
49
50 potentially higher pedagogical knowledge, preferred rubrics that were more flexible and
51
52 customizable to their context.
53
54
55
56
57
58
59
60
61
62
63
64
65

Limitations

This study had several limitations. First, use of the AMT website as the platform to recruit participants allows generalization of the findings to people who are familiar with information technologies, as they are more likely to search for and use OER. However, the participants of this study were not randomly selected, but instead on a “first come first serve” basis from the AMT website. Thus, these participants cannot be regarded as random representatives of the population, which may limit the generalizability of this study.

Second, there were potential limitations in the data collection process. When collecting survey data, AMT and Qualtrics allowed flexibility in how participants completed the evaluation task. For example, Qualtrics logs showed that some participants finished the task in several hours in one continuous session, while other participants finished the task over several days and multiple sessions; some finished the evaluation chronologically, while some did not. These differences could affect participants’ ratings, and thus the results of the study.

Finally, there were flaws in the data analysis process. In particular, all three rubrics in this study contained “N/A” as one point in their rating scales. Here, a participant’s selection of “N/A” is not the same as missing data, where a participant could choose “prefer to not respond.” Thus, in our analysis, we coded “N/A” as “-1” and included this value in the calculation of ICC, which may affect results. Another limitation was the violation of the independence assumption in some analyses. In this study, each participant evaluated 20 OER, and the ratings from each participant may be related.

Conclusion

Findings from this study reveal how people apply quality rubrics for evaluating OER and

1
2
3
4 how they perceive their utility. Overall, study results revealed that most indicators showed good
5
6 reliability, the foundation of an effective rubric. Study results also identified specific indicators
7
8 that participants perceived as more useful, including “Clarity” (GEM), “Quality of explanations”
9
10 (Achieve), “Appropriateness” (GEM), “Motivation” (GEM) and “Learning goal alignment”
11
12 (LORI).
13
14

15
16 Study results also revealed a few indicators that were less applicable or produced low
17
18 reliability. For example, participants frequently selected “N/A” when using indicator 4 (quality
19
20 of technological interactivity). During interviews, participants noted that many OER did not
21
22 include interactive components, which made this indicator less applicable. Additionally,
23
24 participants achieved the lowest reliability values on indicator 17 (reusability). Interviews
25
26 revealed that participants were confused about the definition of this indicator.
27
28
29
30

31 Findings from this study provide suggestions for rubric developers by identifying
32
33 indicators with low reliability or low applicability. For items with low reliability, developers can
34
35 choose to provide more clear definitions or provide detailed guidance on using these items. For
36
37 items with low applicability, developers can choose to remove or replace them with more
38
39 applicable items.
40
41
42

43 Finally, results from this study revealed that one audience’s application and perception of
44
45 rubrics may not simply generalize to another. In particular, our results showed differences
46
47 between teachers and non-teachers in terms of their application and perception of rubrics. Thus,
48
49 developers need to consider the target audience, especially with regards to prior knowledge and
50
51 expertise, when developing rubrics for evaluation OER quality. In particular, when creating
52
53 rubrics for teachers, developers should consider that teachers may already possess rich PK and
54
55 PCK, and thus rubrics that contain only brief instructions and allow more freedom in application
56
57
58
59
60
61
62
63
64
65

1
2
3
4 may sometimes be more helpful.
5
6
7
8

9 **Acknowledgements**

10
11 This research is partially supported by Utah State University. Portions of this research
12 were previously presented at the American Educational Research Association Annual Meeting
13
14 (AERA 2016) in Washington, DC. We thank Drs. Anne Diekema and Andy Walker for their
15
16
17
18
19 valuable input.
20
21
22

23 **References**

- 24
25
26 About TPT. (2016). Retrieved from <https://www.teacherspayteachers.com/About-Us>
27
28 Abramovich, S., & Schunn, C. (2012). Studying teacher selection of resources in an ultra-large
29 scale interactive system: Does metadata guide the way? *Computers & Education*, 58(1),
30 551-559.
31
32 Abramovich, S., Schunn, C. D., & Correnti, R. J. (2013). The role of evaluative metadata in an
33 online teacher resource exchange. *Educational Technology Research and Development*,
34 61(6), 863-883.
35
36 Achieve (2011). *Rubrics for evaluating open education resource (OER) objects*. Washington,
37 DC. Retrieved from <http://www.achieve.org/publications/achieve-oer-rubrics>
38
39 Andrade, H., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical*
40 *Assessment, Research & Evaluation*, 10 (5), 1–11.
41
42 Atkins, D. E., Brown, J. S., & Hammond, A. L. (2007). *A review of the open educational*
43 *resources (OER) movement: Achievements, challenges, and new opportunities*. Retrieved
44 from: <http://www.hewlett.org/uploads/files/ReviewoftheOERMovement.pdf>
45
46 Bolton, F. C. (2006). Rubrics and adult learners: Andragogy and assessment. *Assessment Update*,
47 18(3), 5-6.
48
49 Brand-Gruwel, S., Wopereis, I., & Vermetten, Y. (2005). Information problem solving by
50 experts and novices: Analysis of a complex cognitive skill. *Computers in Human*
51 *Behavior*, 21(3), 487-508.
52
53 Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in*
54 *Psychology*, 3(2), 77-101. Retrieved from <http://dx.doi.org/10.1191/1478088706qp063oa>
55
56
57
58
59
60
61
62
63
64
65

- Bundsgaard, J., & Hansen, T. I. (2011). Evaluation of learning materials: A holistic framework. *Journal of Learning Design*, 4(4), 31-44.
- Campbell, A. (2005). Application of ICT and rubrics to the assessment process where professional judgement is involved: the features of an e - marking tool. *Assessment & Evaluation in Higher Education*, 30(5), 529-537.
- Chismar, W. G., & Wiley-Patton, S. (2003). Does the extended technology acceptance model apply to physicians? In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences* (pp. 8-15). Big Island, HI: Institute of Electrical and Electronics Engineers (IEEE).
- Chong, A., & Romkey, L. (2017). Testing Inter-Rater Reliability in Rubrics for Large Scale Undergraduate Independent Projects. In *Proceedings of the Canadian Engineering Education Association*, Halifax, Nova Scotia.
- Colton, A. B., & Sparks-Langer, G. M. (1993). A conceptual framework to guide the development of teacher reflection and decision making. *Journal of Teacher Education*, 44(1), 45-54.
- Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: Sage.
- Custard, M., & Sumner, T. (2005). Using machine learning to support quality judgments. *D-Lib Magazine*, 11(10). Retrieved from <http://www.dlib.org/dlib/october05/custard/10custard.html>
- Davis, E. A., & Krajcik, J. S. (2005). Designing educative curriculum materials to promote teacher learning. *Educational Researcher*, 34(3), 3-14.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319-340.
- Duncan, A., (2012). *Arne Duncan introduces the Why Open Education Matters Video Competition*. Retrieved from: <https://www.youtube.com/watch?v=8SdrhGrcvsk&feature=youtu.be>
- East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing*, 14(2), 88-115.
- Fitzgerald, M., Lovin, V., & Branch, R. M. (2003). The gateway to educational materials: An evaluation of an online resource for teachers and an exploration of user behavior. *Journal of Technology & Teacher Education*, 11(1), 21-51.
- Gardner, R. C. (2011). *Psychological statistics using SPSS for Windows*. Upper Saddle River, NJ: Prentice Hall.

- Gligora Marković, M., Kliček, B., & Plantak Vukovac, D. (2014). The effects of multimedia learning materials quality on knowledge acquisition. In *Proceedings of 23rd International Conference on Information Systems Development* (pp. 140–149). Croatian: Varaždin.
- Hallgren, K. A. (2012). Computing interrater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), 23-24.
- Harris, J., Grandgenett, N., & Hofer, M. J. (2010). Testing a TPACK-based technology integration assessment rubric. *Society for Information Technology and Teacher Education*.
- Haughey, M., & Muirhead, B. (2005). Evaluating learning objects for schools. *E-Journal of Instructional Sciences & Technology*, 8(1). Retrieved from http://ascilite.org.au/ajet/e-jist/docs/vol8_no1/fullpapers/eval_learnobjects_school.htm
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130-144.
- Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, 38(1), 23-31.
- Kalyuga, S., & Renkl, A. (2010). Expertise reversal effect and its instructional implications: Introduction to the special issue. *Instructional Science*, 38, 209-215.
- Kammerer, Y., Bråten, I., Gerjets, P., & Strømsø, H. I. (2013). The role of Internet-specific epistemic beliefs in laypersons' source evaluations and decisions during Web search on a medical issue. *Computers in Human Behavior*, 29, 1193-1203.
- King, J. (2016). *U.S. Department of Education Recognizes 14 States and 40 Districts Committing to #GoOpen with Educational Resources*. Retrieved from: <http://www.ed.gov/news/press-releases/us-department-education-recognizes-13-states-and-40-districts-committing-goopen-educational-resources>
- Kurilovas, E., Birenienė, V., & Serikovienė, S. (2011). Methodology for evaluating quality and reusability of learning objects. *Electronic Journal of e-Learning*, 9(1), 39-51.
- Leary, H., Recker, M., Walker, A., & Recker, M., Wetzler, P., Sumner, T., & Martin, J. (2011). Automating Open Educational Resources Assessments: A Machine Learning Generalization Study. In *Proceedings of the Joint Conference on Digital Libraries* (pp. 283-286), New York: ACM
- Legris, P., Ingham, J., & Collerette, P. (2003). Why do people use information technology? A critical review of the technology acceptance model. *Information & management*, 40(3), 191-204.
- Livingston, C., & Borko, H. (1989). Expert-novice differences in teaching: A cognitive analysis and implications for teacher education. *Journal of Teacher Education*, 40(4), 36-42.

- 1
2
3
4 Maloch, B., Flint, A. S., Eldridge, D., Harmon, J., Loven, R., Fine, J. C., ... Martinez, M. (2003).
5 Understandings, beliefs, and reported decision making of first-year teachers from
6 different reading teacher preparation programs. *The Elementary School Journal*, 103,
7 431-457.
8
9
10 McMartin, F., McKenna, A., & Youssefi, K. (2000). Scenario assignments as assessment tools
11 for undergraduate engineering education. *IEEE Transactions on Education*, 43(2), 111-
12 119.
13
14
15 Mooi, E., & Sarstedt, M. (2011). *A concise guide to market research: The process, data, and*
16 *methods using IBM SPSS statistics*. Berlin/Heidelberg, Germany: Springer.
17
18 Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability.
19 *Practical Assessment, Research & Evaluation*, 7(10), 1-11.
20
21
22 Murphy, A. (2013). Open educational practices in higher education: institutional adoption and
23 challenges. *Distance Education*, 34(2), 201-217.
24
25 Nesbit, J., Belfer, K., & Leacock, T. (2007). *Learning Object Review Instrument (LORI), Version*
26 *1.5*. Retrieved from <http://www.transplantedgoose.net/gradstudies/educ892/LORI1.5.pdf>
27
28
29 Newell, J. A., Dahm, K. D., & Newell, H. L. (2002). Rubric development and inter-rater
30 reliability issues in assessing learning outcomes. *Chemical Engineering Education*, 36(3),
31 212-215.
32
33
34 Oakleaf, M. (2009). Using rubrics to assess information literacy: An examination of
35 methodology and interrater reliability. *Journal of the American Society for Information*
36 *Science and Technology*, 60(5), 969-983.
37
38
39 Oksa, A., Kalyuga, S., & Chandler, P. (2010). Expertise reversal effect in using explanatory
40 notes for readers of Shakespearean text. *Instructional Science*, 38, 217-236.
41
42 Opfer, D., Kaufman, J., & Thompson, L. (2017) Implementation of K–12 State Standards for
43 Mathematics and English Language Arts and Literacy: Findings from the American
44 Teacher Panel. Santa Monica, CA: RAND Corporation, 2016. Retrieved from
45 https://www.rand.org/pubs/research_reports/RR1529-1.html.
46
47
48 Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon
49 mechanical turk. *Judgment & Decision Making*, 5, 411-419.
50
51
52 Parkes, K. A. (2006). *The effect of performance rubrics on college-level applied studio grading*
53 *(Order No. 3277079)*. Available from ProQuest Dissertations & Theses Global.
54 (305310537). Retrieved from
55 <http://ezproxy.lib.utah.edu/docview/305310537?accountid=14677>
56
57
58 Porcello, D., & Hsi, S. (2013). Crowdsourcing and curating online education resources. *Science*,
59 341, 240-241. Retrieved from <http://www.sciencemag.org/content/341/6143/240.full>
60
61
62
63
64
65

- Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, 35(4), 435-448.
- Reisslein, J., Atkinson, R. K., Seeling, P., & Reisslein, M. (2006). Encountering the expertise reversal effect with a computer-based environment on electrical circuit analysis. *Learning and Instruction*, 16(2), 92-103.
- Remillard, J. T. (2005). Examining key concepts in research on teachers' use of mathematics curricula. *Review of Educational Research*, 75(2), 211-246.
- Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15(1), 18-39.
- Richter, T., & McPherson, M. (2012). Open educational resources: Education for the world?. *Distance Education*, 33(2), 201-219.
- Saldaña, J. (2012). *The coding manual for qualitative researchers*. Thousand Oaks, CA: Sage.
- Shavelson, R. J., & Stem, P. (1981). Research on teachers' pedagogical thoughts, judgments, decisions, and behavior. *Review of Educational Research*, 51, 455-498.
- Shoukri, M. M., Asyali, M. H., & Donner, A. (2004). Sample size requirements for the design of reliability study: review and new results. *Statistical Methods in Medical Research*, 13, 251-271.
- Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2), 4-14.
- Stellmack, M. A., Konheim-Kalkstein, Y. L., Manor, J. E., Massey, A. R., & Schmitz, J. A. P. (2009). An assessment of reliability and validity of a rubric for grading APA-style introductions. *Teaching of Psychology*, 36(2), 102-107.
- Sumner, T., Khoo, M., Recker, M., & Marlino, M. (2003). Understanding Educator Perceptions of "Quality" in Digital Libraries. In Proceedings of the Joint Conference on Digital Libraries (pp. 269-279). New York: ACM.
- United Nations Educational, Scientific and Cultural Organization (UNESCO). (2016). *Open Educational Resources*. Retrieved from <http://www.unesco.org/new/en/communication-and-information/access-to-knowledge/open-educational-resources/>
- Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, 25(2), 195-217.
- Vargo, J., Nesbit, J. C., Belfer, K., & Archambault, A. (2003). Learning object evaluation: computer-mediated collaboration and inter-rater reliability. *International Journal of Computers and Applications*, 25(3), 198-205.

1
2
3
4 Venkatesh, V., Morris, M. G., & Ackerman, P. L. (2000). A longitudinal field investigation of
5 gender differences in individual technology adoption decision-making processes.
6 *Organizational Behavior & Human Decision Processes*, 83(1), 33-60.

7
8
9 Wetzler, P., Bethard, S., Leary, H., Butcher, K., Zhao, D., Martin, J. S., & Sumner, T. (2013).
10 Characterizing and predicting the multi-faceted nature of quality in educational web
11 resources. *Transactions on Interactive Intelligent Systems*, 3(3), 1-25.

12
13
14 Yuan, M., & Recker, M. (2015). Not all rubrics are equal: A review of rubrics for evaluating the
15 quality of Open Educational Resources. *International Review of Research in Open &*
16 *Distance Learning*, 16(5). Retrieved from
17 <http://www.irrodl.org/index.php/irrodl/article/view/2389>
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65