

Hanna Eide Solstad

A comparison of manual and automated quality assessment of Open Educational Resources and their reliability

Master's thesis in Master of Technology in Computer Science

Supervisor: Sofia Papavlasopoulou

June 2022



Norwegian University of
Science and Technology

Hanna Eide Solstad

A comparison of manual and automated quality assessment of Open Educational Resources and their reliability

Master's thesis in Master of Technology in Computer Science
Supervisor: Sofia Papavlasopoulou
June 2022

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

Abstract

The fourth Sustainable Development Goal is to: "*Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all*". UNESCO considers Open Educational Resources (OER) vital in achieving this. OERs are educational material shared under an open license permitting free access, use, adaption, and redistribution under few restrictions. The OER movement has regained attention, with COVID-19 forcing millions to study from home. However, there are many challenges to continued growth. One of the most crucial challenges is quality control. Current approaches are mainly built on manual reviews, which are time-consuming and expensive.

This thesis proposes a white-box algorithm that combines theoretical quality knowledge with measurable metrics to give a quality score. The algorithm was developed for the educational resource type Interactive Videos created with the framework H5P. I performed a comparative study of the algorithm and the most adopted approach: manual reviews. 23 H5P users were recruited to perform 107 manual reviews of 57 OERs. The manual reviews scored different quality factors, two overall scores, and could add a comment for each resource. The data were then used to find the degree of agreement between the two methods and their reliability.

The result was low to moderate degree of agreement between the manual reviews and algorithm scores. That means that the algorithm can be a suitable approach in certain cases, but mostly as an addition to other methods. However, the most crucial finding was the low reliability of the manual reviews. The reviews were highly subjective and this has significant implications for this study and all research using reviews as a data source. Future studies need to continue to work on automated approaches but consider how they can be evaluated correctly.

Sammen drag

FNs bærekraftsmål nummer fire er å: "*Sikre inkluderende, rettferdig og god utdanning og fremme muligheter for livslang læring for alle*". UNESCO ser på åpent læringsinnhold (OER (Open Educational Resources)) som en viktig del i å oppnå dette. OER er læringsinnhold som er delt under en åpen lisens som tillater fri tilgang, bruk, endring eller redistribusjon under få eller ingen vilkår. OER bevegelsen har gjenvunnet oppmerksomhet med COVID-19 som gjorde det nødvendig å studere hjemmefra for millioner. På tross av denne effekten er det mange utfordringer for videre vekst. En av de viktigste er kvalitetskontroll. Metodene som brukes nå er hovedsakelig bygget på manuell evaluering som er for tidkrevende og kostbare.

Denne avhandlingen foreslår en hvit box ("white box") algoritme som kombinerer teoretisk kvalifikasjonskunnskap og tilgjengelige metrikker og beregner en kvalitetsscore basert på disse. Algoritmen var utviklet for innholdstypen "Interactive Video" i rammeverket H5P. For å evaluere nytten av algoritmen ble den sammenlignet med den mest brukte metoden: manuelle evalueringer. De manuelle evalueringene hadde ulike kvalitetsfaktorer, samt to overordnede faktorer deltakerne skulle gi en score til. I tillegg kunne de velge å legge til en kommentar. Resultatene fra evalueringen ble så brukt til å finne hvor stor overensstemmelse de metodene har og hvor pålitelige de er.

Resultatet var en lav til moderat grad av overensstemmelse mellom de manuelle vurderingene og algoritmeresultatene. Det betyr at algoritmen nyttig i visse situasjoner, men for det hovedsakelig som i tillegg til andre metoder. Det mest avgjørende funnet var imidlertid den lave påliteligheten til de manuelle vurderingene. Vurderingene var svært subjektive, og dette har betydelige implikasjoner for denne studien og all forskning som bruker manuelle evalueringer som datakilde. Videre studier må fortsette å utvikle automatiske evalueringsmetoder, men ta i betraktning hvordan de best kan evalueres.

Acknowledgements

There are several people that helped me along this thesis. I want to thank my supervisor Sofia Papavlasopoulou for always answering my questions, providing feedback and especially assisting in narrowing the scope. This thesis have been written in cooperation with Joubel AS and I want to thank Svein-Tore Griff With for always motivating me and contributing making the data collection possible. I also want to give a special thanks to all participants that were willing to give of their time by conducting reviews of OER. Last, but not least do I want to greatly thank my friends and family for their invaluable support this period.

Contents

Abstract	iii
Sammendrag	v
Acknowledgements	vii
Contents	ix
Figures	xi
Tables	xiii
Code Listings	xv
1 Introduction	1
1.1 Motivation	1
1.2 Research questions	3
1.2.1 RQ1: What is the degree of agreement between a white-box algorithm and manual reviews in assessing the quality of H5P's Interactive Video?	3
1.2.2 RQ2: How reliable are the manual reviews?	3
1.3 Outline	3
2 Background and related work	5
2.1 Open Educational Resources	5
2.1.1 Definition of Open Educational Resource	5
2.1.2 The rationale for OER	6
2.1.3 History of OER	7
2.1.4 Stakeholders in OER	7
2.2 Quality of OER	8
2.2.1 Aggregation of quality metrics proposed	11
2.3 H5P and the H5P Content Hub	12
2.4 Related work	15
2.4.1 Automated approaches to measuring quality	15
2.5 Summary and challenges with current approaches	18
3 Development	19
3.1 Development tools	19
3.2 Data extraction	20
3.3 Feature scores	20
3.3.1 Metadata added by the user	20
3.3.2 Accessibility metrics	21
3.3.3 Video analysis	22

3.3.4	Interactions	22
3.3.5	Text analysis	23
3.4	One score	23
4	Methodology	25
4.1	Research method	25
4.2	Research design	25
4.2.1	Data collection	26
4.2.2	Survey	27
4.2.3	Data analysis	28
4.2.4	Review data	28
4.2.5	Algorithm testing	31
5	Results	33
5.1	Reliability of reviews	33
5.1.1	Reliability of survey questions	33
5.1.2	Distributions	35
5.1.3	Repeatability	36
5.1.4	Background	40
5.1.5	Comments received	44
5.2	Comparison to algorithm	46
5.2.1	Metadata classification	48
5.2.2	Feature scores correlation	48
6	Discussion and conclusion	51
6.0.1	Implications	53
6.0.2	Limitations	54
6.0.3	Future work	55
6.0.4	Conclusion	55
	Bibliography	57
A	JSON of a H5P interactive video	63
B	Information letter and consent form	95
C	Additional statistics	99
D	Survey	107

Figures

2.1	Taxonomy for DER	9
5.1	Correlation between average(Q1-Q10) and Q12	34
5.2	Bland Altman plot of Q12 and average(Q1-Q10)	35
5.3	Distributions average(Q1-Q10) and overall	36
5.4	Distributions metrics	37
5.5	Correlations plot for averages for each reviewer	38
5.6	Bland Altman plot for agreement between reviewers for average(Q1-Q12)	39
5.7	Correlations plot for Q11 and Q12 for each reviewer	39
5.8	Bland Altmann agreement between reviewers on Q12	40
5.9	Distribution of high and low score (Q5) divided by review experience	43
5.10	Boxplot gender differences average(Q1-Q12)	43
5.11	Distribution of high and low score (Average Q1-Q12) divided by genders	44
5.12	$Alg_{weighted}$ Bland Altman plot with average(Q1-Q10), average(Q1-Q12) and Q12	46
5.13	$Alg_{unweighted}$ Bland Altman plot with average(Q1-Q10), average(Q1-Q12) and Q12	47
5.14	$Alg_{unweighted_v2}$ Bland Altman plot with average(Q1-Q10), average(Q1-Q12) and Q12	47

Tables

2.1	Technical factors in frameworks	12
2.2	Pedagogical factors in frameworks	13
3.1	WCAG scores [63].	21
3.2	Importance weights used in $alg_{weighted}$	24
4.1	Survey questions	29
5.1	Correlation between metrics and the overall questions Q11 and Q12	34
5.2	Mean, median and standard deviation for metrics	37
5.3	Correlation of metrics between two reviewers	38
5.4	Age distribution	41
5.5	Educational background	41
5.6	Occupation distribution	41
5.7	OER use	42
5.8	H5P use	42
5.9	Correlations $Alg_{unweighted}$	47
5.10	Correlations $alg_{unweighted.v2}$	48
5.11	Explanation of feature scores used in correlation measures	48
5.12	Highest feature score for each question	49
C.1	Shapiro-Wilk test of normality metrics	99
C.2	Shapiro-Wilk test of normality averages	100
C.3	The significant ($p < 0.05$) metric correlations for average(Q1-Q10)	100
C.4	The significant ($p < 0.05$) metric correlations for average(Q1-Q12)	100
C.5	The significant ($p < 0.05$) metric correlations for Q1	101
C.6	The significant ($p < 0.05$) metric correlations for Q2	101
C.7	The significant ($p < 0.05$) metric correlations for Q3	102
C.8	The significant ($p < 0.05$) metric correlations for Q4	102
C.9	The significant ($p < 0.05$) metric correlations for Q5	102
C.10	The significant ($p < 0.05$) metric correlations for Q6	103
C.11	The significant ($p < 0.05$) metric correlations for Q7	103
C.12	The significant ($p < 0.05$) metric correlations for Q8	103
C.13	The significant ($p < 0.05$) metric correlations for Q9 (none)	103
C.14	The significant ($p < 0.05$) metric correlations for Q10	104

C.15 The significant ($p < 0.05$) metric correlations for Q11	104
C.16 The significant ($p < 0.05$) metric correlations for Q12	105

Code Listings

2.1	H5P Image params	14
2.2	H5P single choice params	14

Chapter 1

Introduction

1.1 Motivation

The concept of Open Educational Resource (OER) is an emerging phenomenon influenced and related to movements like Open Source Software (OSS) and Open Access (OA) [1]. It has recently gained more attention due to the COVID-19 pandemic, with school closures that impacted 80% of students worldwide and will change the future of education. [2]. Even though the topic of digital learning has regained attention recently, the concept of OER is far from new; established 20 years ago [3]. The term OER and its different definitions and constraints will be further elaborated on in the Chapter 2. However, in short, OER can be defined as different kinds of educational material, everything from a quiz to an entire course, that are free to use by anyone under few constraints. It is hard to know the exact number of users and resources because they are spread among various projects, initiatives, and repositories. Additionally, they are in different formats, languages, and geographical areas. However, we know for sure that the number is growing fast [1]. The largest OER repository is the Merlot project which currently has over 98 000 resources and 192 000 registered users [4], and this is only one of many projects.

With the growing number of shared resources, several challenges exist, some new, some old. With the transition from mainly text-book resources to more digital resources, several concerns about the quality and trust of the resources arise. Hylén [1] present "*Quality assurance*" as one of the three main challenges for OER, and that will be the topic of this thesis.

Quality assurance for OER can be a topic in all phases, from the decision to create a resource to someone reusing it [5]. This thesis will focus on quality assurance when a resource is shared. Today, different repositories have different approaches to presenting quality information to users. The main approaches are user reviews, and peer reviews [6] following either formal or informal evaluation methods. Unfortunately, this manual evaluation is very time-consuming and requires many users to be reliable [5]. A peer-review process can be possible for projects governed by universities, but it becomes a challenge for repositories de-

pendent on volunteering users. In contrast to e-commerce which relies on users' ratings, it is harder to find users motivated to review in OER [7]. The result is a lack of quality information which makes it hard for the users to find the best-suited resources.

Therefore, there is a need for an evaluation that is more cost- and time-effective. A suggested solution for this problem is to create a method for automatic evaluation. Since usage data measure popularity rather than quality [8] and favor the oldest resources, the evaluation should instead use characteristics of the resources. The characteristics could be the metadata provided by the author or metrics like text length, the number of images, or other data available. With this approach, all resources can get a quality score immediately after sharing. This score can assist the user in selecting the right resource.

Automated quality measurement can not replace all other approaches, and human evaluation will still be an important factor, especially in ensuring trust [9]. However, it can save time and lower the evaluation cost, making a repository less dependent on a large number of reviewers. An automated approach will also give an objective evaluation and not depend on a single user's view. It can also provide an initial quality score that can assist in selecting which resources to review manually.

An added benefit of a method that can measure the quality using intrinsic metrics is the possibility for the contributors to get their resources evaluated even before sharing. By this the creator can improve the resource apriori to sharing [6]. For this to be possible, the quality result needs to be explainable and measure the factors that makes a high-quality resource. Even though certain factors can be correlated to an observed quality, like the author [10], this is not a factor that imply that a resource is high-quality. Instead Cechinel *et al.* [11] suggest a white-box models that gives understandable and interpretable results.

One way to automatically assess the quality of resources is to build machine learning models based on resources classified as high, and low-quality [10–12]. This approach can be efficient, but by itself, it is impossible to make any claims of cause and effect by the measured factors [13]. Given the high variation of resources, another challenge is that one would need a high number of heterogeneous resources reviewed by many people. The results can easily be hard to generalize and restricted to a specific data set. Since they are not explainable, they can also not be used to guide the creator.

Another way to look at the problem is by evolving from the traditional quality factors in different frameworks. The advantage is that the metrics can be explainable and directly connected to the quality. This thesis will try this approach which has been suggested [8], but not tested.. The algorithm will combine the measurable metrics with quality factors. There are, of course, challenges with this approach because all quality factors are not instantly measurable; for instance, how is engagement measured? It will try to overcome this challenge by using what is available and see if that is enough to give a usable prediction.

1.2 Research questions

This thesis will investigate the possibility of using a white box algorithm based on quality factors. The motivation shows a need for a cost-effective way to create quality information about an OER. Novel metrics can be gathered using new data collection methods enabled by the resource type studied. The thesis will bridge a research gap by using these metrics and connecting them to theoretical frameworks.

No "true value" of quality exists, but users can evaluate how well fitted a resource is for its purpose or how well it is made based on defined quality criteria. Voluntary participants will conduct manual reviews, and I will estimate the reviews' reliability. The thesis will therefore contribute by developing a new type of algorithm, creating new insight into the perception of quality and reliability of manual reviews. To test the potential of this algorithm, the two research questions presented below shall be answered.

1.2.1 RQ1: What is the degree of agreement between a white-box algorithm and manual reviews in assessing the quality of H5P's Interactive Video?

The white-box algorithm is proposed as a new quality assessment method. Since it can not be compared with the "true" value, the most adopted approach will be used: manual reviews. In the comparison of the two methods, the agreement degree is the central question, an estimate based on the difference in each resource evaluated. A high degree of agreement indicate that the algorithm is suitable for replacing manual reviews.

1.2.2 RQ2: How reliable are the manual reviews?

When comparing the algorithm with manual reviews, the result's significance depends on the reliability of the manual reviews. Reliability is concerned with "*if one's findings will be found again*" [14]. For the applicability of this thesis' results the reliability is very important, as Hanneman [15] states "*If one or both methods do not give repeatable results, assessment of agreement between methods is meaningless.*".

1.3 Outline

Chapter 1 introduces the research topic and motivation, the research questions, and the thesis structure.

Chapter 2 presents the background on OER, frameworks for evaluating the quality of OER, and related work on automated approaches to estimate it.

Chapter 3 presents the development of an algorithm to assess the quality and the development tools used.

Chapter 4 presents the research methodology and design.

Chapter 5 presents the findings from an analysis of the collected data.

Chapter 6 discuss the results, present a conclusion, and proposes future work.

Chapter 2

Background and related work

2.1 Open Educational Resources

2.1.1 Definition of Open Educational Resource

An educational resource can take many forms, from an educational text to an entire course. The most used definition of Open Educational Resource (OER) is that "*OER are teaching, learning, and research resources that reside in the public domain or have been released under an intellectual property license that permits their free use or re-purposing by others*" [16]. This broad definition does not limit OER to e-learning but includes printed materials. However, only digital resources are in the scope of this thesis.

Another term often used with OER is Open CourseWare (OCW), defined as "a free and open digital publication of high-quality university-level educational materials. These materials are organized as courses and often include course planning materials and evaluation tools, as well as thematic content" [3]. OCW can be seen as a subset of OER.

In addition to OER and OCW, Learning Object(LO) is a term in related literature. *Learning objects* can be defined as "*educational resources that can be employed in technology-supported learning*" [17]. Therefore, the practical use of these resources is very similar and, in some literature, used synonymously. The critical difference between the two terms is that the license type. The license of Learning Objects is not specified, and it do not need to be under open access like an OERs. Without this licence type the focus is less on reusing resources, a critical part of the OER movement. Papers on Learning Objects can therefore be used when considering these differences.

As the license is theoretically the only fundamental difference between OER and other educational resources [3] it is worth discussing the practical implications. A necessary part of an OER is allowing reusing and a license type that is commonly used is the creative commons (CC) which lets people reuse, alter and redistribute Butcher [3]. Different CC versions let the author choose how open the resource will be. This enables the author to permit usage, but under certain

limitations, like attribution and restricting commercial use. It gives a controlled way of sharing content that lets the author keep their right. The Creative Commons license was founded in 2001 [16] as an answer to the growing Open Access movement.

Apart from the theoretical differences between LO and OER, there are other perspectives on what the term "Open" means and entails. "By Walker, "open" is described as "convenient, effective, affordable, and sustainable and available to every learner and teacher worldwide" Sir John Daniel speaks of "the 4 As: accessible, appropriate, accredited, affordable" [1]. This brings out another important point about OER it is not only about price or licensing but also other kinds of availability. Here accessibility is an essential factor, and it is possible to argue that an educational resource is not fully open unless it is available to everyone, including people with different functionality levels [18].

As shown with the different terms above, learning objects and OER can be very diverse, and it, therefore, makes sense to categorize or divide them according to their properties. The IEEE Learning Object Metadata standard classifies a learning object according to three properties which are aggregation level, interactive type, and resource type [19]. Aggregation types refer to the level of granularity going from level 1 being the atomic level to level 4 being a collection of courses. An object can have three interactive types: expositive, meaning information from the object to the learner; active, meaning information to and from the learner; and mixed, which combines the two. Resource types include exercise, questionnaire, diagram, figure, graph, index, slide, table, narrative, text, exam, experiment, problem, and self-assessment [19].

2.1.2 The rationale for OER

The term OER was first used at a UNESCO meeting in 2002, and afterward, they published recommendations on the use of OER. One of the reasons for OERs importance is that UNESCO consider it to be vital in achieve Sustainable Development Goal 4 (SDG4) "to ensure inclusive and equitable quality education and promote lifelong learning opportunities for all"[20]. The application of open licenses to educational materials can give access to learning resources to everyone, everywhere [20].

The result is that high-quality learning materials can become available to more people. At the same time, the cost for high-quality content can be reduced, and reusing content means that the resource quality can keep improving for each iteration [3]. That will happen without effort but is an opportunity that OER creates [3].

OER should pass geographic, and language barriers[3]. UNESCO, therefore, recommends that the different stakeholders continue working to develop OER both the development of strategies and frameworks, and technical development.

2.1.3 History of OER

The concept of OER was enabled by legal changes (Creative Commons licenses) and digital development. The term OER was first coined at a UNESCO Forum on Open CourseWare for Higher Education in Developing Countries held in 2002 [3]. The OER journey began when MIT started to share its courses, calling them Open CourseWare (OCW). They shared the equivalent of the courses openly, mostly in .pdf formats [16]. In the years after this, distance learning gradually transformed from being centered around traditional lectures to becoming resource-learning [3]. Digital resources also became relevant not only for those learning at distance, but for everyone.

Digital development has made it possible for many more people to access and share content and enabled many new forms of learning objects. In 2007 computing and communication infrastructure was seen as one of the major challenges for OER [16]. Today that is a less challenge in most countries even though there is a *digital divide* and many still do not have internet access. [21].

The shift and increased use of OER is not only a digital or legal one but also a pedagogically, meaning how we define teaching and learning is changed. It is worth noting that the shift to digital learning resources not necessarily are pedagogically innovative or increases the learning outcome. The outcome largely depends on the resource[22]. It is therefore important to present quality information.

The rise of OER has been hugely driven by universities and country-specific OER projects, which have had their platform for sharing their courses [16], rather than large geographically independent projects. This has led to a challenge with sustaining repositories over a longer period and, naturally, also the growth [23].

2.1.4 Stakeholders in OER

Since the stakeholders are very important for defining quality, presenting the main stakeholders in an OER is crucial. Camilleri *et al.* [5] divide them into four groups: Policymaker level, Management, and administration level, Educational level (teachers, professors, curriculum designers, and others at that level), Teaching and learning levels (learners, students, tutors, teachers). For the purpose of the algorithm the most relevant groups are the educational level and the teaching and learning level. The learning level will be the end-user, but the person selecting a resource may vary depending on the level of granularity. While it is typical that a learner directly chooses an open course, lower granularity elements are more often chosen by the educational level and integrated into their course. This thesis focus mainly on the granularity level 1-3 where the teacher mainly selects the resource. Since the teacher typically sees the quality information and the learner use the resource both are important stakeholders in this case.

2.2 Quality of OER

To measure the quality, we would need some form of a universal definition of what quality is. Harvey and Green [24] Points out how hard it is to give a clear definition of quality as it means different things to different stakeholders but suggests five dimensions: "*excellent, perfect, fit for purpose, cost-efficient, and/or transformative.*" Kawachi [25] points out that the dimension most relevant for OER is *Fitness for Purpose*, which will be the definition used here. *Fitness for purpose* can be defined as "*Satisfying the aims or reasons for producing the item, according to the judgments of the various stakeholders - particularly the consumers*" [24]. The most important stakeholders will be the educational, teaching, and learning levels described in the previous section on the relevant stakeholders.

Measuring the quality of an OER is challenging since the resources are diverse, and the stakeholders have different perspectives. The users expect OER to have reasonable quality, but the perception quality is highly subjective [26]. The quality perspective can be diverse also between expert reviewers and users. Sanz-Rodriguez *et al.* [27] and Cechinel and Sánchez-Alonsor [28] found only a low correlation between the rating of expert reviewers and users. Even though this highlights the subjectivity many metrics and instruments exist to quantify the quality.

There is no common standard for evaluating OER [29] so different frameworks for evaluating the quality of OER and Learning objects will be presented. While these are primarily meant for manual evaluation, it does not mean that some aspects could apply to automated processes. The framework DER Quality Assessment is an example of a framework similar to the ones used for manual evaluation. However, it is created to be used in an automated process.

A evaluation method for Learning Objects is the Learning Object Review Instrument (LORI) [22] which was also used by the repository eLERA [28]. In their definition of a Learning Object, both traditional paper-based learning objects and the emerging interactive and different digital tools are considered. It is developed by Vargo *et al.* [19] and the latest version, LORI 1.5, has nine items. A scale from 1 to 5 is used for each evaluation metric, with 5 being the best score.

For LORI 1.5, the nine items are:

1. Content Quality: Accuracy, balanced presentation of ideas, appropriate level of detail.
2. Alignment among learning goals, activities, assessments, and learner characteristics.
3. Feedback and Adaptation: Adaptive content or feedback driven by differential learner input or learner modeling.
4. Motivation: Ability to motivate and interest an identified population of learners.
5. Presentation Design: Design of visual and auditory information for enhanced learning and efficient mental processing.
6. Interaction Usability: Ease of navigation, predictability of the user interface, and quality of the interface help features.

7. Accessibility: Design of controls and presentation formats to accommodate disabled and mobile learners.
8. Reusability: Ability to use in varying learning contexts and with learners from differing backgrounds.
9. Standards Compliance: Adherence to international standards and specifications. [30]

Vargo *et al.* [19] showed that LORI can give reliable assessment and that a collaborative assessment can improve the reliability and that other factors to improve the quality are prior training and expertise of reviewers.

Another set of quality criteria for OER resources is proposed by Scheunemann *et al.* [8]. It is worth noting that the term Digital Educational Resource used here includes both Open Educational Resource and Learning object, but excludes non-digital resources. They suggest a set of criteria that can be evaluated both manually and automatically in future research. Their motivation is to create a set of criteria that do not depend on user popularity, which is not necessarily the same as quality [8]. The criteria are presented as a graph split into a pedagogical perspective and a technical perspective which can be seen in 2.1

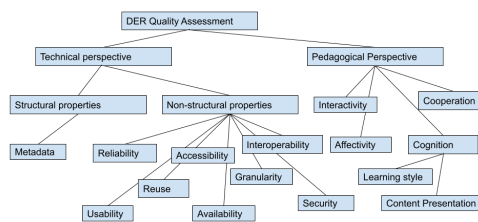


Figure 2.1: Taxonomy for DER [8]

A more complex framework for the quality of OER is presented in Romero-Pelaez *et al.* [31]: Quality4OER. Their quality framework is based on Peláez *et al.* [32] which presents six quality metrics:

- Development, distribution, and licensing models
- Academic rank
- Presentation to the user
- Evaluation and support material
- Technological requirements and interoperability
- Accessibility

Their proposed framework consists of the six quality metrics over five maturity levels. The base level is that the resource is available on the web and has an open license to be an OER. With the other four levels the six quality metrics become more and more complete and using higher standards.

Almendro and Silveira [26] uses the concepts of Quality Assurance from software and transfer them to OER testing. They define a framework to generate a reliability index for a resource based on how many successfully executed tests. Even though they point out that OER resources are heterogeneous, they divide the different testing dimensions into pedagogical, content, and technical. For these dimensions, the test considered relevant are: Unitary, integration, operational, positive-negative, regression, black-box, white-box, functional, usability, performance, load, user acceptance, volume, stress, configuration, installation, and security.

A framework that has been created mainly as a guideline to creators of OER is the TIPS framework [25]. This framework consists of 38 criteria divided into:

- Teaching and learning processes
- Information and material content
- Presentation product and format
- System technical and technology

This framework works as a checklist to guide the creator. This perspective means that it is not something that is used directly in the evaluation of resources. An example of a criterion that is hard to transform for evaluation use is "*Monitor the completion rate, student satisfaction and whether the student recommends your OER to others*". In contrast, others could be more easily transformed into an evaluation criterion like "*Use a learner-centered approach*". All of these criteria are validated both by teachers and OER experts. They also provide a view of the quality of OER on three levels. Presented as: "*(i) the upper-most level-1 of the repository containing the internationalised OER that have been standardised by OER experts and like a textbook are almost context-free, (ii) the intermediate level-2 of readily adaptable OER, and then (iii) the ground level-3 of the fully localised OER used by actual students.*" [25]. This current framework aims to achieve level 2 and is an important aspect of the important sharing and reusability of OER.

A quality model proposed in Mayrberger *et al.* [33] is presented in Zawacki-Richter *et al.* [29]. The model has a short and long version, the long giving a more detailed way of evaluating. It uses two main categories pedagogical and technical. The pedagogical dimension consists of eight factors divided into content and instructional design. While it for the technical dimension is seven factors, sorted under accessibility and usability.

While there are many proposed frameworks for the evaluation, different methods are also used in repositories. MERLOT stores metadata and a reference to the educational resource and can be considered a refractory [6]. In MERLOT, a combination of peer reviews and user reviews is employed. While the user reviews do not have clear criteria, the criteria for peer reviews are:

1. Quality of Content
2. Potential Effective as a Teaching Tool
3. Ease of use. [6]

By Kumar *et al.* [34] these three criteria were linked to LORI to unite these

two methods, as shown below.

Quality of content:

- Content Quality
- Reusability

Potential Effective as a Teaching Tool:

- Feedback & Adaption
- Motivation
- Learning Goal Alignment
- Presentation Design

Ease of use:

- Interaction Usability
- Accessibility

Other:

- Standard compliance.

The repository Graphite use three learning dimensions:

1. Engagement
2. Pedagogy
3. Supports [6]

Another perspective on the quality of OER comes from the important stakeholders: teachers. Clements and Pawlowski [9] analyzed teacher view on OER quality. Trust is found to be essential for teachers to find and use resources in this paper. The trust can be either in the resource itself or in the person or organization that created the resource. They also researched what is important for teachers when choosing resources and had 146 respondents. One question was about what the quality of a resource means for users, and the top answers are listed below:

- 83% describes it as a good use of multimedia. Especially relevant as this might be hard to produce themselves.
- 80% defines it as scientifically correct.
- 79%: fit their lessons or curriculum
- 69%: interoperability between LOR and LMSs
- 55%: source is an organization with a good reputation
- 17%: their own quality strategy.

2.2.1 Aggregation of quality metrics proposed

In table 2.1 and table 2.2 the different quality metrics that have been suggested in the frameworks are aggregated. The factors are divided into a technical and pedagogical dimension suggested by Scheunemann *et al.* [8] and Mayrberger *et al.* [33]. Some, of course, have different wording, but the general meaning of

Quality factor	DER	IQOER	LORI	Quality4All	OERTrust	Other
Technical	X	X			X	Tips
Usability	X	X(Also as: "Structure, navigation and orient- ation")	Interaction usability		X	Merlot: Ease of use
Accessibility	X	X	X	X		
Reusability	X	X	X			
Metadata	X		Standard compli- ance			
Interoperability	X			X		
Reliability	X	X			X	
Licence		X		X		
Granularity	X					
Availability	X					
Security	X				X	
Portability						

Table 2.1: Technical factors in frameworks

the factor is the same. The most agreed-upon technical factors are usability, accessibility, and reusability. Content and student support are the most important pedagogical factors.

2.3 H5P and the H5P Content Hub

H5P is the framework for Open Educational Resources used in this thesis, and it is worth describing it and some of its properties. H5P is a tool to create, share and reuse HTML5 content online and is widely used, especially in educational institutions. H5P itself is an Open Source tool created and launched in 2014 [35] and was in 2021 used in over 200 000 sites and by more than 200 000 000 users [36].

H5P consists of an interactive content creator that lets the creator choose from 52 resource types [35] with different characteristics and structures. The metrics related to the quality of a resource have been seen as related to the type of resource [11, 13, 37] and short introduction to the types of resource will given here. For each content type, the aggregation level and resource type are always set while the interactive type for some content types be set by the author [19]. The content types are categorized into three categories "Larger Resources", "Tasks", and "Other". The category of "Tasks" are content types that all have the Interactive Type of Active. From this category, the user can select typical question types like "Multiple choice", "Fill in the Blanks", "Drag and Drop" and many others. The "Other" group is heterogeneous resource types. Some are solely *Expositive* like "Chart", "Collage",

Quality factor	DER	IQOER	LORI	Quality4All	OERTrust	TIPS	Other
Pedagogical	X	X			X	X	Graphite
Presentation	X	X	X	X		X	
Alignment		X	X	X		X	Merlot, Teachers
Content		X	X		X		Merlot
Student support		X	Feedback & adoption	X			Merlot, Graphite
Instructional design						X	Teachers
Interactivity	X	X				X	
Engagement			X				Graphite
Academic foundation		X		X			
Cooperation	X	X					
Affectivity	X						
Cognition	X						
Learning style	X						
Applicability		X					
Media		X					Teachers

Table 2.2: Pedagogical factors in frameworks

and "Accordion", and some that do not fit into the other categories like "Audio Recorder" and "Personality Quiz". "Larger Resources" resembles an aggregation of level 2 or 3 [19] and is a content type that lets you include other lower-level contents from "Tasks" and "Other" and, in some cases, other "Larger Resources". A larger resource could, for instance, be a collection of quizzes or an Interactive Book, which can be solely expositive by having images, videos, text, and video. It can also be solely active by having a wide range of questions or, most commonly, a combination.

H5P content is created by using the interactive editor. It can be shared by linking to the page where it is hosted, downloading it as a .h5p file, or copying by utilizing the local storage within pages that support this. This standardized system enables the reuse and re-purpose, which is one of the main ideas behind OER [16]. With this system, H5P resources can easily be modified after creation. The resource is rendered based on a JSON which provides a reliable way of knowing what content consists of and makes it readable for a computer. All text, images, and interactions can easily be extracted and analyzed.

In code 2.1, the typical structure for an image is shown. Here the path, size, copyright information, and the alt text are shown.

Similarly, a single choice question will have the structure shown in code 2.2. The whole JSON for a sample H5P content can be seen in the appendix A. For a "Larger Resource," there would be a hierarchical structure with many sub-content

```
"params": {
  "contentName": "Image",
  "file": {
    "path": "images/file-5eecbc7ecafb0.jpg#tmp",
    "mime": "image/jpeg",
    "copyright": {
      "license": "U"
    },
    "width": 2028,
    "height": 444
  },
  "alt": "Strawberries header"
},
```

Code listing 2.1: H5P Image params

```
{
  "answers": [
    "<p>Brittany,&nbsp;France</p>\n",
    "<p>London,&nbsp;UK</p>\n",
    "<p>Vienna, Austria</p>\n",
    "<p>Lima, Peru</p>\n"
  ],
  "question": "<p>The very first garden strawberry was grown in:</p>\n",
  "subContentId": "0a5f12b1-c025-4a19-9771-d31d58511438"
},
```

Code listing 2.2: H5P single choice params

specified by the library and settings.

The H5P OER HUB is a repository that lets users share the created content [38]. The sharing metadata is Title, Licence, Language, Discipline, Level, Author, Tags, Description, and Screenshots. Title, Licence, and Language are mandatory fields. This standard correlates with many other repositories and is similar to the presented work of Tavakoli *et al.* [12], where the metadata fields are title, description, subjects, level, language, time required, and access abilities.

The user searching for resources in the H5P OER HUB can type in a search word, order the search on the newest or most popular, and apply a set of filters. Those are discipline, content type, license, language, level, and if the resource is reviewed. It is important to note that the "reviewed" here does not mean that the content has been reviewed in a qualitative sense, only that it is not considered spam or violating the regulations. In addition to the H5P OER HUB, many other repositories also let users share their H5P resources. They all use a different set of metadata.

Regarding accessibility in H5P, most content types are fulfill WCAG requirements, which means that most resources inherit this quality. It is, however, still possible to make a inaccessible resource. This happens if an author changes the color style and text size or does not provide alt text for an image.

2.4 Related work

2.4.1 Automated approaches to measuring quality

There are many approaches when it comes to the evaluation of educational resources and the prediction of high-quality resources. One way to categorize the different quality rankings we can use is explicit, implicit and characteristically [27]. Explicit data is the result of evaluations from users and experts. Implicit uses data from the material like counts of downloads, visits, bookmarks, and other analytics derived from how users interact with a resource. Characteristical is descriptive information about the characteristics of the resource that could come from the metadata or how the resource is built. While explicit and implicit data comes from usage data, the characteristical data is available as soon as the resource is created.

Explicit and implicit data can be used in recommender systems. Different ways are content-based, collaborative filter, knowledge-based, and hybrid recommender systems [39]. There are, however, different challenges with these. One of the main challenges is the lack of reviews [5, 34]. After analyzing reviews in the MERLOT repository Cechinel *et al.* [13] found that of a sample of 20,506 items, only 12,65% had peer review, 12,24% user review, and 3,38% had both. Other typical problems with collaborative filtering are cold start, sparsity, first rater bias, and popularity bias [34]. This generally means that the recommendation will favor already popular resources. In contrast, new possible high-quality resources will not be discovered and consequently not reviewed. For these reason in this

thesis is on creating an algorithm using the characteristic metrics.

The research presented here is not necessarily meant to replace the evaluation methods that use implicit and explicit data. Nevertheless, they can be an additional feature that could overcome some of the presented challenges. There are also several advantages with an automated evaluation as it is inexpensive, time-saving, and can be conducted at or before the publication time [6]. A possible future advantage is that the contributor can view the results before posting and make improvements before sharing [40].

Metadata

Several studies focus on measuring the quality of metadata [41] [42], and in Tavakoli *et al.* [12] a connection between the metadata quality and the quality of the actual resource is made. High-quality metadata is essential for identifying resources, selecting resources, and acquiring resources [41] and is also a part of the LORI framework 1.3 [19].

A study that uses the Metadata-based Approach is Tavakoli *et al.* [12]. This paper uses 8,887 OERs from SkillsCommons to analyze and create a model. Their approach was to use the metadata provided by the user and then have a scoring method. They used a set of manually quality-controlled OERs as a reference. The metadata fields present were used to create a scoring model. For title, description, and subjects the highest score was if it was close to the mean, while 0 was if it was not present. Level, length, language, and accessibility had 1 or 0 if present or not. They achieved an accuracy of 94,6%. Applying the same model to YouTube videos, the videos classified as high quality generally had higher ratings [43].

An approach built on top of Tavakoli *et al.* [12] is presented in Elias *et al.* [44]. They suggest an OER recommender system that considers the accessibility quality aspect. The system takes in the users' needs to give them result which fulfills their requirements. The returned result consists of a quality prediction using the metadata approach [12] and a 28-dimensional vector describing the required accessibility properties. In this, they used 1500 OERs, and 100 of the recommended OERs received an average score from accessibility experts. The results were also tested using manual and automatic accessibility evaluation tools, and most passed this test.

Other metrics

In Cechinel *et al.* [13] the statistical profiles of rated objects in MERLOT are created and used for classifying objects. They use 35 metrics that are based on similar research in related fields [10, 45–52]. The 35 metrics used are in the categories: link measures, text measures, graphic interactive and Multimedia measures, site architecture measures, and evaluation metadata. The metrics were collected using a crawler going 2-level deep. The research was based on 6470 learning objects, but only 1765 were used since they excluded the ones without a review. Based on these analyses, the result varied between different disciplines and materials.

However, the most significant value was within images (size and number), which was important for peer reviewers and user reviews. Based on these correlations found during the analysis, they separated good and poor resources with 91,49% accuracy and statistically significant at 99%. This was significantly higher than separating between good and average. One of the drawbacks of this study is the use of a crawler which will not give accurate information if the whole object is a script. They opt for creating different profiles of granularity, discipline, and material type [13].

A study that is built on top of Cechinel *et al.* [13] and Cechinel and Sánchez-Alonsor [28] is Cechinel *et al.* [37]. Here, the objective is to find the lower-level and easily quantifiable measures related to a learning object's quality. The work uses the same data collected in Cechinel *et al.* [13] and, using these metrics, builds a classification tree, dividing the data into training, validation, and testing. For each subset of discipline and resource type, accuracy, sensitivity, specificity, kappa, MAE, and nodes are measured. The result here is that few metrics, in general, can describe the quality of a resource. They also show that the most important metrics vary between resource types and disciplines. Their suggested direction further investigates the threshold values and other metrics like readability measures.

In a study done by Bethard *et al.* [10] they created a machine-learning algorithm to do a quality assessment. This algorithm was based on assessing quality indicators by experts and non-experts that assessed quality factors and content. Based on the preliminary study, they found the seven most predictive indicators of accepting/rejecting a resource. Accepting a resource means that experts believe the resource quality is high enough to be shown. The resulting indicators were: has prestigious sponsor, content is appropriate for age range, has sponsored, identifies learning goals, has instructions, identifies age range and organized for learning goals. In this study, 1000 resources were included and annotated with these properties. For some metrics, the model assessed their presence with an accuracy of over 80%. The result of these especially shows how easy it is to predict the acceptance of a resource by a prestigious sponsor and sponsor, having a correlation of 0.905 and 0.858 with accept/reject, meaning that the creator is an easy predictor, but a factor that can also be excluding to new creators. All the factors were predicted using natural language processing, and the library used was Digital Library for Earth System Education.

Another similar approach is used by Cechinel *et al.* [11]. Here the goal was to classify LOs as *good* or *not good*, for resources in MERLOT and Connexions. They used 35 metrics in four measurement classes: *link measures*, *text measures*, *graphic*, *interactive*, and *multimedia Measures*, and *Site Architecture Measures*. They gathered 20,582 LO in Merlot, and 2076 rated by experts were used. They divided the data into subsets based on resource type and used ANN to create models. The highest accuracy was for a subset in Merlot(75%) and Connexions (50% and 53%). The base reference for Connexions was the elements with just one endorsement and those with two or more. Most models across subsets and repositories better classified low-quality resources rather than high quality.

2.5 Summary and challenges with current approaches

These evaluation methods presented can typically be divided into peer-review and public review. The peer-review process is similar to the one used when scientific papers are assessed and are therefore conducted by a group of experts [28]. It is considered a formal process with requirements of expertise of the reviewers and is typically slow and following a specified process. Public reviews, on the other hand, are open to everyone without any requirements of knowledge and are typically performed fast without any specific process [28]. These are analogous to reviews conducted in public marketplaces like Amazon or eBay or communities like IMDb or YouTube. Previous studies have found little correspondence between user and expert reviews [27, 28]. The results from Sanz-Rodriguez *et al.* [27] Cechinel and Sánchez-Alonsor [28] are also not directly transferable since they compare guided peer reviews with just unguided scores for the public reviews.

The empirical studies presented above showed many ways to measure a resource's quality automatically. The approaches used data collection of either or other metrics and looked at the factors presented in high-quality resources. The general results were that automated approaches predicted the quality of a resource with relatively high accuracy. These results are, however, based on limited data. These approaches can be an addition to the use of reviews, especially as the reliability of the reviews can be low when it is sparsely [5].

Another limitation of the empirical research results is that they do not include a theoretical foundation supporting the metrics. With this foundation, the results could be more universally applicable and translatable to different resource types with other metrics.

Furthermore, some approaches suffer from a general problem in gathering the metrics. As most metrics are specified to a specific repository, there is no way of applying these results directly to another to test the result. Another challenge is that gathering general metrics of a resource is complicated. MERLOT, for instance, does not store its resources but links to the actual page, making it hard to get correct data. With a crawler, the website of the resource is measured, but the actual resource might not be measured, leading to false results [13].

With the empirical research presented here, there is no evidence between the causality of these metrics and the quality of a resource. As most of this research also focuses on lower-level metrics, most of these metrics used are not explainable as a valid reason for the resource being of high quality. A clear example is that metadata can be used to predict quality [12], with a clear correlation between the metadata quality and the resource quality. Even though metadata can be considered one quality factor [19] this is probably not enough to indicate a causality saying that metadata is the only qualifying factor. However, the research shows some relation between a resource with high-quality metadata and high quality itself, which suggest that the topic should be explored more.

Chapter 3

Development

The main objective of this thesis is to develop an algorithm to automatically measure the quality of OERs and this chapter will provide the implementation details.

The use case for this algorithm is to take in one or more OERs and return a score from 0 to 1, with 1 meaning high quality and 0 low quality. This score should reflect the overall quality of the OER as defined in Chapter 2. The result could be used to sort a search result or be presented to the creator.

The algorithm developed in this thesis builds on the suggestion from Scheunemann *et al.* [8] of combining measurable quality criteria and, by that, enabling an automatic approach. Using this idea the algorithm calculates different feature scores and add them together to create one score. These feature scores are calculated from the available metrics of the OERs. A feature score could, for instance, be the availability of metadata or the percentage of images having an alternative text. It could also be more advanced like quality analysis of video or text analysis.

For the algorithm to be valuable, there are some different requirements. The score needs to be analogous to the users' perceptions and be discriminative. An essential requirement when developing this algorithm was that each feature score should be explainable and connected to a quality criterion, in contrast to earlier research [10, 11]. Doing this increase the chance of the result being independent on the specific data set and increases the reliability. It also makes it possible to provide specific suggestions for improvements of a resource.

3.1 Development tools

A set of different tools and packages were used in the development. *Python* was chosen as the programming language because it is fast and flexible and the many packages available makes it well-suited for scientific and engineering code [53]. A set of different packages were utilized in this development:

- *Numpy* for multidimensional arrays [54]
- *beautifulsoup4* to parse HTML [55]
- *skvideo* for video analysis [56]

- *OpenCV* to get duration of videos [57]
- *Youtube dl* to get the YouTube videos [58].
- *Textstat* for text analysis [59]
- *Language_tool_python* for grammar and spelling check [60]

GitHub was used for the hosting of the code and to ensure version control and Visual Studio Code was chosen as an IDE.

3.2 Data extraction

All scoring functions were based on the OER metrics, and in the preprocessing phase, these were extracted. H5P resources follow a typical format, but there are variations for each content type. The input format is a JSON (see appendix A) following the structure described in Chapter 2. Supplementary media files are provided in a folder.

From the JSON, all interactions, media, and assets were extracted. They were then categorized into *images*, *text elements*, *links*, *shapes*, *tables*, *navigation*, *bookmarks*, and *tasks*. Different methods were needed to access the media within interactive elements and all text because of the structural differences. In addition to these metrics, metadata was extracted. It was collected from the respective repository and documented in a CSV format.

3.3 Feature scores

In this section the different feature scores will be presented. The selection is based on using the available metrics in the H5P Interactive Video to measure factors considered important for the quality by frameworks. The most relevant feature scores were included in the proposed algorithm, while some were only used in a correlation analysis.

3.3.1 Metadata added by the user

Both Tavakoli *et al.* [12] and Bethard *et al.* [10] showed a clear connection between metadata quality and resource quality. High-quality metadata itself is crucial for localizing resources and providing insight of how to use. It is an easy metric to measure, but it can be challenging to compare results because different metadata are available in different repositories. Given the good results in Tavakoli *et al.* [12] and Tavakoli *et al.* [43], their scoring functions were used. The metadata in their case were *title*, *description*, *subjects*, *level*, *language*, *time required* and *accessibilities*. The two scoring functions most important for quality prediction were chosen. The Availability score shown in 3.1 measure the availability of different metadata based on their importance. Expanding on this, the Norm score in 3.1 rank the title, description, and subjects based on their closeness to the mean of their distribution.

$$avail_score(o) = \sum_{k=fields} norm_import_rate(k) \quad (3.1)$$

$$norm_score(o) = \sum_{k=fields} norm_import_rate(k) * rating(o,k) \quad (3.2)$$

3.3.2 Accessibility metrics

One of the basic ideas about OER is to make it available to everyone. Accessibility can be defined as "usability of a product, service, environment or facility by people with the widest range of capabilities" [61]. Accessibility is tightly connected to OER since SDG4 concerns equity and ensuring education for everyone, especially including people with disabilities [62]. Although accessibility is crucial for OER and mentioned in many frameworks, it is not included in the quality control in major repositories for OERs today [62]. It was therefore included as a feature score.

One of the most widely accepted standards is WCAG 2.0, created by Web Accessibility Initiative (WAI) [62]. The next version WCAG 3.0, which is still a draft, presents an accessibility scoring model [63]. Each element passes or fails a test, and the total rating is based on the percentage of passes. The ratings were transformed into a score between 0 and 1 to become a score with the result presented below 3.1. It was impossible to measure critical errors with the available metrics, so the scoring was based solely on the alternative text. The percentage of alternative text was found by extracting all images and interactions with images and checking that the field for alt-text was not empty. 3.1

Rating	Criteria	Score
0	Less than 60% of all images have appropriate text alternatives OR there is a critical error in the process	0
1	60% - 69% of all images have appropriate text alternatives AND no critical errors in the process	0.25
2	70%-79% of all images have appropriate text alternatives AND no critical errors in the process	0.5
3	80%-94% of all images have appropriate text alternatives AND no critical errors in the process	0.75
4	95% to 100% of all images have appropriate text alternatives AND no critical errors in the process.	1

Table 3.1: WCAG scores [63]

3.3.3 Video analysis

The algorithm was implemented for Interactive Video, and therefore, the video and its metadata were available metrics. In an Interactive Video created with H5P, the video can be from YouTube or uploaded. To be able to evaluate all resources, youtube-dl [58] was used to download the YouTube videos.

The video itself can be considered an essential part of the content of an interactive video. Both *presentation of the content* and in general *content* are part of many quality frameworks (see 2.2). Therefore the viideo algorithm from Mittal *et al.* [64] was used to get a quality score. This algorithm measures video distortion without reference videos, thus fitting the use case. The python implementation from sk-video [56] was used. Because of the computational constraints, the parameter was the first 2000 frames of each video. In the returned scores, a higher number indicated high distortion and hence worse quality. Therefore, the score was reversed by taking $1 - score$ to use the same scale as the other feature scores.

3.3.4 Interactions

In an Interactive Video, different interactions can be added. These interactions can be expositive like images, text elements, links, shapes, tables, active like tasks, or navigational like bookmarks and or navigational links. They will then be shown at the specified time.

These interactions make the video interactive and, therefore, an integral part of the resource. As shown in 2.2 interactivity is part of several frameworks, and results suggest that active learning can give better results than passive [65]. Also, empirical results support this by the number of scripts and applets being correlated with high-quality resources [13].

Clear navigation is a factor in Mayrberger *et al.* [33] and media is the most important factor in the opinion of teachers [9]. For media, both Cechinel *et al.* [13] and Cechinel *et al.* [11] showed that this is a metric correlated to high rated resources. Other affiliated metrics like the size of images were also associated.

Although the research supports the importance of these interactions metric correlated with quality, it is unclear if the quality keeps increasing with more elements. Therefore, different scoring models were implemented to find the best-suited and for further analysis.

1. $Num = NumberOfResources$
2. $Time = NumberOfResources \div resourceLength$
3. $Decreasing_5 =$ A decreasing score added for first 5 elements following:
[0.35, 0.25, 0.2, 0.15, 0.05]
4. $Decreasing_{10} =$ A decreasing score added for first 10 elements following:
[0.20, 0.20, 0.15, 0.10, 0.10, 0.075, 0.075, 0.05, 0.025, 0.025]
5. $Combination_5 = Decreasing_5 + 7 * (NumberOfResources - 5) \div resourceLength$

The rationale for scoring models 2 and 3 is that the interactions are important quality factors [8, 13, 25, 33], but nothing indicates that the value increases with

increasing number after a certain point. The specific numbers were created by giving highest importance to the first resources and lowering the importance with specified intervals and the sum adding up to 1. Scoring model 5 is an experimental test to see if it can be valuable to combine approaches 2 and 3.

For each of the different elements, these five scoring models were calculated as well as for collections of elements. In the initially tested algorithm, images, navigation, bookmarks, text, tasks, and all combined were implemented using scoring model two: *Decreasing₅*. The other scoring models were used in the data analysis studying correlated metrics.

3.3.5 Text analysis

To do the text analysis all text elements were extracted from the different interactions, not only the actual text elements. They were all added to a list and for each text element a grammatical/spelling score and the Flesch reading easiness score [66]. For the Flesch reading score the implementation from [59] were used. The Flesch score gives an indication of how hard a text is to read and therefore which levels it is suitable for. It was not included in the algorithm, because resource level were seldom stated in the metadata, but it was included for the metric correlation tests. The average from the Flesch score on all text elements were used.

To find the errors in the text *language_tool* [60] were used and the number of error for each text were calculated. To get a relative score the numbers of errors were divided on the number of words. The average of this were taken and then normalized. To get a score where 1 was the least errors this score was subtracted from 1 and the result can be seen in equation 3.3.

$$error_score = 1 - normalised\left(\sum_{i=1}^{num(text)} \frac{1}{n} * \frac{num(errors)}{num(words)}\right) \quad (3.3)$$

3.4 One score

The proposed feature scores needed to be combined into one overall quality score. Before that, all scores were normalized using min-max normalization, then selected and added. Two versions of the algorithm were therefore proposed. One was giving a weight of the each feature based on how often it was mentioned in quality frameworks. Each occurrence in a framework were counted and the sum was divided by the number of frameworks as shown in equation 3.4. The result were then normalized by dividing on the sum of importance to give a score between 0 and 1. The weights are shown in table 3.2.

$$alg_{weighted} = \sum_{i=1}^{num(feature_scores)} feature_score * importance(k) \quad (3.4)$$

Feature	Scoring model	Importance
Metadata	avail_score	2/5
Accessibility	WCAG scoring model	4/5
Video	Viideo score	5/5
Tasks	<i>Decreasing</i> ₅	3/5
Media	<i>Decreasing</i> ₅	1/5
Bookmarks	<i>Decreasing</i> ₅	1/10
Navigation	<i>Decreasing</i> ₅	1/10
Text elements	<i>Decreasing</i> ₅	2/5
Combination_all	<i>Decreasing</i> ₅	1/5
Text_analysis_all	Reversed error-average	4/5

Table 3.2: Importance weights used in $alg_{weighted}$

Secondly an unweighted algorithm was proposed with equal weighted on all factors. During testing it was found that the video score contributed negatively to the correlation and a third unweighted version without this were therefore created. There was not enough reviews to do a regression with parts of the data to find weights.

The metrics that were not part of the proposed algorithms were also calculated for the resources. They were thereafter used to do a detailed correlation analysis that could contribute to future work.

Chapter 4

Methodology

This chapter will present the research method and research design created to answer the research questions.

4.1 Research method

This thesis aims to compare a manual and an automated quality assessment method. To find their level of agreement they had to assess the quality of a sample of collected OERs.

The manual reviews could either be gathered from existing sources like Merlot/SkillCommons [6, 11–13, 28, 37, 44], Connexions [11], or be collected for the research itself [10, 19]. With the choice of resource type, no evaluations were conducted, so the data needed to be generated for the research. Although gathering the data is more time- and resource-consuming it enables evaluation of the reviews reliability.

To perform the manual assessment a set of participants had to be recruited and their quality assessment of resources collected. Scoring different quality criteria is a well-adopted approach to assess the overall quality and was therefore chosen [29]. A survey was determined as the best method to efficiently collect this data. Additionally, an open-ended question was included to explain the numeric results and collect more data on perceived quality.

For the automated approach, the data was collected by running the algorithm with the set of resources. The result was then a quality score for each resource.

The chosen approach was, thus, a mixed method, utilizing both qualitative and quantitative methods. This method allowed for easy analysis and simultaneously more insight into quality perception.

4.2 Research design

A preliminary literature review on OER quality resulted in the Research Questions presented in Chapter 1. To answer these, the research design consisted of a dif-

ferent phases. As described in chapter 3 an algorithm was developed based on the relevant literature. Then the a survey using a mixed method approach were conducted. Using the results from the algorithm and the survey a data analysis were performed.

Before developing the algorithm the target resources had to be selected. In comparable studies like Cechinel *et al.* [11] the resource metrics have been collected using a crawler. To avoid these limitations a resource type with more available metrics was chosen. The result was the H5P resource that is rendered based on a JSON which makes metrics extraction easy and reliable. Another advantage is that H5P have a set of predefined resource types. This aligns well with the proposal from Cechinel *et al.* [13] that different profiles should be created for different resources.

Given the limited number of reviews the algorithm was developed for one single resource type. The resource type needed to have enough resource from different authors to ensure heterogeneity. It should also have a wide range of metrics, granularity level of 3 [19], meaning it consists of sub-resources. Based on these criteria, Interactive Video [67] was chosen.

4.2.1 Data collection

The first data that had to be collected was the OERs. A weaknesses with sampling already reviewed resources in Merlot[4] is that only resources of a certain quality are peer-reviewed. When using these reviews the "poor" resources might not be the actual low quality resources[13]. To overcome this weakness the sampled resources should as heterogeneous as possible and chosen randomly. The resources were collected from four different sources; H5P Hub [38], LibreStudio [68], Catalogue[69], Ontario [70]. A maximum of two resources from the same author were collected.

To obtain as valuable data as possible, it was decided to obtain the reviews from either two groups: OER experts or a user group. Since these groups are stakeholders [5] they are qualified to assess the "*fitness for the purpose*", in contrast to the general public. It was not possible to find experts to conduct the reviews within the time frame so a user group was selected.

The participants came from a group of beta-testers in the H5P community. This group tests typically new features, and have experience using educational resources created with H5P. It was chosen because of their general interest in developing H5P, their experience with educational resources, and since they could (or are) potential users of an OER repository. Although many of them are experienced users, there are no requirements to be part of this group, so no assumptions about their competence can be made.

To recruit the subjects, an invitation email was sent to the group to inform them about the project and invite them to take part. Since name and email address had to be collected for the review process NSD (Norwegian centre for research data) was notified and all participants signed a consent form (see appendix B).

The survey was conducted using the tool Nettskjema.no that provides secure data handling.

To provide flexibility and get the highest number of responses possible, the participants choose how many resources they wanted to review. Each participants were then assigned resources from different repositories in a randomized manner. To evaluate the reliability of the reviews each resource were assigned to at least two participants.

4.2.2 Survey

To observe differences between reviewers of different backgrounds and expertise, the surveys started by collecting some demographic data. As Cechinel and Sánchez-Alonsor [28] showed, there is little correlation between the peer reviews and the user reviews, and the background is therefore relevant for the result and validity of the reviews. Clements and Pawlowski [9] also found that the perceptions of quality differ in different groups. Relevant occupation groups were therefore selected from ISCO-08 [71] and *creator of educational content* added additionally. Other sampled questions were age, gender, educational background, and experience with Open Educational Resources. They were also asked for their specific experience with OER, review of OER, and experience with H5P. The whole survey can be seen in appendix D.

Since the competence of the participants were unknown, this had to be reflected in the selection of survey questions. To conduct enough reviews, it was impossible to do any training, and all questions needed to be self-explanatory. Without knowing either the resource topic or the reviewer's experience, the resource facts, standard compliance or accessibility could not be evaluated.

There are no adopted standard for OER evaluation. Chapter 2 presented many frameworks and instruments like LORI [30], IQOER [29], Quality4OER [31], and TIPS [25]. However, they are too time-consuming in their whole format and require expertise on OER standards and accessibility [27]. Some of the questions were also not suited for this use case. The solution was to use questions from different sources representing the most frequent mentioned quality aspects and modify some of them to simplify the process.

The factors were selected based on the comparison of different review tools in Chapter 2. The most mentioned pedagogical factors were presentation, alignment, content, student support, instructional design, interactivity and engagement. From these student support were excluded since it assumes an actual use case and instructional design because of the lack of expertise. From the technical factors usability, accessibility, interoperability, reliability, licence, reusability and metadata were the most frequently mentioned. Accessibility and licence were excluded based on competence and reliability and interoperability because those are factors the authors can not control within H5P. Based on these essential aspects, formulations were selected from the frameworks. Some of them were modified before use to ensure they were suited for an audience without any review exper-

ience.

This gave the factors; presentation, alignment, content, engagement, interactivity, usability, reusability and metadata. In addition two overall questions were added. Q11 measure the Net Promoter Score [72], while Q12 let the user give their overall score. All the questions are shown in table 4.1

For all questions, a Likert scale from 1 to 7 were used, where 1 means completely disagrees while 7 means completely agrees. The 7 point scale were chosen because it has a greater potential of giving more reliable results compared to a 5 point [73].

4.2.3 Data analysis

For the data analysis some general practices were used. The results were considered significant if the p-value were less than 0.05 which is a general practice and for some analysis only those were reported. For the data analysis three types of question data were used. One was the use of one question directly. This was to do a more detailed analysis of a certain quality factor. Then two different averages were used; Q1-Q10 and Q1-Q12. The reason for testing both these averages is because the first one only takes into account the specified metrics while Q1-Q12 also include two solely subjective quality factors and give a different insight. Often Q12 was compared with those two average because this metric were less effected by the questions and quality factors used in this study and therefore more transferable to other research. Many of the analysis were correlation tests and for that Pearson was generally preferred because it measures the linear relation. Since much of the data did not satisfy a normality test Kendall τ_b [74] were in many cases used instead. It measures how strongly two variables are monotonously related. Since these two correlation test do not measure the same, results from two different test were not directly compared. All correlation test were made using `cor.test()` in R.

To answer the research questions the data analysis were two-fold. One part was analyzing the reviews to assess it's reliability and the other comparing it to to the algorithm results.

4.2.4 Review data

The reliability of the reviews can be divided into three categorized; survey questions, repeatability and background influence.

A way to test the reliability of the survey question is to assess the test's internal consistency. This can estimate if a test jointly measure the same construct [75], here the OER quality. The Coefficient Alpha [76] was calculated using the implementation in the R package LTM [77].

Another aspect of the survey questions are how correlated the different metrics are with the overall questions (Q11, Q12). The correlation between the different factors and the overall quality can give insight to the factors perceived as most important by the reviewers. The correlation between the Q1-Q10 and Q11 and

Question	Based on	Category
Q1. The intended learning outcomes are made clear to the learners. The content, learning activities, tasks, and assessment presented are consistent with these learning outcomes.	IQOER: Instructional design/alignment [33]	Alignment
Q2. The resource include a variety of self-assessments such as multiple-choice, concept questions, and comprehension tests.	C-23 [25] Kawachi	Interactivity
Q3. The resource contains interactive elements that can be used by the learners to independently perform constructive or manipulative actions.	IQOR: Usability/interactivity [33]	Interactivity
Q4. The structure is simple and clear. Learners can stop the learning sequence at any time. All learning content (previously presented) can be accessed at any time.	IQOR: Usability/-structure [33]	Usability
Q5. All texts and graphics are easy to read. The interface always responds quickly to learner input.	IQOR: Usability/design and readability [33]	Presentation
Q6. The interactions are understandable and easy to use.	6. Interaction Usability LORI [30]	Usability
Q7. The metadata allows others to effectively use that information to search and evaluate the resource's relevance.	9. Standards compliance [6]	Metadata
Q8. The contents are presented in such a generic way that they can be used in other contexts without much effort.	Content/reusability of content [33]	Reusability
Q9. The visual and auditory information is presented in a clear, concise, and coherent way, taking care with sound quality.	C-44 [25]	Presentation
Q10. The resource motivates and is able to hold learners's interest.	Engagement Graphite [6]	Engagement
Q11. I am likely to recommend this resource to a friend or colleague.	Net Promoter Score [72]	Overall
Q12. This is a high quality learning resource.		Overall

Table 4.1: Survey questions

Q12 were measured using Kendall τ_b [74] (Shapiro-Wilk: Q12: $p=3.65e-05$, Q11: $p=1.61e-05$). Additionally a Bland Altman [78] plot was used to assess the level of agreement between the overall of metric Q1-Q10 and Q12. The mean, median and standard deviation was also calculated for all metrics and averages to give show the trends for the questions.

Repeatability can be defined as "*The degree to which the same method produces the same results on repeated measurements*" [15]. In this case the measurement is manual reviews and the topic investigated is repeating reviews of the same resource with a different person. Since quality is a subjective matter a full agreement is not expected, but the level of agreement can indicate how representative one review is for the general opinion and the level of subjectivity. It is important to note that there is not two distinctive groups of reviewers that share any characteristics. The relevant difference is only between two reviews on each resource.

To test this level of agreement the correlation between the two reviewers were tested for all metrics and their average. Because the metrics distributions did not satisfy the test of normality (as seen in table C.1) Kendall τ_b [74] test was used. For the averages a Person's r were also used since these satisfied the test of normality (see C.2). Thereafter an one-way Intraclass correlation [79] test were conducted as used in similar research earlier [19]. The implementation from irr[80] were used measuring "*agreement*". A Bland Altman plot [81] is a way to measure the level of agreement and was also used.

The third factor tested was the influence was the background of the reviewers. The demographic data collected were age, occupation, education, OER experience, H5P experience and reviewer experience. Gender and reviewing experience were analysed because they could be split into only two groups with enough samples in each group. Because of lack of normality in each groups a the R implementation (wilcox.test()) of Wilcox un-paired test were used [74, 82]. Additionally a χ^2 test were conducted by dividing the score into *low* and *high* by the median.

In addition to the ordinal metrics the participants could add a textual comment. All of the comments were read and were first used to understand the differences between reviewers on the same resources. After analysing the comments three hypothesis for why some resources had a high difference (>3 (50%) for Q12 or average Q1-Q10) score were created. Comments could fit multiple explanations and categorized as more than one. The three hypothesis were that the participants considered different factors, disagreeing on the scale or that one of the reviewers did not understand a specific resource.

Secondly the comments were used to find which factors were most important for the participants. Here each comments were read and counted for each factor it mentioned. The factors mentioned more than 4 times were presented in the results.

4.2.5 Algorithm testing

The three versions of the algorithm presented in Chapter 3 were all tested against average(Q1-Q10), average(Q1-Q12) and Q12. Only resources having at least two reviewers were used and the average between the two reviews were used.

For all comparisons, correlation test were conducted using Kendall's τ_b , [74]. Correlation is not sufficient by itself [81] and Bland Altman was therefore also used [78].

On hypothesis was that the resources with most divergent algorithm and review result had disagreeing reviewers. This hypothesis was tested by splitting the resources into quartiles based on the absolute value of the difference between algorithm and average(Q1-Q12). The highest and lowest quartiles were then compared. The comparison value was the disagreement between reviewers. This was tested with a two-sided t-test using the implementation in R.

In addition to testing the algorithms proposed in this paper the metadata model proposed in Tavakoli *et al.* [12] were tested. It was tested to find how new data affect a result. This is a classification algorithm so the resources were divided into *high* and *low* quality. Two classification were tested; dividing by higher or lower than median and by a fixed value of 4 (median value of scale). The confusion matrix, f-value and accuracy scores were then compared by the results in Tavakoli *et al.* [12].

Because there were a limited number of reviews all of these were used to test the proposed algorithm, rather than using parts to make a prediction model. The risk would be to overfit the model, rather than testing a generalized result. To provide insight into some other metrics could be more suitable than the ones used in the algorithm correlation tests were finally conducted. These can provide details to future work within the field.

Chapter 5

Results

In this chapter, the results will be presented. It will start with RQ2 concerning the reviews' reliability and then the result of comparing those to the algorithm.

5.1 Reliability of reviews

A total of 26 participants were recruited to perform reviews, and 23 completed. The result was 107 reviews conducted. The goal was to get all of the resources evaluated at least twice. Because some participants could not conduct the reviews assigned, the result was one resource reviewed thrice, 48 twice, and eight once.

5.1.1 Reliability of survey questions

First, a reliability test was performed for the 12 questions resulting in a Cronbach's alpha of 0.93. This result is considered high reliability by being above both the generally accepted cutoff value is .80 and .90 better for important decisions [75].

The second aspect of the survey questions was how correlated the metrics are with the overall quality and this is shown in table 5.1. The two overall questions Q11 and Q12 had a strong association ($\tau=.82$, $p=2.2e^{-16}$). They have a mean difference of 0.06 with Q11 being the highest on average. For the other metrics the most correlated with the overall quality (Q11 and Q12) are Q9 (Q12: $\tau=.61$, $p=1.46e-15$, Q11: $\tau=.58$, $p=1.32e-14$) and Q10 (Q12: $\tau=.64$, $p=2.2e-16$, Q11: $\tau=.69$, $p=2.2e-16$) measuring presentation and engagement. The three lowest still showed a moderate positive association and were Q2 (Q12: $\tau=.34$, $p=7.3e-06$, Q11: $\tau=.41$, $p=3.60e-08$), Q3 (Q12: $\tau=.38$, $p=5.06e-07$, Q11: $\tau=.39$, $p=2.39e-07$) and Q7 (Q12: $\tau=.39$, $p=2.87e-07$, Q11: $\tau=.42$, $p=2.44e-08$) that measuring interactivity(Q2,Q3) and metadata(Q7). The average of Q1-Q10 had the highest association with the overall (Q12: $\tau=.68$, $p=2.2e-16$, Q11: $\tau=.70$, $p=2.2e-16$).

In figure 5.1 the resources are plotted with scores from average(Q1-Q10) being the x-value, while Q12 the y-value. Additionally a line shows a fitted linear model displaying a linear trend.

Metric	τ Q12	P-value Q12	τ Q11	P-value Q11
Q1	0.55	1.48e-13	0.56	6.46e-14
Q2	0.34	7.3e-06	0.41	3.60e-08
Q3	0.38	5.06e-07	0.39	2.39e-07
Q4	0.51	2.49e-11	0.48	3.38e-10
Q5	0.57	7.47e-14	0.57	6.38e-14
Q6	0.45	1.93e-09	0.48	1.62e-10
Q7	0.39	2.87e-07	0.42	2.44e-08
Q8	0.53	2.47e-12	0.53	1.70e-12
Q9	0.61	1.46e-15	0.58	1.32e-14
Q10	0.64	2.2e-16	0.69	2.2e-16
Q11	0.82	2.2e-16		
Average Q1-Q10	0.68	2.2e-16	0.70	2.2e-16

Table 5.1: Correlation between metrics and the overall questions Q11 and Q12

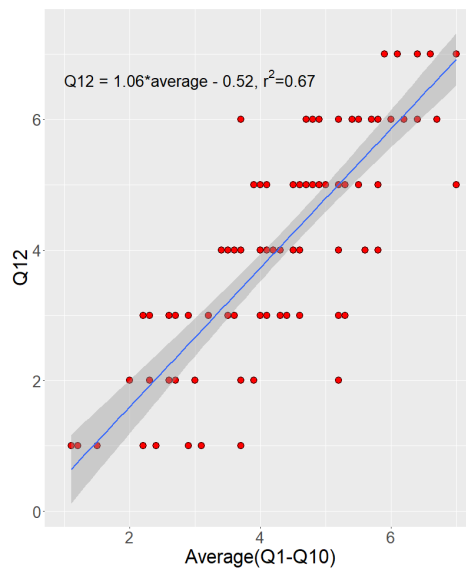


Figure 5.1: Correlation between average(Q1-Q10) and Q12

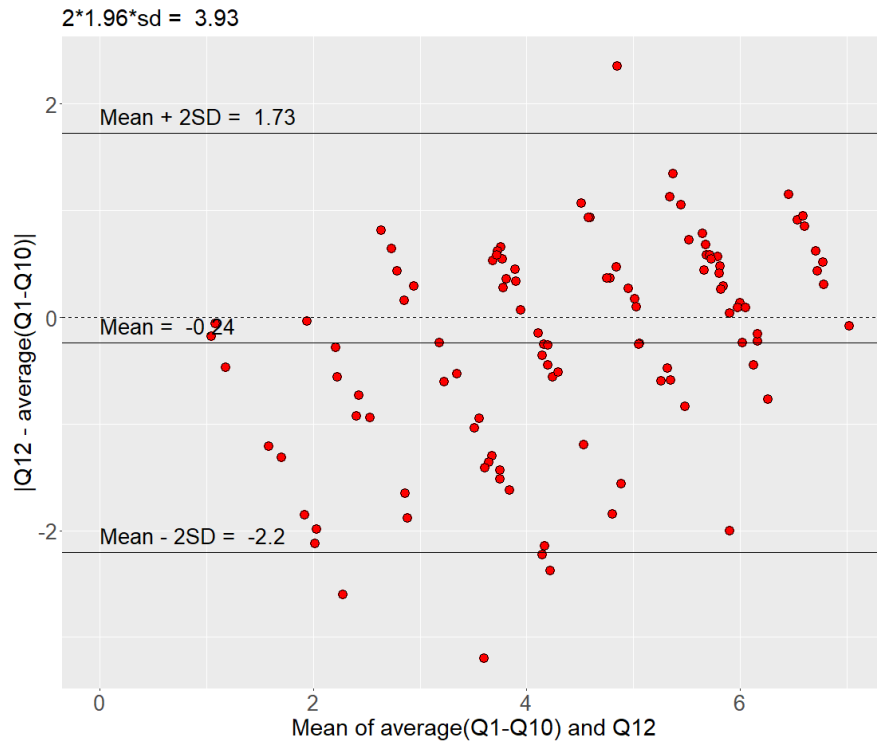


Figure 5.2: Bland Altman plot of Q12 and average(Q1-Q10)

To measure the degree of agreement between the average (Q1-Q10) and Q12, a Bland Altman [78] plot was also used. Figure 5.2 shows that the average difference between average(Q1-Q10) and Q12 is -0.24. This means that the overall score is generally 4% lower than the average(Q1-Q10). 95% of the points fall between -2.20 and 1.73.

5.1.2 Distributions

In 5.3 the distributions of the average (Q1-Q10) and the overall score are presented, which have quite different distributions. As seen in the distribution and also tested by the Shapiro-Wilk normality test, the a normal distribution can not be excluded for average(Q1-Q10) ($p\text{-value} = 0.064 > 0.05$), while Q12 is non-normal ($p=3.65e-05$).

An interesting observation is also the distributions of the other metrics shown in figure 5.4. They all have a non-normal distribution with p -values of the Shapiro-Wilk test in table C.1. In table 5.2 the mean, median, and standard deviation are presented for all the metrics. Q9 (presentation) and Q4 (usability) have the highest means and medians (Q9: 5.07, 6, Q4: 5.02, 6). The lowest values are found for Q2 (interactivity) (3.84, 4), Q3 (interactivity) (3.88, 4), and Q7 (metadata) (3.87, 4). Q2, Q3, Q7, Q11, and Q12 have a median of 4, the same as the median

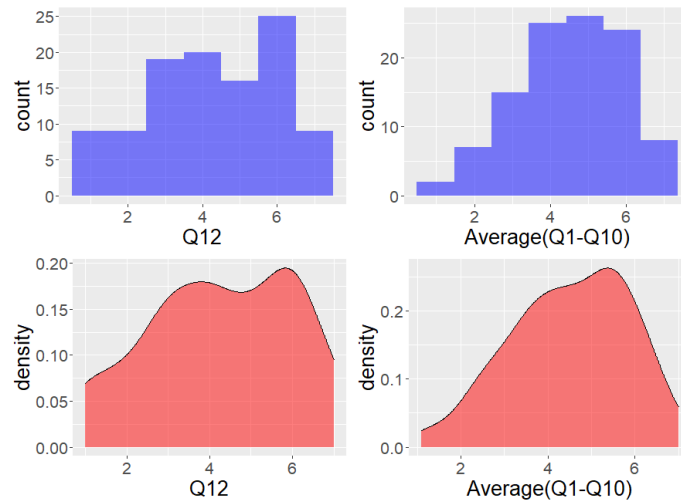


Figure 5.3: Distributions average(Q1-Q10) and overall

of the scale from 1 to 7.

5.1.3 Repeatability

The repeatability test can be split into averages, overall, and all metrics. These metrics have different use cases making making the each comparison nessecarly. Testing the repeatability can also be seen as a test of how subjective the reviewers are.

The average of different sets of questions for the two recipients was compared, and no significant correlation (Q1-Q10: $r=.11$ $p=.42$, Q1-Q12: $r=.10$ $p=0.47$) was found. In figure 5.5 the points are plotted with the reviewers on each axis, and it is hard to see any clear correlation.

Another way to test the agreement is by Bland-Altman analysis. In figure 5.6 the difference between the average(Q1-Q12) for the two participants is plotted. Since there are not two specific groups of reviewers, the absolute value of the difference is used. The mean difference is 1.53 with 95% of the samples within 0 and 3.46 ($1.96 * sd$) since the absolute of the difference cannot be less than 0. These results account for 26% of the scale (1-7) for the mean difference and 57.7% difference for the 95%-interval.

In figure 5.7 the scores on Q12 and Q11 for the two reviewers are plotted and a linear regression line added. It can be observed that the values have a high spread and many elements with high differences. No significant correlation was found for either Q11 or Q12 ($p=.56$, $p=.87$). Out of 49 resources (reviewed by at least two), only 5 (10%) had the same score for Q12 by both reviewers. For Q11, even fewer (3.8%) had the same score.

A Bland Altman plot was also made for Q12 and is shown in figure 5.8. This shows that the mean absolute difference was 2.04, and 95% of the samples had

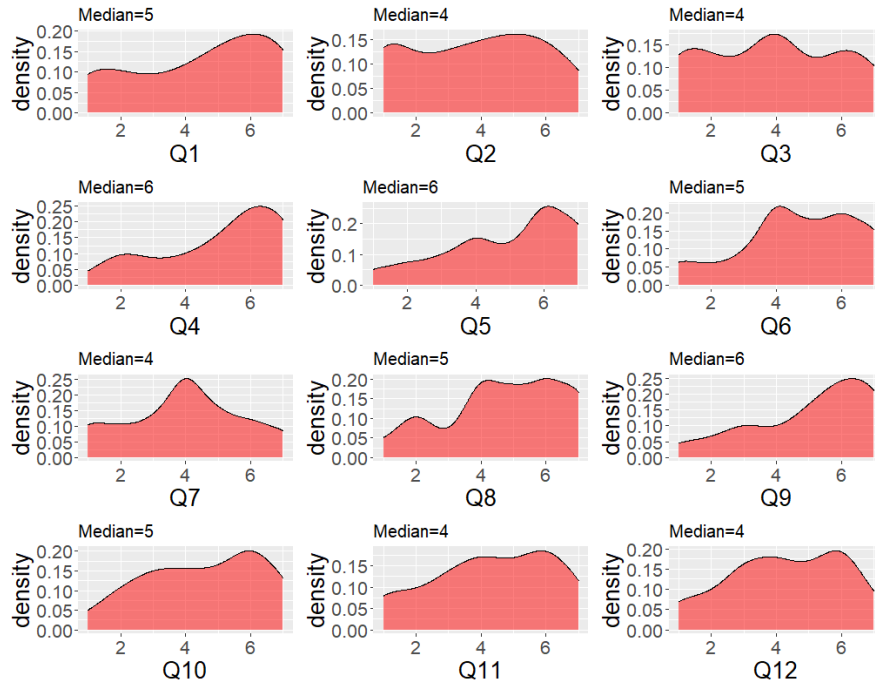


Figure 5.4: Distributions metrics

Metric	Mean	Median	SD
Q1	4.49	5.00	2.06
Q2	3.84	4.00	1.99
Q3	3.88	4.00	2.02
Q4	5.02	6.00	1.84
Q5	4.91	5.00	1.81
Q6	4.66	5.00	1.75
Q7	3.98	4.00	1.77
Q8	4.72	5.00	1.77
Q9	5.07	6.00	1.82
Q10	4.51	5.00	1.77
Q11	4.33	4.00	1.85
Q12	4.27	4.00	1.75
Average Q1-Q10	4.51	4.00	1.35
Average Q1-Q11	4.49	4.55	1.37
Average Q1-Q12	4.47	4.50	1.39

Table 5.2: Mean, median and standard deviation for metrics

Metric	Kendall's τ	P-value	Pearson's r	P-value
Q1	0.02	0.86		
Q2	0.28	0.01		
Q3	0.20	0.07		
Q4	-0.07	0.54		
Q5	0.11	0.35		
Q6	0.10	0.38		
Q7	-0.12	0.27		
Q8	-0.18	0.11		
Q9	0.12	0.29		
Q10	0.02	0.83		
Q11	0.07	0.56		
Q12	-0.02	0.87		
Average Q1-Q10	0.06	0.53	0.12	0.42
Average Q1-Q11	0.05	0.59	0.11	0.44
Average Q1-Q12	0.04	0.70	0.11	0.48

Table 5.3: Correlation of metrics between two reviewers

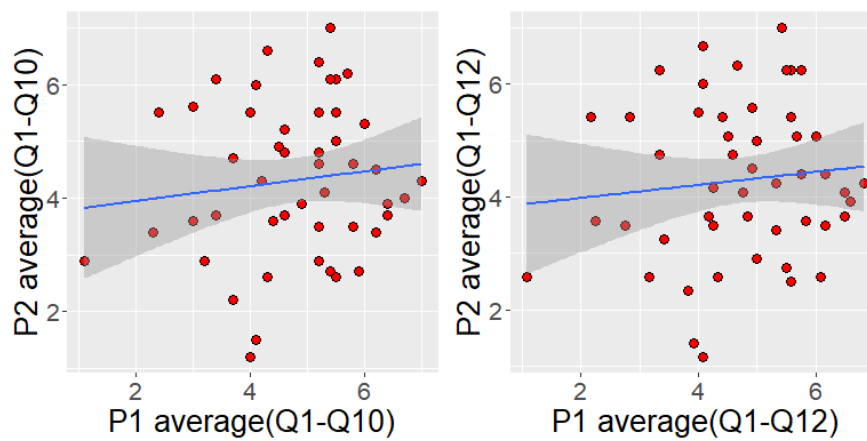


Figure 5.5: Correlations plot for averages for each reviewer

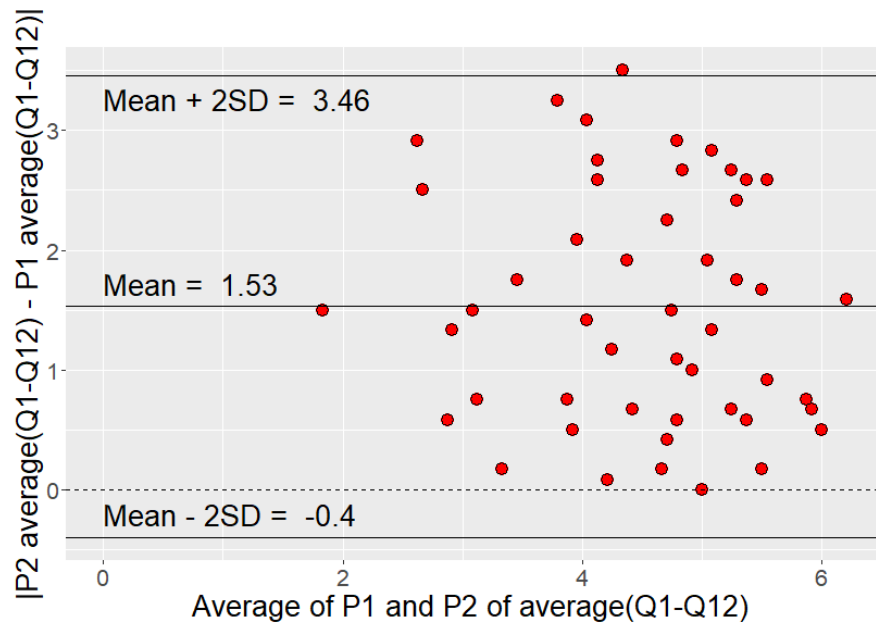


Figure 5.6: Bland Altman plot for agreement between reviewers for average(Q1-Q12)

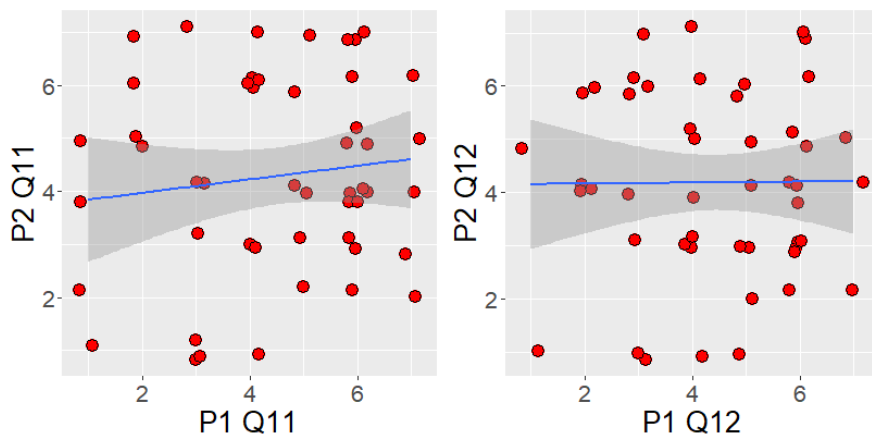


Figure 5.7: Correlations plot for Q11 and Q12 for each reviewer(jitter of 0.2 is used to show overlapping points)

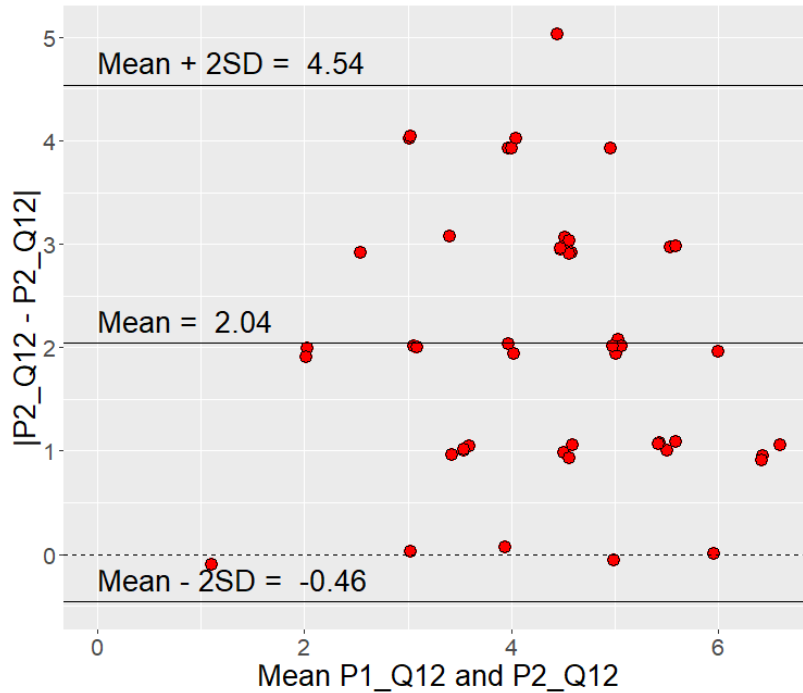


Figure 5.8: Bland Altman agreement between reviewers on Q12. (Jitter is used to display the overlapping points)

between 0 and 4.54 differences (absolute value is always >0). Given that the maximal difference is 6 (1-7) scale, the average difference is 34% and the 95% interval 75.6%. This is 8% and 17% higher than the difference between the reviewers for average(Q1-Q12).

Correlation tests were also run for the other questions to find the most agreed on and can be seen in table 5.3. Only one had a significant results ($p\text{-value} < 0.05$) and that was Q2 ($\tau = 0.28, p = 0.01$) measuring the interactivity of the resource. Q3 which also measures interactivity was the one closest to a significant value ($\tau = 0.20, p = 0.07$).

Additionally, a one-way Intraclass Correlation test was conducted. The result was $ICC = .09$ with a $p\text{-value}$ of $.027$, which is considered poor reliability [83].

5.1.4 Background

The reviewers' background is a factor that could potentially impact the reviews, as seen in earlier research [9, 27, 28]. The sample consisted of 13 females and 10 males. The median age group was 45-54, and the full distribution can be seen in table 5.4. The most common educational background was a Master's degree (see table 5.5) and all but 1 had some degree. The most common occupation was *creator of educational material* (34.5%) followed by *other teaching professional* (26.1%).

Age group	N	Percentage
25-34	1	4.35%
35-34	3	13.0%
45-54	10	43.5%
55-64	7	30.4%
65-74	2	8.70%
Sum	23	100%

Table 5.4: Age distribution

Completed education	N	Percentage
Some college, no degree	1	4.35%
Associate degree (e.g. AA, AS)	2	8.70%
Bachelor's degree (e.g. BA, BS)	8	34.8 %
Master's degree (e.g. MA, MS, MEd)	10	43.5 %
Doctorate or professional degree (e.g. MD, DDS, PhD)	2	8.70 %
Sum	23	100%

Table 5.5: Educational background

91 % of participants stated that they had experience with OER, and the mean years of experience were 11.57 and the median 10 years. The most common way they had used it was as a creator (78.3%) and as a end user (56.5%).

All participants reported experience with H5P. The average years of experience were 3.65, and the median was 3. Same as for OER, the most common ways they had used H5P were as a creator (95.7%), but following were reusing resources (43.5%) and sharing resources (43.5%).

When it comes to the experience with reviewing resources 8 (34.8%) of the participants had some form of experience. The average number of reviews conducted (for the group with experience) was 102.88, median 74.5 and standard deviation of 105.99.

One of the potential influencing parts of the reviewers' background was their

Occupation group	N	Percentage
Creator of educational material	8	34.8 %
Other teaching professional	6	26.1%
Information and communications technology professional	3	13.0 %
Other	2	8.70%
University and higher education teacher	2	8.70%
Secondary education teacher	1	4.35%
Prefer not to answer	1	4.35%
Sum	23	100%

Table 5.6: Occupation distribution

OER use	N	Percentage
Creator	18	78.3%
Reused resource(s)	12	52.2%
End user (learner)	13	56.5 %
Shared resource(s)	11	47.8%
Other	5	21.7%

Table 5.7: OER use

H5P use	N	Percentage
Creator	22	95.7%
Reused resource(s)	10	43.5%
End user (learner)	6	26.1 %
Shared resource(s)	10	43.5%
Other	5	21.7%

Table 5.8: H5P use

experience with reviewing. 34.8% of the participants had experience with reviewing, and someone with experience reviewed 30.84% of the resources. This gave a disbalance between the two groups. Wilcoxon un-paired test were used to compare the average for each question and the average of the questions for each group. Only Q5 had a significant ($p < 0.05$) result ($W = 823.5$, $p = 0.006288$) which measures the presentation of the resource. None of the others had a p -value less than 0.10.

Performing a χ^2 test showed the same result, with Q5 being the only one with a significant result ($X\text{-squared} = 13.42$, $df = 1$, $p = 0.0002$). The distribution of a high and low score on Q5 for the two groups can be seen in figure 5.9. It shows that experienced reviewers gave a *low score* more often than the other group.

The second background groups that were studied were the genders. Two groups were created with *females* and *males* having 13(56.52%) and 10(43.48) participants in each. This resulted in 61 resources (57.01%) for the female group and 46 resources (43.00%) for the males. Wilcoxon's un-paired test for all metrics showed a significant difference for Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q8, Average Q1-Q12, Average Q1-Q11, and Average Q1-Q12. The values for the average(Q1-Q12) was $W = 1899$, $p\text{-value} = 0.002$. In Figure 5.10 a boxplot of the average(Q1-Q12) group after gender is shown. While the standard deviation is quite similar (female: 1.39, male: 1.17), the average of the female group was 4.84 and 4.07 for the male group.

A χ^2 test had significant results for Q3(10.2,0.001) , Q4(6.61,0.01), Q8 (7.04, 0.008), Average Q1-Q10 (6.61,0.01), Average Q1-Q11(6.61, 0.01) and Average Q1-Q12 (7.80, 0.005). In figure 5.11 it is shown that the resources reviewed by males had a greater proportion of "low" rating than the resources reviewed by females. 65.21% of the male reviewed resources had a *low score*, while for the



Figure 5.9: Distribution of high and low score (Q5) divided by review experience



Figure 5.10: Boxplot gender differences average(Q1-Q12), (Jitter used to show overlapping points)

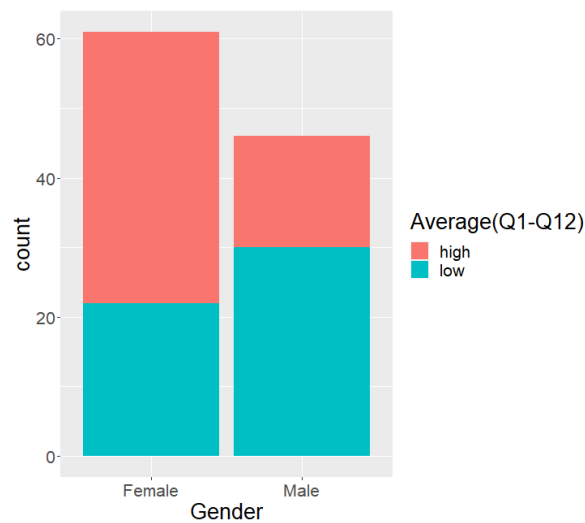


Figure 5.11: Distribution of high and low score (Average Q1-Q12) divided by genders

resources reviewed by females, the percentage was 36.07%.

5.1.5 Comments received

In addition to the question, the users could add a comment. 14 reviewers commented on all resources they reviewed, five on some and four on none. The result was a comment for 69 (64.49%) of the resources.

18 (36.73%) of the resources satisfied had a difference between the reviewers for Q12 or average(Q1-Q10) of 3 or higher. The comments were used to investigate potential reasons. Twelve of these resource had comments from both reviewers and 16 from at least one of the reviewers. One hypothesis was that the reviewers consider different factors to make a high-quality resource. These are characterized by the two reviewers commenting on different factors. An example of this is this is a resource that got a score of 3 and 6 on Q12. The review with the lowest score had the following: *"Keywords were lacking. The intro was pretty general about the content. Mostly this is a narrated PowerPoint, it's not very engaging. I didn't think the question type fit the content for the fill in the blank. Perhaps matching would have been better."* The review with a score of 6, on the other hand, had this: *"At the beginning of video, to say what to do when the purple circle pops up!"*. Here the one giving a lower score possibly judge the resource on different factors. Nine of the resources can be characterized within this category.

Another hypothesis is that the reviewers mainly agree on which factors make a high-quality resource but disagree on the scale. That means that one of the reviewers is more "strict" in evaluating. These comments are characterized by the one scoring highest mentioning faults but still giving it a higher score. An example

of this is shown in a resource where both mentioned the lack of interactivity. P1 wrote *"There were no interactions even when this would be useful (e.g. hyperlinks). A lot of information was given without testing understanding. I would have liked to have been able to turn the subtitles off."* P2 wrote: *"It might be more motivating to add a free text box to ask users their existing/prior methods of using references to collect that info. I'm not sure about the claim that all scholarly published papers are 'credible' but this may be the norm for that particular institution. It would be useful to add examples of what journals/articles look like in the library search facility."* Both reviewers wrote about the lack of interactions, but P1 gave a score of 2 on Q12 while P2 gave a score of 6. 6 resources fit into this category. A third hypothesis was that one of the reviewers did not understand how the resources were to be used, therefore giving a low score. An example is this comment: *"I miss the content. what are we going to learn? what do we need? calculator? show the elaborated solution as feedback / solution.."*. A total of 5 resources were categorized as lack of understanding.

Another aspect was the high difference between the average of the selected metrics (Q1-Q10) and the overall score given on Q12. 9 resources had an absolute difference higher than 2. A possible explanation for this could be metrics not covered by the question Q1-Q10. An example of this was a resource mentioning accessibility.

In addition to using the comments to explain differences between reviewers or between average and overall scores, they can give some general insight into how resources can be improved. The most mentioned issue with the resources was the complete lack of or few interactive elements/tasks. This was mentioned for 16 of the resources in 19 comments. Another general issue was difficulty reading or completing a task because the video did not pause (10 resources, 10 comments). Two related issues were that the learning outcome was unclear (9 resources, 10 comments) and the lack of metadata (7 resources, 8 comments). A common problem was usability (8 resources, 11 comments), often because the interactions or the tasks were misconfigured (6 resources, 6 comments). Usability issues could be using the touchscreen or some tasks were impossible to do. Misconfigured tasks were tasks where the provided answer was not correct or only one option for a quiz. Some reviewers (6 resources, 6 comments) also mentioned that the resource type chosen was not the best suited. The lack of subtitles was mentioned for (6 resources, 6 comments), while (2 resources, 2 comments) mentioned accessibility more generally. Other fairly common issues were confusing or unnecessary content (5 resources, 5 comments), problems with navigation (4 resources, 4 comments), overall confusing resources (4 resources, 4 comments), bookmarks issues (4 resources, 4 comments), missing or bad graphics (4 resources, 4 comments) and lack of engagement (4 resources, 4 comments).

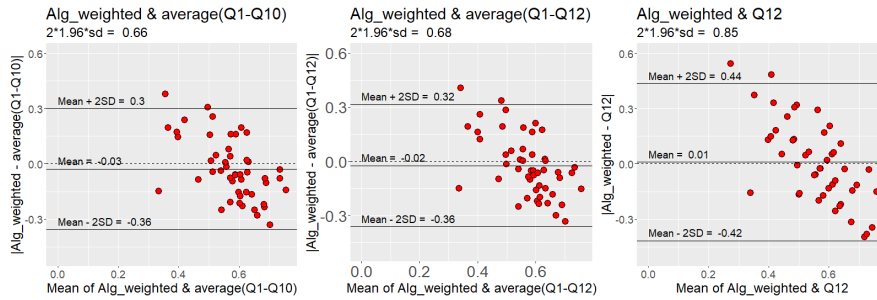


Figure 5.12: $Alg_{weighted}$ Bland Altman plot with average(Q1-Q10), average(Q1-Q12) and Q12. (All scales are equal on all three plots)

5.2 Comparison to algorithm

First, the weighted algorithm presented in Chapter 3 was tested. The results from the weighted algorithm did not satisfy the Shapiro-Wilk normality test ($p=0.04$), so the Kendall correlation test was used, and no significant correlations were found (Q12: $p=.88$, Q1-Q12: $p=.46$, Q1-Q10: $p=.379$). The Bland Altman plot is shown in figure 5.12. This shows best results compared with average(Q1-Q10) (mean=-.03, [-.36,.30]) followed by average(Q1-Q12) (mean=-.02, [-.36, .32]) and Q12 (mean=.01, [-.42,.44]). The mean difference is low (<0.05) for all comparisons, but the width of the 95% interval spans over a large part of the scale (66%, 68%, 86%).

Secondly, an unweighted version was tested. As shown in figure 5.13 this version had a smaller window of difference but a higher mean difference. Also here highest agreement was with with average(Q1-Q10) (mean=-.17, [-.36,.11]) followed by average(Q1-Q12) (mean=-.17, [-.46, .13]) and Q12 (mean=-.14, [-.52,.25]). These spans account for 56%, 59%, and 77%. Even though the mean differences are higher for this version of the algorithm, if this is consistent, it can be adjusted.

Since this algorithm passed the Shapiro-Wilk test of normality ($p=0.64$) as well as the averages (Q1-Q12: $p=0.14$, Q1-Q10: $p=0.15$) and Q12 ($p=0.26$), Pearson’s correlation test was conducted. This resulted in a significant correlation for all the tests as shown in table 5.9. There were only a small difference (0.01) between the average of Q1-Q10($r = .44, p=.002$) and Q1-Q12 ($r = .43, p = .002$) while Q12 were significant lower ($r = .30, p = .04$) corresponding with the findings from the Bland Altman plot.

Shapiro: 0.87

To make sure none of the feature scores included contributed negatively, removal of all were tested. This showed better results without the Viideo score, and a version of the algorithm without this was tested. A Bland Altman plot is shown in figure 5.14. The means were the same as $alg_{unweighted}$ (-.17, -.17,-.14), but the

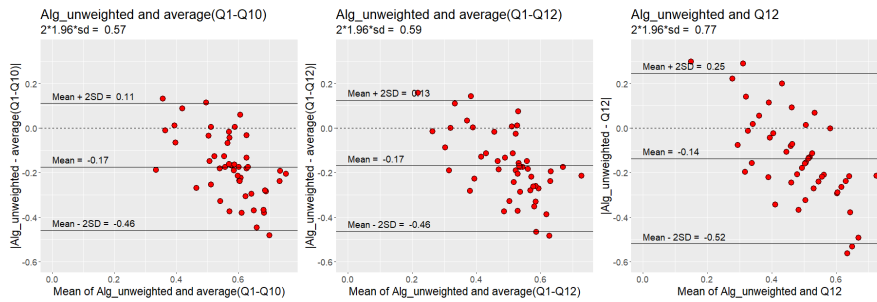


Figure 5.13: $Alg_{unweighted}$ Bland Altman plot with average(Q1-Q10), average(Q1-Q12) and Q12. (All scales are equal on all three plots)

Test	Pearson’s r	p-value
Average (Q1-Q10)	0.44	0.002
Average (Q1-Q12)	0.43	0.002
Q12	0.30	0.04

Table 5.9: Correlations $Alg_{unweighted}$

intervals slightly smaller $([-.45, .1], [-.45, .56], [-.51, .23])$ spanning 55%, 56% and 74% of the full scale.

All the correlation tests were significant with Average (Q1-Q10) and Average(Q1-Q12) having a high correlation $(r = .51, p = 2e - 04)$ and Q12 a bit lower $(r = 0.38, p = 7e - 03)$.

A t-test was performed to find if the reviewers agreed less on the resource with a high difference between the algorithm and reviews. First, the resources were split into quartiles based on the difference between the average(Q1-Q12) and the algorithm score (unweighted without video). Then the reviewer’s subjectivity was

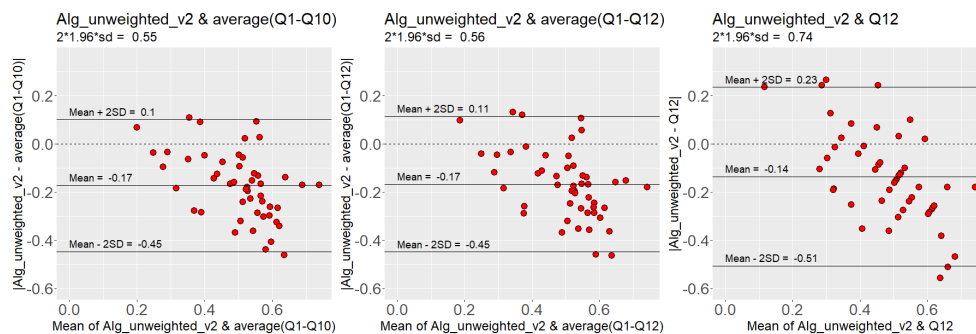


Figure 5.14: $Alg_{unweighted,2}$ Bland Altman plot with average(Q1-Q10), average(Q1-Q12) and Q12. (All scales are equal on all three plots)

Test	Pearson's r	p
Average (Q1-Q10)	0.51	2e-04
Average (Q1-Q12)	0.51	2e-04
Q12	0.38	7e-03

Table 5.10: Correlations $alg_{unweighted.v2}$

Metric	Explanation
Navigation	An image, text, question or hotspot that let the user navigate to another point in the video
Bookmarks	Used to indicate a new topic, an important interaction or an event within the video. Let the user navigate to that point by using the bookmark. [67]
Links	A web address
metadata_availability_score	Metadata availability score presented in chapter 3
Interactivity	A task the user can do
Combination_mnb	Uses total count of media, navigation and bookmarks
Combination_imnb	Uses total count of tasks, media, navigation and bookmarks
Combination_imnb	All types of interactions

Table 5.11: Explanation of feature scores used in correlation measures

evaluated using the absolute difference of the two scores on the same resource. The t-test showed no significant difference in subjectivity for the resources with high or low agreement between algorithm and reviews.

5.2.1 Metadata classification

To understand more about how transferable results from one study are to another, the classifier from Tavakoli *et al.* [12] was tested for the resources and reviews collected in this thesis. First, the resources were divided into "With control" and "Without control" based on their average(Q1-Q12) being above or below the median. For this application, the achieved accuracy was 49% with an F1 score of 20% for "With control" and 62% for "Without control".

Secondly, a test was made dividing the two classes based on being below or above a fixed value (4, median of scale). The result was an accuracy of 36% and an F1 score of 29% for "With control" and 42% for "Without control".

5.2.2 Feature scores correlation

All feature scores, also the ones unused in the algorithm were tested for correlation with the different question and averages. All the significant correlations can be found in appendix C. The names of the feature scores are explained in table 5.11.

Metric	Feature score	Correlation r	p-value
Average (Q1-Q10)	<i>combination_{imnb_decreasing5}</i>	0.31	0.001
Average (Q1-Q12)	<i>combination_{imnb_decreasing5}</i>	0.32	0.001
Q1	<i>navigation_{num}</i>	0.33	0.01
Q2	<i>interactivity_{decreasing_5}</i>	0.47	1.7e-05
Q3	<i>combination_{all_decreasing5}</i>	0.32	0.004
Q4	<i>navigation_{num}</i>	0.3	0.02
Q5	<i>text_{num}</i>	0.31	0.01
Q6	<i>combination_{imnb_decreasing5}</i>	0.31	0.01
Q7	<i>combination_{imnb_decreasing5}</i>	0.28	0.01
Q8	<i>text_{num}</i>	0.26	0.03
Q9	-	-	-
Q10	<i>combination_{imnb_decreasing5}</i>	0.34	0.02
Q11	<i>links_{time}</i>	0.33	0.01
Q12	<i>links_{num}</i>	0.35	0.005

Table 5.12: Highest feature score for each question

For many of them they are followed with a scoring model. This is the scoring model explained in section 3.3.4.

From table 5.12, we can see that the question with the highest correlated feature score is Q2 (0.47), with a significantly higher correlation r than the others. The feature score correlated with most questions is *combination_{imnb_decreasing5}* (Q3, Q6, Q7, Q10). From all correlation matrices, some similarities can be derived. Different combination scores were correlated with most questions (Q1, Q2, Q3, Q5, Q6, Q10, Q11, Q12). Navigation scores were correlated with four questions (Q1, Q4, Q11, Q12), followed by links (Q4, Q11, Q12). Interactivity scores were only correlated with Q2 and Q3, both questions measuring interactivity. Text elements scores were correlated with Q5 (presentation) and Q8 (reusability). Metadata scores were only correlated with two questions, Q1 (alignment) and Q12 (overall), and not significantly correlated with Q7 (Metadata). Finally, bookmark scores were only correlated with Q1.

Chapter 6

Discussion and conclusion

This thesis aimed to find the degree of agreement between measuring OER quality with manual reviews and a proposed algorithm. 107 manual reviews were collected from 23 participants and scores from these were compared to the algorithm scores. The reliability of the manual reviews was thoroughly evaluated to understand the implications of this result. The following chapter will discuss the interpretations, implications, limitations, and suggestions for future work based on these results. Finally, a conclusion follows.

Reliability measures were conducted, categorized as survey questions, repeatability, and background impact. While the survey questions show high inner consistency, the repeatability of the surveys was low. Reviews of the same resources conducted by two separate persons had a low level of agreement, low or no correlation, and low intraclass correlation. Possible explanations for this are the lack of guidance in the review process or just that quality is highly subjective, as pointed out by Almendro and Silveira [26]. The comments showed that the answer probably is a combination of users considering different factors important, valuing them differently, and some had problem understand the resources.

Even though the general agreement between reviewers was low, some questions had a higher correlation. The reason might be that some factors are more objective, easier to measure or have more clear phrasing. Since few comparative studies of evaluation methods are conducted [29], this is a field with many uncertainties. An example of higher correlation was measuring interactivity.

The background was the third reliability factor evaluated. In Sanz-Rodriguez *et al.* [27] and Cechinel and Sánchez-Alonsor [28] significant differences were found between different types of reviews conducted. Sanz-Rodriguez *et al.* [27] proposes that this is caused by the competence differences or users being more concerned with ease of use than experts. In contrast, in this thesis, a difference between users with or without reviewing experience was only found for one question. With these results, the differences found in Sanz-Rodriguez *et al.* [27] can be explained just the fact that different people have conducted them or that the way resources were reviewed differently. In contrast to reviewing background, between females and males, a significant difference was observed. The difference was that females

generally gave higher scores, but other factors could also have caused this result.

In addition to analyzing the reviews' reliability, the reviews were also a source of data on quality perception. Different insights can be found from the comments, questions correlated with an overall score (Q11, Q12), and feature scores correlations. From comparing the overall questions with the metrics measured automatically, the number of links, navigation, and combined numbers of media, navigation, and bookmarks was most important, corresponding to the result of Cechinel *et al.* [13]. The correlation tests with the overall score suggested that presentation (Q9) and engagement (Q10) are the most significant factors. From the comments, the most mentioned factors are lack of interactivity (often tasks), that the video should pause for interactions and that the learning outcome was unclear. Presentation, which in Q9 includes visual and audio quality and engagement, is also mentioned in the comments but significantly less frequently.

A possible explanation for the difference between the comments and the correlation with average scores can be found in the median values. Interactivity (Q2, Q3) had lower median values than presentation (Q9) and engagement (Q10). A hypothesis is that the comments generally focus on what should be improved, and since presentation and engagement generally scores relatively high, there is no need to comment on it. A similar connection can also be seen with the metadata question (Q7).

The central hypothesis was that automated assessment of quality could replace manual reviews. The best results were achieved with the unweighted version without the video score. The reason the unweighted version performed better might be because the mentions in frameworks does not directly lead to importance. It is also a difference in how discriminative the metrics are. For instance, most of the resources received the highest score on accessibility giving this feature a low discriminate power. The reasons for the negative correlation with the video score is that it measures distortion, which might not be a suitable measure for video quality in this case.

The degree of agreement for the unweighted algorithm without video compared with the manual reviews were generally low, the 95% interval covering 56% of the scale at best. This is approximately the same degree as between two reviewers for the average(Q1-Q12) which was 57.7%. However the a moderate linear correlation were found between the algorithm score and reviews. For all the algorithm versions, the difference was lower compared with the average(Q1-Q10) than the average(Q1-Q12) and significantly better than with Q12. A reason can be that the algorithm is better at predicting quality-specific factors rather than an overall more subjective score or that that the averages better reflect the actual quality. The general agreement were lower than expected based on similar research [10–12, 43]. There is however several important differences. The other research does classification rather than comparing two scores and used review data to build models based on them. That makes the predictions more adjusted for the OERs of a specific repository.

An example of the challenge of building models based on data from one repos-

itories can be seen in the results using the classifier from Tavakoli *et al.* [12]. Contrary to the results achieved in Tavakoli *et al.* [43], the classifier did not perform well with new data. The accuracy was approximately 50%, meaning it performs as well as classifying at random. The reason might be because some of the metadata was unavailable or that the association between high-quality metadata and high-quality OER is not as strong as believed. There is also a difference between how the resources were classified as high or low quality. In Tavakoli *et al.* [12] the resources were classified as high quality if they went through manual quality control, and the qualifications for passing are not specified.

Bethard *et al.* [10] and Cechinel *et al.* [13] suggest that some metrics might correlate more with higher quality resources. The results from the feature scores correlations support this with the number of different interactions most correlated. It is also clear that some of the questions are more closely connected to a measurable metric. Most clearly, this is shown with the interactivity question.

6.0.1 Implications

The results of the low repeatability of the reviews need to be considered when using review data in research, both for validation or creating models. Especially when quality data come from repositories where the reliability can not be controlled or measured. Given the low repeatability, reviews or rating from a single person can be considered unreliable, and building models on these lower the general applicability. An example was how much the results varied when testing an old classifier with new data [12]. It also has implications for the repositories that use review data to indicate the quality or rank resources.

Given the high variability of OER, one would need a high number of different resources and also many controlled and validated reviews. Therefore, these results challenge the use of review data (or other user-generated quality remarks) uncritically to create prediction models. It strengthens the suggestion of using white-box models [11] and theory rather than unreliable data.

The low correlation has been shown previously in Sanz-Rodriguez *et al.* [27] and Cechinel and Sánchez-Alonsor [28], but the focus was mainly on different review types rather than the subjective view of humans. The subjectivity results suggest that cooperation methods like the ones used in Vargo *et al.* [19] should be considered since they can increase the reliability.

The insights learned about quality perception can also give some practical implications. The comments provided insights into how the resource could be improved, and the results could be incorporated into authoring tools to create higher-quality resources. This could, for instance, be to pause the video when interactions are shown automatically.

The algorithm may be suited for some usage dependent on the need for precision. The degree of agreement is not high enough to exclude resources or say that a specific resource has the highest quality. The difference could be acceptable for certain kinds of applications. It can be a better alternative compared with no

form of ranking. The subjectivity of the reviews also shows the quality of using an algorithm for quality assessment since it will give the same result every time. The quality of knowing what the result is based on and objectivity can also compensate for the lack of exact agreement with reviews.

6.0.2 Limitations

In this thesis, the approach was tested only on one resource type, H5P's Interactive Video, limiting the results. The implications are probably less applicable the more different a resource is from this. Another limitation is that the algorithm did not consider the discipline of a resource as suggested by Cechinel *et al.* [11, 13].

One of the limitations of this study was the group of reviewers. Almost all of them had completed higher education (that might be representative of the user group), had long experience with OER, and were recruited from a beta tester group. Many of them were also creators of educational material themselves, influencing the view on quality. The relatively low number of participants also led to a limited number of resources being reviewed. This affects the generability and significance of the results. Having even more than two reviewers per resource could probably also improve the results.

Another limitation was how the reviews were conducted. The participants were given a link to the resource and filled out a survey. They were not given any training on how to conduct reviews, therefore both the way they did it and how much time they used differed. There is little knowledge about the reliability and validation of different review instruments and questions [29] and it is therefore also uncertain how much the results would have varied with other questions. In the comparison with the algorithm the metrics average, but it is reasonable to believe some questions are more important [34] and that they should be weighted differently.

A limitation of the data analysis conducted is how correlation can be interpreted. The correlation between all questions and algorithmic metrics was measured, but the significance should not be overestimated. Many of the metrics most highly correlated with the reviews did not contain many samples. For instance, only 10% of the resources had a link. As Cechinel *et al.* [13] pointed out, correlation is not causation, but it is a first step to identifying important variables.

A part of the reliability analysis concerned the reviewers' background. The reviewers' backgrounds were unknown beforehand, and resources were therefore not assigned based on that. This also means that different backgrounds were not compared for each resource; only the means were compared. The implication is that the difference between the genders and the lack of difference between review experiences could be based on the resources assigned rather than the background.

RQ2 was asked to evaluate the significance of the algorithm results. The low reliability means that the algorithm developed can neither be said to be suited to replace manual reviews or not. Even though steps were taken, like only comparing with the average of two reviews it is not possible to judge it based on the data

produced in this thesis. This is formulated by Hanneman [15]: "*If one or both methods do not give repeatable results, assessment of agreement between methods is meaningless.*" and shows the importance of RQ2. With manual quality reviews full agreement will never be achieved, but the results can become more reliable with more reviews which at a certain point will diverge to a score.

The results are also generally limited by the data analysis conducted. Performing other analyses could have led to different results or insights.

6.0.3 Future work

Future studies need to gather more empirical data on how quality is perceived and how reliable and valid reviews can be conducted. Reliability should always be a concern and taken into consideration when user-generated data is used. Comparative studies on frameworks should be conducted, and the number of reviewers needed for reliable reviews must be established. To do this, reviews from experts and users should be evaluated, and collaboration should be considered since it can increase the quality [19]. Since so many factors impact a resource's quality, more resources need to be evaluated. Usage data could also be used as an indicator of quality and combined with reviews as suggested by Sanz-Rodriguez *et al.* [27].

Future work can extend the algorithm with more different resource types and make the necessary adjustments. The feature scores can also be further extended. Text analysis can become more advanced. With better metadata, it can take into consideration the level a resource is meant for. It can also avoid considering words from other languages as spelling errors if a resource is meant for language training. The review comments suggest that testing for misconfiguration of tasks and if the video pauses for interactions can improve the assessment.

Based on the result, future work should be cautious about using review data and instead focus on learning more about which measurable factors influence the quality of OER. Based on this algorithms can iteratively find more and better metrics to measure quality factors. This will probably lead to a lower agreement on a certain data set, but can give higher reproducibility. Finding how different scores should be weighted also needs more research.

Another suggestion is to use the algorithm to improve resources in the authoring or sharing phase. Instead of using the full score, the different parts of the scores can be shown so the creator can improve the quality before sharing. Feature scores most suited for this are accessibility, interactivity, text analysis, and metadata.

6.0.4 Conclusion

This thesis can be seen as giving an alternative to the quality prediction based on review data, taking a more theoretical approach. It can undoubtedly give advantages since does not depend on any review data and all scores are explainable and understandable. Three versions of a white-box algorithm that measure the quality

were proposed and 107 reviews were collected. That gave valuable data used to further extend the knowledge on OER quality. The degree of agreement between the manual reviews and algorithm was lower than expected, indicating that it needs more development to be a viable alternative to manual reviews. However the reviews does not represent a ground truth.

By administering reviews, the reviews' reliability could be calculated. The results were that it is challenging to use manual reviews uncritical. The reliability results are valuable and have not been the topic of any papers testing classification models. It shows the need for an automated approach, if not to replace manual reviews, but to supplement, giving a more objective alternative. Future work should continue on the work to create automated approaches and increase the knowledge of OER quality and how to reliably measure it.

Bibliography

- [1] J. Hylén, ‘Open educational resources: Opportunities and challenges,’ 2021.
- [2] *United nations educational, scientific and cultural organization. 1.37 billion students now home as covid-19 school closures expand, ministers scale up multimedia approaches to ensure learnin continuity.* [Online]. Available: <https://en.unesco.org/news/137-billion-students-now-home-covid-19-school-closures-expand-ministers-scale-multimedia>.
- [3] N. Butcher, *A basic guide to open educational resources (OER)*. Commonwealth of Learning (COL); 2015.
- [4] C. S. U. L. Beach, *Merlot*. [Online]. Available: <https://www.merlot.org/merlot> (visited on 13/05/2022).
- [5] A. F. Camilleri, U. D. Ehlers and J. Pawlowski, *State of the art review of quality issues related to open educational resources (OER)*. Luxembourg: Publications Office of the European Union, 2014.
- [6] C. Cechinel and X. Ochoa, ‘A brief overview of quality inside learning object repositories,’ in *Proceedings of the XV International Conference on Human Computer Interaction*, 2014, pp. 1–7.
- [7] K. Clements, J. Pawlowski and N. Manouselis, ‘Open educational resources repositories literature review–towards a comprehensive quality approaches framework,’ *Computers in human behavior*, vol. 51, pp. 1098–1106, 2015.
- [8] S. Scheunemann, A. Brandão and D. Brauner, ‘Towards defining quality criteria for digital educational resources in distance learning,’ in *2018 IEEE World Engineering Education Conference (EDUNINE)*, IEEE, 2018, pp. 1–4.
- [9] K. Clements and J. Pawlowski, ‘User-oriented quality for oer: Understanding teachers’ views on re-use, quality, and trustjcal_450,’ 2011.
- [10] S. Bethard, P. Wetzer, K. Butcher, J. H. Martin and T. Sumner, ‘Automatically characterizing resource quality for educational digital libraries,’ in *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, 2009, pp. 221–230.
- [11] C. Cechinel, S. da Silva Camargo, M.-Á. Sicilia and S. Sánchez-Alonso, ‘Mining models for automated quality assessment of learning objects.,’ *J. Univers. Comput. Sci.*, vol. 22, no. 1, pp. 94–113, 2016.

- [12] M. Tavakoli, M. Elias, G. Kismihók and S. Auer, 'Quality prediction of open educational resources a metadata-based approach,' in *2020 IEEE 20th international conference on advanced learning technologies (ICALT)*, IEEE, 2020, pp. 29–31.
- [13] C. Cechinel, S. Sánchez-Alonso and E. García-Barriocanal, 'Statistical profiles of highly-rated learning objects,' *Computers & Education*, vol. 57, no. 1, pp. 1255–1269, 2011.
- [14] S. B. Merriam, 'N of i?: Issues of validity and reliability in,' *PAACE Journal of lifelong learning*, vol. 4, pp. 51–60, 1995.
- [15] S. K. Hanneman, 'Design, analysis, and interpretation of method-comparison studies,' *AACN advanced critical care*, vol. 19, no. 2, pp. 223–234, 2008.
- [16] D. E. Atkins, J. S. Brown and A. L. Hammond, *A review of the open educational resources (OER) movement: Achievements, challenges, and new opportunities*. Creative common Mountain View, 2007, vol. 164.
- [17] R. McGreal, 'Learning objects: A practical definition,' *International Journal of Instructional Technology and Distance Learning (IJITDL)*, vol. 9, no. 1, 2004.
- [18] S. Kocdar and A. Bozkurt, 'Supporting learners with special needs in open, distance, and digital education,' in *Handbook of Open, Distance and Digital Education*, Springer, 2022, pp. 1–16.
- [19] J. Vargo, J. C. Nesbit, K. Belfer and A. Archambault, 'Learning object evaluation: Computer-mediated collaboration and inter-rater reliability,' *International Journal of Computers and Applications*, vol. 25, no. 3, pp. 198–205, 2003.
- [20] D. M. Grath, 'Certified copy of the recommendation on open educational resources (OER),' p. 61,
- [21] T. Dreesen, S. Akseer, M. Brossard, P. Dewan, J.-P. Giraldo, A. Kamei, S. Mizunoya and J. S. Ortiz, 'Promising practices for equitable remote learning: Emerging lessons from covid-19 education responses in 127 countries,' 2020.
- [22] Y. Akpınar, 'Validation of a learning object review instrument: Relationship between ratings of learning objects and actual learning outcomes,' *International Journal of Doctoral Studies*, vol. 4, no. 4, pp. 291–302, 2009.
- [23] S. Mishra, 'Open educational resources: Removing barriers from within,' *Distance education*, vol. 38, no. 3, pp. 369–380, 2017.
- [24] L. Harvey and D. Green, 'Defining quality,' *Assessment & evaluation in higher education*, vol. 18, no. 1, pp. 9–34, 1993.
- [25] P. Kawachi, *Quality assurance guidelines for open educational resources: Tips framework*, 2014.

- [26] D. Almendro and I. F. Silveira, 'Quality assurance for open educational resources: The oer trust framework,' *International Journal of Learning, Teaching and Educational Research*, vol. 17, no. 3, pp. 1–14, 2018.
- [27] J. Sanz-Rodriguez, J. M. M. Dodero and S. Sánchez-Alonso, 'Ranking learning objects through integration of different quality indicators,' *IEEE transactions on learning technologies*, vol. 3, no. 4, pp. 358–363, 2010.
- [28] C. Cechinel and S. Sánchez-Alonso, 'Analyzing associations between the different ratings dimensions of the merlot repository,' *Interdisciplinary Journal of E-Learning and Learning Objects*, vol. 7, no. 1, pp. 1–9, 2011.
- [29] O. Zawacki-Richter, W. Müskens and V. I. Marín, 'Quality assurance of open educational resources,' in *Handbook of Open, Distance and Digital Education*, Springer, 2022, pp. 1–19.
- [30] J. C. Nesbit and J. Li, 'Web-based tools for learning object evaluation,' in *International conference on education and information systems: Technologies and Applications*, 2004, pp. 21–25.
- [31] A. Romero-Pelaez, V. Segarra-Faggioni, N. Piedra and E. Tovar, 'A proposal of quality assessment of oer based on emergent technology,' in *2019 IEEE Global Engineering Education Conference (EDUCON)*, IEEE, 2019, pp. 1114–1119.
- [32] A. R. Peláez, N. P. Pullaguari and E. T. Caro, 'Quality model proposal for educational material production in oer sites,' in *2011 IEEE Global Engineering Education Conference (EDUCON)*, IEEE, 2011, pp. 1074–1080.
- [33] K. Mayrberger, O. Zawacki-Richter and W. Müskens, 'Qualitätsentwicklung von oer,' *Hamburg: Universität Hamburg*, 2018.
- [34] V. Kumar, J. Nesbit, P. Winne, A. Hadwin, D. Jamieson-Noel and K. Han, 'Quality rating and recommendation of learning objects,' in *E-Learning Networked Environments and Architectures*, Springer, 2007, pp. 337–373.
- [35] *About the project*. [Online]. Available: <https://h5p.org/about-the-project> (visited on 06/06/2022).
- [36] *We help the world create better content faster*. [Online]. Available: <https://joubel.com/#about-h5p-04> (visited on 06/06/2022).
- [37] C. Cechinel, S. d. Silva Camargo, S. Sánchez-Alonso and M.-Á. Sicilia, 'On the search for intrinsic quality metrics of learning objects,' in *Research Conference on Metadata and Semantic Research*, Springer, 2012, pp. 49–60.
- [38] Joubel, *The h5p oer hub*. [Online]. Available: <https://h5p.org/oer-hub-coming#progress> (visited on 13/04/2022).
- [39] M. Erdt, A. Fernandez and C. Rensing, 'Evaluating recommender systems for technology enhanced learning: A quantitative survey,' *IEEE Transactions on Learning Technologies*, vol. 8, no. 4, pp. 326–344, 2015.

- [40] F. van Assche and R. Vuorikari, 'A framework for quality of learning resources,' in *Handbook on quality and standardisation in E-learning*, Springer, 2006, pp. 443–456.
- [41] T. Trippel, D. Broeder, M. Durco and O. Ohren, 'Towards automatic quality assessment of component metadata,' in *LREC 2014: 9th International Conference on Language Resources and Evaluation*, 2014, pp. 3851–3856.
- [42] X. Ochoa and E. Duval, 'Quality metrics for learning object metadata,' in *Ed-Media + innovate learning*, Association for the Advancement of Computing in Education (AACE), 2006, pp. 1004–1011.
- [43] M. Tavakoli, M. Elias, G. Kismihók and S. Auer, 'Metadata analysis of open educational resources,' in *LAK21: 11th International Learning Analytics and Knowledge Conference*, 2021, pp. 626–631.
- [44] M. Elias, M. Tavakoli, S. Lohmann, G. Kismihok and S. Auer, 'An oer recommender system supporting accessibility requirements,' in *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 2020, pp. 1–4.
- [45] M. Custard and T. Sumner, 'Using machine learning to support quality judgments,' *D-Lib Magazine*, vol. 11, no. 10, pp. 1082–9873, 2005.
- [46] M. Meyer, A. Hannappel, C. Rensing and R. Steinmetz, 'Automatic classification of didactic functions of e-learning resources,' in *Proceedings of the 15th ACM international conference on Multimedia*, 2007, pp. 513–516.
- [47] E. Mendes, W. Hall and R. Harrison, 'Applying metrics to the evaluation of educational hypermedia applications,' *Journal of Universal Computer Science*, vol. 4, no. 4, pp. 382–403, 1998.
- [48] J. E. Blumenstock, 'Size matters: Word count as a measure of quality on wikipedia,' in *Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 1095–1096.
- [49] A. Stefani, B. Vassiliadis and M. Xenos, 'On the quality assessment of advanced e-learning services,' *Interactive Technology and Smart Education*, 2006.
- [50] M. A. Sicilia, E. García, C. Pagés, J. J. Martinez and J. M. Gutierrez, 'Complete metadata records in learning object repositories: Some evidence and requirements,' *International Journal of Learning Technology*, vol. 1, no. 4, pp. 411–424, 2005.
- [51] M. Y. Ivory and M. A. Hearst, 'Statistical profiles of highly-rated web sites,' in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2002, pp. 367–374.
- [52] E. García-Barriocanal and M. Á. Sicilia, 'Preliminary explorations on the statistical profiles of highly-rated learning objects,' in *Research Conference on Metadata and Semantic Research*, Springer, 2009, pp. 108–117.

- [53] T. E. Oliphant, 'Python for scientific computing,' *Computing in science & engineering*, vol. 9, no. 3, pp. 10–20, 2007.
- [54] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke and T. E. Oliphant, 'Array programming with NumPy,' *Nature*, vol. 585, pp. 357–362, 2020. DOI: 10.1038/s41586-020-2649-2.
- [55] L. Richardson, *Beautifulsoup4*, <https://github.com/wention/BeautifulSoup4>, 2004.
- [56] tgoodall, *Sk-video 1.1.10*, <https://github.com/scikit-video/scikit-video>, 2008.
- [57] openCV, *Opencv – 4.5.5*, <https://github.com/opencv/opencv>, 2021.
- [58] R. G. Gonzalez, *Youtube-dl*, <https://github.com/ytdl-org/youtube-dl>, 2016.
- [59] S. Bansal, *Textstat*, <https://github.com/shivam5992/textstat>, 2016.
- [60] J. Morris, *Language_tool_python2.7.1*, https://github.com/jxmorris12/language_tool_python, 2022.
- [61] 'Ergonomics of human-system interaction — Part 171: Guidance on software accessibility,' International Organization for Standardization, Geneva, CH, Standard, Jul. 2008.
- [62] X. Zhang, A. Tlili, F. Nascimbeni, D. Burgos, R. Huang, T.-W. Chang, M. Jemni and M. K. Khribi, 'Accessibility within open educational resources and practices for disabled learners: A systematic literature review,' *Smart Learning Environments*, vol. 7, no. 1, pp. 1–19, 2020.
- [63] J. Spellman, R. B. Montgomery, S. Lauriat and M. Cooper, *W3c working draft*. [Online]. Available: <https://www.w3.org/TR/2021/WD-wcag-3.0-20211207/> (visited on 27/05/2022).
- [64] A. Mittal, M. A. Saad and A. C. Bovik, 'A completely blind video integrity oracle,' *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289–300, 2015.
- [65] J. Michael, 'Where's the evidence that active learning works?' *Advances in physiology education*, 2006.
- [66] R. Flesch, 'How to write plain english,' *University of Canterbury*. Available at http://www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml. [Retrieved 5 February 2016], 1979.
- [67] *Interactive video tutorial*. [Online]. Available: <https://h5p.org/tutorial-interactive-video> (visited on 06/06/2022).

- [68] LibreStudio, *Librestudio*. [Online]. Available: <https://studio.libretexts.org/> (visited on 13/04/2022).
- [69] S. R. Tadinada, *H5pcatalogue*. [Online]. Available: <https://h5pcatalogue.in/h5ppview> (visited on 13/04/2022).
- [70] eCampusOntario, *Ecampusontario h5p studio*. [Online]. Available: <https://h5pstudio.ecampusontario.ca/> (visited on 13/04/2022).
- [71] H. B. Ganzeboom, 'International standard classification of occupations isco-08 with isei-08 scores,' *Version of July*, vol. 27, p. 2010, 2010.
- [72] F. Reichheld, *The ultimate question 2.0 (revised and expanded edition): How net promoter companies thrive in a customer-driven world*. Harvard Business Review Press, 2011.
- [73] A. Joshi, S. Kale, S. Chandel and D. K. Pal, 'Likert scale: Explored and explained,' *British journal of applied science & technology*, vol. 7, no. 4, p. 396, 2015.
- [74] M. Hollander, D. A. Wolfe and E. Chicken, *Nonparametric statistical methods*. John Wiley & Sons, 2013, vol. 751.
- [75] R. K. Henson, 'Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha,' *Measurement and evaluation in counseling and development*, vol. 34, no. 3, pp. 177–189, 2001.
- [76] L. J. Cronbach, 'Coefficient alpha and the internal structure of tests,' *psychometrika*, vol. 16, no. 3, pp. 297–334, 1951.
- [77] N. Carchedi, *Cronbach.alpha: Cronbach's alpha*. [Online]. Available: <https://www.rdocumentation.org/packages/lrm/versions/1.2-0/topics/cronbach.alpha> (visited on 06/02/2022).
- [78] J. M. Bland and D. Altman, 'Statistical methods for assessing agreement between two methods of clinical measurement,' *The lancet*, vol. 327, no. 8476, pp. 307–310, 1986.
- [79] P. E. Shrout and J. L. Fleiss, 'Intraclass correlations: Uses in assessing rater reliability,' *Psychological bulletin*, vol. 86, no. 2, p. 420, 1979.
- [80] M. Gamer, J. Lemon, M. M. Gamer, A. Robinson and W. Kendall's, 'Package 'irr', *Various coefficients of interrater reliability and agreement*, vol. 22, 2012.
- [81] K. Van Stralen, F. Dekker, C. Zoccali and K. Jager, 'Measuring agreement, more complicated than it seems,' *Nephron Clinical Practice*, vol. 120, no. 3, pp. c162–c167, 2012.
- [82] D. F. Bauer, 'Constructing confidence sets using rank statistics,' *Journal of the American Statistical Association*, vol. 67, no. 339, pp. 687–690, 1972.
- [83] C. A. Bobak, P. J. Barr and A. J. O'Malley, 'Estimation of an inter-rater intraclass correlation coefficient that overcomes common assumption violations in the assessment of health measurement scales,' *BMC medical research methodology*, vol. 18, no. 1, pp. 1–11, 2018.

Appendix A

JSON of a H5P interactive video

Attached is an example of an Interactive Video created with H5P to show the structure and its metrics.


```
1 {
2   "interactiveVideo": {
3     "video": {
4       "startScreenOptions": {
5         "title": "Interactive Video",
6         "hideStartTitle": false
7       },
8       "textTracks": {
9         "videoTrack": [
10          {
11            "label": "Subtitles",
12            "kind": "subtitles",
13            "srcLang": "en"
14          }
15        ]
16      },
17      "files": [
18        {
19          "path": "videos/files-62404d6937a40.mp4",
20          "mime": "video/mp4",
21          "copyright": {
22            "license": "U"
23          }
24        }
25      ]
26    },
27    "assets": {
28      "interactions": [
29        {
30          "x": 3.4310221586847747,
31          "y": 1.924730216981253,
32          "width": 10,
33          "height": 10,
34          "duration": {
35            "from": 0,
36            "to": 10
37          },
38          "libraryTitle": "Image",
39          "action": {
40            "library": "H5P.Image 1.1",
41            "params": {
42              "contentName": "Image",
43              "alt": "asdasd"

```

```
44     },
45     "subContentId": "9f5dbbb9-11b7-4e51-982a-bf998b58db23"
46     ,
47     "metadata": {
48       "contentType": "Image",
49       "license": "U",
50       "title": "Untitled Image"
51     }
52   },
53   "visuals": {
54     "backgroundColor": "rgba(0,0,0,0)",
55     "boxShadow": true
56   },
57   "pause": false,
58   "displayType": "poster",
59   "buttonOnMobile": false,
60   "goto": {
61     "url": {
62       "protocol": "http:\\\\/"
63     },
64     "visualize": false,
65     "type": ""
66   },
67   "label": ""
68   {
69     "x": 36.498491224206795,
70     "y": 48.96195919755794,
71     "width": 10,
72     "height": 10,
73     "duration": {
74       "from": 0,
75       "to": 10
76     },
77     "libraryTitle": "Label",
78     "action": {
79       "library": "H5P.Nil 1.0",
80       "params": {
81
82       },
83     },
84     "subContentId": "5bdcfcbc-30a0-4e5f-87b7-48e656afe2cd"
85     ,
86     "metadata": {
87       "contentType": "Label",
```

```
86         "license": "U"
87     }
88 },
89     "label": "Lorem ipsum dolor sit amet...",
90     "pause": false,
91     "displayType": "button",
92     "buttonOnMobile": false
93 },
94 {
95     "x": 41.1722659042173,
96     "y": 37.21145086163756,
97     "width": 10,
98     "height": 10,
99     "duration": {
100         "from": 0,
101         "to": 10
102     },
103     "libraryTitle": "Text",
104     "action": {
105         "library": "H5P.Text 1.1",
106         "params": {
107             "text": "<p>Text text</p>\n"
108         },
109         "subContentId": "b0032b13-e7ab-4e0d-afba-21c0954f1e5e"
110     },
111     "metadata": {
112         "contentType": "Text",
113         "license": "U",
114         "title": "Untitled Text"
115     }
116 },
117     "pause": false,
118     "displayType": "button",
119     "buttonOnMobile": false,
120     "visuals": {
121         "backgroundColor": "rgb(255, 255, 255)",
122         "boxShadow": true
123     },
124     "goto": {
125         "url": {
126             "protocol": "http:\\\\\\"
127         },
128         "visualize": false,
129         "type": "timecode"
```

```

129     },
130     "label": "<p>Text label</p>\n"
131 },
132 {
133     "x": 28.60411899313501,
134     "y": 14.762516046213094,
135     "width": 10,
136     "height": 10,
137     "duration": {
138         "from": 0,
139         "to": 10
140     },
141     "libraryTitle": "Table",
142     "action": {
143         "library": "H5P.Table 1.1",
144         "params": {
145             "text": "<table class=\"h5p-table\">\n\t<thead>\n\t\t
146                 <tr>\n\t\t\t\t<th scope=\"col\">&nbsp;</th>\n\t\t\t
147                 <th scope=\"col\">Heading Column 2</th>\n\t\t\t
148                 </tr>\n\t\t</thead>\n\t\t<tbody>\n\t\t\t<tr>\n\t\t\t\t
149                 <td>Row 1 Col 1</td>\n\t\t\t\t<td>Row 1 Col 2</
150                 td>\n\t\t\t\t</tr>\n\t\t\t<tr>\n\t\t\t\t<td>Row 2 Col 1
151                 </td>\n\t\t\t\t<td>Row 2 Col 2</td>\n\t\t\t\t</tr>\n
152                 \t</tbody>\n</table>\n"
153         },
154         "subContentId": "99fe9bc4-430a-4b39-8f6b-87b34915f1db"
155     },
156     "metadata": {
157         "contentType": "Table",
158         "license": "U",
159         "title": "Untitled Table"
160     }
161 },
162 "pause": false,
163 "displayType": "button",
164 "buttonOnMobile": false,
165 "visuals": {
166     "backgroundColor": "rgb(255, 255, 255)",
167     "boxShadow": true
168 },
169 "label": "<p>Table lable</p>\n"
170 },
171 {
172     "x": 34.33201345814927,

```

```
165     "y": 55.85265171698755,
166     "width": 10,
167     "height": 10,
168     "duration": {
169       "from": 0,
170       "to": 10
171     },
172     "libraryTitle": "Link",
173     "action": {
174       "library": "H5P.Link 1.3",
175       "params": {
176         "linkWidget": {
177           "protocol": "http:\\\\\/",
178           "url": "Link url"
179         },
180         "title": "Link title"
181       },
182       "subContentId": "54234c18-7147-4c24-b80e-3b32f99123dc"
183     },
184     "metadata": {
185       "contentType": "Link",
186       "license": "U"
187     },
188     "displayType": "poster",
189     "visuals": {
190       "backgroundColor": "rgba(0,0,0,0.5)",
191       "boxShadow": true
192     },
193     "pause": false,
194     "buttonOnMobile": false
195   },
196   {
197     "x": 67.51962646769357,
198     "y": 35.30914763717604,
199     "width": 10,
200     "height": 10,
201     "duration": {
202       "from": 0,
203       "to": 10
204     },
205     "libraryTitle": "Image",
206     "action": {
207       "library": "H5P.Image 1.1",
```

```
208     "params": {
209         "contentName": "Image",
210         "alt": "Image alt text",
211         "title": "Image hover text"
212     },
213     "subContentId": "6ee87f81-0289-4095-9dfd-d9b5eaf08ab9"
214     ,
215     "metadata": {
216         "contentType": "Image",
217         "license": "U",
218         "title": "Untitled Image"
219     },
220     "visuals": {
221         "backgroundColor": "rgba(0,0,0,0)",
222         "boxShadow": true
223     },
224     "pause": false,
225     "displayType": "poster",
226     "buttonOnMobile": false,
227     "goto": {
228         "url": {
229             "protocol": "http:\\\\\\"
230         },
231         "visualize": false,
232         "type": ""
233     },
234     "label": ""
235 },
236 {
237     "x": 45.76659038901602,
238     "y": 27.599486521181,
239     "width": 10,
240     "height": 10,
241     "duration": {
242         "from": 0,
243         "to": 10
244     },
245     "libraryTitle": "Statements",
246     "action": {
247         "library": "H5P.Summary 1.10",
248         "params": {
249             "intro": "Choose the correct statement.",
250             "overallFeedback": [
```



```
251     {
252         "from": 0,
253         "to": 100
254     }
255 ],
256 "solvedLabel": "Progress:",
257 "scoreLabel": "Wrong answers:",
258 "resultLabel": "Your result",
259 "labelCorrect": "Correct.",
260 "labelIncorrect": "Incorrect! Please try again.",
261 "alternativeIncorrectLabel": "Incorrect",
262 "labelCorrectAnswers": "Correct answers.",
263 "tipButtonLabel": "Show tip",
264 "scoreBarLabel": "You got :num out of :total points"
265 ,
266 "progressText": "Progress :num of :total",
267 "summaries": [
268     {
269         "subContentId": "b5f42a0a-f858-479f-bef3-9a6b04d
270             9f3e6",
271         "summary": [
272             "<p>First statement</p>\n",
273             "<p>Second statement</p>\n"
274         ],
275         "tip": "<p>Tip statement</p>\n"
276     }
277 ],
278 "subContentId": "429c98cc-6a78-4649-98cf-99ce5a459a0a"
279 ,
280 "metadata": {
281     "contentType": "Summary",
282     "license": "U",
283     "title": "Untitled Summary"
284 }
285 },
286 "pause": false,
287 "displayType": "button",
288 "buttonOnMobile": false,
289 "adaptivity": {
290     "correct": {
291         "allowOptOut": false,
292         "message": ""
293     }
294 },
```

```
292     "wrong": {
293         "allowOptOut": false,
294         "message": ""
295     }
296 },
297 "label": "<p>Statement label</p>\n"
298 },
299 {
300     "x": 64.30847643166148,
301     "y": 20.85776436992543,
302     "width": 10,
303     "height": 10,
304     "duration": {
305         "from": 0,
306         "to": 10
307     },
308     "libraryTitle": "Single Choice Set",
309     "action": {
310         "library": "H5P.SingleChoiceSet 1.11",
311         "params": {
312             "choices": [
313                 {
314                     "subContentId": "522d590f-aaab-455e-8dbc-6e3fbd6
315                         714be",
316                     "question": "<p>First Question</p>\n",
317                     "answers": [
318                         "<p>First alternative</p>\n",
319                         "<p>Second alternative</p>\n"
320                     ]
321                 },
322                 {
323                     "subContentId": "b003c052-0d57-4c50-bcb7-b6cca02
324                         f03ba",
325                     "question": "<p>Second question</p>\n",
326                     "answers": [
327                         "<p>First alternative</p>\n",
328                         "<p>Second alternative</p>\n"
329                     ]
330                 }
331             ],
332             "overallFeedback": [
333                 {
334                     "from": 0,
335                     "to": 100
336                 }
337             ]
338         }
339     }
340 }
```

```

334     }
335   ],
336   "behaviour": {
337     "autoContinue": true,
338     "timeoutCorrect": 2000,
339     "timeoutWrong": 3000,
340     "soundEffectsEnabled": true,
341     "enableRetry": true,
342     "enableSolutionsButton": true,
343     "passPercentage": 100
344   },
345   "l10n": {
346     "nextButtonLabel": "Next question",
347     "showSolutionButtonLabel": "Show solution",
348     "retryButtonLabel": "Retry",
349     "solutionViewTitle": "Solution list",
350     "correctText": "Correct!",
351     "incorrectText": "Incorrect!",
352     "muteButtonLabel": "Mute feedback sound",
353     "closeButtonLabel": "Close",
354     "slideOfTotal": "Slide :num of :total",
355     "scoreBarLabel": "You got :num out of :total
356       points",
357     "solutionListQuestionNumber": "Question :num",
358     "allyShowSolution": "Show the solution. The task
359       will be marked with its correct solution.",
360     "allyRetry": "Retry the task. Reset all responses
361       and start the task over again."
362   }
363 },
364 "subContentId": "5209af2f-0e55-4e23-9577-d121022d48f5"
365 ,
366 "metadata": {
367   "contentType": "Single Choice Set",
368   "license": "U",
369   "title": "Untitled Single Choice Set"
370 }
371 },
372 "pause": false,
373 "displayType": "button",
374 "buttonOnMobile": false,
375 "adaptivity": {
376   "correct": {
377     "allowOptOut": false,

```

```

374     "message": ""
375   },
376   "wrong": {
377     "allowOptOut": false,
378     "message": ""
379   },
380   "requireCompletion": false
381 },
382 "label": "<p>Single choice set label</p>\n"
383 },
384 {
385   "x": 68.62044317369549,
386   "y": 83.40497606918763,
387   "width": 10,
388   "height": 10,
389   "duration": {
390     "from": 0,
391     "to": 10
392   },
393   "libraryTitle": "Multiple Choice",
394   "action": {
395     "library": "H5P.MultiChoice 1.14",
396     "params": {
397       "media": {
398         "disableImageZooming": false
399       },
400       "answers": [
401         {
402           "correct": false,
403           "tipsAndFeedback": {
404             "tip": "<p>MPC tip</p>\n",
405             "chosenFeedback": "<div>MPC if selected</div>\n",
406             "notChosenFeedback": "<div>MPC if not selected</div>\n"
407           },
408           "text": "<div>MPC option</div>\n"
409         },
410         {
411           "correct": false,
412           "tipsAndFeedback": {
413             "tip": "",
414             "chosenFeedback": "",
415             "notChosenFeedback": ""

```

```
416         },
417         "text": "<div>MPC option</div>\n"
418     }
419 ],
420 "overallFeedback": [
421     {
422         "from": 0,
423         "to": 100
424     }
425 ],
426 "behaviour": {
427     "enableRetry": true,
428     "enableSolutionsButton": true,
429     "enableCheckButton": true,
430     "type": "auto",
431     "singlePoint": false,
432     "randomAnswers": true,
433     "showSolutionsRequiresInput": true,
434     "confirmCheckDialog": false,
435     "confirmRetryDialog": false,
436     "autoCheck": false,
437     "passPercentage": 100,
438     "showScorePoints": true
439 },
440 "UI": {
441     "checkAnswerButton": "Check",
442     "submitAnswerButton": "Submit",
443     "showSolutionButton": "Show solution",
444     "tryAgainButton": "Retry",
445     "tipsLabel": "Show tip",
446     "scoreBarLabel": "You got :num out of :total
447         points",
448     "tipAvailable": "Tip available",
449     "feedbackAvailable": "Feedback available",
450     "readFeedback": "Read feedback",
451     "wrongAnswer": "Wrong answer",
452     "correctAnswer": "Correct answer",
453     "shouldCheck": "Should have been checked",
454     "shouldNotCheck": "Should not have been checked",
455     "noInput": "Please answer before viewing the
456         solution",
457     "a11yCheck": "Check the answers. The responses
458         will be marked as correct, incorrect, or
459         unanswered."
```

```

456         "allyShowSolution": "Show the solution. The task
           will be marked with its correct solution.",
457         "allyRetry": "Retry the task. Reset all responses
           and start the task over again."
458     },
459     "confirmCheck": {
460         "header": "Finish ?",
461         "body": "Are you sure you wish to finish ?",
462         "cancelLabel": "Cancel",
463         "confirmLabel": "Finish"
464     },
465     "confirmRetry": {
466         "header": "Retry ?",
467         "body": "Are you sure you wish to retry ?",
468         "cancelLabel": "Cancel",
469         "confirmLabel": "Confirm"
470     },
471     "question": "<p>MPC question</p>\n"
472 },
473     "subContentId": "094f8f9f-a0de-416f-b0b2-351fe2952e05"
474     ,
475     "metadata": {
476         "contentType": "Multiple Choice",
477         "license": "U",
478         "title": "Untitled Multiple Choice"
479     },
480     "pause": false,
481     "displayType": "button",
482     "buttonOnMobile": false,
483     "adaptivity": {
484         "correct": {
485             "allowOptOut": false,
486             "message": ""
487         },
488         "wrong": {
489             "allowOptOut": false,
490             "message": ""
491         },
492         "requireCompletion": false
493     },
494     "label": "<p>MPC label</p>\n"
495 },
496 {

```

```
497     "x": 69.25528231102004,  
498     "y": 24.97560626270041,  
499     "width": 10,  
500     "height": 10,  
501     "duration": {  
502         "from": 0,  
503         "to": 10  
504     },  
505     "libraryTitle": "True/False Question",  
506     "action": {  
507         "library": "H5P.TrueFalse 1.6",  
508         "params": {  
509             "media": {  
510                 "disableImageZooming": false  
511             },  
512             "correct": "true",  
513             "behaviour": {  
514                 "enableRetry": true,  
515                 "enableSolutionsButton": true,  
516                 "enableCheckButton": true,  
517                 "confirmCheckDialog": false,  
518                 "confirmRetryDialog": false,  
519                 "autoCheck": false  
520             },  
521             "l10n": {  
522                 "trueText": "True",  
523                 "falseText": "False",  
524                 "score": "You got @score of @total points",  
525                 "checkAnswer": "Check",  
526                 "submitAnswer": "Submit",  
527                 "showSolutionButton": "Show solution",  
528                 "tryAgain": "Retry",  
529                 "wrongAnswerMessage": "Wrong answer",  
530                 "correctAnswerMessage": "Correct answer",  
531                 "scoreBarLabel": "You got :num out of :total  
532                     points",  
533                 "a11yCheck": "Check the answers. The responses  
534                     will be marked as correct, incorrect, or  
535                     unanswered.",  
536                 "a11yShowSolution": "Show the solution. The task  
537                     will be marked with its correct solution.",  
538                 "a11yRetry": "Retry the task. Reset all responses  
539                     and start the task over again."  
540             },  
541             "l10n": {  
542                 "trueText": "True",  
543                 "falseText": "False",  
544                 "score": "You got @score of @total points",  
545                 "checkAnswer": "Check",  
546                 "submitAnswer": "Submit",  
547                 "showSolutionButton": "Show solution",  
548                 "tryAgain": "Retry",  
549                 "wrongAnswerMessage": "Wrong answer",  
550                 "correctAnswerMessage": "Correct answer",  
551                 "scoreBarLabel": "You got :num out of :total  
552                     points",  
553                 "a11yCheck": "Check the answers. The responses  
554                     will be marked as correct, incorrect, or  
555                     unanswered.",  
556                 "a11yShowSolution": "Show the solution. The task  
557                     will be marked with its correct solution.",  
558                 "a11yRetry": "Retry the task. Reset all responses  
559                     and start the task over again."  
560             }  
561         }  
562     }  
563 }
```

```

536     "confirmCheck": {
537         "header": "Finish ?",
538         "body": "Are you sure you wish to finish ?",
539         "cancelLabel": "Cancel",
540         "confirmLabel": "Finish"
541     },
542     "confirmRetry": {
543         "header": "Retry ?",
544         "body": "Are you sure you wish to retry ?",
545         "cancelLabel": "Cancel",
546         "confirmLabel": "Confirm"
547     },
548     "question": "<p>True false question</p>\n"
549 },
550 "subContentId": "cf3827ae-a761-42e6-b2e3-23833bf22716"
551 ,
552 "metadata": {
553     "contentType": "True\False Question",
554     "license": "U",
555     "title": "Untitled True\False Question"
556 }
557 },
558 "pause": false,
559 "displayType": "button",
560 "buttonOnMobile": false,
561 "adaptivity": {
562     "correct": {
563         "allowOptOut": false,
564         "message": ""
565     },
566     "wrong": {
567         "allowOptOut": false,
568         "message": ""
569     },
570     "requireCompletion": false
571 },
572 "label": "<p>True false label</p>\n"
573 {
574     "x": 68.62044317369549,
575     "y": 5.774190650943759,
576     "width": 10,
577     "height": 10,
578     "duration": {

```



```
579     "from": 0,
580     "to": 10
581   },
582   "libraryTitle": "Fill in the Blanks",
583   "action": {
584     "library": "H5P.Blanks 1.12",
585     "params": {
586       "media": {
587         "disableImageZooming": false
588       },
589       "text": "<p>Task description</p>\n",
590       "overallFeedback": [
591         {
592           "from": 0,
593           "to": 100
594         }
595       ],
596       "showSolutions": "Show solution",
597       "tryAgain": "Retry",
598       "checkAnswer": "Check",
599       "submitAnswer": "Submit",
600       "notFilledOut": "Please fill in all blanks to view
        solution",
601       "answerIsCorrect": "&#039;:ans&#039; is correct",
602       "answerIsWrong": "&#039;:ans&#039; is wrong",
603       "answeredCorrectly": "Answered correctly",
604       "answeredIncorrectly": "Answered incorrectly",
605       "solutionLabel": "Correct answer:",
606       "inputLabel": "Blank input @num of @total",
607       "inputHasTipLabel": "Tip available",
608       "tipLabel": "Tip",
609       "behaviour": {
610         "enableRetry": true,
611         "enableSolutionsButton": true,
612         "enableCheckButton": true,
613         "autoCheck": false,
614         "caseSensitive": true,
615         "showSolutionsRequiresInput": true,
616         "separateLines": false,
617         "confirmCheckDialog": false,
618         "confirmRetryDialog": false,
619         "acceptSpellingErrors": false
620       },
621       "scoreBarLabel": "You got :num out of :total points"
```

```

622         '
        "allyCheck": "Check the answers. The responses will
            be marked as correct, incorrect, or unanswered."
        ,
623         "allyShowSolution": "Show the solution. The task
            will be marked with its correct solution.",
624         "allyRetry": "Retry the task. Reset all responses
            and start the task over again.",
625         "allyCheckingModeHeader": "Checking mode",
626         "confirmCheck": {
627             "header": "Finish ?",
628             "body": "Are you sure you wish to finish ?",
629             "cancelLabel": "Cancel",
630             "confirmLabel": "Finish"
631         },
632         "confirmRetry": {
633             "header": "Retry ?",
634             "body": "Are you sure you wish to retry ?",
635             "cancelLabel": "Cancel",
636             "confirmLabel": "Confirm"
637         },
638         "questions": [
639             "<p>H5P content may be edited using a *browser\
                web-browser:Something you use every day*.<\
                >\n"
640         ]
641     },
642     "subContentId": "abbedc3a-a3f0-46d6-91e6-5f5a2ae27c4d"
        ,
643     "metadata": {
644         "contentType": "Fill in the Blanks",
645         "license": "U",
646         "title": "Untitled Fill in the Blanks"
647     }
648 },
649 "pause": false,
650 "displayType": "button",
651 "buttonOnMobile": false,
652 "adaptivity": {
653     "correct": {
654         "allowOptOut": false,
655         "message": ""
656     },
657     "wrong": {

```

```
658         "allowOptOut": false,
659         "message": ""
660     },
661     "requireCompletion": false
662 },
663 "label": "<p>Fill in the blanks label</p>\n"
664 },
665 {
666     "x": 75.47868325601954,
667     "y": 69.10812915700615,
668     "width": 10,
669     "height": 10,
670     "duration": {
671         "from": 0,
672         "to": 10
673     },
674     "libraryTitle": "Drag and Drop",
675     "action": {
676         "library": "H5P.DragQuestion 1.13",
677         "params": {
678             "scoreShow": "Check",
679             "submit": "Submit",
680             "tryAgain": "Retry",
681             "scoreExplanation": "Correct answers give +1 point.
682                 Incorrect answers give -1 point. The lowest
683                 possible score is 0.",
684             "question": {
685                 "settings": {
686                     "size": {
687                         "width": 620,
688                         "height": 310
689                     },
690                     "background": {
691                         "path": "images/background-628d2b258af5e.jpg"
692                     },
693                     "mime": "image/jpeg",
694                     "copyright": {
695                         "license": "U"
696                     },
697                     "width": 821,
698                     "height": 961
699                 }
700             }
701         }
702     }
703 },
704     "task": {
```

```
699     "elements": [
700       {
701         "x": 72.58064516129032,
702         "y": 22.58064516129032,
703         "width": 5,
704         "height": 5,
705         "dropZones": [
706
707         ],
708         "type": {
709           "library": "H5P.AdvancedText 1.1",
710           "params": {
711             "text": "<p>Drag and drop task</p>\n"
712           },
713           "subContentId": "aa1a8153-5f68-42fb-8b10-3
714             9ba46441e03",
715           "metadata": {
716             "contentType": "Text",
717             "license": "U",
718             "title": "Untitled Text"
719           }
720         },
721         "backgroundOpacity": 100,
722         "multiple": false
723       }
724     ],
725     "dropZones": [
726       {
727         "x": 43.54838709677419,
728         "y": 43.54838709677419,
729         "width": 5,
730         "height": 2.5,
731         "correctElements": [
732
733         ],
734         "showLabel": false,
735         "backgroundOpacity": 100,
736         "tipsAndFeedback": {
737           "tip": "<p>Drag and drop dropzone tip</p>
738             >\n",
739           "feedbackOnCorrect": "DnD correct match",
740           "feedbackOnIncorrect": "DnD incorrect
741             match"
742         }
743       }
744     ]
745   }
746 }
```

```
740         "single": false,
741         "autoAlign": false,
742         "label": "<div>Drag and drop dropzone</div>
              >\n"
743     }
744 ]
745 }
746 },
747 "overallFeedback": [
748     {
749         "from": 0,
750         "to": 100
751     }
752 ],
753 "behaviour": {
754     "enableRetry": true,
755     "enableCheckButton": true,
756     "showSolutionsRequiresInput": true,
757     "singlePoint": false,
758     "applyPenalties": true,
759     "enableScoreExplanation": true,
760     "dropZoneHighlighting": "dragging",
761     "autoAlignSpacing": 2,
762     "enableFullScreen": false,
763     "showScorePoints": true,
764     "showTitle": true
765 },
766 "grabbablePrefix": "Grabbable {num} of {total}.",
767 "grabbableSuffix": "Placed in dropzone {num}.",
768 "dropzonePrefix": "Dropzone {num} of {total}.",
769 "noDropzone": "No dropzone.",
770 "tipLabel": "Show tip.",
771 "tipAvailable": "Tip available",
772 "correctAnswer": "Correct answer",
773 "wrongAnswer": "Wrong answer",
774 "feedbackHeader": "Feedback",
775 "scoreBarLabel": "You got :num out of :total points"
776     ,
777 "scoreExplanationButtonLabel": "Show score
              explanation",
778 "allyCheck": "Check the answers. The responses will
              be marked as correct, incorrect, or unanswered."
              ,
779 "allyRetry": "Retry the task. Reset all responses"
```

```
        and start the task over again.",
779     "localize": {
780         "fullscreen": "Fullscreen",
781         "exitFullscreen": "Exit fullscreen"
782     }
783 },
784 "subContentId": "b0a8ad9b-9330-422e-8f0c-69230abe816c"
,
785 "metadata": {
786     "contentType": "Drag and Drop",
787     "license": "U",
788     "title": "Untitled Drag and Drop"
789 }
790 },
791 "pause": false,
792 "displayType": "button",
793 "buttonOnMobile": false,
794 "adaptivity": {
795     "correct": {
796         "allowOptOut": false,
797         "message": ""
798     },
799     "wrong": {
800         "allowOptOut": false,
801         "message": ""
802     },
803     "requireCompletion": false
804 },
805 "label": "<p>Drag and drop label&nbsp;&lt;/p>\n"
806 },
807 {
808     "x": 43.48721704698908,
809     "y": 67.40837276188152,
810     "width": 10,
811     "height": 10,
812     "duration": {
813         "from": 0,
814         "to": 10
815     },
816     "libraryTitle": "Mark the Words",
817     "action": {
818         "library": "H5P.MarkTheWords 1.9",
819         "params": {
820             "overallFeedback": [
```

```
821     {
822         "from": 0,
823         "to": 100
824     }
825 ],
826 "checkAnswerButton": "Check",
827 "submitAnswerButton": "Submit",
828 "tryAgainButton": "Retry",
829 "showSolutionButton": "Show solution",
830 "behaviour": {
831     "enableRetry": true,
832     "enableSolutionsButton": true,
833     "enableCheckButton": true,
834     "showScorePoints": true
835 },
836 "correctAnswer": "Correct!",
837 "incorrectAnswer": "Incorrect!",
838 "missedAnswer": "Answer not found!",
839 "displaySolutionDescription": "Task is updated to
840     contain the solution.",
841 "scoreBarLabel": "You got :num out of :total points"
842 ,
843 "allyFullTextLabel": "Full readable text",
844 "allyClickableTextLabel": "Full text where words can
845     be marked",
846 "allySolutionModeHeader": "Solution mode",
847 "allyCheckingHeader": "Checking mode",
848 "allyCheck": "Check the answers. The responses will
849     be marked as correct, incorrect, or unanswered."
850 ,
851 "allyShowSolution": "Show the solution. The task
852     will be marked with its correct solution.",
853 "allyRetry": "Retry the task. Reset all responses
854     and start the task over again.",
855 "taskDescription": "<p>MTW task description</p>\n",
856 "textField": "<p>The correct words are marked like
857     this: *correctword*, an asterisk is written like
858     this: *correctword***.</p>\n"
859 },
860 "subContentId": "fa54c7d3-dbf5-46a9-ba20-17862b86ffdf"
861 ,
862 "metadata": {
863     "contentType": "Mark the Words",
864     "license": "U",
```

```

855     "title": "Untitled Mark the Words"
856   }
857 },
858 "pause": false,
859 "displayType": "button",
860 "buttonOnMobile": false,
861 "adaptivity": {
862   "correct": {
863     "allowOptOut": false,
864     "message": ""
865   },
866   "wrong": {
867     "allowOptOut": false,
868     "message": ""
869   },
870   "requireCompletion": false
871 },
872 "label": "<p>Mark the w Label</p>\n"
873 },
874 {
875   "x": 10.85105805794784,
876   "y": 95.06843674189189,
877   "width": 10,
878   "height": 10,
879   "duration": {
880     "from": 0,
881     "to": 10
882   },
883   "libraryTitle": "Drag the Words",
884   "action": {
885     "library": "H5P.DragText 1.8",
886     "params": {
887       "taskDescription": "<p>Drag the words into the
888         correct boxes</p>\n",
889       "overallFeedback": [
890         {
891           "from": 0,
892           "to": 100
893         }
894       ],
895       "checkAnswer": "Check",
896       "submitAnswer": "Submit",
897       "tryAgain": "Retry",
898       "showSolution": "Show solution",

```



```
898     "dropZoneIndex": "Drop Zone @index.",
899     "empty": "Drop Zone @index is empty.",
900     "contains": "Drop Zone @index contains draggable
901               @draggable.",
902     "ariaDraggableIndex": "@index of @count draggables."
903     ,
904     "tipLabel": "Show tip",
905     "correctText": "Correct!",
906     "incorrectText": "Incorrect!",
907     "resetDropTitle": "Reset drop",
908     "resetDropDescription": "Are you sure you want to
909                             reset this drop zone?",
910     "grabbed": "Draggable is grabbed.",
911     "cancelledDragging": "Cancelled dragging.",
912     "correctAnswer": "Correct answer:",
913     "feedbackHeader": "Feedback",
914     "behaviour": {
915       "enableRetry": true,
916       "enableSolutionsButton": true,
917       "enableCheckButton": true,
918       "instantFeedback": false
919     },
920     "scoreBarLabel": "You got :num out of :total points"
921     ,
922     "allyCheck": "Check the answers. The responses will
923                 be marked as correct, incorrect, or unanswered."
924     ,
925     "allyShowSolution": "Show the solution. The task
926                          will be marked with its correct solution.",
927     "allyRetry": "Retry the task. Reset all responses
928                  and start the task over again.",
929     "textField": "H5P content may be edited using a *
930                  browser:What type of program is Chrome?*.\\nH5P
931                  content is *interactive\\+Correct! \\-Incorrect,
932                  try again!*"
933   },
934   "subContentId": "e88f7373-287a-44d3-9ee9-f13384c5098f"
935   ,
936   "metadata": {
937     "contentType": "Drag the Words",
938     "license": "U",
939     "title": "Untitled Drag the Words"
940   }
941 },
```

```

930     "pause": false,
931     "displayType": "button",
932     "buttonOnMobile": false,
933     "adaptivity": {
934       "correct": {
935         "allowOptOut": false,
936         "message": ""
937       },
938       "wrong": {
939         "allowOptOut": false,
940         "message": ""
941       },
942       "requireCompletion": false
943     },
944     "label": "<p>Drag qords label</p>\n"
945   },
946   {
947     "x": 4.576659038901601,
948     "y": 41.7201540436457,
949     "width": 10,
950     "height": 10,
951     "duration": {
952       "from": 0,
953       "to": 10
954     },
955     "libraryTitle": "Crossroads",
956     "action": {
957       "library": "H5P.GoToQuestion 1.3",
958       "params": {
959         "choices": [
960           {
961             "text": "Choice 1",
962             "ifChosenText": "If chosen text"
963           },
964           {
965             "text": "Choice 2",
966             "ifChosenText": "If chosen text"
967           }
968         ],
969         "continueButtonLabel": "Continue",
970         "text": "Crossroad question"
971       },
972       "subContentId": "9e011e46-96ce-448c-a0ea-4df23dfffc37d"

```

```
973     "metadata": {
974         "contentType": "Crossroads",
975         "license": "U"
976     }
977 },
978 "pause": true,
979 "displayType": "poster",
980 "buttonOnMobile": false
981 },
982 {
983     "x": 6.864988558352403,
984     "y": 20.539152759948653,
985     "width": 10,
986     "height": 10,
987     "duration": {
988         "from": 0,
989         "to": 10
990     },
991     "libraryTitle": "Navigation Hotspot",
992     "action": {
993         "library": "H5P.IVHotspot 1.2",
994         "params": {
995             "destination": {
996                 "type": "timecode",
997                 "url": {
998                     "protocol": "http:\\\\\\"
999             }
1000         },
1001         "visuals": {
1002             "shape": "rectangular",
1003             "backgroundColor": "rgba(255, 255, 255, 0)",
1004             "pointerCursor": true,
1005             "animation": false
1006         },
1007         "texts": {
1008             "showLabel": false,
1009             "labelColor": "rgb(0, 0, 0)",
1010             "alternativeText": "Go to alt"
1011         }
1012     },
1013     "subContentId": "84351fe5-0deb-445a-9efe-81a4a674e421"
1014     ,
1015     "metadata": {
1016         "contentType": "Navigation Hotspot",
```

```
1016         "license": "U"
1017     }
1018 },
1019     "pause": false,
1020     "displayType": "poster",
1021     "buttonOnMobile": false
1022 },
1023 {
1024     "x": 79.1488940697372,
1025     "y": 57.02315838473176,
1026     "width": 10,
1027     "height": 10,
1028     "duration": {
1029         "from": 0,
1030         "to": 10
1031     },
1032     "libraryTitle": "Free Text Question",
1033     "action": {
1034         "library": "H5P.FreeTextQuestion 1.0",
1035         "params": {
1036             "placeholder": "Free text placeholder",
1037             "maxScore": 1,
1038             "isRequired": false,
1039             "i10n": {
1040                 "requiredText": "required",
1041                 "requiredMessage": "This question requires an
1042                     answer",
1043                 "skipButtonLabel": "Skip Question",
1044                 "submitButtonLabel": "Answer and proceed",
1045                 "language": "en"
1046             },
1047             "question": "Free text question"
1048         },
1049         "subContentId": "69aea51d-37f7-4158-81aa-aa966b5544c4"
1050     },
1051     "metadata": {
1052         "contentType": "Free Text Question",
1053         "license": "U",
1054         "title": "Untitled Free Text Question"
1055     }
1056 },
1057     "pause": false,
1058     "displayType": "button",
1059     "buttonOnMobile": false,
```

```
1058     "label": "<p>Free text question label</p>\n"
1059   },
1060   {
1061     "x": 29.735525375268047,
1062     "y": 5.132613911950008,
1063     "width": 10,
1064     "height": 10,
1065     "duration": {
1066       "from": 1.194,
1067       "to": 11.193999999999999
1068     },
1069     "libraryTitle": "Label",
1070     "action": {
1071       "library": "H5P.Nil 1.0",
1072       "params": {
1073
1074     },
1075     "subContentId": "44ed3bed-6683-4fdd-99e7-8d5d9e6dbd01"
1076
1077     ,
1078     "metadata": {
1079       "contentType": "Label",
1080       "license": "U"
1081     }
1082   },
1083   "label": "<p>LAbelLorem ipsum dolor sit amet...</p>\n",
1084   "pause": false,
1085   "displayType": "button",
1086   "buttonOnMobile": false
1087 }
1088 ],
1089 "endscreens": [
1090   {
1091     "time": 14.976,
1092     "label": "0:14 Submit screen"
1093   }
1094 ]
1095 },
1096 "summary": {
1097   "task": {
1098     "library": "H5P.Summary 1.10",
1099     "params": {
1100       "intro": "<p>Choose the correct statement. Summary&nbsp;
1101         ;</p>\n",
1102       "summaries": [
```

```
1100     {
1101         "subContentId": "4d4866e5-3401-4d00-b8f3-643b652fe42
1102             d",
1103         "tip": "",
1104         "summary": [
1105             "<p>Statement 1</p>\n",
1106             "<p>Statement 2</p>\n"
1107         ]
1108     },
1109     "overallFeedback": [
1110         {
1111             "from": 0,
1112             "to": 100
1113         }
1114     ],
1115     "solvedLabel": "Progress:",
1116     "scoreLabel": "Wrong answers:",
1117     "resultLabel": "Your result",
1118     "labelCorrect": "Correct.",
1119     "labelIncorrect": "Incorrect! Please try again.",
1120     "alternativeIncorrectLabel": "Incorrect",
1121     "labelCorrectAnswers": "Correct answers.",
1122     "tipButtonLabel": "Show tip",
1123     "scoreBarLabel": "You got :num out of :total points",
1124     "progressText": "Progress :num of :total"
1125 },
1126 "subContentId": "ffe5cdc8-e92f-4520-a019-3d8dd9aec6ef",
1127 "metadata": {
1128     "contentType": "Summary",
1129     "license": "U",
1130     "title": "Untitled Summary"
1131 }
1132 },
1133 "displayAt": 3
1134 }
1135 },
1136 "override": {
1137     "autoplay": false,
1138     "loop": false,
1139     "showBookmarksMenuOnLoad": false,
1140     "showRewind10": false,
1141     "preventSkipping": false,
1142     "deactivateSound": false
```

```
1143 },
1144 "l10n": {
1145     "interaction": "Interaction",
1146     "play": "Play",
1147     "pause": "Pause",
1148     "mute": "Mute, currently unmuted",
1149     "unmute": "Unmute, currently muted",
1150     "quality": "Video Quality",
1151     "captions": "Captions",
1152     "close": "Close",
1153     "fullscreen": "Fullscreen",
1154     "exitFullscreen": "Exit Fullscreen",
1155     "summary": "Open summary dialog",
1156     "bookmarks": "Bookmarks",
1157     "endscreen": "Submit screen",
1158     "defaultAdaptivitySeekLabel": "Continue",
1159     "continueWithVideo": "Continue with video",
1160     "playbackRate": "Playback Rate",
1161     "rewind10": "Rewind 10 Seconds",
1162     "navDisabled": "Navigation is disabled",
1163     "sndDisabled": "Sound is disabled",
1164     "requiresCompletionWarning": "You need to answer all the
        questions correctly before continuing.",
1165     "back": "Back",
1166     "hours": "Hours",
1167     "minutes": "Minutes",
1168     "seconds": "Seconds",
1169     "currentTime": "Current time:",
1170     "totalTime": "Total time:",
1171     "singleInteractionAnnouncement": "Interaction appeared:",
1172     "multipleInteractionsAnnouncement": "Multiple interactions
        appeared.",
1173     "videoPausedAnnouncement": "Video is paused",
1174     "content": "Content",
1175     "answered": "@answered answered",
1176     "endcardTitle": "@answered Question(s) answered",
1177     "endcardInformation": "You have answered @answered questions,
        click below to submit your answers.",
1178     "endcardInformationOnSubmitButtonDisabled": "You have answered
        @answered questions.",
1179     "endcardInformationNoAnswers": "You have not answered any
        questions.",
1180     "endcardInformationMustHaveAnswer": "You have to answer at
        least one question before you can submit your answers.",
```

```
1181     "endcardSubmitButton": "Submit Answers",
1182     "endcardSubmitMessage": "Your answers have been submitted!",
1183     "endcardTableRowAnswered": "Answered questions",
1184     "endcardTableRowScore": "Score",
1185     "endcardAnsweredScore": "answered",
1186     "endCardTableRowSummaryWithScore": "You got @score out of
        @total points for the @question that appeared after
        @minutes minutes and @seconds seconds.",
1187     "endCardTableRowSummaryWithoutScore": "You have answered the
        @question that appeared after @minutes minutes and
        @seconds seconds."
1188   }
1189 }
```


Appendix B

Information letter and consent form

Attached is the information letter and consent form sent to the participants.

Are you interested in taking part in the research project “Measuring the quality of OER automatically using characteristics”?

This is an inquiry about participation in a research project where the main purpose is to research the use of characteristic metrics to automatically evaluate Open Educational Resources (OERs). In this letter we will give you information about the purpose of the project and what your participation will involve.

Purpose of the project

This project is a master's thesis conducted as a part of a master's in computer sciences at Norwegian University of Science and Technology. The purpose of this is to research a method for automated quality reviews of Open Educational Resources. This research has the goal of finding if characteristic metrics of an Open Educational Resource can be used for this purpose and if an algorithm using this could provide a quality assessment that can be a good supplement to manual reviews. The resources used for this research will be resources created with the tool H5P (<https://h5p.org/>). To have empirical data to evaluate there is a need for manually reviewed resources to compare the results with.

The research questions are as follows:

- Are the current automated approaches applicable to other repositories?
- Which metrics are most correlated with a high quality OER?
- How many metrics are needed to get a high accuracy estimation of the quality of an OER?
- Is it possible to gain good accuracy on a range of quality or is the use limited to differentiating good and poor-quality resources?
- Do some of the metrics have a high correlation with some of the properties in LORI?

Who is responsible for the research project?

Norwegian University of Science and Technology (NTNU) is the institution responsible for the project.

The supervisor at NTNU is Sofia Papavlasopoulou.

The project is also done in cooperation with Joubel AS with the contact person being Svein-Tore Griff With.

Why are you being asked to participate?

You have been selected to participate in this research because you can perform manual reviews of Open Educational Resources.

What does participation involve for you?

The participation involves filling out an online survey with questions for a quality review of one or more Open Educational Resources. In addition, some background questions about your experience with Open Educational Resources and profession will be asked.

Participation is voluntary

Participation in the project is voluntary. If you chose to participate, you can withdraw your consent at

any time without giving a reason. All information about you will then be made anonymous. There will be no negative consequences for you if you chose not to participate or later decide to withdraw.

Your personal privacy – how we will store and use your personal data

We will only use your personal data for the purpose(s) specified in this information letter. We will process your personal data confidentially and in accordance with data protection legislation (the General Data Protection Regulation and Personal Data Act).

The only people having access to the personal data are the student, supervisor, and the contact person in the company in which this project is done in cooperation with.

I will replace your name and contact details with a code. The list of names, contact details and respective codes will be stored separately from the rest of the collected data.

To collect the data Nettskjema will be used as a provider.

The participants will not be recognized in the publications, but notion of the type of occupation and experience Open Educational Resources and if relevant the experience with manual review.

What will happen to your personal data at the end of the research project?

The project is scheduled to end on July 31st, 2022. After this period, all data will be anonymised, and no further personal information will be kept.

Your rights

So long as you can be identified in the collected data, you have the right to:

- access the personal data that is being processed about you
- request that your personal data is deleted
- request that incorrect personal data about you is corrected/rectified
- receive a copy of your personal data (data portability), and
- send a complaint to the Data Protection Officer or The Norwegian Data Protection Authority regarding the processing of your personal data

What gives us the right to process your personal data?

We will process your personal data based on your consent.

Based on an agreement with Norwegian University of Science and Technology, Data Protection Services has assessed that the processing of personal data in this project is in accordance with data protection legislation.

Where can I find out more?

If you have questions about the project, or want to exercise your rights, contact:

- Norwegian University of Science and Technology via Hanna Eide Solstad, which is the student at hannaes@stud.ntnu.no, or the supervisor Sofia Papavlasopoulou at spapav@ntnu.no.
- Our Data Protection Officer: Thomas Helgesen, by email: thomas.helgesen@ntnu.no
- Data Protection Services, by email: (personverntjenester@sikt.no) or by telephone: +47 53 21 15 00.

Yours sincerely,

Hanna Eide Solstad, Student

Sofia Papavlasopoulou, supervisor

Consent form

I have received and understood information about the project Measuring the quality of OER automatically using characteristics and have been given the opportunity to ask questions. I give consent:

to participate in an online survey

I give consent for my personal data to be processed until the end date of the project, approx. July 31st, 2022.

(Signed by participant, date)

Appendix C

Additional statistics

In this appendix some statistics that did not fit in the results chapter are presented. First are the results from Shapiro-Wilk test of normality for metrics and averages. Following are the correlation matrices between each question and feature scores.

Reviewer	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
P1	5e-4	0.003	0.003	3e-5	5e-5	0.003	0.003	0.001	0.001	0.001	0.001	0.004
P2	9e-05	9e-4	0.001	0.001	0.001	0.001	0.02	0.002	0.002	2e-4	0.008	0.01

Table C.1: Shapiro-Wilk test of normality metrics

Reviewer	Q1-Q10	Q1-Q11	Q1-Q12
P1	0.19	0.27	0.19
P2	0.70	0.59	0.70
All	0.064	0.056	0.049

Table C.2: Shapiro-Wilk test of normality averages

Metric	Correlation r	p-value
combination_imnb_decreasing_5	0.31	0.001
combination_all_decreasing_5	0.29	0.01
combination_all_num	0.26	0.01
combination_all_decreasing_10	0.26	0.01
combination_all_combined_5	0.26	0.01
combination_imnb_num	0.25	0.02
combination_imnb_decreasing_10	0.25	0.02
combination_imnb_combined_5	0.25	0.01
combination_all_time	0.22	0.02

Table C.3: The significant ($p < 0.05$) metric correlations for average(Q1-Q10)

Metric	Correlation r	p-value
combination_imnb_decreasing_5	0.32	0.003
combination_all_decreasing_5	0.28	0.01
combination_imnb_num	0.26	0.01
combination_imnb_decreasing_10	0.26	0.01
combination_imnb_combined_5	0.26	0.01
combination_all_num	0.26	0.01
combination_all_decreasing_10	0.26	0.01
combination_all_combined_5	0.26	0.01
links_num	0.24	0.04
links_time	0.24	0.04
links_decreasing_5	0.24	0.04
links_decreasing_10	0.24	0.04
links_combined_5	0.24	0.04
combination_imnb_time	0.21	0.04
combination_all_time	0.21	0.03

Table C.4: The significant ($p < 0.05$) metric correlations for average(Q1-Q12)

Metric	Correlation r	p-value
navigation_num	0.33	0.01
navigation_time	0.33	0.01
navigation_decreasing_5	0.33	0.01
navigation_decreasing_10	0.33	0.01
navigation_combined_5	0.33	0.01
bookmarks_decreasing_5	0.3	0.01
bookmarks_num	0.29	0.02
bookmarks_time	0.29	0.02
bookmarks_decreasing_10	0.29	0.02
bookmarks_combined_5	0.29	0.02
metadata_availability_score	0.29	0.01
combination_mnb_num	0.26	0.03
combination_mnb_decreasing_5	0.26	0.03
combination_mnb_decreasing_10	0.26	0.03
combination_mnb_combined_5	0.26	0.03
combination_mnb_time	0.25	0.03
combination_imnb_decreasing_5	0.22	0.04

Table C.5: The significant ($p < 0.05$) metric correlations for Q1

Metric	Correlation r	p-value
interactivity_decreasing_5	0.47	2e-05
interactivity_num	0.44	4e-05
interactivity_decreasing_10	0.44	5e-05
interactivity_combined_5	0.44	4e-05
combination_all_decreasing_5	0.44	8e-05
combination_all_num	0.42	8e-05
combination_all_decreasing_10	0.42	8e-05
combination_all_combined_5	0.42	7e-05
interactivity_time_based	0.37	0.0004
combination_all_time	0.33	0.002
combination_imnb_decreasing_5	0.31	0.01
combination_imnb_num	0.28	0.01
combination_imnb_decreasing_10	0.28	0.01
combination_imnb_combined_5	0.28	0.01
combination_imnb_time	0.22	0.03

Table C.6: The significant ($p < 0.05$) metric correlations for Q2

Metric	Correlation r	p-value
combination_all_decreasing_5	0.32	0.004
interactivity_decreasing_5	0.31	0.006
combination_all_time	0.31	0.003
combination_all_num	0.3	0.006
combination_all_decreasing_10	0.3	0.005
combination_all_combined_5	0.3	0.005
interactivity_num	0.28	0.01
interactivity_time_based	0.28	0.01
interactivity_decreasing_10	0.28	0.01
interactivity_combined_5	0.28	0.01
combination_imnb_decreasing_5	0.25	0.02
combination_imnb_combined_5	0.22	0.04

Table C.7: The significant ($p < 0.05$) metric correlations for Q3

Metric	Correlation r	p-value
navigation_num	0.3	0.02
navigation_time	0.3	0.02
navigation_decreasing_5	0.3	0.02
navigation_decreasing_10	0.3	0.02
navigation_combined_5	0.3	0.02
links_num	0.28	0.02
links_time	0.28	0.02
links_decreasing_5	0.28	0.02
links_decreasing_10	0.28	0.02
links_combined_5	0.28	0.02

Table C.8: The significant ($p < 0.05$) metric correlations for Q4

Metric	Correlation r	p-value
text_num	0.31	0.01
text_decreasing_5	0.31	0.01
text_decreasing_10	0.31	0.01
text_combined_5	0.31	0.01
text_time	0.3	0.01
combination_all_decreasing_5	0.24	0.04
combination_all_time	0.22	0.04

Table C.9: The significant ($p < 0.05$) metric correlations for Q5

Metric	Correlation r	p-value
combination_imnb_decreasing_5	0.31	0.01
combination_all_decreasing_5	0.29	0.01
combination_imnb_decreasing_10	0.28	0.01
combination_imnb_combined_5	0.28	0.01
combination_imnb_num	0.27	0.01
combination_all_num	0.26	0.02
combination_all_combined_5	0.26	0.01
combination_all_decreasing_10	0.25	0.02
combination_all_time	0.24	0.02
combination_imnb_time	0.22	0.04

Table C.10: The significant ($p < 0.05$) metric correlations for Q6

Metric	Correlation r	p-value
combination_imnb_decreasing_5	0.28	0.01
combination_imnb_combined_5	0.25	0.02
combination_imnb_num	0.24	0.03
combination_imnb_decreasing_10	0.24	0.03

Table C.11: The significant ($p < 0.05$) metric correlations for Q7

Metric	Correlation r	p-value
text_num	0.26	0.03
text_decreasing_10	0.26	0.03
text_combined_5	0.26	0.03
text_decreasing_5	0.25	0.04

Table C.12: The significant ($p < 0.05$) metric correlations for Q8

Metric	Correlation r	p-value
--------	---------------	---------

Table C.13: The significant ($p < 0.05$) metric correlations for Q9 (none)

Metric	Correlation r	p-value
combination_imnb_decreasing_5	0.34	0.002
combination_imnb_combined_5	0.33	0.003
combination_imnb_num	0.32	0.004
combination_imnb_decreasing_10	0.31	0.004
combination_all_decreasing_10	0.3	0.006
combination_all_combined_5	0.3	0.005
combination_all_num	0.29	0.01
combination_all_decreasing_5	0.28	0.01
combination_imnb_time	0.27	0.01
combination_all_time	0.26	0.01
links_num	0.25	0.04
links_time	0.25	0.04
links_decreasing_5	0.25	0.04
links_decreasing_10	0.25	0.04
links_combined_5	0.25	0.04

Table C.14: The significant ($p < 0.05$) metric correlations for Q10

Metric	Correlation r	p-value
links_time	0.33	0.007
combination_imnb_decreasing_5	0.32	0.004
links_num	0.32	0.008
links_decreasing_5	0.32	0.008
links_decreasing_10	0.32	0.008
links_combined_5	0.32	0.008
combination_imnb_combined_5	0.3	0.006
combination_all_num	0.3	0.005
combination_all_decreasing_5	0.3	0.007
combination_all_decreasing_10	0.3	0.005
combination_all_combined_5	0.3	0.01
combination_imnb_num	0.29	0.01
combination_imnb_decreasing_10	0.29	0.01
navigation_num	0.25	0.04
navigation_time	0.25	0.04
navigation_decreasing_5	0.25	0.04
navigation_decreasing_10	0.25	0.04
navigation_combined_5	0.25	0.04
combination_all_time	0.22	0.03

Table C.15: The significant ($p < 0.05$) metric correlations for Q11

Metric	Correlation r	p-value
links_num	0.35	0.005
links_time	0.35	0.004
links_decreasing_5	0.35	0.005
links_decreasing_10	0.35	0.005
links_combined_5	0.35	0.005
combination_mnb_num	0.29	0.02
combination_mnb_decreasing_10	0.29	0.02
combination_mnb_combined_5	0.29	0.02
navigation_num	0.28	0.03
navigation_time	0.28	0.03
navigation_decreasing_5	0.28	0.03
navigation_decreasing_10	0.28	0.03
navigation_combined_5	0.28	0.03
combination_mnb_time	0.28	0.02
combination_mnb_decreasing_5	0.28	0.02
metadata_availability_score	0.24	0.04
combination_imnb_decreasing_5	0.23	0.04

Table C.16: The significant ($p < 0.05$) metric correlations for Q12

Appendix D

Survey

Following is the survey that all the participants answered. The questions *Review*, *Rating* and *Do you want to add any additional comment* were showed again for additional resources as long as the participant kept answering *yes* to "*Do you have more resources to review*".

[Check the form for accessibility violations](#)

Review of H5Ps

Page 1

Mandatory fields are marked with a star *

Background information

What gender do you identify as? *

- Male
- Female
- Other
- Prefer not to answer

What is your age? *

- 18-24
- 25-34
- 35-44
- 45-54
- 55-64
- 65-74
- 75-84
- 85 or older
- Prefer not to answer

Which of the following best describes your current occupation? *

- Unemployed
- Student
- Science and engineering professional
- Information and communications technology professional
- University and higher education teacher
- Vocational education teacher
- Secondary education teacher
- Primary school or early childhood teacher
- Other teaching professional
- Librarians, archivists, or curator
- Creator of educational material
- Other
- Prefer not to answer

What is the highest degree or level of school you have completed? *

- Less than a high school diploma
- High school degree or equivalent (e.g. GED)
- Some college, no degree
- Associate degree (e.g. AA, AS)
- Bachelor's degree (e.g. BA, BS)
- Master's degree (e.g. MA, MS, MEd)
- Doctorate or professional degree (e.g. MD, DDS, PhD)
- Prefer not to answer


Do you have any experience with Open Educational Resources? *

Open Educational Resource (OER) can be defined as:

"OER are teaching, learning, and research resources that reside in the public domain or have been released under an intellectual property license that permits their free use or re-purposing by others"

- Yes
- No


How many years of experience do you have? *

 This element is only shown when the option "Yes" is selected in the question "Do you have any experience with Open Educational Resources?"



Value

In which way have you used Open Educational Resources? *


 This element is only shown when the option "Yes" is selected in the question "Do you have any experience with Open Educational Resources?"

- Creator
- Reused resource(s)
- End user (learner)
- Shared resource(s)
- Other

Do you have any experience with H5P content (<https://h5p.org/>)? *

- Yes
- No


How many years of experience do you have? *

 This element is only shown when the option "Yes" is selected in the question "Do you have any experience with H5P content (<https://h5p.org/>)?"

0	1	2	3	4	5	6	7	8	9	10
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Value

In which way have you used H5P? *

 This element is only shown when the option "Yes" is selected in the question "Do you have any experience with H5P content (<https://h5p.org/>)?"


- Creator
- Reused resource(s)
- End user (learner)
- Shared resource(s)
- Other

Do you have any experience with reviewing Open Educational Resources? *

Reviewing it is referred to as the task of evaluating the quality of an OER

- Yes
- No

Approximately, how many resources have you reviewed? *

 This element is only shown when the option "Yes" is selected in the question "Do you have any experience with reviewing Open Educational Resources?"

0 100 200 300 400 500 600 700 800 900 1000



Value

 Page break

Page 2

Mandatory fields are marked with a star *

In the email, you have received a list of contents. In these questions, the general word resource is used instead of content, but the meaning is the same and refers to one H5P content. Here you can rate each of them. After each review, you can continue reviewing the next by answering "Yes" to the question "Do you have more resources to review". When you are finished you submit by pressing "Send".

(1) Review

(1) What is the id of the resource you are currently reviewing? *

The ID can be found in the document that was shared with you.

(1) Rating

Please rate the resource by indicating how much you agree or disagree with the following statements.

	1 (Strongly disagree)	2	3	4- (Neither agree or disagree)	5	6	7 (Strongly agree)
1. The intended learning outcomes are made clear to the learners. The content, learning activities, tasks, and assessment presented are consistent with these learning outcomes. *	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2. The resource include a variety of self-assessments such as multiple-choice, concept questions, and comprehension tests. *	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3. The material contains interactive elements that can be used by the learners to independently perform constructive or manipulative actions. *	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4. The structure is simple and clear. Learners can stop the learning sequence at any time. All learning content (previously presented) can be accessed at any time. *	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

5. All texts and graphics are easy to read. The interface always responds quickly to learner input. *

6. The interactions are understandable and easy to use. *

7. The metadata allow others to effectively use that information to search and evaluate the resource's relevance. *

8. The contents are presented in such a generic way that they can be used in other contexts without much effort. *

9. The visual and auditory information is presented in a clear, concise, and coherent way, taking care with sound quality. *

10. The resource motivate and is able to hold learners's interest. *

11. I am likely to recommend this resource to a friend or colleague. *

12. This is a high quality learning resource. *

(1) Do you want to add any additional comment?

(2) Do you have more resources to review?

Yes

No

i This element is only shown when the option "Yes" is selected in the question "(2) Do you have more resources to review?"

(2) Review

