

Análise quantitativa da concordância entre pares de avaliadores de recursos digitais educacionais: um estudo de caso com dados do repositório Merlot

Mayara Sousa Stein¹, Cristian Cechinel¹, Vinicius F. C. Ramos¹

¹Centro de Ciências, Tecnologias e Saúde (CTS)
Universidade Federal de Santa Catarina (UFSC)
Araranguá – SC – Brasil

stein.mayara@gmail.com, contato@cristiancechinel.pro.br, v.ramos@ufsc.br

Abstract. *Community evaluation (through comments and ratings) is one of the most common strategies adopted by learning resources repositories to assess quality of the resources. In these cases, the final rating of a given resource is calculated by the average value of all ratings given by this community. Despite its wide adoption and consolidation, this strategy normally hides the distribution of the ratings and the nuances of the assessments given from users with different backgrounds and disciplines of interest. The present paper analyzes the inter-rater agreement between pairs of reviewers of the Merlot repository and using the Intra-class correlation coefficient (ICC). The results show a very high disagreement between reviewers of the same resources.*

Resumo. *Dentre as estratégias comumente adotadas pelos repositórios de recursos educacionais para a garantia da qualidade dos recursos está a avaliação dos mesmos pela própria comunidade de usuários (por meio de comentários e notas). A apresentação final da qualidade do recurso (nota) é então realizada a partir do cálculo da média geral das pontuações fornecidas por essa comunidade. Apesar da consolidação dessa estratégia e da sua larga adoção, a mesma acaba escondendo a distribuição das pontuações obtidas e as nuances dessas avaliações com respeito às características específicas dos diferentes tipos de usuários avaliadores (nível educacional, atuação profissional, assuntos de interesse). O presente artigo apresenta uma análise quantitativa da concordância entre pares de avaliadores de recursos digitais educacionais do repositório Merlot, utilizando o coeficiente ICC (Intra-class correlation coefficient) para medição do grau de concordância. Os resultados apontam para uma alta discordância entre os avaliadores.*

1. Introdução

Os repositórios de recursos educacionais digitais (ou repositórios de objetos de aprendizagem - ROA) são os sistemas que possibilitam o armazenamento (físico ou do metadado) dos recursos e que fornecem as funcionalidades essenciais para seu compartilhamento, manutenção e utilização. O aumento da quantidade de recursos educacionais digitais dentro dos diferentes repositórios existentes exige com que essas plataformas adotem estratégias para a garantia da qualidade dos recursos armazenados. Considerando que os

ROA normalmente atraem comunidades de usuários que compartilham dos mesmos interesses [Han et al. 2008], uma das estratégias comumente utilizada para a garantia da qualidade dos recursos é aproveitar a colaboração e as interações dessas comunidades de prática no sentido de avaliar e opinar (impressões de uso e qualidade) sobre os diferentes recursos armazenados [Clements et al. 2015]. A garantia da qualidade dos recursos é uma tarefa extremamente complexa por envolver diferentes aspectos e dimensões (qualidade do conteúdo, usabilidade, facilidade de uso, etc) [Cechinel et al. 2011] além de possuir fundamental importância dentro dos repositórios. Essa importância se dá uma vez que as informações de qualidade (pontuações) são utilizadas pelos algoritmos de ranqueamento dessas plataformas para posicionar os recursos após uma busca realizada pelos usuários, pelos algoritmos de recomendação (quando os mesmos estão implementados), além de servir como referência para os usuários que estão buscando algum tipo de material específico [Downes 2017].

Apesar dessa estratégia de avaliação pela comunidade já ser amplamente consolidada e adotada pelos principais ROA existentes [Clements et al. 2015], a apresentação da qualidade de cada recurso acaba sendo realizada por meio de uma nota (pontuação) final única que é calculada pela média de todas as notas fornecidas pela comunidade de usuários. Esse formato de apresentação omite a distribuição das notas efetivamente fornecidas por toda a comunidade, e esconde informações referentes a como os diferentes tipos de grupos de usuários pontuaram um determinado recurso. Por exemplo, um determinado recurso educacional para a área de matemática pode ter sido muito bem avaliado por usuários que sejam professores de matemática, porém mal avaliado por usuários de outras áreas. Ou pode ter sido muito bem avaliado pelos professores que o utilizaram, porém mal avaliado pelos estudantes que participaram das aulas. Essas nuances referentes as avaliações dos diferentes grupos simplesmente desaparecem na apresentação da qualidade dos mesmos dentro dos ROA. Nessa linha, uma pesquisa anterior [Cechinel and Sánchez-Alonso 2011] já demonstrou que existe considerável discordância com relação a qualidade dos recursos entre a comunidade de usuários do Merlot e a comunidade de especialistas avaliadores (*peer-reviewers*).

O presente artigo apresenta uma abordagem quantitativa para medir a concordância entre pares de usuários avaliadores com relação a qualidade dos recursos educacionais avaliados pelos mesmos dentro do repositório Merlot. A metodologia utilizada pode ser replicada por outros repositórios que possuam recursos educacionais avaliados pela comunidade de usuários e os resultados podem ser utilizados para propor novas formas de apresentação da avaliação de qualidade realizada pela comunidade de usuários, auxiliando, por exemplo, no projeto de interface dos ROA. O artigo se propõe a responder as seguintes perguntas de pesquisa:

- **P1:** Qual a concordância geral entre pares de avaliadores de recursos educacionais digitais ?
- **P2:** Qual a concordância entre pares de avaliadores considerando suas preferências por determinadas disciplinas e suas áreas de atuação profissional?

O restante do artigo está estruturado da seguinte maneira. A seção 2 apresenta uma visão geral sobre como é realizada a avaliação da qualidade de recursos digitais educacionais dentro dos ROA e a seção 3 discute brevemente algumas medidas de consistência existentes. A seção 4 descreve os dados utilizados e a metodologia adotada no estudo e a

seção 5 discute os principais resultados encontrados. Finalmente, a seção 6 apresenta as considerações finais e possibilidades de trabalhos futuros.

2. Avaliação da qualidade dentro de repositórios e o Merlot

As abordagens de avaliação da qualidade utilizadas nos ROA são classificadas por [Clements et al. 2015] em três categorias. A primeira categoria abrange as normas de abordagem genérica da qualidade que fornece um padrão ou procedimento independente do domínio como por exemplo o ISO 9000, o que garante que as organizações consigam um selo de aprovação por meio de avaliação externa [Ehlers and Pawlowski 2006]. A segunda categoria envolve as abordagens de qualidade específica, fornecendo ao ROA procedimentos aprimorados por tecnologia (tais como, revisão por especialistas, benchmarking de qualidade, entre outros)[Pawlowski 2007]. A terceira categoria é a que utiliza algum instrumento específico de qualidade em que recursos são avaliados pela comunidade de usuários [Atenas and Havemann 2014].

Essa terceira categoria (utilização de instrumentos de qualidade específicos) é aquela que mais se popularizou entre os ROA para assegurar a qualidade de seus recursos [Clements et al. 2015], sendo que as avaliações podem ser realizadas por meio de critérios de qualidade previamente definidos e estruturados, ou de maneira livre (sem nenhuma estrutura ou critério de qualidade definido). Na classificação de avaliação de recursos digitais em ROA proposta por [Hylén 2006], essa abordagem é considerada aberta (todas as pontuações e comentários são possíveis de serem visualizados no ROA) e descentralizada (não há controle nem gerenciamento direto por parte do ROA, e todas as avaliações são realizadas de maneira livre pelos usuários). Exemplos de repositório que utilizam essa abordagem são: OpenStat Cnx¹, Common Sense Education², Edmodo Spotlight³, o Merlot⁴ e a Plataforma MEC de recursos educacionais digitais⁵.

No caso específico do repositório Merlot (*The Multimedia Educational Resource for Learning and Online Teaching*), além da avaliação por parte da comunidade de usuários (por meio de comentários e pontuações), o ROA também utiliza um modelo de revisão por pares especialistas, além de permitir a organização dos recursos em coleções pessoais e conceder um prêmio (selo) de qualidade para os recursos mais bem pontuados e comentados pelos especialistas e pela comunidade (Merlot classic awards)[Kutluca 2010]. Todas essas estratégias são complementares e utilizadas pelo ROA para o ranqueamento dos recursos nas buscas realizadas pelos usuários.

O presente trabalho lida diretamente com as avaliações dos recursos educacionais digitais dentro do repositório Merlot e que foram realizadas pela comunidade de usuários, portanto pertencendo a terceira categoria de estratégia de avaliação mencionada, e sendo classificada como uma abordagem de avaliação aberta e descentralizada.

3. Concordância e confiabilidade entre avaliadores

Existem basicamente dois termos principais utilizados para trabalhar com o cálculo da consistência entre as pontuações dadas por diferentes avaliadores (às vezes chamados de

¹<https://cnx.org/browse>

²<https://www.commonsense.org/education>

³<https://spotlight.edmodo.com>

⁴<https://www.merlot.org>

⁵<https://plataformaintegrada.mec.gov.br/>

juízes), sendo eles a *concordância* e a *confiabilidade*. Muitas vezes esses termos são utilizados de maneira intercambiável e generalizada, no entanto confiabilidade e concordância de avaliadores possuem algumas diferenças importantes de serem mencionadas [Matos 2014]. A *confiabilidade* refere-se à medida da consistência entre os avaliadores seja na classificação ou na posição relativa das classificações de desempenho/notas fornecidas. Por outro lado, a *concordância* é a medida em que dois ou mais avaliadores que usam a mesma escala de classificação e avaliam os mesmos materiais fornecem as mesmas notas, ou seja a concordância mede a frequência com que os avaliadores atribuem exatamente a mesma classificação/pontuação para os recursos/objetos pontuados(avaliados)[Tinsley and Weiss 2000].

Uma das medidas de concordância bastante conhecida e utilizada para variáveis binárias é o coeficiente Kappa de Cohen que permite a avaliação simultânea de no máximo dois avaliadores [Matos 2014]. O coeficiente Kappa é um método estatístico que considera a concordância ao acaso, ou seja calcula a proporção de concordância entre os juízes após subtrair a proporção de concordância devido ao acaso [Fonseca et al. 2014]. Outra medida de concordância existente (mas que também pode ser utilizada para medir confiabilidade) é o ICC (Intra-class correlation) que permite o cálculo da concordância entre avaliadores para dados quantitativos e comparando a variabilidade de diferentes classificações do mesmo objeto com a variação total em todas as classificações dos objetos[Koo and Li 2016]. Outros métodos existentes para o cálculo da consistência entre avaliadores são o Kappa Ponderado (concordância) e Alpha de Krippendorff (confiabilidade e concordância)[Matos 2014].

No presente trabalho utilizou-se o ICC para medir a concordância nas avaliações entre pares de avaliadores. Considerando que o ICC pode ser interpretado de diferentes maneiras na literatura, aqui utilizamos a interpretação de [Joseph L. Fleiss, Bruce Levin 1981] que sugere que: valores de ICC menores que 0,4 são indicativos de uma concordância baixa ou pobre, valores de ICC entre 0,4 e 0,75 indicam uma concordância moderada ou satisfatória, e valores de ICC maiores ou iguais a 0,75 indicam uma concordância excelente entre avaliadores. Para escolher a fórmula do ICC foi utilizado o método de decisão para correlação intraclass apresentado no trabalho de [Matos 2014]. Existem basicamente três modelos de ICC existentes: **1) efeito aleatório unidirecional** - cada sujeito/objeto é avaliado por um conjunto diferente de avaliadores que foram escolhidos aleatoriamente), **2) efeito aleatório bidirecional** - os avaliadores são selecionados aleatoriamente de acordo com suas características semelhantes e **3) efeito misto bidirecional** - considera apenas os avaliadores que foram selecionados. O cálculo do ICC também é dividido em dois tipos, sendo que o primeiro tipo utiliza o valor médio dos avaliadores (média de K avaliadores) e o segundo tipo utiliza apenas a medida de um avaliador.

O presente trabalho utilizou o ICC de efeito misto bidirecional já que existe o interesse de calcular o ICC para cada par de avaliadores que possuem conjuntos de objetos comumente avaliados e que a escolha desses pares não é aleatória (e sim um consenso). O tipo utilizado foi o valor médio de K avaliadores, sendo assim possível utilizar o modelo de decisão ICC (3, K) (ou aleatório de duas vias - *two-way random*). De acordo com [Koo and Li 2016] a fórmula do ICC (3,K) é definida da seguinte maneira: $Y_{ij} = \mu + S_j + T_i + e_{ij}$ yonde $i = 1, \dots, k; j = 1, \dots, n$. A fórmula geral do ICC (3,K) foi

redefinida para ser utilizada no algoritmo desenvolvido, sendo definido como fórmula computacional. A fórmula computacional utilizada para calcular o ICC utiliza o $MS_B = \frac{k}{n-1} \sum_{j=1}^n (\bar{y}_{.j} - \bar{y}_{..})^2$ onde $\bar{y}_{.j}$ é a média das avaliações do avaliador j e $\bar{y}_{..}$ é a média de todas as avaliações. E $MS_E = \frac{k}{(n-1)(k-1)} \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$, na qual, $\bar{y}_{i.}$ é a média das avaliações de i documentos e y_{ij} é o valor da avaliação i do avaliador j . Então a fórmula computacional utilizada foi $\frac{MS_B - MS_E}{MS_B}$. A variável MS_E refere-se ao quadrado médio do erro e MS_B ao quadrado médio entre os sujeitos e o K é a quantidade de avaliadores dividida pelas medições.

4. Materiais e métodos

Os dados utilizados nesse estudo foram extraídos do repositório Merlot (ano 2015), que possuía um total de 11.823 avaliações de recursos educacionais realizadas pelos usuários, para os mais de 45.000 recursos cadastrados nas diferentes áreas (ciência e tecnologia, humanas, negócios, matemática, artes, educação). Juntamente com as identificações dos recursos educacionais, foram coletadas suas avaliações e as identificações (ids) dos usuários avaliadores, além de suas características tais como a disciplina de atuação preferencial (ciência e tecnologia, humanas, negócios, etc) e a sua atuação profissional (estudante, professor, bibliotecário, etc).

Das 11.823 avaliações coletadas, foram excluídas todas aquelas pertencentes a avaliadores que possuíam apenas uma única avaliação resultando em um total de 1.452 avaliações. Com esse conjunto, foram selecionados os pares de usuários que avaliaram os mesmos recursos simultaneamente resultando em 179 pares de usuários que avaliaram um total de 925 recursos digitais. A Figura 1 apresenta a distribuição da quantidade de recursos avaliados simultaneamente por pares (duplas) de avaliadores. Como é possível observar, apenas um par de usuários avaliou simultaneamente um total de 15 recursos idênticos (maior quantidade de recursos avaliados simultaneamente por uma mesma dupla de avaliadores). Desse total, 39 pares de usuários que avaliaram três ou menos recursos simultaneamente (≤ 3) foram excluídos da amostra por não serem consideradas quantidades confiáveis para cálculos de concordância (amostras pequenas de 3 ou menos recursos podem ocasionar o acaso [Koo and Li 2016]). Ainda, 38 pares de usuários em que as avaliações consistiam somente de comentários (sem nota quantitativa) também foram excluídos, resultando em uma amostra de 102 pares de avaliadores para o estudo realizado.

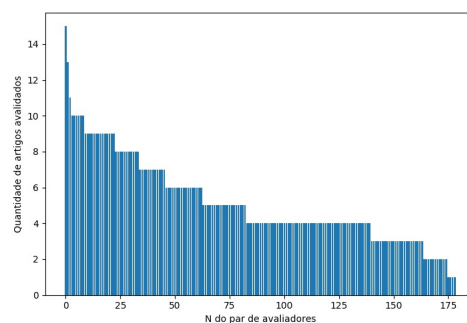


Figura 1. Quantidade de materiais avaliados por pares de usuários

Os pares de usuários avaliadores e as pontuações dadas simultaneamente aos mes-

mos recursos foram extraídos da base de dados por meio de script desenvolvido em Python. Após a extração dos dados, o coeficiente ICC foi calculado para cada um dos pares e foram geradas estatísticas descritivas das concordâncias calculadas considerando principalmente dois aspectos dos avaliadores: 1) disciplina de atuação preferencial, e 2) classe de atuação profissional.

5. Resultados e discussão

5.1. Concordância considerando as disciplinas de atuação preferencial

Usuários são registrados no Merlot com uma série de informações associadas, como a sua disciplina de atuação preferencial, sua atuação profissional, além de outros dados pessoais. As áreas de atuação preferencial são de maneira geral idênticas as categorias de disciplinas em que os recursos educacionais são cadastrados, sendo elas: ciência e tecnologia, ciências sociais, humanas, negócios, matemática, artes e educação. Essas categorias primárias, por sua vez são subdivididas em subcategorias secundárias (por exemplo, ciência e tecnologia é subdividida em Biologia, Física, Química, Tecnologia da Informação, entre outras). Nesse trabalho, consideramos somente a categoria de disciplina principal (ou primária)

A Tabela 1 apresenta os percentuais de concordância dos pares de avaliadores considerando as situações em que ambos pertencem a uma área de atuação preferencial (mesma disciplina primária) e quando pertencem a áreas de atuação preferencial distintas, além da concordância geral sem considerar a área de atuação preferencial. Como pode ser visto na Tabela, 79,41% dos pares de avaliadores possuem uma concordância pobre ($ICC < 0,4$) enquanto 13,72% possuem concordância satisfatória ($0,4 \leq ICC < 0,7$) e apenas 6,87% possuem uma concordância excelente ($ICC \geq 0,7$). O percentual de concordâncias satisfatória e excelente entre os pares de avaliadores que pertencem a uma mesma área de atuação preferencial é levemente maior do que os percentuais de concordância entre os pares de avaliadores com áreas de atuação distintas. Enquanto no primeiro caso os percentuais de concordância satisfatória e excelente são de 15,15% e 7,58%, no segundo cenário esses percentuais caem para 11,1% e 5,5%.

Tabela 1. Concordância geral por áreas primárias iguais e diferentes

Tipo	Pobre	Satisfatória	Excelente	Total
Avaliadores de áreas diferente	83,4% (30)	11,1% (4)	5,5% (2)	36
Avaliadores de áreas iguais	77,27% (51)	15,15% (10)	7,58% (5)	66
Geral	79,41% (81)	13,72% (14)	6,87% (7)	102

A Tabela 2 detalha os percentuais de concordância para pares de avaliadores pertencentes a uma mesma área de atuação preferencial (disciplina primárias iguais). Apenas três áreas de atuação (ou disciplinas) possuem pares de avaliadores de uma mesma área de atuação preferencial e que realizaram avaliações simultâneas, sendo elas: Humanas, Ciências e Tecnologia e Educação. A disciplina que possuiu um melhor grau de concordância entre os pares foi a de Ciências e Tecnologia com 15,2% de concordância satisfatória e 8,4% de concordância excelente.

Tabela 2. Concordância por áreas primárias iguais

Áreas iguais	Pobre	Satisfatória	Excelente	Total
Humanas	75% (3)	25% (1)		3
Ciências e Tecnologia	76,4% (45)	15,2% (9)	8,4% (5)	59
Educação	100% (3)			3
Total de pares de avaliadores	77,27% (51)	15,15% (10)	7,58% (5)	66

A Tabela 3 apresenta o detalhamento dos percentuais de concordância para os pares de usuários avaliadores pertencentes a áreas de atuação de preferência distintas. Como já comentado, os percentuais de concordância estão concentrados em um ICC classificado como pobre. As exceções são as seguintes combinações áreas de atuação: 1) Educação x Ciência e Tecnologia, 2) Negócios X Humanas, e 3) Negócios X Ciência e Tecnologia.

Tabela 3. Concordância por áreas primária diferentes

Áreas diferentes	Pobre	Satisfatória	Excelente	Total
Humanas x Ciência e Tecnologia	100% (2)			2
Educação x Ciências Sociais	100% (2)			2
Educação x Ciência e Tecnologia	83,3% (10)	16,7% (2)		12
Ciência e Tecnologia x Ciências Sociais	100% (1)			1
Artes x Ciência e Tecnologia	100% (1)			1
Humanas x Ciências Sociais	100% (1)			1
Negócios x Humanas	50% (1)		50% (1)	2
Negócios x Ciência e Tecnologia	57,10% (4)	28,50 (2)	14,4% (1)	7
Educação x Serviços de apoio acadêmico	100% (2)			2
Educação x Humanas	100% (4)			4
Artes x Humanas	100% (1)			1
Matemática estatística x Ciências e Tecnologia	100% (1)			1
Total de pares de avaliadores	83,4% (30)	11,1% (4)	5,5% (2)	36

A base de dados analisada apresentou uma grande discordância entre as avaliações dadas pelos pares de avaliadores. Como mostram as Tabelas 6, 2 e 3 essa discordância é levemente maior quando os usuários avaliadores pertencem a áreas de atuação preferencial distintas.

5.2. Concordância considerando a atuação profissional

A segunda perspectiva de concordância avaliada foi com relação a atuação profissional dos avaliadores. Exemplos de categorias de atuação profissional são: professor não universitário (*teacher*), estudante, profissional universitário (*faculty professional* - professor universitário, pesquisador), bibliotecário, consultor, e administrador. A Tabela 4 apresenta os percentuais de concordância para avaliadores que possuem a mesma atuação profissional, atuações profissionais distintas e a concordância geral.

Tabela 4. Concordância considerando as áreas de atuação profissional

	Pobre	Satisfatória	Excelente	Total
Mesma atuação profissional	75,80% (47)	14,51% (9)	9,69% (6)	62
Atuação profissional distinta	85% (34)	12,5% (5)	2,5% (1)	40
Total de pares de avaliadores	79,41% (81)	13,72% (14)	6,87% (7)	102

Como mostra na Tabela 5, todos os pares de avaliadores professores não universitários e funcionários apresentaram uma concordância pobre ($ICC < 0,4$). Por outro lado, os avaliadores estudantes apresentaram um percentual de concordância maior, sendo que a soma dos percentuais de concordância satisfatório e excelente totaliza 28,31%, sendo esse grupo aquele que mais concorda nas avaliações realizadas.

Tabela 5. Concordância considerando a mesma área de atuação profissional

Mesma atuação profissional	Pobre	Satisfatória	Excelente	Total
Professor não universitário	100% (8)			8
Estudante	71,69% (38)	16,98% (9)	11,33% (6)	53
Funcionário	100% (1)			1
Total de pares de avaliadores	75,80% (47)	14,51% (9)	9,69% (6)	62

A Tabela 6, mostra a concordância entre avaliadores de áreas de atuação profissional distintas. Como pode ser visto, 82,5% dos pares apresentam uma concordância pobre entre as avaliações dadas. Com exceção de apenas 1 caso (Professor não universitário X Consultor), todos os demais casos em que existe alguma concordância satisfatória ou excelente, essa concordância envolve a classe Estudante.

Tabela 6. Concordância considerando áreas de atuação profissional distintas

Atuação profissional		Pobre	Satisfatória	Excelente	Total
Professor não universitário	Profissional universitário	100% (6)			6
Estudante	Funcionário	87,5% (7)	12,5% (1)		8
Estudante	Professor não universitário	100% (3)			3
Estudante	Profissional universitário	77,78% (7)	22,22% (3)		10
Estudante	Administrador	75% (3)	25% (1)		4
Funcionário	Profissional universitário	100% (4)			4
Professor não universitário	Funcionário	100% (1)			1
Profissional universitário	Bibliotecário	100% (1)			1
Professor não universitário	Consultor	50% (1)		50% (1)	2
Bibliotecário	Professor não universitário	100% (1)			1
Total de pares de avaliadores		85% (34)	12,5% (5)	2,5% (1)	40

6. Considerações finais

A presente pesquisa utilizou dados recolhidos do repositório Merlot para medir a concordância entre pares de avaliadores que realizaram avaliações simultâneas sobre um mesmo conjunto de recursos educacionais digitais. Para a medição da concordância entre pares de avaliadores foi utilizado a medida ICC (Intra-class correlation) e posteriormente

calculadas as estatísticas descritivas para cada uma das classes de concordância existentes (pobre, satisfatória e excelente) e de acordo com as diferentes perspectivas que se desejava medir (concordância geral, concordância considerando as disciplinas de atuação preferencial, e concordância considerando a atuação profissional dos avaliadores). De acordo com o conhecimento dos autores, esse é o primeiro estudo que realiza uma avaliação de concordância entre avaliadores de recursos educacionais digitais utilizando dados de um repositório existente e consolidado. Após as análises é possível responder as perguntas de pesquisa inicialmente propostas.

P1: Qual a concordância geral entre pares de avaliadores de recursos educacionais digitais? **R1:** Os resultados mostraram que os pares de avaliadores possuem forte discordância com relação a qualidade dos recursos educacionais digitais que avaliaram simultaneamente. Dos 102 pares avaliados, 79,41% possuem uma concordância pobre, 13,72% possuem uma concordância satisfatória e apenas 6,87% possuem uma concordância excelente.

P2: Qual a concordância entre pares de avaliadores considerando suas preferências por determinadas disciplinas e suas áreas de atuação profissional? **R2:** As estatísticas descritivas dos ICCs calculados apontam que existe um leve aumento da concordância entre os pares de avaliadores quando ambos pertencem a uma mesma área de atuação preferencial (mesma preferência por uma categoria de disciplina ou assunto curricular). Sobre a concordância considerando a área de atuação dos avaliadores, o grupo de avaliadores que mais apresenta concordância em suas avaliações é o de estudantes (com aproximadamente 28% dos pares com concordância satisfatória ou excelente). De maneira geral, o único grupo de avaliadores que concorda de maneira satisfatória ou excelente com outros grupos também é o grupo de avaliadores de estudantes (exceção para o caso específico dos pares formados pelos grupos professor não universitário e consultor).

Os resultados encontrados indicam a possibilidade de aprimoramento nas formas de apresentação da qualidade dos recursos educacionais digitais dentro de repositórios, de maneira que as diferentes opiniões sobre a qualidade dos mesmos apareçam de forma mais transparente/detalhada para os usuários. Uma possibilidade, por exemplo, seria a apresentação da pontuação geral da qualidade de cada recurso (como já vem sendo feito) juntamente com as pontuações de qualidade fornecidas pelos diferentes tipos de avaliadores de acordo com suas características (disciplinas de atuação, área de atuação profissional). Os repositórios poderiam também personalizar a apresentação da qualidade de um recurso para um determinado usuário considerando especificamente as suas características em comparação com as características dos diferentes tipos de avaliadores e as pontuações que os mesmos deram aos recursos.

Uma das limitações do presente trabalho é a pequena quantidade de pares de avaliadores analisada e a consequente esparsidade das combinações de pares de avaliadores. Em contrapartida, cabe ressaltar que o Merlot é um dos repositórios mais disseminados atualmente e que conta com uma das maiores quantidades de recursos, usuários e avaliações fornecidas pela sua comunidade. Nesse sentido, ainda que a quantidade de pares possa ser considerada pequena, ela é representativa de um dos maiores e mais importantes ROA atualmente existentes. Como trabalho futuro propõe-se a continuação da análise de concordância utilizando outras medidas como a correlação entre as médias das avaliações dadas pelos avaliadores de cada uma das categorias de disciplina, permitindo

a verificação de concordância entre as médias dos grupos.

Referências

- Atenas, J. and Havemann, L. (2014). Questions of quality in repositories of open educational resources: a literature review. *Research in Learning Technology*, 22.
- Cechinel, C. and Sánchez-Alonso, S. (2011). Analyzing associations between the different ratings dimensions of the merlot repository. *Interdisciplinary Journal of E-Learning and Learning Objects*, 7(1):1–9.
- Cechinel, C., Sánchez-Alonso, S., and García-Barriocanal, E. (2011). Statistical profiles of highly-rated learning objects. *Computers & Education*, 57(1):1255–1269.
- Clements, K., Pawlowski, J., and Manouselis, N. (2015). Open educational resources repositories literature review - Towards a comprehensive quality approaches framework. *Computers in Human Behavior*, 51:1098–1106.
- Downes, S. (2017). Models for Sustainable Open Educational Resources. *Interdisciplinary Journal of e-Skills and Lifelong Learning*, 3(September):029–044.
- Ehlers, U.-D. and Pawlowski, J. M. (2006). *Handbook on Quality and Standardisation in E-Learning*.
- Fonseca, R., Silva, P., and Silva, R. (2014). Acordo inter-juízes: O caso do coeficiente kappa. *Laboratório de Psicologia*, 5(1):81–90.
- Han, P., Kortemeyer, G., Kramer, B. J., and von Prummer, C. (2008). Exposure and support of latent social networks among learning object repository users. *Journal of Universal Computer Science*, 14(10):1717–1738.
- Hylén, J. (2006). Open educational resources: Opportunities and challenges. *Proceedings of Open Education*, 4963.
- Joseph L. Fleiss, Bruce Levin, M. C. P. (1981). Statistical Methods for Rates and Proportions. *Wiley Series in Probability and Statistics*.
- Koo, T. K. and Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of chiropractic medicine*, 15(2):155–63.
- Kutluca, A. (2010). MULTIMEDIA EDUCATIONAL RESOURCE FOR LEARNING AND ONLINE TEACHING (MERLOT) (Website Review). *The Turkish Online Journal of Distance Education*, 11.
- Matos, D. A. S. (2014). Confiabilidade e concordância entre juízes: aplicações na área educacional. *Estudos em Avaliação Educacional*, 25(59):298–324.
- Pawlowski, J. M. (2007). The quality adaptation model: Adaptation and adoption of the quality standard iso/iec 19796-1 for learning, education, and training. *Journal of Educational Technology & Society*, 10(2):3–16.
- Tinsley, H. E. A. and Weiss, D. J. (2000). Interrater reliability and agreement. In *Handbook of applied multivariate statistics and mathematical modeling.*, pages 95–124. Academic Press, San Diego, CA, US.