Indicadores de gestão na revisão por pares: confiabilidade da revisão recíproca anônima de propostas de mestrado

Marta Denisczwicz[*]

Vinícius Medina Kern[**]

Resumo A revisão por pares é pobre em indicadores, apesar de sua importância. Este artigo apresenta indicadores de confiabilidade da revisão recíproca anônima de propostas de mestrado. Doze alunos de mestrado em Ciência da Informação deram pareceres anônimos sobre propostas de colegas em 7 itens avaliativos, usando escala Likert-6, sendo a confiabilidade calculada através de índices utilizados na literatura. A confiabilidade não é uma medida absoluta de qualidade, mas o cálculo desses indicadores permite estudar sistematicamente a qualidade da revisão por pares.

Palavras-chave Revisão por pares, Confiabilidade, Concordância entre revisores, Coeficiente de correlação intraclasse, Revisão por pares na aprendizagem.

Peer review indicators: reliability of a reciprocal anonymous review of masters proposals

Abstract Peer review lacks indicators, which is incongruent with its importance. This article reports the reliability of mutual, anonymous review of masters research proposals. Twelve masters students in Information Science refereed anonymously their peers' research proposals according to 7 evaluation items, using a Likert-6 scale. Reliability was expressed as intraclass correlation indices between 0.500 and 0.202, with 9 positive and 3 negative values - lower than usual in professional processes. Reliability is not an absolute measure of quality, but calculating reliability allows for a systematic study of the quality of peer review.

Keywords Peer review; Reliability; Agreement among referees; Intraclass correlation coefficient; Peer review in learning.

[*] Graduanda do curso de Biblioteconomia pela Universidade Federal de Santa Catarina. Bolsista de Iniciação Científica PIBIC/CNPq. Universidade Federal de Santa Catarina. Endereço: Rua Cônego Bernardo, 44, ap. 3 - Bairro Trindade - CEP 88036-570, Florianópolis – SC. Telefone: (48) 8812-0594. E-mail: marta.denisczwicz@gmail.com

[**] Doutor. Professor do Departamento de Ciência da Informação. Universidade Federal de Santa Catarina, Centro de Ciências da Educação, Departamento de Ciência da Informação Endereço: Campus Universitário – Trindade.– CEP: 88010-970 - Florianópolis, SC – Brasil. Telefone: (48) 3721-2970. E-mail: kern@cin.ufsc.br

Introdução

A revisão por pares é o principal mecanismo de controle de qualidade na maioria das disciplinas científicas (BORNMANN, 2011). Segundo Mulligan, Hall e Raphael (2013), a revisão por pares exerce tamanha autoridade no meio cientifico que se torna possível observar que pesquisas disseminadas em domínio público, mas que não passam pelo processo de revisão por pares de revistas científicas, acabam por não serem vistas com bom olhos por boa parte da comunidade cientifica.

Ainda assim, a revisão por pares vem sendo duramente criticada como método de avaliação. As críticas apontam sua ineficácia para detectar fraude, plágio, baixa qualidade ou erro, bem como a pouca confiabilidade, ao tolerar grandes disparidades entre revisões de um mesmo artigo ou proposta de pesquisa, além de ser complacente com a mediocridade e bloquear ideias de pesquisa originais (HORROBIN, 1974, 1996, 2001; STEHBENS, 1999; SEATON, 1996; BENOS et al., 2007; CASADEVALL; FANG, 2009).

Apesar dos defeitos reconhecidos, não há modelos alternativos amplamente efetivos e a revisão por pares continua sendo o principal método de avaliação. Essa perspectiva de continuidade sugere que há mérito na busca de seu aperfeiçoamento. Jennings (2006), por exemplo, defende a criação de indicadores quantitativos de gestão da revisão por pares.

Entre os principais debates sobre a revisão por pares está a questão da divulgação (pública ou apenas para o autor) ou não do nome do revisor (KHAN, 2010) e a busca por maior confiabilidade no processo, que Hacket & Chubin (2003) traduzem como um alto nível de concordância entre os revisores. Bornmann (2011) afirma que a concordância é rara e a confiabilidade é, portanto, precária.

Não há outros relatos na literatura brasileira sobre confiabilidade da revisão por pares, nem relatos internacionais sobre a confiabilidade da revisão por pares feita por pesquisadores em formação, como é o caso no presente artigo. Os trabalhos relacionados ao presente incluem o estudo de Yankulov e Couto (2012), que geraram estatísticas dos escores de avaliação atribuídos pelos alunos em uma disciplina de graduação em genética molecular durante cinco anos, concluindo que há grande variabilidade, mas com competência dos graduandos para detectar desempenho abaixo do padrão. Há muitas outras aplicações educacionais da revisão por pares, algumas apontadas por Kern et al. (2009), mas o cálculo de indicadores do processo é incomum e os de confiabilidade, inexistentes. Há muitos estudos publicados sobre a confiabilidade da revisão por pares editorial ou por agências de fomento, com destaque para as revisões desses estudos feitas por Weller (2002) e por Bornmann (2011). No Brasil, destacam-se os estudos sobre o processo de avaliação de artigos, especialmente em revistas (COSTA, 1996; CASTRO; NEGRÃO; ZAHER, 1996; MUELLER, 1997, PESSANHA, 1998, STUMPF, 2008, PAVAN; STUMPF, 2009).

Neste artigo, apresentamos os resultados da mensuração da confiabilidade da revisão por pares mútua e anônima entre estudantes de mestrado, com suas propostas de dissertação. Comparamos os resultados com indicadores de confiabilidade em certames profissionais de revisão por pares editorial ou em processos seletivos de fomento. Discutimos brevemente o significado dos índices de confiabilidade e a conveniência ou não de obter índices altos, bem como a importância desses cálculos para melhorar a gestão da revisão por pares.

Na próxima seção, esboçamos uma descrição das principais virtudes e falhas da revisão por pares. Na seção seguinte, abordamos a confiabilidade na revisão por pares, ou concordância entre os revisores. Descrevemos os procedimentos metodológicos para o cálculo de indicadores de confiabilidade da revisão de propostas de mestrado por pares, apresentamos os resultados e discutimos sua qualidade e as perspectivas do uso desse tipo de cálculo para compreender melhor a qualidade da revisão por pares e melhorar sua gestão.

Virtudes e falhas da revisão por pares

O sistema de revisão por pares age como um filtro, através de uma avaliação criteriosa dos manuscritos que poderão ser publicados. Esta avaliação permite que especialistas no assunto forneçam um parecer que poderá contribuir para a melhoria do artigo (JENNINGS, 2006). É um trabalho que vai além da mera aceitação ou rejeição de manuscritos. Ele ajuda os autores com correções, e as críticas podem auxiliar em trabalhos futuros (STUMPF, 2008). Ela também ajuda a aprimorar algumas seções dos manuscritos (MULLIGAN; HALL; RAPHAEL, 2013).

Este método de avaliação, além de contribuir para a construção do conhecimento científico, é também responsável por definir os rumos que a ciência deve seguir (DAVYT GARCÍA; VELHO, 2000). Os defensores deste sistema a veem como o instrumento mais eficaz para fazer a seleção crítica (BORNMANN, 2011).

Apesar de ser considerada como o principal método de avaliação, a revisão por pares recebe muitas críticas devido a uma série de problemas relatados na literatura. As críticas surgem de tempo em tempo, realimentando as antigas e originando novas (DAVYT GARCÍA; VELHO, 2000). E como existe uma crescente demanda, ano após ano, por manuscritos a serem publicados, gerando a necessidade de mais revisores, os questionamentos sobre a eficácia da revisão por pares também aumentam (MULLIGAN; HALL; RAPHAEL, 2013).

Esses questionamentos estão relacionados principalmente com o modo tradicional de como a revisão é feita, chamado de duplo-cego, no qual preserva-se o anonimato de ambas as partes. Mulligan, Hall e Raphael (2013) creem que o processo duplo-cego é considerado mais eficaz, pois eliminaria vieses e faria com que o revisor se concentre mais na qualidade do manuscrito.

Alguns autores creem que por trás do anonimato da revisão duplo-cega escondem-se "[...] preconceitos, comportamento não ético e incompetência" (MUELLER, 1997, p. 3). Perante essas razões, surgem algumas alternativas para tentar melhorar o processo, como a revisão cega simples, no qual o revisor conhece o autor, e a revisão aberta, em que os nomes são revelados. Mas, embora existam propostas de mudanças, a revisão por pares tradicional continua sendo a que prevalece "nos mais altos níveis acadêmicos" (YANKULOV; COUTO, 2012, p. 161).

O processo de revisão por pares carrega consigo muitas falhas, mas continua sendo o único método de avaliação na ciência. É uma atividade "carregada de subjetividade" (STUMPF, 2008, p. 21), e sendo uma atividade humana, os revisores apresentam visões diferentes sobre um mesmo artigo (EGGHE; BORNMANN, 2013). A subjetividade, no entanto, não implica que o processo não possa ou não deva ser gerido, inclusive com o uso de indicadores, como os de confiabilidade, discutidos a seguir.

Indicadores de confiabilidade da revisão por pares

A confiabilidade, na revisão por pares, é entendida como a concordância entre revisores (HACKET; CHUBIN, 2003). Weller (2002) revisou a literatura sobre medidas de confiabilidade da revisão por pares e encontrou 40 estudos sobre concordância entre revisores realizados entre 1970 e 1994. As medidas estatísticas usadas incluem o *r* de *Finn*, o *Kappa de Fleiss* e o coeficiente de *Kendall*, entre outros menos usados. O mais usado é, com larga margem, o Coeficiente de Correlação Intraclasse (ICC) de Shrout & Fleiss (1979). Variável entre -1 e +1 (BORNMANN, 2011), esse índice parte da suposição de que os avaliadores usam métricas semelhantes, com variância homogênea (SHROUT; FLEISS, 1979). Bornmann (2011) cita a revisão de Weller e a estende, com indicadores de concordância resultantes de pesquisas de concordância de revisores de revistas (11 estudos), de conferências (2 estudos) e agências de fomento (2 estudos), mas não especifica o tipo de indicador usado em cada estudo (embora indique que todos usam algum *Kappa* ou algum ICC).

O ICC, com 12 estudos, embasa o dobro da quantidade de pesquisas que usaram as outras 3 medidas citadas por Weller (2002). Dessa forma, o ICC é adotado largamente pela comunidade científica, embora a revisão de Weller não registre qual dos seis tipos de ICC descritos por Shrout & Fleiss (1979) foi adotado em cada estudo.

As primeiras pesquisas que utilizaram o coeficiente de correlação intraclasse mediram a homogeneidade da herdabilidade (FISHER, 1970). Fisher foi o primeiro a introduzir o conceito de correlação intraclasse e a interpretar o termo "classe" dos coeficientes de correlação intraclasse.

McGraw e Wong (1996) argumentam que o uso do termo "classe" por Fisher é preferível, no sentido de que enfatiza a correlação entre medidas que constituem uma classe, pois têm em comum a mesma métrica e variância, embora mesmo Fisher tenha sido ambíguo em algumas situações em que se referiu a objetos de medida como "classes". McGraw e Wong denunciam uma "confusão genuína" nessa questão semântica.

O cálculo do ICC é feito por meio de uma razão de variâncias (LAUREANO, 2011). Ele é uma das medidas mais utilizadas para estudos de confiabilidade. Esse coeficiente possui 6 versões, que podem gerar resultados diferentes se aplicados aos mesmos dados. Os pesquisadores devem estar atentos a isso, pois, na literatura, às vezes isso não fica claro. Dessa forma, podem ocorrer divergências nos resultados finais (SHROUT; FLEISS, 1979).

Procedimentos metodológicos

Para esta pesquisa, elegemos a medida ICC(2,1) ou ICC(*agreement*) de Shrout & Fleiss, na qual identificamos mais evidências de ser a medida apropriada, embora também tenhamos gerado indicadores com a medida ICC(3,1) ou ICC(*consistency*). O ICC (2,1) é utilizado quando todos os participantes são avaliados pelos avaliadores disponíveis, que são assumidos como um subconjunto de todos os avaliadores possíveis, medindo o grau de concordância absoluta, assumindo a permutabilidade dos avaliadores. É baseado na análise de variância (ANOVA) de duas vias de efeitos aleatórios. Já o ICC (3,1) é utilizado quando todos os participantes são

avaliados por avaliadores que se supõe constituir toda a população de avaliadores. É baseado na análise de variância (ANOVA) modelo misto, com os avaliadores tratados como um efeito fixo.

Os dados coletados são escores atribuídos pelos pares em pareceres sobre propostas de pesquisa de mestrado. O processo fez parte de uma disciplina obrigatória de mestrado em Ciência da Informação, voltada para a revisão e aperfeiçoamento das propostas, em formato semelhante ao de um exame de qualificação de mestrado. Doze alunos atuaram como autores e revisores anônimos das propostas de seus pares, sendo que os autores não foram revisores das próprias propostas.

Cada proposta recebeu de 9 a 11 pareceres, num total de 117 pareceres. Quatro pareceres foram eliminados por conter escores em branco, o que inviabiliza o cálculo de ICC. Com isso, o número de pareceres por proposta baixou para de 8 a 10, num total de 113 (média de 9,4 pareceres por proposta).

Cada parecer constava de 7 itens avaliados numa escala Lickert com 6 pontos detectando a variação da concordância. Continha, também, uma avaliação geral textual, mas apenas os escores foram usados no nosso estudo. A Figura 1 mostra o formulário usado para a elaboração de cada parecer, usando uma planilha de cálculo.

Figura 1: Formulário de revisão de propostas de mestrado em Ciência da Informação

UFSC/PGCIN - Relatório de revisão anônima por pares sobre proposta de dissertação - PGCIN - Seminários de Pesquisa

IDENTIFICAÇÃO DO(A) REVISOR(A) - não será publicada

Nome

INFORMAÇÕES GENÉRICAS SOBRE A PROPOSTA DE DISSERTAÇÃO REVISADA

Autor:

Título: (Preencha com as primeiras palavras do título - isto é uma dupla verificação da identidade da proposta revisada)

SEÇÃO I - VISÃO GERAL

Resuma a proposta em uma sentença ou parágrafo.

SEÇÕES II E III - AVALIAÇÃO OBJETIVA E COMENTÁRIOS

Posicione o cursor sobre a célula laranja na coluna A e atribua um escore, conforme a escala a seguir. Caso tenha comentários específicos sobre o item, escreva na caixa maior, logo após o item avaliado.

1	Discordo totalmente
2	Discordo parcialmente / mais discordo que concordo
3	Sem certeza, mas tendo a discordar
4	Sem certeza, mas tendo a concordar
5	Concordo parcialmente / mais concordo que discordo
6	Concordo totalmente

Problema e objetivos: Existe um problema de pesquisa relevante e claramente enunciado, bem contextualizado na literatura, com objetivo geral coerente com o problema e objetivos específicos mensuráveis, atingíveis, realistas, realizáveis no tempo disponível, específicos e consistentes entre si.

Comentários?

Fundamentação: Motiva o tópico de pesquisa, descreve os conceitos-chave, inclui apenas o que é essencial para a pesquisa, revisa a literatura relevante na área e (quando for o caso) áreas correlatas.

Comentários?

Abordagem metodológica: Os procedimentos metodológicos são escolhas adequadas para abordar o problema enunciado e atingir os objetivos, bem como são claros e completos a ponto de assegurar a repetibilidade da pesquisa.

Comentários?

Resultados esperados e conclusões: Os resultados antevistos são claros e a provável repercussão da pesquisa na área é excelente, compatível com uma pesquisa de mestrado.

Referências: As referências reúnem as obras mais relevantes, inclusive relatos de resultados de pesquisa recente, com presença nula ou minoritária de literatura cinzenta.

Comentários?

Linguagem e estilo: O texto da proposta faz bom uso da língua portuguesa. É possível compreender a partir da leitura (e não da imaginação do leitor ou depois de ler). Há sentenças introdutórias e de transição que facilitam acompanhar o encadeamento de idéias, que segue um fluxo lógico. O Resumo resume e a Introdução introduz.

Comentários?

Formatos: A proposta respeita as convenções estabelecidas para sua apresentação, adotando as normas indicadas.

Comentários?

Sobre a proposta em geral (Abaixo: Se desejar, faça comentários sobre a proposta como um todo e outras observações).

Fonte: Os autores

.

Para dispor de termos de comparação com os indicadores de confiabilidade a calcular para as propostas de mestrado, colhemos dados de uma segunda disciplina, esta optativa, que admite matrículas isoladas, de mestrado e doutorado em programa interdisciplinar, com 21 autores-revisores e 11 artigos escritos em dupla (exceto um, individual). Usamos o mesmo sistema de avaliação. Nesse caso, cada aluno foi autor de um artigo em dupla e revisor individual de 2 artigos de colegas. Tivemos 9 artigos com 4 pareceres e 2 artigos com 3 pareceres.

Em ambos os processos, colhemos os dados usados no cálculo de indicadores de confiabilidade apresentados mais adiante. Também buscamos dados de ICC de processos profissionais de publicação científica ou fomento, presentes na literatura (WELLER, 2002; BORNMANN, 2011), para poder comparar com os processos estudantis.

Os dados dos pareceres foram armazenados em um gerenciador de bancos de dados *MySQL*, por meio do qual foi possível gerar um conjunto de dados que foi exportado para planilhas de cálculo e, daí, importado pelo pacote estatístico SPSS. Os ICC foram calculados usando a opção de cálculo de análise de confiabilidade, coeficiente de correlação intraclasse, com intervalo de confiança de 95%.

O processo de revisão de propostas de mestrado por pares precedeu, na disciplina em questão, sessões de apresentação pública, com banca examinadora, de forma semelhante a um exame de qualificação de mestrado. O processo foi conduzido pelo segundo autor deste artigo em 2011, numa disciplina obrigatória de pós-graduação em Ciência da Informação voltada para consolidar os projetos dos alunos (n=12). Em disciplina anterior, os alunos haviam estudado a epistemologia e o estado da pesquisa na área, bem como métodos, técnicas e instrumentos de pesquisa predominantes. Haviam apresentado, também, uma primeira versão da proposta.

Os alunos foram instruídos sobre o papel do revisor, com orientações compatíveis com as de Smith (1990). Cada aluno deu parecer sobre as propostas de todos os colegas, exceto os que apresentavam sua proposta no mesmo dia (para liberar cada apresentador para concentrar-se em sua proposta e apresentação). Cada proposta teve pareceres anônimos entregues com antecedência às 5 sessões de apresentação.

Cada apresentação contou com a presença dos colegas. As revisões dos pares foram devolvidas ao candidato, posteriormente à apresentação. Cada revisão abordou os seguintes pontos, atribuindo escores de 1 a 6 de acordo com a escala:

Discordo totalmente;

Discordo parcialmente / mais discordo que concordo;

Sem certeza, mas tendo a discordar;

Sem certeza, mas tendo a concordar;

Concordo parcialmente / mais concordo que discordo;

Concordo totalmente

A aplicação dessa escala foi feita considerando as seguintes afirmações sobre os 7 itens avaliativos:

- Problema e objetivos: Existe um problema de pesquisa relevante e claramente enunciado, bem contextualizado na literatura, com objetivo geral coerente com o problema e objetivos específicos mensuráveis, atingíveis, realistas, realizáveis no tempo disponível, específicos e consistentes entre si.

- Fundamentação: Motiva o tópico de pesquisa, descreve os conceitos-chave, inclui apenas o que é essencial para a pesquisa, revisa a literatura relevante na área e (quando for o caso) áreas correlatas.

- Abordagem metodológica: Os procedimentos metodológicos são escolhas adequadas para abordar o problema enunciado e atingir os objetivos, bem como são claros e completos a ponto de assegurar a repetibilidade da pesquisa.

- Resultados esperados e conclusões: São claros e a provável repercussão da pesquisa na área é excelente, compatível com uma pesquisa de mestrado.

- Referências: Reúnem as obras mais relevantes, inclusive relatos de resultados de pesquisa recente, com presença nula ou minoritária de literatura cinzenta.

- Linguagem e estilo: O texto da proposta faz bom uso da língua portuguesa. É possível compreender a partir da leitura (e não da imaginação do leitor ou depois de ler). Há sentenças introdutórias e de transição que facilitam acompanhar o encadeamento de ideias, que segue um fluxo lógico. O Resumo resume e a Introdução introduz.

- Formatos: A proposta respeita as convenções estabelecidas para sua apresentação, adotando as normas indicadas.

Havia espaço para comentários e recomendações por escrito para cada ponto avaliado, bem como para a proposta no todo. Esse rol de pontos de avaliação foi consolidado a partir de uma discussão prévia com a participação dos demais docentes do programa de pós-graduação.

Nesse cenário, dentre os 6 tipos de ICC descritos por Shrout e Fleiss (1979), adotamos o ICC(2,1) ou ICC (*agreement*), ou ainda "concordância absoluta", que considera que os pareceristas são selecionados a partir de uma população maior, e cada parecerista avalia cada proposta. Ainda assim, calculamos também o ICC (3,1) ou ICC (*consistency*), que considera que cada proposta é avaliada pelos mesmos pareceristas, que são os únicos que interessam. Na verdade, nenhuma classe coincide perfeitamente com a situação, embora o ICC (2,1) pareça o mais usado e adequado segundo nossa apreciação a partir da literatura.

Resultados

Nesta seção são apresentados os resultados deste estudo.

A Tabela 1 apresenta as medidas de confiabilidade dos pareceres que foram coletados no processo de revisão por pares aplicado a propostas de mestrado em Ciência da Informação, com valores ICC (2,1) ou ICC (*agreement*) e ICC(3,1) ou ICC(*consistency*) para cada proposta.

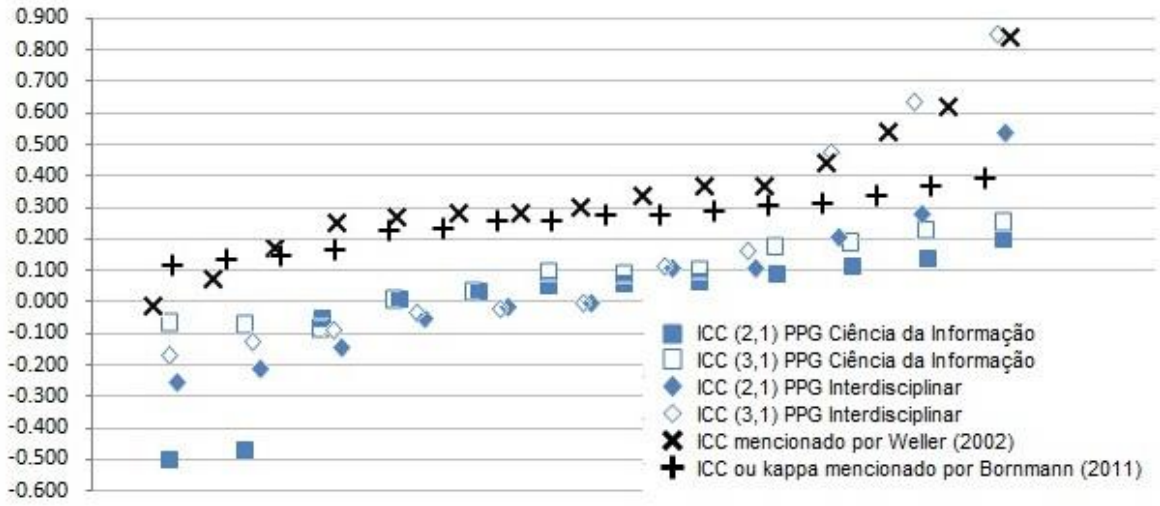
Tabela 1: - Índices de correlação intraclasse ICC(2,1) e ICC(3,1) para revisões de propostas de mestrado

N. da proposta de mestrado	1	2	3	4	5	6	7	8	9	10	11	12	mediana
Revisões válidas	10	10	10	9	10	10	9	8	9	9	9	9	9
ICC (agreement)	-0,500	-0,470	-0,050	0,011	0,034	0,054	0,062	0,066	0,090	0,116	0,143	0,202	0,058
ICC (consistency)	-0,062	-0,066	-0,078	0,011	0,041	0,100	0,094	0,106	0,180	0,194	0,231	0,261	0,097

Fonte: Elaboração própria, com resultados dos dados coletados na pesquisa (2011)

A Figura 2 apresenta os valores dos ICC de pareceres de mestrandos em Ciência da Informação, com valores ICC (2,1) e ICC (3,1) para cada proposta. Ordenamos os valores do menor para o maior, da esquerda para a direita, apenas para melhor vislumbrar o conjunto. O mesmo foi feito com os valores de ICC encontrados na avaliação de artigos de pós-graduação interdisciplinar e os ICC relatados nas revisões de Weller (2002) e de Bornmann (2011), sobre estudos de confiabilidade de pareceres de pesquisadores profissionais.

Figura 2: ICC dos pareceres de mestrandos em Ciência da Informação e outros



Fonte: Elaboração própria, com resultados dos dados coletados na pesquisa (2011)

O ICC (2,1), medida de confiabilidade mais usual, variou entre -0,500 e 0,202, com 9 valores positivos e 3 negativos para as revisões de propostas de mestrado em Ciência da Informação e

entre -0,254 e 0,540, com 5 valores positivos, um zero e 4 negativos para as revisões multidisciplinares de artigos de disciplina de pós-graduação.

Os estudos mencionados por Weller (2002, p. 186-187) apontam valores de ICC no intervalo [-0,15; 0,84], com mediana 0,30. Pode-se perceber a predominância de valores superiores de ICC nos certames profissionais em relação aos certames estudantis. A comparabilidade, no entanto, é precária, uma vez que não há indicação de qual ICC foi usado nos valores coligidos por Weller (2002) e por Bornmann (2011). Neste último, sequer é claro se a medida é um ICC ou um Kappa (κ), outra medida de confiabilidade. Entre os certames estudantis, há predominância de valores do ICC maiores para os artigos do que para as propostas de mestrado.

Discussão

Dentre as 12 medidas de correlação intraclasse encontradas na revisão de propostas de mestrado em Ciência da Informação, 9 são positivas, compatíveis com as medidas encontradas no levantamento de Weller (2002) sobre estudos de confiabilidade da revisão por pares. Isso significa que, nesses casos, os acadêmicos alcançaram um nível de concordância compatível com os verificados entre pareceristas que são pesquisadores profissionais. A mediana dos índices de confiabilidade é de 0,058 para ICC (*agreement*) ou 0,097 para ICC (*consistency*) – índices relativamente baixos, mas ainda assim positivos, que revelam predomínio da concordância sobre a falta de concordância.

Os índices alcançados pelos mestrandos em Ciência da Informação são inferiores aos apontados por Weller (2002) e por Bornmann (2011), como se poderia esperar ao comparar alunos de pós-graduação a pesquisadores profissionais. No entanto, trata-se de uma medida pontual – para um certame de revisão por pares.

Outra questão é o pressuposto – sem sustentação teórica, segundo Weller (2002) – de que é positivo haver índices de confiabilidade altos. Entre os que consideram positivo, Marsh, Bond e Jayasinghe (2007, p. 33), sustentam que a "falta de confiabilidade em avaliações pelos pares é, talvez, a fraqueza mais importante do processo de revisão por pares".

Uma agência de fomento australiana obteve um incremento nos índices médios de concordância por meio da alocação de um conjunto de propostas de pesquisa a um mesmo grupo de avaliadores (JAYASINGHE; MARSH; BOND, 2006). Esse experimento foi possível porque a agência sistematizou as informações derivadas do processo de revisão a ponto de poder calcular com eficiência os índices de confiabilidade. Essa sistematização, no entanto, não é prática comum entre editores de revistas científicas e gestores de processos de alocação de verbas de fomento.

Por outro lado, há quem defenda a conveniência da baixa confiabilidade. Wood, Roberts e Howeel (2004) arguem que, se dois revisores concordam, não se pode saber se é porque ambos foram capazes de fazer uma avaliação "correta" ou o motivo foi outro.

Além de apresentar indicadores de confiabilidade para certames específicos, esta pesquisa tem como objetivo de longo prazo a sistematização desse tipo de cálculo estatístico para poder, num futuro não distante, realizar análises longitudinais para eventos e periódicos científicos. Embora não haja pesquisas evidenciando correlação entre a confiabilidade e a qualidade de um processo

de revisão por pares (ou seja, não se pode afirmar que maior confiabilidade é melhor), é possível afirmar que um processo com maior confiabilidade apresenta maior concordância entre revisores e, portanto, maior homogeneidade e menor dependência de revisores específicos, o que parece desejável. Da mesma forma, parece desejável saber os índices de confiabilidade para um evento ou periódico – e isso atualmente não existe de forma sistemática na literatura ou mesmo nas páginas web de periódicos e agências de fomento, revelando a oportunidade da criação sistemática de índices de confiabilidade, objetivo de longo prazo desta pesquisa. Portanto, persistir na pesquisa sobre confiabilidade da revisão por pares parece meritório.

Considerações finais

Neste artigo, apresentamos os índices de confiabilidade de pareceres de mestrandos em Ciência da Informação sobre as propostas de mestrado de seus pares. Os índices referem-se a um certame pontual. É o primeiro artigo brasileiro a tratar da confiabilidade da revisão por pares e o primeiro relato internacional a exibir indicadores de confiabilidade da revisão por pares entre pesquisadores em formação. O propósito mais amplo da iniciativa é fomentar uma cultura de crítica acadêmica entre os pós-graduandos (KERN et al., 2009).

Os ICC são predominantemente positivos, mas não tão altos como costumam ser em certames de revisão por pesquisadores profissionais, o que correspondeu à nossa expectativa. De qualquer forma, a confiabilidade não é uma medida absoluta de qualidade. De fato, as revisões de Weller (2002) e de Bornmann (2011) citam opiniões diversas – o alto nível de concordância é visto como "indicador de alta qualidade do processo" (BORNMANN, 2011, p. 209), mas alguns cientistas o veem como mau sinal. Ainda assim, o cálculo sistemático desses indicadores permite estudar sistematicamente a qualidade da revisão por pares.

Os resultados são preliminares. As próximas etapas da pesquisa incluem a construção de indicadores de confiabilidade longitudinais, tanto para os processos acadêmicos discutidos aqui quanto para outros fóruns, como revistas científicas. Já há acordo com a editoria de uma revista da área de Ciência da Informação. Se a falta de métricas de gestão é uma das falhas da revisão por pares, pretendemos contribuir para minorar esse defeito.

Artigo recebido em 15/02/2013 e aprovado em 08/03/2013.

Referências

BENOS, D. J. et al. The ups and downs of peer review. *Advances in physiology education,* v. 31, n. 1, p. 145-152, 2007.

BORNMANN, L. Scientific peer review. *Annual Review of Information Science and Technology,* v. 45, n. 1, p. 199-245, 2011.

CASADEVALL, A.; FANG, F. C. Is peer review censorship?. *Infection and Immunity,* v. 77, n. 4, p. 1273-1274, 2009.

CASTRO, R. C. F.; NEGRÃO, M. B.; ZAHER, C. R. Procedimentos editoriais na avaliação de artigos para publicação em periódicos de ciência da saúde da América Latina e Caribe. *Ciência da Informação,* v. 25, n. 3, p. 352-356, 1996.

COSTA, S. M. S. Controle de qualidade em periódicos científicos eletrônicos disponibilizados na Internet: a questão do julgamento pelos pares. *Revista de Biblioteconomia de Brasília,* v. 20, n. 2, p. 227-236, 1996.

DAVYT GARCÍA, A.; VELHO, L. A Avaliação da Ciência e a Revisão por Pares: passado e presente. Como será o Futuro?. *História, Ciência, Saúde – Manguinhos,* v. 7, n. 1, p. 93-116, mar./jun. 2000.

EGGHE, L.; BORNMANN, L. Fallout and Miss in journal peer review. *Journal of Documentation ,* 2013. No prelo.

FISHER, R. A. *Statistical methods for research workers.* New York: Hafner Press, 1970. 362 p.

HACKETT, E. J.; CHUBIN, D. E. *Peer review for the 21st century:* applications to education research. Washington, 2003.

HORROBIN, D. F. Peer review of grant applications: a harbinger for mediocrity in clinical research?. *The Lancet,* v. 348, p. 1293-1295, 1996.

_____. Referees and research administrators: barriers to scientific research?. *British Medical Journal,* v. 2, p. 216, 1974.

_____. Something rotten at the core of science?. *Trends in Pharmacological Sciences,* v. 22, n. 2, p. 51-52, Feb. 2001.

JAYASINGHE, U. W.; MARSH, H. W.; BOND, N. A new reader trial approach to peer review in funding research grants: an australian experiment. *Scientometrics,* v. 69, n. 3, p. 591-606, 2006.

JENNINGS, C. G. Quality and value: the true purpose of peer review: what you can't measure, you can't manage: the need for quantitative indicators in peer review. *Nature,* 2006. Nature Peer Review Debate.

KERN, V. M. et al. Growing a peer review culture among graduate students. In: TATNALL, A.; JONES, A. (Org.) *Education and technology for a better world.* New York: Springer, 2009. p. 388-397.

KHAN, K. Head to head: is open peer review the fairest system?: no. *British Medical Journal,* v. 341, 2010.

LAUREANO, G. H. da C. *Coeficiente de correlação intraclasse:* Comparação entre métodos de estimação clássicos e bayesianos. 2011. 69 f. Trabalho de Conclusão de Curso (Graduação em

Estatística) - Instituto de Matemática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2011. Disponível em: <http://www.lume.ufrgs.br/bitstream/handle/10183/36714/000818152.pdf?sequence=1>. Acesso em: 29 jan. 2013.

MARSH, H.; BOND, N.; JAYASINGHE, U. Peer review process: assessments by applicant-nominated referees are biased, inflated, unreliable and invalid. *Australian Psychologist,* v. 42, n. 1, p. 33-38, 2007.

MCGRAW, K. O.; WONG, S. P. Forming inferences about some intraclass correlation coefficients. *Psychological Methods,* v. 1, n. 1, p. 30-46, 1996.

MUELLER, S. P. M. A seleção de artigos científicos para publicação em revistas brasileiras: um levantamento de práticas e procedimentos adotados pelas revistas científicas brasileiras financiadas pelo CNPq e INEP, 1995-1996. *Revista de Biblioteconomia de Brasília,* v. 21, n. 2, p. 229-250, 1997.

MULLIGAN, A.; HALL, L.; RAPHAEL, E. Peer review in a changing world: An international study measuring the attitudes of researchers. *Journal of the American Society for Information Science and Technology,* v. 6, n. 1, p. 132-161, 2013.

PAVAN, C.; STUMPF, I. R. C. Avaliação pelos pares nas revistas brasileiras de ciência da informação: procedimentos e percepções dos atores. *Encontros Bibli,* v. 14, n. 28, p. 73-92, 2009.

PESSANHA, C. Critérios editoriais de avaliação científica: notas para discussão. *Ciência da Informação,* v. 27, n. 2, p. 226-229, 1998.

SEATON, A. V. Blowing the whistle on tourism referees. *Tourism Management,* v. 17, n. 6, p. 397-399, 1996.

SHROUT, P. E.; FLEISS, J. L. Intraclass correlations: use in assessing rater reliability. *Psychological Bulletin,* v. 86, n. 2, p. 420-428, 1979.

SMITH, A. J. The task of the referee. *Computer,* v. 23, n. 4, p. 65-71, 1990.

STEHBENS, W. E. Basic philosophy and concepts underlying scientific peer review. *Medical Hypotheses,* v. 52, n. 1, p. 31-36, 1999.

STUMPF, I. R. C. Avaliação pelos pares nas revistas de comunicação: visão dos editores, autores e avaliadores. *Perspectivas em Ciência da Informação,* v. 13, n. 1, p. 18-32, 2008.

WELLER, A. C. Editorial peer review: its strengths and weaknesses. *Information Today,* 2002. 342 p.

WOOD, M.; ROBERTS, M.; HOWELL, B. The reliability of peer reviews of papers on information systems. *Journal of Information Science,* v. 30, n. 1, p. 2-11, 2004.

YANKULOV, K.; COUTO, R. Peer review in class: metrics and variations in a senior course. *Biochemistry and Molecular Biology Education,* v. 40, n. 3, p. 161-168, 2012.