Relatório Técnico da DAV

AVALIAÇÃO DE ENSINO E PESQUISA

Ministério da Educação
Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

RELATÓRIO TÉCNICO DAV 1/2018

Publicação seriada que divulga os resultados de estudos e pesquisas desenvolvidos pela Diretoria de Avaliação da CAPES.

As publicações da DAV estão disponíveis para download gratuito no formato PDF.
Acesse: www.capes.gov.br

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR
DIRETORIA DE AVALIAÇÃO
COORDENAÇÃO GERAL DE NORMATIZAÇÃO E ESTUDOS
DIVISÃO DE ESTUDOS E PESQUISAS[1]
Gabriela da Rocha Barbosa[2]
Angélica Guedes Dantas[3]

AVALIAÇÃO DE ENSINO E PESQUISA

[1] Agradecemos os comentários e sugestões de Rita de Cássia Barradas Barata, Diretora de Avaliação da Capes, e de Sérgio Oswaldo de Carvalho Avellar, Coordenador-Geral de Acompanhamento e Avaliação do Mestrado Profissional (CGNE) e Analista em Ciência e Tecnologia da Capes.
[2] Chefe da Divisão de Estudos e Pesquisas e Analista em Ciência e Tecnologia da Capes.
[3] Analista em Ciência e Tecnologia da Capes.

1.Introdução

Metodologias de Avaliação são instrumentos cada vez mais valorizados na gestão de programas e políticas de governo. Sua principal atribuição é responder questões que evidenciam se um projeto ou conjunto de atividades estão no caminho certo para produzir os resultados esperados (RUEGG; FELLER, 2003).

É a partir do final da Segunda Guerra Mundial, quando da inauguração de uma nova dinâmica nas relações entre ciência, tecnologia e sociedade, que diferentes sistemas ou vertentes de avaliação foram desenvolvidos de forma ordenada, principalmente nos Estados Unidos e no Reino Unido (CHELIMSKY, 2006; LEEUW; FURUBO, 2008).

A atividade de avaliação atinge seu auge durante o estado de bem-estar social para identificar se as políticas públicas, implementadas em larga escala neste período, foram efetivas no enfrentamento dos problemas sociais (CHELIMSKY, 2006; ROSSI; LIPSEY; FREEMAN, 1998).

Dessa forma, a avaliação de programas/políticas sociais se apoiará principalmente no conhecimento desenvolvido em diferentes áreas das Ciências Sociais como Sociologia, Antropologia, Psicologia, Ciência Política e Economia. (CHELIMSKY, 2006; LEEUW, FURUBO, 2008; ROSSI; LIPSEY; FREEMAN, 1998). Por ser realizada por cientistas sociais, a avaliação se confundirá com a própria atividade de pesquisa. Vale ressaltar que programas sociais com componentes de avaliação também serão adotados em países em desenvolvimento, com destaque na América Latina para a área da Saúde Pública e Educação (ROSSI; LIPSEY; FREEMAN, 1998).

É em resposta às experiências empíricas da atividade de avaliação que se desenvolverá a discussão teórica com o intuito de embasar a escolha de instrumentos metodológicos. A atividade de avaliação se expandirá para o campo da administração e da análise política, tornando-se também uma atividade política e gerencial (ROSSI; LIPSEY; FREEMAN, 1998; SHADISH JR.; COOK; LEVITON, 1991).

Nos anos 1980, ferramentas e métodos de avaliação passam a ser utilizados para legitimar e melhorar o desempenho dos serviços públicos. A ampliação dos métodos de avaliação coincide com o processo de reforma na gestão pública dos serviços, conhecida como New Public Management. Essas ideias surgem inicialmente no Reino Unido, e se espalham posteriormente aos demais países membros da Organização para Cooperação e Desenvolvimento Econômico (OCDE), diante do acirramento da concorrência internacional e do crescente déficit econômico dos países desenvolvidos (POWER, 1999; SANDERSON, 2001).

A busca por redução de custos e produtividade faz com que se adote cada vez mais os princípios da organização empresarial na gestão pública. No bojo deste processo, assiste-se à diminuição do papel do Estado na regulação das atividades

sociais e econômicas por meio de iniciativas flexibilizadoras e da privatização de serviços, cada vez mais baseados em relações contratuais.

Essa maior autonomia e flexibilização na gestão dos serviços públicos levou à criação de ferramentas indiretas de controle que visam supervisionar a gestão dos serviços a distância. Cresce em importância o desenvolvimento de sistemas de monitoramento de desempenho e a realização de auditorias e inspeções, a tal ponto que, no início do século XXI, diversos autores chamam a atenção para o nascimento de uma nova cultura ou ideologia organizacional de avaliação (DAHLER-LARSEN, 2012; LEEUW; FURUBO, 2008; POWER, 1999).

Power (1999) acredita que os processos de auditoria ganham força também por conta de seu poder enquanto ideia; ou seja, por ser um sistema de conhecimento capaz de restaurar a confiança da sociedade diante da crescente percepção de riscos em diferentes áreas da vida[4]. Dessa forma, uma das características da avaliação é fornecer a garantia (assurance) de que as coisas irão funcionar da forma desejada (LEEUW; FURUBO, 2008; POWER, 1999).

Nesse sentido, a atividade de avaliação passa a servir tanto para direcionar a alocação mais eficiente dos recursos quanto para informar a sociedade e stakeholders sobre os resultados alcançados pelos investimentos públicos, no sentido do termo accountability[5].

Vale ressaltar que a promoção da accountability requer o desenvolvimento de relações formais visíveis e sujeitas a validação independente de forma a restaurar a confiança nos serviços ofertados, o que, por sua vez, encorajou o crescimento de um tipo específico de avaliação baseado na mensuração do desempenho por meio de padrões (standards) que devem ser alcançados, ou de esforços e produtos (inputs/outptus) que devem ser gerados e medidos por meio de indicadores (SANDERSON, 2001).

A disseminação de tais práticas propiciou a utilização de métricas de desempenho por organizações empresariais e governamentais como ferramentas de controle, de forma a direcionar comportamentos desejados a partir de mecanismos de recompensa ou punição (REBORA; TURRI, 2013).

Dos anos 1990 em diante, apesar da redução generalizada dos gastos dos governos em políticas sociais, a discussão em torno da avaliação de investimentos

[4] Conclusão apoiada no trabalho de Ulrich Beck (1992) "Sociedade de risco", que analisa a sociedade industrial e suas transformações, principalmente quanto a percepção dos efeitos indesejados provenientes dos avanços científicos e tecnológicos.

[5] A ideia é que, por meio da divulgação de relatórios anuais, as organizações públicas e privadas passam a prestar contas à sociedade, que, de posse de tais informações, estariam "empoderadas" para tomar decisões e escolher os melhores serviços.

[6] Iniciativas apoiadas na teorização do modelo tripla hélice desenvolvido por Loet e Etzkowitz (1996) e que tenta explicar a dinâmica do processo de inovação tecnológica por meio da transformação do conhecimento produzido pela interação não linear entre universidades, indústrias e governo.

em pesquisa e desenvolvimento (P&D) ganha força com a valoração crescente dos processos de inovação pelos países, principalmente por meio da colaboração entre universidade-indústria-governo[6], o que fomenta a discussão teórica da avaliação das atividades de pesquisa.

2.Como Avaliar?

Podemos descrever a avaliação como uma atividade que se utiliza de métodos, critérios e instrumentos com o propósito de mensurar, reconhecer, informar ou atribuir valor a algo ou alguém (ROSSI; LIPSEY; FREEMAN, 1998). A avaliação constitui-se, portanto, como um processo complexo, difícil de ser padronizado.

Neste sentido, não há uma única melhor maneira de se avaliar. Esta constatação é fruto de longa discussão teórica no campo da Educação, Filosofia e Administração Pública, discussão essa que pavimentou a evolução da teoria da avaliação situando sua importância na definição das escolhas metodológicas.

Vale ressaltar que não se trata de teorias descritivas, mas prescritivas no sentido de dizer quando, como e por que alguns métodos devem ser usados e outros não, bem como a forma com que os diferentes métodos podem ser combinados (CARDEN; ALKIN, 2012, p. 103; SHADISH JR.; COOK; LEVITON, 1991).

Carden e Alkin (2012) descrevem o desdobramento teórico da avaliação em 3 diferentes ramos, que denominaram "evaluation theory tree". Assim, mostram como, ao longo da história, surgiram as abordagens teóricas que priorizam os aspectos metodológicos da avaliação, aquelas que passarão a dar ênfase aos valores e interesses atribuídos aos processos e resultados e, finalmente, aquelas que atribuem importância ao uso que será dado aos esforços da avaliação.

Os autores ressaltam que estas abordagens não devem ser vistas de forma independente, uma vez que as diferentes visões se relacionam na construção de um framework de avaliação.

Mertens (2015) chama a atenção para o fato de que diferentes perspectivas teóricas no campo da avaliação relacionam-se a premissas filosóficas sobre a maneira como entendemos a realidade e o processo de construção do conhecimento. Essas diferentes visões poderiam ser descritas na forma de diferentes paradigmas orientadores da pesquisa e da avaliação[7], todavia sem negar a possibilidade de características de um campo teórico aparecerem em diferentes paradigmas.

Esta ressalva é válida, visto que o entendimento de que o paradigma determina o método alimentou, e em alguns casos ainda alimenta, discussões teóricas sobre qual seria o melhor método de avaliação. Neste sentido, as controvérsias em torno da utilização de abordagens qualitativas versus quantitativas geralmente está associada a métodos de pesquisas específicos e seus respectivos paradigmas.

[7] Que a autora denomina como pós-positivista, pragmático, construtivista e transformativo.

Uma das razões para o entendimento de que a aproximação com determinada premissa filosófica leva necessariamente à escolha entre abordagens insere-se na crença de que dados qualitativos são necessariamente subjetivos, e dados quantitativos, necessariamente objetivos. A dualidade entre objetivo/subjetivo apoia-se na definição de que tudo que é subjetivo é "influenciado pelo julgamento humano" ou serve para "medir sentimentos e crenças", mas evidências numéricas também estão imbuídas de premissas e valores (REINHARDT; COOK, 1979, p. 12; WEISS, 1998).

Assim, historicamente, observa-se que havia uma tendência nas atividades de avaliação de políticas sociais, principalmente na Sociologia e na Psicologia, à utilização de métodos de pesquisa experimental ou quase experimental[8], tendo em vista sua natureza "objetiva" e passível de controle, como único método, portanto, capaz de testar hipóteses sobre a realidade da política. A adoção de um enfoque qualitativo nas metodologias de avaliação ganha força a partir da constatação de que o objetivo precípuo da avaliação é valorar, julgar. Esta constatação insere a subjetividade do avaliador no processo de avaliação (CARDEN; ALKIN, 2012, p. 104). Teóricos subsequentes abrirão a caixa-preta da avaliação para além do avaliador, incorporando os significados e interesses de diferentes atores[i].

A possibilidade de se misturar atributos de determinado paradigma com perspectivas que se apoiam em ambos os métodos coincidirá com o nascimento de vertentes teóricas, dentro do campo da avaliação, cujo propósito está além da descoberta da verdade, aproximando-se de uma avaliação de natureza pragmática e propositiva.

Assim, encontram-se teóricos preocupados com o uso das informações que estão sendo geradas e que direcionam a avaliação para a tomada de decisão por gestores públicos. Há ainda aqueles que entendem como propósito da avaliação a busca por valores que ultrapassem a eficiência incluindo a promoção da justiça social e dos direitos humanos, principalmente em se tratando de avaliações de políticas públicas. Essa perspectiva chama atenção para questões éticas e de responsabilidade social ao entender como papel dos avaliadores o questionamento de dimensões críticas geradoras de assimetrias de poder na sociedade (MERTENS, 2015, p. 81).

De forma geral, o que se observa no campo teórico é a constatação da natureza contingente do método em relação à pergunta a ser respondida (DONALDSON; LIPSEY, 2006).

[8] "O delineamento experimental consiste em determinar um objeto de estudo, selecionar as variáveis que seriam capazes de influenciá-lo, definir as formas de controle e de observação dos efeitos que a variável produz no objeto" (GIL, 2008, p. 51).

2.1 Melhores práticas de avaliação

É a partir do relato de experiências em avaliação de políticas sociais e de programas de inovação que se pretende esboçar brevemente as melhores práticas de avaliação, bem como apontar suas principais dificuldades e críticas.

De forma geral, o primeiro passo fundamental para identificar o desempenho de um programa ou política pública é definir o seu escopo. Isso passa por definir o que se está fazendo e por quê, ou seja, por identificar quais os problemas que direcionam a criação do programa/política a ser avaliado (MCLAUGHLIN; JORDAN, 1999; RUEGG, FELLER, 2003; WEISS, 1998).

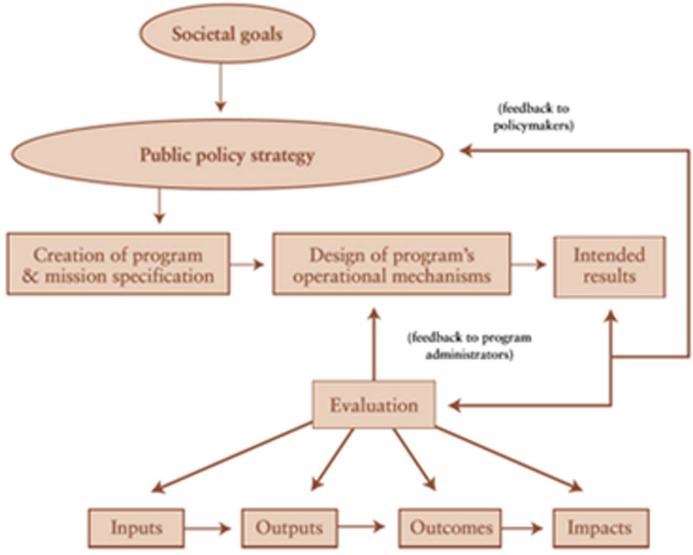
Definida a missão do programa, é preciso relacioná-la às questões que nortearão a avaliação. Neste sentido, a avaliação pode perguntar se os resultados esperados foram alcançados, mas também *como* foi alcançado e *por quê*, o que implica situar a avaliação como instrumento de conhecimento e não apenas de controle.

A definição das perguntas já delimita escolhas relativas à forma como os resultados serão mensurados, os fatores de contexto que precisam ser medidos e analisados e, com isso, quais instrumentos e técnicas serão utilizadas. É fundamental nesse processo a participação daqueles que farão uso das informações (*stakeholders*) de forma a compartilhar a visão e as expectativas de como o programa deverá funcionar (ROSSI; LIPSEY; FREEMAN, 1998; WEISS, 1998).

Esse entendimento do processo é o que os autores denominam de avaliação dirigida pela teoria (*theory-driven evaluation*), ou seja, pautada na importância de se pensar nos conceitos que embasam a implementação do programa (DONALDSON; LIPSEY, 2006; ROSSI; LIPSEY; FREEMAN, 1998).

Uma de suas abordagens é o modelo lógico da avaliação (*evaluation logic model*). Esta abordagem entende que uma boa prática de avaliação envolve construir, conjuntamente à definição das perguntas, o desenho ou diagrama da avaliação. Este diagrama ilustrará como o programa de avaliação deverá funcionar, servindo de guia para o desenvolvimento e prática da avaliação (Figura 1) (MCLAUGHLIN; JORDAN, 1999; RUEGG, FELLER, 2003).

Figura 1 – Modelo lógico de Avaliação



Fonte: Ruegg & Feller, (2003).

Em um programa de avaliação abrangente, ou seja, que passe por todas as fases da criação de uma política de governo, o ideal é combinar diferentes métodos e abordagens de forma a entender a relação entre esforço gerado (*inputs*), produtos (*outputs*), resultados (*outcomes*) e impactos (*impacts*), bem como casar os distintos enfoques de avaliação em cada fase do programa, que podem ser descritas da seguinte maneira:

• *Ex ante*: aquela que ocorre anteriormente à implantação do programa/projeto como parte do planejamento estratégico, de forma a analisar as opções disponíveis e estabelecer o melhor curso de ação para o alcance dos resultados esperados.

• Acompanhamento/monitoramento: realizada durante a execução do programa/projeto, serve para avaliar o progresso da avaliação, envolve aspectos de gestão do programa, ou seja, analisa quais procedimentos devem ser adotados, etc. Neste sentido, direciona alterações necessárias.

• *Ex post*: realizada após a execução do projeto para identificar os resultados alcançados e os possíveis impactos.

É esta construção conjunta que auxilia a determinar como a política está progredindo no alcance dos resultados, ou seja, trazendo conhecimento que direcionará decisões futuras, bem como captando efeitos não previstos ou indesejados no processo (RUEGG; FELLER, 2003).

O planejamento da avaliação não deve perder o foco do objetivo que se quer alcançar. Quanto mais abrangente o objetivo, mais ele demandará a combinação de diferentes instrumentos e métodos de pesquisa.

Link e Vonortas (2013, p. 5) apresentam uma classificação dos métodos de avaliação em programas de pesquisa e desenvolvimento quanto à sua natureza econômica, não econômica, híbrida ou guiada por dados.

De forma geral, pode-se descrever os métodos econômicos, como aqueles utilizados para analisar o desempenho de um programa ou política, baseado em medidas econômicas e de lucratividade, o que inclui preocupações em torno da alocação de recursos a nível da firma ou de projetos. Exemplos: análise econômica de impacto, métodos econométricos, análise de custo-benefício. Da mesma forma, há os métodos não-econômicos, empregados para avaliar a qualidade ou mérito de projetos, bem como a melhor alocação de recursos no atendimento a objetivos públicos. Exemplos: revisão por pares; análise de percepção pública (*research value mapping*), modelagem lógica (busca relações causais entre elementos como inputs, atividades, outputs e influências ambientais), etc.

Os métodos híbridos normalmente envolvem a análise de resultado/impacto em suas diferentes dimensões (social, ambiental, econômica, etc.). Exemplos: sociometria, análise de redes sociais, análise histórica (observação direta ou análise estatística de dados secundários para traçar comportamento social, econômico, etc.). Já os métodos guiados por dados partem da análise de informações provenientes de bancos de dados variados ou se dão por meio do levantamento de informações que podem ser analisados estatisticamente. Exemplos: dados bibliométricos, análise de patentes, *surveys* e estudos de caso.

Todavia o desenho da avaliação deve levar em conta a quantidade de recursos financeiros e de tempo disponível para sua execução, o que leva, na maior parte das vezes, à concentração de esforços em apenas uma das fases do programa.

Nesse sentido, as avaliações tendem a direcionar seus métodos para a obtenção de métricas que facilitem a sua operacionalidade em termos de custo e de tempo. Cabe, portanto, diferenciar a análise de resultados baseados em diferentes tipos de métricas.

De forma geral, podem ser utilizadas nas avaliações: métricas de *inputs*, *outputs*, *outcomes* e de análise de impacto. Métricas de *input* medem os esforços empreendidos em um projeto em termos de recursos financeiros, humanos, etc., enquanto as de *output* representam o resultado imediato deste esforço. Já os resultados de médio e longo prazo, que possuem relação direta com o objetivo do projeto, são chamados de *outcomes*.

A análise de impacto é uma medida de resultado diferente, que supera o escopo dos objetivos estabelecidos incluindo resultados não imaginados, positivos ou negativos.

De acordo com Salles Filho (2011, p. 1), a análise de indicadores de *input/output* em alguns casos pode, no máximo, ser utilizada como proxies de impacto, mas "impacto é uma medida diferente de resultado. Impacto é o efeito ou consequência que o resultado traz".

Para exemplificar, podemos pensar em uma agência de fomento que tem como um de seus objetivos a capacitação docente. Neste caso, uma métrica de *input* corresponderia ao montante de recursos investidos na formação dos alunos; enquanto uma métrica de *output* corresponderia ao número de alunos titulados. O resultado (*outcome*), por sua vez, passaria por verificar quantos destes egressos atuam enquanto docentes. Já uma análise de impacto envolveria a análise de múltiplas variáveis, como o impacto desta formação na geração de renda, na qualidade da educação e da pesquisa, etc.

É consenso que a avaliação de impacto é uma tarefa difícil, sendo dificultada metodologicamente por fatores como o longo tempo de conclusão dos efeitos do que se quer avaliar e a dificuldade em se estabelecer conexões de causalidade devido, dentre outros fatores, à natureza assimétrica da distribuição dos resultados de impacto (LINK; VONORTAS, 2013, p. 5). Além disso, há a dificuldade em mensurar impactos de forma simultânea, em sua multidimensionalidade econômica, social, ambiental, etc. (SALLES FILHO, 2011).

Ao lidar com tantas variáveis, avaliações de impacto demandam a combinação de diferentes técnicas e métodos de pesquisa (*a mixed method approach*) com prevalência no emprego de métodos de pesquisa experimental ou quase-experimental quando se deseja identificar relações de causa e efeito.

Desta forma, frequentemente, avaliações de projetos ou políticas são baseadas apenas nos inputs/outputs gerados, já que estes trazem consigo características de objetividade, sendo facilmente transformados em métricas de desempenho e indicadores.

Todavia diversos autores chamam atenção para que não se perca de vista as limitações inerentes a estes indicadores, de forma a não reduzir as possibilidades de compreensão da realidade que eles tentam traduzir.

Thiel e Leeuw (2002), por exemplo, ressaltam a importância de se entender o significado dos indicadores de forma a observar o acontecimento, intencional ou não, do que se denomina *performance paradox*, ou seja, a deterioração dos indicadores a partir da sua fraca correlação com o desempenho[9].

Tal processo de deterioração pode ocorrer em função de diversos fatores, como quando:

- O desempenho melhora e o indicador não consegue mais identificar o desempenho ruim, o que os autores denominam de "aprendizado positivo";
- o avaliado manipula a avaliação, pois sabe quais indicadores são utilizados

[9] Termo criado por Meyer & Gupta (1994), em *The performance paradox.*

ou não, direcionando o seu esforço somente para aquele aspecto que está sendo avaliado. Esse "aprendizado perverso" pode levar a um processo de paralisia e falta de inovação;

• o fator seleção ocorre, ou seja, quando há a substituição dos avaliados que possuem um desempenho inferior por aqueles que apresentam melhor desempenho, o que faz o indicador perder o seu poder de discriminação.

• Por fim, há o que os autores denominam de supressão: quando as diferenças de desempenho são ignoradas.

Os indicadores também dizem pouco quando estão baseados em premissas que perderam a validade ou que precisam ser questionadas. Para Leeuw & Furubo (2008), avaliações com enfoque em desempenho tendem a reduzir as possibilidades de questionar as premissas sobre as quais as políticas foram criadas. Assim, questões fundamentais que respondem onde as políticas foram bem-sucedidas ficam sem resposta.

Os autores dão o exemplo de avaliações de campanha de convencimento na qual os indicadores apontam o recebimento das informações por uma população, mas não conseguem identificar a mudança de comportamento gerada pela campanha, ou seja, apenas confirmam as premissas de como uma determinada política funciona.

Quando a avaliação é feita nesses termos, há uma tendência de se buscar maximizar os indicadores de forma descolada dos objetivos da política de governo, levando ao que Bohte e Meier (2000) denominam de *goal displacement*[ii].

O que se observa é que os efeitos da avaliação não estão limitados a ganhos de produtividade e de eficiência. Em muitos casos, os avaliadores se tornam os políticos de fato, sendo necessário entender os efeitos práticos da avaliação no comportamento dos avaliados e na ressignificação das atividades públicas[iii] (DAHLER-LARSEN, 2012; POWER, 2005, p. 335).

Diante de tais limitações, alguns autores ressaltam a importância de ao longo do tempo se repensar e reformular os indicadores utilizados. Ademais reconhecem a importância de abordagens mais participativas e qualitativas, disseminando responsabilidade e desenvolvendo capacidade para avaliação entre o público alvo da avaliação e demais interessados. Isto asseguraria a definição de objetivos apropriados, que não podem perder de vista que o objetivo da avaliação de políticas sociais não é apenas medir eficiência, mas também promover accountability, justiça e equidade (ROSSI; LIPSEY; FREEMAN, 1998; SANDERSON, 2001; THIEL; LEEUW, 2002).

3.Sistemas De Avaliação

Neste item, iremos expandir o foco de análise da atividade de avaliação de políticas ou programas de governo para a análise de sistemas de avaliação que

compreendem setores inteiros ou áreas de atividade não necessariamente restritas a fronteira dos países, mas podendo envolver a participação de organizações supranacionais.

Apesar das especificidades locais de cada sistema de avaliação, a ideia é que estes compartilhem uma lógica de avaliação e de produção de informações, visto que possuem o mesmo objeto de análise e o alcance de objetivos semelhantes.

De acordo com Leeuw e Furubo (2008), para configurar-se enquanto um sistema de avaliação, a atividade deve prever a produção de um tipo específico de conhecimento fruto de uma perspectiva epistemológica compartilhada e acordada entre seus agentes e direcionada para orientar processos decisórios. As atividades devem possuir um arranjo permanente, com certo volume e periodicidade, e envolver a participação e compartilhamento de responsabilidades entre mais de uma organização, principalmente com a organização usuária da avaliação.

A seguir, pretende-se descrever a forma como se organiza a avaliação das atividades de ensino e de pesquisa científica a partir da perspectiva de sistemas de avaliação esboçada aqui brevemente. Para tanto, foram selecionados alguns países, dentre os quais: Espanha, Reino Unido, Finlândia, Estados Unidos, França, México, Colômbia e Chile. A escolha se deu pela disponibilidade de acesso à documentação de forma online, pela posição geográfica dos países e pela diferença nos propósitos de avaliação.

3.1.Sistema de acreditação de ensino e pesquisa

A partir dos anos 1980, diversos autores relacionam a expansão do sistema de acreditação de ensino à diminuição do papel do Estado na regulamentação dos serviços públicos. Este processo, conhecido como New Public Management, foi liderado pelo Reino Unido e posteriormente disseminado aos demais países membros da OCDE.

Neste processo, as instituições de ensino ganham maior autonomia e passam a estabelecer sistemas gerenciais internos de qualidade, de forma a responder as demandas do mercado. A acreditação transforma-se assim no principal mecanismo de controle de qualidade da formação acadêmica e profissional em cada país.

O grau de autonomia e desregulamentação das instituições de ensino varia entre os países. Em alguns deles, o processo de acreditação é centralizado em órgãos vinculados ao Estado; em outros casos, organizações privadas ligadas a distintas áreas profissionais e do conhecimento podem realizar a tarefa. Da mesma forma, o processo de acreditação nem sempre é obrigatório; alguns países facultam à instituição a busca por esse reconhecimento junto a agências acreditadoras. Nos casos em que a acreditação é obrigatória, ela pode levar ao descredenciamento dos cursos que ficam impedidos de continuar suas atividades.

As unidades de avaliação também são variadas, podendo ser acreditados programas e/ou instituições de ensino públicos e privados, departamentos ou apenas títulos (degrees) de graduação ou de pós-graduação. Cursos de doutorado também são acreditados, todavia, na maior parte dos países, estes também são avaliados a partir de suas atividades de pesquisa.

De forma geral, o processo de acreditação envolve avaliação ex ante (submissão de auto avaliação) e ex post (avaliação para acreditação). Geralmente, a acreditação possui um prazo de validade, sendo necessário processos de revalidação cuja periodicidade varia em cada país.

A realização de autoavaliação geralmente envolve a resposta a um questionário e a apresentação de documentação que corrobore as informações fornecidas. De forma geral, contém perguntas sobre o corpo discente, currículo, administração, facilidades e suporte institucional.

As propostas são submetidas para avaliação por pares, geralmente organizadas em comissões que podem ou não ser subdivididas entre as distintas áreas do conhecimento. As comissões são formadas por pessoas da academia, governo, indústria e diferentes setores privados. Pode contar também com a presença de estudantes.

Na maioria dos casos ocorrem visitas in loco, nas quais uma comissão realiza entrevistas e verifica a veracidade das informações declaradas. Ao final, é produzido um relatório de avaliação que permite ao programa corrigir possíveis falhas e melhorar processos.

Em alguns casos, há o processo de acompanhamento do programa antes de ele submeter o seu título para revalidação. A ideia de avaliação de acompanhamento ganha ares de avaliação formativa[iv], já que visa garantir e aprimorar a qualidade do programa.

Nesse sentido, vale ressaltar a simbiose crescente entre sistemas de acreditação e sistemas de avaliação de desempenho com objetivo de assegurar a qualidade do ensino (LEEUW; FURUBO, 2008). Como visto anteriormente, avaliações baseadas em desempenho utilizam-se de padrões (standards) que devem ser alcançados ou de esforços e produtos (inputs/outptus) que devem ser gerados e medidos por meio de indicadores (SANDERSON, 2001).

Pode-se dizer que esse processo ganha força a partir do final da década de 1990, quando da implementação de um modelo único de avaliação de qualidade do ensino para os países europeus. A criação deste sistema foi celebrada pela Declaração de Bolonha, assinada em 1999, e até aquele momento circunscrita aos países da União Europeia[10].

A declaração prevê a criação de um sistema de ensino comum aos países

[10] Ao longo dos anos, diferentes países assinaram a declaração. Atualmente, mais de 40 países integram o sistema.

integrantes por meio da convergência de suas políticas de ensino superior. Dentre seus principais objetivos está a ampliação da competitividade do sistema europeu de ensino superior e a promoção da empregabilidade e mobilidade de estudantes e pesquisadores europeus.

De forma a atestar o resultado de aprendizagem do aluno e permitir a sua comparação internacional, faz-se necessária a criação de um sistema nacional de avaliação e certificação que garanta a qualidade do ensino. Este sistema baseia-se na criação de uma estrutura de titulação dividida em 3 ciclos – graduação, mestrado e doutorado, cada qual com seu sistema de créditos –, bem como a criação e difusão de critérios e metodologias para a avaliação da qualidade dos programas.

O acordo europeu para a educação superior (EHEA) prevê que este sistema deve incluir: definição de responsabilidade das equipes e instituições envolvidas; avaliação de programas ou instituições (incluindo avaliação interna e avaliação externa); participação de estudantes e a publicação dos resultados; criação de um sistema de acreditação; certificação ou comparação de procedimentos; participação internacional, cooperação e networking (MARCO ESPAÑOL DE CUALIFICACIONES DE EDUCACIÓN SUPERIOR, 2014).

A premissa fundamental do framework dos países europeus integrantes de Bolonha é que as qualificações são concedidas com base na realização demonstrada dos resultados de aprendizado (learning outcomes). Assim, em vez de trazer o enfoque da acreditação na comparação dos esforços empregados (inputs) pelos países em termos de quantidade de anos de estudo ou quantidade de créditos, a análise recai em dados que buscam demonstrar o conhecimentos e as habilidades adquiridas pelo discente, sua aplicação em consonância com as demandas do mercado de trabalho, sua capacidade em realizar julgamentos, etc.

O acordo traz expectativas de aprendizado e habilidades genéricas que correspondem a cada ciclo de formação, mas que não possuem a intenção de serem prescritivos nem exaustivos.

O framework, portanto, não prescreve como deve ser a organização interna de um programa acadêmico. As universidades é que são responsáveis por estabelecer e manter a qualidade de seus programas. Assim, cada instituição irá demonstrar a forma como ocorrerá o progresso acadêmico e intelectual dos discentes a partir dos pressupostos de aprendizado que considera necessário.

A qualidade do programa é auditada de forma indireta pelas agências de acreditação, já que estas verificarão a existência e funcionamento do sistema interno de garantia de qualidade (internal quality assurance systems) pelas instituições. Este sistema será submetido a auditoria e, por sua vez, garantirá a acreditação das instituições.

Na maior parte das vezes, as universidades são obrigadas a dar transparência

dos resultados desta avaliação interna nos sites institucionais dos programas. Os procedimentos que garantem a qualidade dos programas geralmente envolvem a disseminação de informações quanto a: mobilidade; inserção dos egressos no mercado de trabalho; satisfação dos alunos; inclusão social de minorias, etc.

Iniciativas recentes de países latino-americanos caminham na direção de vincular cada vez mais a criação de sistemas de acreditação com sistemas de avaliação de desempenho com foco na qualidade, principalmente de programas de doutorado, modalidade de ensino em expansão na maior parte dos países da região.

Assim como os demais países europeus, estes chamam atenção para a utilização do sistema de qualidade como forma de ampliar a colaboração e a mobilidade de estudantes e pesquisadores, bem como priorizar áreas de pesquisa estratégica para os países da região.

Todavia há ainda uma grande assimetria na região quanto à oferta de cursos de pós-graduação, e o tamanho dos sistemas de pós-graduação destes países acompanha as escolhas metodológicas e operacionais de avaliação destes. Desta forma, diferentemente do Brasil, que já possui certa tradição na acreditação e avaliação de desempenho de programas de pós-graduação, alguns países como Chile, Colômbia e México há pouco iniciaram esse processo.

De forma geral, o processo de acreditação na maior parte dos países latino-americanos é voluntário, seguindo a mesma lógica operacional descrita aqui anteriormente como processos de autoavaliação e revisão por pares informada por indicadores de input/output.

O Quadro 1, em anexo, traz informações comparativas quanto ao processo de acreditação na Espanha, Estados Unidos, Finlândia, França, Reino Unido, Colômbia, Chile e México. O Quadro 3 reúne os principais critérios utilizados pelos países no processo de acreditação.

Cabe ressaltar a iniciativa recente do Reino Unido de adoção de avaliação de desempenho acadêmico nas atividades de ensino por meio da elaboração de um novo framework de avaliação do ensino, o Teaching Excellence Framework (TEF), que tem como unidade de análise as universidades. O objetivo da avaliação é informar os estudantes sobre os melhores cursos, reconhecer e premiar o ensino de excelência e promover a melhoria do ensino aproximando-o das habilidades demandadas pelo mercado de trabalho.

A avaliação do TEF está em fase piloto, e o primeiro teste foi realizado em meados de 2016. Ela previu a utilização de métricas e de informações submetidas

[11] Dentre as características constam a modalidade de ensino ofertado (integral ou parcial), número e características étnicas, de sexo, idade, domicílio e informações econômicas dos alunos que frequentam a instituição, de forma a não penalizar universidades quando da análise do rendimento escolar dos alunos.

pela universidade no formato de autoavaliação, que é encaminhada para análise por um painel de especialistas que se utilizam também de informações de contexto. Utilizou-se também de indicadores provenientes da Pesquisa Nacional do Estudante (National Student Survey) quanto à qualidade do ensino e ao suporte acadêmico, o que vem suscitando diversas controvérsias conforme aponta Hall (2017).

As universidades foram comparadas em clusters a partir da utilização de três critérios principais: qualidade do ensino; ambiente de aprendizado e aprendizado do aluno; e resultados[11]. A avaliação priorizou a utilização de métricas para medir resultado e impacto do ensino no aprendizado. Vale ressaltar a incorporação da avaliação de egressos, utilizando, por exemplo, percentual de alunos empregados ou que continuam estudando, e aqueles que se encontram em trabalhos superqualificados ou que estavam estudando por mais de seis meses após a titulação (DEPARTMENT FOR BUSINESS, INNOVATION AND SKILLS, 2016).

A avaliação não é obrigatória e não está diretamente vinculada ao fomento. Todavia universidades de língua inglesa que são bem avaliadas no sistema (recebendo os conceitos de universidades ouro, prata ou bronze) recebem autorização para aumentar o valor das suas taxas de mensalidade. Incentivo que, além do prestígio, estimula a competitividade entre as instituições A seguir, é apresentado o sistema de avaliação de desempenho para fomento da pesquisa. Este assume características diferentes do sistema de avaliação de ensino, tendo em vista que a atividade de pesquisa, desenvolvida principalmente em programas de doutorado, está atrelada aos sistemas de inovação dos países.

Neste processo, as universidades, que já acumulam uma dupla identidade em torno das atividades de ensino e pesquisa, passarão por transformações importantes em sua missão ao serem cada vez mais submetidas à governança dos sistemas de inovação dos países e de suas ferramentas de controle apoiadas em métricas e indicadores de inovação (HICKS, 2012; REBORA; TURRI, 2013).

3.2.Sistema de avaliação de desempenho da pesquisa

Os procedimentos de avaliação da ciência surgem no bojo da própria atividade como parte do processo de construção do conhecimento científico, uma vez que, para adentrar no domínio público com o status de conhecimento válido, a ciência precisa ser comunicada e avaliada (DAVYT; VELHO, 2000).

Todavia pode-se dizer que é a partir do final dos anos 1980 que começa a se configurar um sistema de avaliação de desempenho da pesquisa, de cunho político e em âmbito nacional, com o objetivo de nortear a distribuição de recursos para pesquisa.

O primeiro país a ter essa iniciativa foi o Reino Unido, a partir da criação do Research Assessment Exercise (RAE), sendo seguido posteriormente por outros

países (HICKS, 2012).

A configuração deste sistema de avaliação da pesquisa acompanhou mudanças importantes no conjunto de crenças, valores e instituições que embasam as atividades de ciência e tecnologia dos países.

Em termos de política científica, o que se observa é que seu surgimento vem responder ao esgotamento do contrato social de Vannevar Bush apoiado no modelo de inovação linear ofertista[v] (VELHO, 2011).

Somam-se a isso mudanças importantes na forma de se fazer ciência no mundo, principalmente em termos de visibilidade da ciência produzida. Entre os anos 1990 e início dos anos 2000, a colaboração internacional dobrou em termos de publicação em coautoria estrangeira[12]. Esse comportamento foi percebido em todas as áreas do conhecimento em maior ou menor grau, chamando a atenção para o que alguns autores consideram ser o surgimento de um "sistema global de ciência" RAMOS; VELHO, 2012).

Nesse processo de intensificação da colaboração internacional, muda-se a forma com a qual governos e os líderes institucionais passam a enxergar o papel das universidades, e o prestígio por elas assumido aparece como fator fundamental na atração de talentos para os países.

É nesse contexto que surgem os rankings de universidades mundiais com o intuito de indicar as instituições de mais alta qualidade no ensino, aprendizado e pesquisa a partir da mensuração de indicadores, com destaque para a produção indexada em bases internacionais e em colaboração.

A busca pela excelência passa paulatinamente a substituir a palavra qualidade, e as avaliações servem cada vez mais para ranquear os melhores entre os bons (VESSURI; GUÉDON; CETTO, 2013). Pode-se dizer que essa é uma das principais diferenças em relação ao sistema de acreditação, já que a intenção está além da garantia de um padrão mínimo de qualidade, mas em promover prestígio e competitividade de forma a estimular a excelência, ou seja, a produção de conhecimento novo.

Soma-se a isso preocupação cada vez maior em avaliar a relevância da ciência financiada com recurso público para o conjunto da sociedade, o que, por sua vez, promove a incorporação de valores "extracientíficos" nas considerações sobre avaliação da pesquisa[13].

É em consonância com essas mudanças que, a partir de 2008, o Reino Unido

[12] Adams (2013) aponta a visibilidade como um dos principais resultados da internacionalização da pesquisa feita pelos países, sendo que nas últimas três décadas os artigos que receberam maior número de citações (FI) são fruto de colaboração internacional.

[13] Assim, aos poucos, o modelo de ciência universal e socialmente neutra dá espaço para a existência de várias ciências ou estilos nacionais de produção de conhecimento. Essa mudança acompanha o surgimento de uma nova visão de ciência, culturalmente situada e construída (VELHO, 2011).

substitui paulatinamente o modelo de avaliação do RAE pelo Research Excellence Framework (REF), cuja primeira atividade de avaliação foi realizada no ano de 2014, e que, dentre outras considerações, amplia o peso para a avaliação de impacto "extracientífico" da pesquisa.

A racionalidade que direciona o desenvolvimento de sistemas de desempenho como o do Reino Unido prevê a concentração de recursos para fomento de pesquisa, destinando-o apenas para os centros de excelência ou de melhor desempenho, de forma a estimular a competição e fazer com que entidades com pior desempenho melhorem (HICKS, 2010).

O Reino Unido lidera, por enquanto, esse modelo de avaliação voltado para a pesquisa de excelência e com o propósito de distribuir recursos de fomento. Outros países paulatinamente se integram a alguns de seus aspectos, mas sem vincular diretamente o desempenho à distribuição de fomento[vi]. Na maioria deles, o objetivo da avaliação é produzir informações para accountability e tomada de decisão organizacional e de políticas públicas.

Esta constatação parte da análise da avaliação da pesquisa realizada em uma amostra de países que incluem Reino Unido, Espanha, Estados Unidos, França e Finlândia (vide informações no Quadro 2 e 4, em anexo).

A partir deste levantamento, observa-se que quanto ao modo de operação, as avaliações são realizadas, em sua maioria, de forma periódica por agências externas ligadas ou não a ministérios de educação e de ciência e tecnologia, ou pelas próprias universidades como parte de seu sistema de avaliação interna.

Na Finlândia, são as universidades que elaboram a avaliação e divulgam os resultados em relatórios de avaliação, geralmente a partir da contratação de especialistas, principalmente para a análise de métricas e indicadores. Na França, desde 2016, as instituições têm a possibilidade de escolher a agência que realizará a avaliação das suas atividades de pesquisa.

A unidade de análise pode ser a universidade, o departamento/programa, a área do conhecimento ou o pesquisador, como é o caso da Espanha, onde são avaliadas as publicações dos professores universitários e pesquisadores públicos para fins de progressão na carreira ou ganhos adicionais de salário.

A avaliação não é obrigatória em todos os países. Nos Estados Unidos, por exemplo, as avaliações das atividades de pesquisa realizadas por agências governamentais, como é o caso do National Research Council, coletam informações divulgadas de forma voluntária pelos programas de pós-graduação.

Na Espanha, também não é obrigatória, salvo para aqueles pesquisadores que desejam concorrer ao recebimento de gratificação adicional (sexênio ou progressão para o cargo de professor titular). Todavia a crítica que mais se encontra é que o que seria um complemento de retribuição foi se transformando em indicador da qualidade do docente/pesquisador.

O que se verifica é uma obrigatoriedade informal, pois, estando ou não atrelada diretamente à obtenção de recursos de fomento, a avaliação promove o julgamento do público e estabelece o prestígio da instituição, o que funciona como estímulo para a competição e consequente adesão aos processos de avaliação (HICKS, 2012).

Em se tratando da mensuração de desempenho, a premissa é que as universidades ou institutos de pesquisa produzam conhecimento ao transformarem recursos financeiros (e outros recursos) em publicações, e na formação de novos pesquisadores (HANSEN, 2010).

Esses recursos e produtos (inputs/outputs) gerados são em boa parte medidos por indicadores. Conforme tipologia criada por Hansen (2010) haveriam 3 tipos de indicadores utilizados na avaliação de pesquisa:

• Indicadores de primeira ordem, cujo objetivo é medir o desempenho por meio de medidas de input, processos, estrutura e/ou resultado[vii];

• indicadores de segunda ordem, que fornecem índices com o objetivo de fornecer medidas simples de efeito, como o fator de impacto de revistas e o índice h[viii];

• e indicadores de terceira ordem, produzidos por meio de avaliação por pares.

De forma geral, o que se observa é que, nos países analisados, as avaliações de desempenho são ex post com critérios de análise voltados para os resultados imediatos (outputs) da pesquisa e, em alguns casos, para a avaliação de impacto "extracientífico" do conhecimento gerado.

Observa-se a predominância na utilização de revisão por pares ou painel de especialistas. No Reino Unido, por exemplo, a avaliação se dá por meio de quatro grandes painéis, subdivididos em 36 áreas do conhecimento (subpainéis).

A revisão por pares geralmente é organizada em comitês/painéis divididos por áreas do conhecimento. Os comitês de avaliação são formados por especialistas externos e independentes que podem ser representantes acadêmicos ou de associações de classe, especialistas estrangeiros e/ou stakeholders. Cada comitê possui liberdade para estabelecer seus critérios e padrões de avaliação que, de forma geral, visam atestar a originalidade do trabalho, o rigor metodológico e o impacto do trabalho sobre a comunidade científica e a sociedade.

Os comitês podem trabalhar também a partir da análise de um relatório de autoavaliação, como é o caso da avaliação realizada na França.

A avaliação realizada por pares pelo National Research Council nos Estados Unidos é a que mais se diferencia das demais, já que a opinião dos especialistas visa estimar o grau de reputação dos departamentos de pesquisa avaliados. A avaliação é utilizada para inferir a qualidade percebida dos programas em cada área do conhecimento. Assim, há o ordenamento dos melhores programas e a identificação indireta das características determinantes da qualidade.

Na maior parte dos casos, a revisão por pares é informada por indicadores, já que a utilização destes traz redução de tempo e de custos para a atividade. Os pares, em alguns casos, exigem a entrega de dados factuais sobre as atividades e resultados alcançados pelas unidades avaliadas (evidence-based assessment).

Observa-se uma predominância de indicadores de input e output, principalmente a contagem do número de publicações, quando o enfoque está voltado para assegurar a qualidade dos programas. Cabe destacar, como exceção, a avaliação realizada pela França, que, desde 2013, busca simplificar seu processo avaliativo. Foram abolidos indicadores que medem a produtividade do pesquisador, mantendo-se apenas indicadores de desempenho organizacional.

Já no Reino Unido, onde o foco da avaliação é a excelência, analisam-se as melhores produções indicadas pelos programas (até quatro produções por docente selecionado para participar da avaliação). Na última avaliação realizada pelo REF, ficou a critério dos subpainéis utilizar a contagem de citações das produções indicadas. Apenas 11 dos 36 painéis utilizaram dados de citação fornecidos pela base Scopus, já que esta possui a melhor cobertura (WANG; VUOLANTO; MUHONEN, 2014).

A avaliação realizada nos Estados Unidos, pela NRC, e na Finlândia também utilizam a contagem de citação como forma de mensurar o impacto das publicações. Vale ressaltar que a análise bibliométrica realizada por algumas universidades da Finlândia contou com pessoal especializado do Centre For Science and Technology Studies (CWTS) da Universidade de Leiden na Holanda (WANG; VUOLANTO; MUHONEN, 2014).

Nas áreas do conhecimento que apresentaram baixa cobertura de indexação nas bases internacionais como Web of Science e Scopus, foram utilizadas bases de dados alternativas ou informações disponibilizadas pelas próprias unidades de avaliação.

Dentre os países analisados, quase não há menção da utilização de indicadores de segunda ordem como fator de impacto de periódicos e índice h, com exceção da Espanha e Finlândia, que citam a utilização de indicadores de impacto de periódicos na avaliação. Entre algumas universidades da Finlândia, há a preocupação em se utilizar o fator de impacto normalizado de forma a obter uma pontuação relativa que permita comparar com o nível médio de citação daquela área do conhecimento (WANG; VUOLANTO; MUHONEN, 2014).

Quanto à análise de indicadores que visam medir o impacto "extracientífico" da pesquisa, cabe destacar a avaliação realizada pelo Reino Unido e pela Finlândia. Nesses países, a sistemática de avaliação envolve a entrega, pelo departamento a ser avaliado, de um estudo de caso no qual são explicitados os benefícios da pesquisa realizada em termos de impactos sociais, culturais, econômicos, etc. Na Finlândia, o estudo de caso é informado por indicadores. As unidades de pesquisa

também são avaliadas pela maneira como desenvolvem estrategicamente suas atividades de forma a alcançar os impactos desejados.

Nos documentos dos demais países analisados, nota-se a utilização de indicadores que visam medir o impacto da pesquisa em termos de inovação ou aplicação do conhecimento. Essa aferição se traduz em métricas que visam contabilizar a produção de patentes e atividades e os recursos externos obtidos por meio da colaboração entre instituições não acadêmicas, da criação de empresas startups, etc.

Ao final de todos os processos de avaliação, são atribuídos conceitos ou notas convertidas em rankings de classificação disponibilizados publicamente. Em países como Reino Unido e Finlândia, o resultado das avaliações traduz o grau de excelência a partir da escala geográfica do impacto produzido (de prestígio nacional a líder mundial), uma vez que o critério se dá por meio da comparação internacional do trabalho realizado.

A seguir, há de se chamar a atenção brevemente para as limitações e críticas dos modelos de avaliação de pesquisa aqui apresentados.

3.2.1.Críticas e limitações

A avaliação da ciência, seus critérios e ferramentas envolvem muitas controvérsias, principalmente quando da utilização cada vez maior de indicadores bibliométricos baseados em citação e produtividade como proxies para avaliar impacto e qualidade da pesquisa.

A premissa que apoia a utilização destes indicadores é de que a comunicação da ciência é componente fundamental ao processo de retroalimentação da produção do conhecimento, e de que quanto mais citações uma publicação recebe mais importante ela deve ser. É baseada nesta premissa que a citação, enquanto visibilidade da ciência produzida, é expressa como sinônimo de qualidade.

Todavia estudos apontam que são diversos os motivos que levam à citação de um artigo, não necessariamente a influência que o trabalho exerceu sobre a pesquisa[ix]. Nesse sentido, pode-se afirmar a importância de se ater aos pressupostos e limitações inerentes a cada indicador quando se pretende utilizá-los para avaliar a qualidade da pesquisa científica.

Pode-se citar como exemplo o fator de impacto de citação (FI) fornecido pela Web of Science. Criado na década de 1960 para auxiliar a busca por informações relacionadas a um tema[14], passou a ser utilizado para indicar a qualidade de periódicos e frequentemente é utilizado para indicar a qualidade dos artigos individuais e de seus autores – prática muito criticada por cientometristas e pela comunidade de pesquisa. Soma-se a isso a sua utilização para comparar áreas do

[14] A partir do pressuposto de que as referências indicam o relacionamento entre documentos.

conhecimento diferentes sem levar em conta os padrões distintos de publicação e de citação de cada uma delas[15].

O índice h, por sua vez, permite a comparação de cientistas em termos de seus impactos científicos mesmo que a quantidade de publicações ou de citações sejam bem diferentes, entretanto há limitações para seu uso. O indicador só possui significado quando comparado na mesma área do conhecimento. O tempo de atividade do pesquisador faz diferença na contabilização, já que o cálculo é influenciado pelo número de publicações e, no caso de publicações escritas em coautoria, há o risco de se inflar artificialmente as citações.

Há inúmeras críticas quanto à utilização de indicadores de citação e produtividade de forma absoluta e determinante da qualidade da pesquisa; dentre elas, a de que os produtos de pesquisa são variados e não se resumem a publicação em periódicos, restrição que criaria um ambiente avesso à inovação, à interdisciplinaridade, etc.

Outros salientam os diversos vieses da avaliação decorrentes do comportamento dos cientistas diante de tal modelo de avaliação de desempenho, quais sejam: práticas de autocitação e de coautoria, de forma a inflar indicadores de desempenho, e a corrida incessante pela publicação em quantidade em detrimento da qualidade.

Logo, a comunidade científica passou a expressar sua preocupação quanto aos efeitos indesejados da utilização desses indicadores para avaliar a qualidade científica. Vale ressaltar iniciativas conhecidas mundialmente como a Declaration on Research Assessment (DORA), lançada por cientistas de São Francisco (EUA) em 2012, seguida, em 2015, pelas publicações do Manifesto de Leiden e de "The Metric Tide" (a maré das métricas), este último feito por uma comissão independente para analisar o papel das métricas na avaliação, tendo como base o resultado do REF em 2014.

No relatório, chega-se à conclusão de que nenhuma métrica pode substituir eficientemente a revisão de pares, apesar das controvérsias em torno de suas limitações e, também, da tentativa de alguns estudos de provar a existência de boa correlação entre o resultado de avaliação obtido por peer review e por bibliometria (BORNMANN; HAUNSCHILD, 2017).

De forma geral, todas as manifestações sinalizam a necessidade da utilização de métricas responsáveis e do estabelecimento de um conjunto de princípios para a utilização de indicadores na avaliação da atividade de pesquisa. De acordo com Wilsdon et al. (2015), métricas responsáveis podem ser entendidas em termos das seguintes dimensões:

[15] Já foi apontado que o FI é influenciado por um pequeno número de artigos muito citados, e que seu valor é influenciado pelo campo de conhecimento, o tipo de documento, o ano da publicação e a janela de citação utilizada no cálculo (WALTMAN, 2016).

• Robustez: basear métricas nos melhores dados possíveis em termos de precisão e escopo.

• Humildade: reconhecer que a avaliação quantitativa deve apoiar – mas não substituir – a avaliação qualitativa e especializada.

• Transparência: manter a coleta de dados e os processos analíticos abertos e transparentes para que aqueles que estão sendo avaliados possam testar e verificar os resultados.

• Diversidade: contabilizar a variação por campo e usar uma série de indicadores para refletir e apoiar a pluralidade de caminhos de pesquisa e de carreira do pesquisador.

• Reflexividade: reconhecer e antecipar os efeitos sistêmicos e potenciais dos indicadores, atualizando-os em resposta.

Cabe destacar que, desde então, surgiram diversas iniciativas que buscam desenvolver indicadores normalizados que possam corrigir ou amenizar o efeito indesejado de determinadas variáveis sobre o resultado, principalmente os que permitem a comparação entre periódicos de diferentes áreas do conhecimento.

Em se tratando de indicadores de impacto de periódicos, essa normalização pode ocorrer ao se verificar o comportamento de citação de cada área do conhecimento (cited side normalization), baseado no fato de que, em algumas áreas do conhecimento, as listas de referência são maiores do que em outras; ou pela quantidade de citações que cada publicação recebe (citing side normalization). Um indicador que se baseia no potencial de citação do periódico é o SNIP, fornecido pela Scopus. Esse indicador se baseia também na reputação e atribui um peso diferente às citações dependendo da sua fonte (WALTMAN, 2016).

Em decorrência principalmente da mudança de enfoque da avaliação de qualidade para a pesquisa de excelência, observam-se também mudanças na oferta de indicadores de impacto. Essa mudança incorpora a ideia de que a produção de resultados de pesquisa inovadoras estaria assentada em pressupostos diferentes e demandaria métricas alternativas e outras fontes de mensuração.

Assim, observa-se uma redução no número de produções analisadas e da utilização da contagem de citação a nível de artigo, principalmente com o advento de novas bases de busca. Dentre elas, pode-se citar os que selecionam apenas as publicações com maior número de citações, como o i10 index fornecido pelo Google Scholar (publicações que receberam pelo menos dez citações), e indicadores normalizados que verificam a proporção de publicações que pertencem às 10% mais citadas de cada área do conhecimento (WALTMAN, 2016).Há também os indicadores que visam medir o impacto "extracientífico" da pesquisa, ou seja, que visa medir a influência da pesquisa em outros setores da sociedade e que, portanto, não passam apenas pela contabilização de produções. Todavia sua utilização

envolve ainda muitas controvérsias. As controvérsias passam pela resistência de cientistas que entendem a exigência de prestar contas de seu trabalho à sociedade como perda de autonomia e autoridade, além da dificuldade metodológica e do custo envolvido neste tipo de avaliação (MARTIN, 2011; HOLBROOK; FRODEMAN, 2011).

Bornmann e Haunschild (2017) chamam atenção para o fato de que ainda não se chegou a um consenso sobre a definição de impacto social da pesquisa, e de que as métricas alternativas disponíveis (altmetrics[16]) possuem ainda um conjunto de limitações grande demais para serem utilizadas de forma confiável.

Apesar da popularização desses indicadores, a avaliação da pesquisa baseada no desempenho individual dos cientistas não é recomendada pelos teóricos, já que atividades de pesquisa são conduzidas por grupos, estes seriam a unidade ideal de avaliação de desempenho (Hicks, 2012, p.254).

Vale ressaltar também críticas a esse modelo de avaliação que privilegia o fomento da pesquisa de excelência em vez da pesquisa "normal"[17]. Vessuri, Guédon & Cetto (2013) chamam atenção para o risco de se privilegiar um conjunto limitado de pesquisadores experientes e de determinadas áreas do conhecimento, promovendo a estagnação ou o declínio da qualidade geral da pesquisa realizada. Tendo em vista que a dinâmica do processo de inovação de cada país é diferente, dificuldade adicional coloca-se para os países latino-americanos que buscam adotar esse modelo de avaliação da pesquisa baseado em prestígio e assentado nos mesmos princípios e instrumentos dos países desenvolvidos.

O fato é que, ao direcionar a pesquisa para a publicação em periódicos indexados internacionalmente com ênfase na publicação em coautoria internacional, os países periféricos tendem a reproduzir a agenda de pesquisa dos países avançados, deixando de lado as especificidades locais e regionais – o que, entre outros motivos, se deve à baixa cobertura de periódicos latino-americanos nas bases, à falta de representantes nos comitês de avaliação, à ausência de linhas de pesquisa que abordem assuntos localmente relevantes e à barreira do idioma (VESSURI; GUÉDON; CETTO, 2013).

[16] Altmetrics fornecem a mensuração de atividade de pesquisa em ambiente online, como as redes sociais. As métricas são baseadas na contabilização de tweets, postagens, compartilhamento, menções, número de downloads, etc.

[17] Termo utilizado por Thomas Khun para descrever as atividades científicas que ocorrem sob um determinado paradigma num campo de conhecimento já estabelecido e que se opõe a produção de conhecimento novo fruto de revolução paradigmática na ciência.

[i] Essa perspectiva relaciona-se, no campo da epistemologia, com o paradigma construtivista que, ao abrir a "caixa-preta" da avaliação, confronta sua pretensa neutralidade, tendo em vista que noções como qualidade e padrões de desempenho são construções sociais. Ademais, o monitoramento desses padrões possui consequências práticas ao impor um tipo de racionalidade e modelar comportamentos.

[ii] Como exemplo, os autores citam a classificação de crimes em agências policiais. Em vez de os policiais serem avaliados com base no nível de segurança pública, eles costumam ser avaliados por sua eficiência na resolução de crimes. Ocorre que os casos criminais de difícil solução acabam considerados infundados e ficam de fora da contabilização estatística, o que leva os agentes a voltarem sua atenção para crimes de fácil resolução (BOHTE; MEIER, 2000, p. 3).

[iii] Quanto aos efeitos práticos ou constitutivos da avaliação, cabe destacar que, ao avaliar uma prática, você invariavelmente a muda (DAHLER-LARSEN, 2015). Martin (2011, p. 250) observa o paralelismo desta afirmação com o "efeito Hawthorne" no campo da administração. No experimento realizado na fábrica da Western Electric (EUA) no início do século XX, foi constatada uma mudança significativa do desempenho dos trabalhadores a partir da mera presença de pesquisadores observando sua execução.

[iv] Vale aqui trazer a diferenciação proposta por Scriven (1966) apud SHADISH JR.; COOK; LEVITON, (1991, p. 59) entre formative e summative evaluation na avaliação de programas educacionais. Na summative evaluation, julga-se o mérito do programa olhando os seus efeitos relevantes. Este julgamento geralmente utiliza-se de referência normativa, ou seja, se dá por meio da comparação com semelhantes. A formative evaluation é a avaliação que se utiliza de critérios iguais, e cujo resultado influencia decisões imediatas que visam melhorar o que ou quem está sendo avaliado.

[v] O contrato social de Vannevar Bush promoveu a política de investimentos maciços pelo governo dos Estados Unidos em ciência e tecnologia a partir do fim da Segunda Guerra Mundial, apoiado na premissa de que bastava investir em ciência na universidade para que esse conhecimento fosse convertido em desenvolvimento. Este modelo esgota-se frente à contestação de tal premissa e à necessidade dos governos em reduzir custos com pesquisa.

[vi] Na maior parte dos países, a avaliação da pesquisa não está diretamente atrelada ao fomento. O funcionamento rotineiro das atividades de formação para pesquisa é garantido por fundos de fomento (block grants) baseados em diferentes critérios de distribuição, como tamanho e quantidade de alunos, que pode ou não utilizar indicadores de atividades de pesquisa.

[vii] Quanto a indicadores de input, a autora cita métricas que buscam mostrar a habilidade dos programas em atrair recursos externos, estudantes e docentes/pesquisadores. Os indicadores de processo envolvem a mensuração de atividades de seminários e conferências, com destaque para composição em mesas de discussão e reuniões/visitas de pesquisadores internacionais. Quanto a indicadores de estrutura: número de docentes/pesquisadores, número de estudantes de doutorado, número de colaborações, grau de reputação e estima e infraestrutura e facilidades de pesquisa. Quanto aos indicadores de resultado estão as publicações, produtos não bibliográficos, número de alunos concluintes de doutorado em relação ao número de matriculados e visibilidade da pesquisa na sociedade, na mídia, etc. Quanto aos indicadores de efeito, estão as citações, prêmios e menções honrosas, empregabilidade dos egressos, transferência de conhecimento (patentes), consultorias e contratos externos.

[viii] O cálculo do fator de impacto envolve a soma de todas as citações recebidas pelo periódico no ano corrente (referente a artigos publicados nos últimos dois anos) dividida pelo número de artigos publicados nos dois anos anteriores. Em função das críticas que consideram o período de dois anos muito curto para o comportamento de publicação e citação de algumas áreas do conhecimento, algumas bases de dados fornecem o FI com janela de citação de 5 anos (WALTMAN, 2016)
O índice h infere impacto e produtividade dos pesquisadores. Seu cálculo pode ser explicitado a partir do seguinte exemplo: um índice h 15 corresponde a um pesquisador que publicou 15 trabalhos que receberam pelo menos 15 citações cada.

[ix] Quanto ao comportamento de citação, Bornmann & Haunschild (2017) citam o trabalho realizado por MacRoberts & MacRoberts entre os anos 1980 e 1990, que identificou três padrões de citação: publicações usadas que não são citadas ou raramente citadas; publicações citadas por meio de fontes secundárias; e aquelas que foram creditadas todas as vezes em que foram usadas.

REFERÊNCIAS

ADAMS, J. The fourth age of research. 30 maio de 2013. Nature, London, v. 497, p. 557-560, 30 maio 2013.

BECK, U. Risk society, towards a new modernity. Londres: Sage, 1992.

BOHTE, J.; MEIER, K. J. Goal displacement: assessing the motivation for organizational cheating. Public Administration Review, New Jersey, v. 60, n. 2, p. 173-182, mar./abr. 2000.

BORNMANN, L.; HAUNSCHILD, R. Does evaluative scientometrics lose its mainfocus on scientific quality by the new orientation towards societal impact? Scientometrics, Dordrecht, v. 110, n. 2, p. 937-943, 2017.

CARDEN, F.; ALKIN, M. C. Evaluation roots: an international perspective. Journal of Multidisciplinary Evaluation, Kalamazoo, v. 8, n. 17, p. 102-118, jan. 2012.

CHELIMSKY, E. The purposes of evaluation in a democratic society. In: SHAW, I. F.; GREENE, J. C., MARK, M. M. The SAGE handbook of evaluation. Thousand Oaks: SAGE publications, 2006. p. 34-55. Disponível em: <https://bit.ly/2uQcPpd>. Acesso em: 5 abr. 2018.

DAHLER-LARSEN, P. Evaluation as a situational or a universal good? Why evaluability assessment for evaluation systems is a good idea, what it might look like in practice, and why it is not fashionable. Scandinavian Journal of Public Administration, Götemburg, v. 16, n. 3, p. 29-46, 2012.

_____. The evaluation society: critique, contestability and skepticism. Spazio Filosofico, Torino, v. 13, p. 21-36, dez. 2015.

DAVYT, A.; VELHO, L. A avaliação da ciência e a revisão por pares: passado e presente: como será o futuro? História, ciências, saúde-Manguinhos, Rio de Janeiro, v. 7, n. 1, p. 93-116, mar/jun. 2000.

DEPARTMENT FOR BUSINESS, INNOVATION AND SKILLS. Teaching excellence framework: technical consultation for year two. London: BIS, 2016. Disponível em: <https://bit.ly/28P9n32>. Acesso em: 5 abr. 2018.

DONALDSON, S. I.; LIPSEY, M. W. Roles for theory in contemporary evaluation practice: developing practical knowledge. In: SHAW, I. F.; GREENE, J. C., MARK, M. M. The SAGE handbook of evaluation. Thousand Oaks: SAGE publications, 2006.

GIL, A. C. Métodos e técnicas de pesquisa social. São Paulo: Ed. Atlas, 2008.

HALL, R. Why should I care about the teaching excellence framework? – explainer. The Guardian, London, 9 jun. 2017. Higher Education Network. Disponível em: <https://bit.ly/2tMX6St>. Acesso em: 5 abr. 2018.

HANSEN, H.F. Performance indicators used in performance-based research funding systems. In: ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. Performance-based funding for public research in tertiary education institutions: workshop proceedings. Paris: OECD Publishing, 2010. p. 53-84.

HICKS, R. Performance-based university research funding systems. Research Policy, Amsterdam, v. 41, n. 2, p. 251-261, mar. 2012.

HOLBROOK, J. B.; FRODEMAN, R. Peer review and the ex ante assessment of societal impacts, Research Evaluation, v 20, n 3, p. 239–246, Sept. 2011.

LEEUW, F. L; FURUBO, J.-E. Evaluation systems: what are they and why study them? SAGE Journals, Thousand Oaks, v. 14, n. 2, p. 157-169, 2008.

LINK, A. N.; VONORTAS, N. S. Introduction to the handbook. In: LINK; A. N.; VONORTAS, N. S. (Eds.). Handbook on the theory and practice of program evaluation. Cheltenham: Ed. Edward Elgar, 2013. p. 1-11.LOET L.; HENRY E.; Emergence of a Triple Helix of university—industry—government relations, Science and Public Policy, v 23, n 5, p. 279–286, Oct. 1996. MARCO ESPAÑOL DE CUALIFICACIONES DE EDUCACIÓN SUPERIOR. Verification of compatibility of MECES (the Spanish qualifications framework for higher education) with the framework for qualifications of the European higher education area (FQ-EHEA). Madrid: MECES, nov. 2014. Disponível em: <https://bit.ly/2GGwDgw>. Acesso em: 5 abr. 2018.

MARTIN, B. R. The research excellence framework and the "impact agenda": are we creating a Frankenstein monster? Research Evaluation, Oxford, v. 20, n. 3, p. 247-254, 2011.

MCLAUGHLIN, J. A.; JORDAN, G. B. Logic models: a tool for telling your program's performance story. Evaluation and Program Planning, Amsterdam, v. 22, n. 1, p. 65-72, 1999.

MERTENS, D. M. Philosophical assumptions and program evaluation. Spazio Filosofico, Torino, v. 13, p. 75-85, fev. 2015.

MEYER, M. W.; GUPTA, V. The performance paradox. In: STAW, B. M.; CUMMINGS, L. L. Research in organizational behavior. Greenwich: JAI Press, 1994. v. 16. p. 309-369.

POWER, M. The audit explosion. 3. ed. London: Demos, 1999.

_____. The theory of the audit explosion. In: FERLIE, E.; LYNN, L. E; POLLITT, C. (Eds.) The Oxford Handbook of Public Management. Oxford: Oxford University Press, 2005. p. 326-344.

RAMOS, M. Y.; VELHO, L. Formação de doutores no brasil: o esgotamento do modelo vigente frente aos desafios colocados pela emergência do sistema global de ciência. Avaliação (Campinas), Sorocaba, v. 18, n. 1, p. 219-246, mar. 2013.

REBORA, G.; TURRI, M. The UK and Italian research assessment exercises face to face. Research Policy, Amsterdam, v. 42, n. 9, p. 1657-1666, 2013.

REINHARDT, C. S.; COOK, T. D. Beyond qualitative versus quantitative methods. In: COOK, T. D.; REINHARDT, C. S. (Eds.) Qualitative and quantitative methods in evaluation research. Beverly Hills: SAGE publications, 1979. p. 7-32.

RUEGG, R.; FELLER, I. A toolkit for evaluating public R&D investment models,

methods, and findings from ATP's first decade. Collingdale: Diane publishing company, 2003.

ROSSI, P. H.; LIPSEY, M. W; FREEMAN, H. E. Evaluation: a systematic approach. 6. ed. Thousand Oaks: SAGE Publications, 1998.

SALLES FILHO, S. Quanto vale o investimento em ciência, tecnologia e inovação? ComCiência, Campinas, n. 129, 2011. Disponível em: <https://bit.ly/2GAzzPF>. Acesso em: 5 abr. 2018.

SANDERSON, I. Performance, management, evaluation and learning in 'modern' local government. Public Administration, Hoboken, v. 79, n. 2, p. 297-313, 2001.

SHADISH JR, W. R.; COOK, T. D; LEVITON, L. C. Foundations of program evaluation: theories of practice. London: SAGE Publications, 1991.

THIEL, S. V.; LEEUW, F. L. The performance paradox in the public sector. Public Performance & Management Review, Thousand Oaks, v. 25, n. 3, p. 267-281, 2002.

VELHO, L. Conceitos de ciência e a política científica, tecnológica e de inovação. Sociologias, Porto Alegre, v. 13, n. 26, p. 128-153, 2011.

VESSURI, H.; GUÉDON, J.-C.; CETTO, A. M. Excellence or quality? Impact of the current competition regime on science and scientific publishing in Latin America and its implications for development. Current Sociology, Thousand Oaks, v. 62, n. 5, p. 647-665, 2013.

WALTMAN, L. A review of the literature on citation impact indicators. Journal of Informetrics, Amsterdam, v. 10, n. 2, p. 365-391, 2016.

WANG, I; VUOLANTO, P.; MUHONEN, R. Bibliometrics in the research assessment exercise reports of Finnish universities and the relevant international perspectives. Tampere: Research centre for knowledge, science, technology and innovation studies school of social sciences and humanities, 2014. Disponível em: <https://bit.ly/2EmDAkS>. Acesso em: 5 abr. 2018.

WEISS, C. H. Planning the evaluation. In: WEISS, C. H. Evaluation: methods for studying programs and policies. 2. ed. Upper Saddle River: Prentice Hall, 1998.

WILSDON, J., et al. The metric tide: report of the independent review of the role of metrics in research assessment and management. Stoke Gifford: Higher education funding council for England, 2015.

CAPES

www.capes.gov.br