Avaliação online feita por pares, auto avaliação e avaliação pelo professor na perspectiva de Bloom

Ribeirão Preto – SP – Abril/2015

José Dutra de Oliveira Neto – USP – dutra@usp.br

Luiz Antonio Titton – USP – titton@usp.br

Investigação Científica (IC)

Educação Superior

Tecnologia Educacional

Relatório de Estudo Concluído

RESUMO

A literatura sobre educação apoia a crença de que habilidades de ordem superior são essenciais na capacitação dos alunos para competir globalmente dentro da profissão contábil. Entre as alternativas de avaliação em ambientes virtuais de aprendizagem (AVA), auto avaliações e por pares são opções implementáveis com baixo custo de recursos materiais e humanos. Algumas inconsistências constatadas podem ser devidas à falta de estrutura teórica nas análises. O objetivo deste estudo é comparar os métodos de avaliação usando taxonomia do domínio cognitivo de Bloom. De uma amostra de 98 alunos de graduação em contabilidade, matriculados numa disciplina de metodologia de pesquisa, foram coletados dados usando diferentes métodos de avaliação em diferentes níveis cognitivos da taxonomia de Bloom com uso do AVA gratuito MOODLE. Os resultados mostram alinhamento estatístico entre os métodos de avaliação. Surgem diferenças no nível superior de Bloom ao agrupar tarefas por essa taxonomia. Estes resultados podem ajudar a explicar as divergências pesquisas anteriores realizados sem estrutura teórica de embasamento e colaboram para novas estratégias para a avaliação, bem como ensino-aprendizagem, especialmente em cursos com um número elevado de estudantes online e potencial para o desenvolvimento de habilidades de ordem superior a partir da experiência de avaliação pelos pares.

Palavras-chave: Avaliação online; rubricas; avaliação por pares; auto avaliação; Bloom

Introdução

A avaliação é um componente de formação e educação que envolve planejamento, discussão, construção de consenso, reflexão, medição, análise e melhoria com base em um objetivo de aprendizagem (BUZZETTO-MORE; ALADE, 2006). A relevância da avaliação emerge da necessidade de fornecer informações de retorno a partir do ensino que são exigidas pelos alunos, pela escola e pela sociedade como parte do processo de ensino e aprendizagem.

Os esforços para melhorar a qualidade da avaliação consomem muito tempo e recursos. Por esta razão, o uso de instrumentos de avaliação que podem promover o pensamento crítico e aprendizagem, enquanto os alunos são avaliados, estão ganhando interesse para serem utilizados nos ambientes virtuais de aprendizagem (AVA). Em MOOCs, o processo de avaliação, crucial em um plano de negócios bem-sucedido, requer o uso intensivo de recursos humanos que o torna muito caro e longo. À medida que cresce o número estudantes, o processo de avaliação se torna um elemento crítico na viabilidade financeira do processo.

Entre os instrumentos disponíveis, a adoção de rubricas se apresenta como a maneira mais rápida e eficiente para avaliar o trabalho dos alunos. Quando a concepção das rubricas está em conformidade com as recomendações da literatura também podem ser ferramentas de ensino que oferecem suporte a aprendizagem dos alunos e no desenvolvimento de habilidades de pensamento sofisticados. Rubricas são particularmente adequadas para *e-learning.* Em vez de escrever comentários repetitivos para estudantes, surge o aumento da qualidade quando as rubricas incorporam esses comentários na forma de padrão desejado, compondo o retorno.

Existem dois tipos principais de rubricas: avaliativa e instrucional (ANDRADE, 2000). Rubricas avaliativas são normalmente utilizadas quando um juízo de qualidade é necessário. Elas são desenvolvidas pelos avaliadores para orientar a análise dos esforços do aluno em um esquema pré-definido que transforma subjetividade e envolve esforços para uma avaliação mais objetiva (ISAACSON; STACY, 2009; MOSKAL, 2000). As rubricas instrucionais adicionam às avaliativas a finalidade de dar retorno sobre o trabalho em

andamento, bem como dar avaliações detalhadas sobre o produto final (ANDRADE, 2000).

Desde que uma alteração na situação do aluno é esperada, supõe-se que o retorno é fornecido em um contexto que permita a redução da deficiência (TARAS, 2005).Portanto, o *feedback* dado requer meios para reduzir a discrepância entre o que é produzido e que é desejado. A premissa fundamental é que os alunos se tornam capazes de desenvolver a capacidade de avaliar o seu trabalho e precisam ter acesso a um alto padrão de qualidade reconhecida para efeitos de comparação. Este argumento implica no desenvolvimento de habilidades de ordem superior por meio de experiências de avaliação autênticas (SADLER-SMITH, 2007). A rubricas formativas suprem essa necessidade, pois utilizam a exposição do *feedback* como parte relevante do critério de avaliação. Neste cenário, a auto avaliação permite ao aluno comparar o produto com um padrão como parte da tarefa - daí resulta a aprendizagem pelo processo de avaliação. Como este mesmo processo é repetido para avaliar tarefas pelos pares também, o processo expande a experiência para mais cada nova rodada de avaliação e aprendizagem em tarefas repetidas mesmo nível do aluno e, assim, a comparação com o padrão de qualidade adotado é acessado mais uma vez.

Observou-se também que o envolvimento fornecido pela revisão por pares desenvolve mais habilidades cognitivas objetivadas que são mais difíceis em uma auto avaliação (CHANG; TSENG; LOU, 2012). Uma simples investigação dos resultados da auto avaliações e por pares, em comparação com o avaliações dos professores, produziu resultados mostrando que há alguma consistência em graus variados (CHANG; LIANG; CHEN, 2013; CHANG; TSENG; LOU, 2012; CHEN, 2010; NAPOLES, 2008a). No entanto, observou-se possíveis problemas relativos ao anonimato e ao tipo de tarefa. Esses dois fatores podem induzir o pesquisador obter conclusões imprecisas.

Uma pesquisa realizada com estudantes que trabalham sobre o mesmo assunto em sala de aula mostrou evidências de diferenças entre os três tipos de avaliação. Verificou-se que a auto avaliações as do professor são consistentes entre si, mas a avaliação por pares mostrou diferenças significativas com relação aos outros dois métodos. (CHANG; LIANG; CHEN, 2013; CHANG; TSENG; LOU, 2012) Um estudo posterior testou uma nova

configuração utilizando o mesmo suporte baseado em web e descobriu que neste novo ambiente os três métodos convergiram para os mesmos resultados (CHANG; TSENG; LOU, 2012). Nestes dois estudos, as avaliações por pares não foram anônimas, uma indicação possível de que a configuração pode ser um elemento a influenciar a convergência dos resultados obtidos. O ambiente colaborativo pode levantar questões sobre a avaliação dos colegas especificamente quanto ao anonimato (CHEN, 2010) pois os alunos podem mostrar-se divididos sobre o impacto entre anonimato e amizade (THEISING; WU; HECK SHEEHAN, 2014). Ao realizar distinção entre tarefas associadas com habilidades de alto nível ou de baixo nível, foi observado maior acurácia nas auto avaliações nas tarefas de desse último tipo em contraste com as primeiras (SADLER; GOOD, 2006).

A análise dos resultados de anteriores comparativos estudos sobre os três tipos de avaliação – auto avaliação, por pares e a do professor - sem considerar o anonimato dos colegas e agrupamento de tarefas pode explicar alguns desses resultados divergentes em estudos anteriores. Uma vez que existe uma diferença aparente entre os resultados associados aos tipos de tarefa, vem a necessidade de usar uma estrutura como marco teórico para fornecer uma base para analisar esta questão. As abordagens da compreensão e raciocínio referem-se a taxonomia proposta por Bloom, que pode ser usado como um modelo para identificar se existem diferenças entre os três métodos de avaliação (auto, por pares e pelo professor) para cada questão e considerando também as diferentes dimensões cognitivas.

A adoção dessa taxonomia não é nova para a análise de avaliações (ANDERSON; KRATHWOHL, 2001; BUZZETTO-MORE; ALADE, 2006; COLLIER-REED, 2011; KARAMUSTAFAOĞLU et al., 2003). Os níveis originais da taxonomia de Bloom são o conhecimento, compreensão, aplicação, análise, síntese e avaliação. Os primeiros são chamados de nível inferior e crescem para os últimos aqueles que são chamados de níveis mais elevados (HUITT, 2011). O nível de taxonomia de Bloom modificado apresentado por Collier-Reed (2011) tem três níveis: Nível 1 com o conhecimento e de coleta de informações; Nível 2 com compreensão, aplicação, entendimento e ser capaz de interpretar os dados; e o mais alto, Nível 3 com a resolução de problemas e

uso do conhecimento e compreensão em novas circunstâncias (PALMER; DEVITT, 2007).

O objetivo deste trabalho é fazer uma análise comparativa da diferença entre as avaliações realizadas pelos estudantes em auto avaliação, dos seus pares e do professor. Isso por meio de avaliação anônima considerando a taxonomia do domínio cognitivo de Bloom. Assim, utilizou-se um quadro teórico conceitual para categorizar as tarefas de acordo com as habilidades representadas neste contexto. As questões de pesquisa são sobre a existência diferenças significativas entre os três métodos de avaliação por tarefa, independente ou não de gênero; e, se ocorrem diferenças nessas abordagens ao se considerar a taxonomia de Bloom, independente ou não de gênero.

Métodos

Os participantes foram 98 alunos de graduação em Ciências Contábeis de uma universidade pública no Brasil. O curso de metodologia de pesquisa apresenta aos alunos na tarefa de escrever artigos científicos e relatórios de pesquisa. O gênero foi testado em métodos de avaliação por pares, que foram aleatoriamente designados pelo Ambiente Virtual de Aprendizagem utilizado (Moodle). O Sistema de Gestão de Aprendizagem (Moodle) foi utilizado para avaliar os três métodos de avaliação, utilizando o módulo chamado *Workshop*. Nesta abordagem particular, os alunos avaliam a sua atividade em comparação com um critério entendido como correto sendo que o mesmo critério é utilizado pelo professor para avaliar a tarefa. Isto dá ao estudante o *feedback* sobre o que se espera com o potencial de promover ou melhorar a aprendizagem (JONSSON; SVINGBY, 2007). O estudo foi realizado por meio de um treinamento inicial com a finalidade de treinar os alunos no processo de preparação, submissão, auto avaliação e pares de forma anônima comparando com um padrão pré-estabelecido. Ao final de 8 semanas em que receberam aulas normais sobre artigos acadêmicos, incluindo uma estrutura padrão para resumos, os alunos receberam um artigo completo e o resumo separadamente. A seções de resumo consideradas como padrão são, nessa ordem, contexto, lacuna, objetivo, materiais e métodos, resultados/discussão e conclusão. O resumo estava em ordem diferente da indicada como o padrão desejado e,

nele, faltavam duas seções - contexto e lacuna. Foram indicadas como tarefas marcar as seções nas cores padronizadas no resumo, usando o editor de textos disponível em sala de aula. A seguir, os alunos deveriam colocar as seções na ordem padrão previamente definida. A seguir, deveriam identificar as seções faltantes e escrever o texto para completa-las. Assim, tem-se que as duas primeiras atividades (classificar e ordenar) correspondem ao nível 1 da Taxonomia de Bloom (conhecimento) e as outras três (identificar partes faltantes e produzir o texto) correspondem ao nível 2 da taxonomia (compreensão). Depois de enviar a tarefa online, os resumos produzidos foram avaliados *online* anonimamente por outros dois colegas anônimos, além da avaliação de seu próprio trabalho em configurações denominadas *double blind peer* e *self-assessment*, respectivamente. O professor realizou a mesma avaliação para todos os alunos utilizando a mesma rubrica. Apesar dos avisos de não identificar o seu próprio trabalho com suas informações pessoais, quatro estudantes deixaram indicações de seus nomes de autoria nos trabalhos. Este valor foi considerado aceitável e sem implicações para esta pesquisa. As rubricas foram criadas com base na literatura usando cinco tarefas com o valor para cada tarefa que varia de um a cinco. A pontuação mínima é de 5 e o máximo é 22. Para cada uma das cinco tarefas as rubricas tiveram valor crescente de 1 (não fez) a 5 (fez corretamente).

Resultados

Os dados foram analisados por meio de Análise de variância (ANOVA). No caso de diferenças significativas fossem identificados, foi realizado o teste de Bonferroni *post*. Ao se verificar se existem diferenças significativas entre os três métodos de avaliação por tarefa, constatou-se que a avaliação do professor é significativamente menor que as avaliações feitas pelos alunos. O anonimato e a avaliação antes de receber o *feedback* dos pares pode ter contribuído para que a avaliação dos alunos fosse maior que a do professor. Esses resultados estão em oposição à pesquisa sem a avaliação por pares duplo-cego, em que a avaliação dos alunos foi realizada depois do *feedback* (CHANG; LIANG; CHEN, 2013; NAPOLES, 2008b).

Com relação a diferenças significativas entre os três métodos de avaliação por tarefa e por sexo também não se encontrou interações estatisticamente significativas. A diferença encontrada entre os métodos de avaliação é a mesma para as três combinações de gênero. Sendo duas avaliações por pares, os avaliadores poderiam ocorrer como Masculino-Masculino (MM), Feminino-Masculino (FM) ou Feminino-Feminimo (FF). Na combinação de MM a avaliação por pares é significativamente menor do que a auto avaliação. Para FM e FF, as avaliações por pares e auto avaliações não têm diferenças significativas. Não foi encontrado em pesquisas anteriores diferenças de gênero na avaliação de diferentes tarefas.

Ao se considerar as tarefas agrupadas pela taxonomia de Bloom, a avaliação pelo professor é significativamente menor do que avaliação por pares e auto avaliação. Nas avaliações por pares e auto avaliações não houve diferença significativa para a tarefa de nível mais baixo e a avaliação por pares é significativamente menor do que a auto avaliação para a tarefa de nível superior. A diferença deve-se ao nível da tarefa. O anonimato e a avaliação antes de receber o *feedback* de pares ou professor podem ter contribuído para o nível mais elevado de entre avaliações por pares e auto avaliações em relação à avaliação pelo professor. Estes resultados estão em oposição à pesquisa anterior sem avaliação duplo-cego em que a avaliação dos alunos foi feita depois de *feedback* (CHANG; LIANG; CHEN, 2013; NAPOLES, 2008b).

Quanto ao aspecto de diferenças significativas entre os três métodos de avaliação por nível de taxonomia de Bloom 1 e 2, por sexo, não existe interação significativa, por isso a diferença encontrada entre os métodos de avaliação é a mesma para as três combinações de gênero para o nível mais baixo e mais elevado de tarefas. Não foram encontradas pesquisas anteriores sobre as diferenças de gênero na avaliação de diferentes tarefas.

Esses achados não são consistentes com pesquisas anteriores sem avaliação duplo-cego, onde a avaliação dos estudantes foi feita depois de *feedback* (CHANG; LIANG; CHEN, 2013; NAPOLES, 2008b).

Quando considerando as dimensões de Bloom ficou demonstrado que a dimensão do Conhecimento não segue a característica geral de que auto avaliações são mais elevadas do que as feitas pelos pares.

Há evidência de favoritismo em auto avaliações dos estudantes se comparadas com as do professor. No entanto, o favorecimento que é convergente na dimensão do conhecimento apresenta-se de forma diferente na dimensão da compreensão da taxonomia de Bloom. Isto leva a considerar a hipótese de que o favoritismo é afetado pelo tipo de tarefa, corroborando o que disse Cho, Lee & Jonassen (2011).

Poderia ser questionado que Cho, Lee & Jonassen (2011) realizaram sua pesquisa com outro procedimento. Entretanto, os resultados obtidos são muito semelhantes aos apresentados pelos pesquisadores mesmo com diferenças metodológicas, permitindo confirmar que o resultado obtido é afetado somente pelo tipo de tarefa.

5. Conclusão e Implicação

Pode-se dizer que não há evidência de convergência entre auto avaliações e as por pares em comparação com as feitas pelo professor. No entanto, notou-se que na dimensão da compreensão ocorreu falta de homogeneidade nos resultados das rubricas dos alunos. As tarefas dessa dimensão são idênticas, variando apenas o objeto de trabalho, ou seja, qual seção do resumo que foi trabalhado pelos alunos. Assim, na dimensão de Bloom da compreensão pode ocorrer uma distorção de convergência, e, portanto, a rigor, as avaliações feitas pelos alunos nessa dimensão podem não ter a mesma configuração que existe na dimensão do conhecimento. Por outro lado, há convergência relativamente uniforme entre auto avaliações e as por pares sobre a dimensão da Bloom de conhecimento demonstrando a tendência para maior favorecimento em que o aluno avalia a si mesmo do que a feita por pares e por tanto superior a feita pelo professor como já identificado em pesquisa anterior. Essas diferenças de convergência observadas entre as duas dimensões não tinham sido previamente identificadas nas pesquisas anteriores citadas. Os resultados reportados corroboram pesquisas anteriores em geral, mas a taxonomia de Bloom apresenta uma nova dimensão para interpretar os dados e permitiu verificar que, em atividades que envolvem a compreensão, as discrepâncias entre as avaliações feitas pelos alunos do seu próprio trabalho e recebidos por seus pares pode não apresentar um comportamento uniforme.

A inserção da taxonomia de Bloom para analisar a precisão das avaliações feitas pelos alunos avança nos conhecimentos em relação aos resultados anteriores pelo fato de mostrar que existem indícios de uma maior precisão nas avaliações em atividades que exigem um maior nível de habilidades cognitivas. O conjunto de resultados neste trabalho indica que a utilização de um modelo teórico de sustentação pode servir como uma base para pesquisas eficazes. A eficiência somativa foi confirmada na primeira dimensão cognitiva de Bloom (conhecimento) e os dados mostraram indícios que na segunda dimensão (compreensão) existem fatores que indicam a necessidade de mais pesquisas. Há uma clara indicação de que esses achados não estão diretamente ligados à forma como a rubrica foi projetada, mas a necessidade de produzi-las considerando um modelo teórico de sustentação não pode ser ignorada.

Referências

ANDERSON, L. W.; KRATHWOHL, D. R. **A Taxonomy for Learning, Teaching and Assessing - A revision of Bloom's taxonomy of educational objectives**. [s.l: s.n.].

ANDRADE, H. G. Using Rubrics to Promote Thinking and Learning. **Educational Leadership**, v. 57, n. 5, p. 13–18, 2000.

BUZZETTO-MORE, N. A.; ALADE, A. J. Best Practices in e-Assessment. **Journal of Information Techology**, v. 5, p. 251–269, 2006.

CHANG, C.-C.; LIANG, C.; CHEN, Y.-H. Is learner self-assessment reliable and valid in a Web-based portfolio environment for high school students? **Computers & Education**, v. 60, n. 1, p. 325–334, jan. 2013.

CHANG, C.-C.; TSENG, K.-H.; LOU, S.-J. A comparative analysis of the consistency and difference among teacher-assessment, student self-assessment and peer-assessment in a Web-based portfolio assessment environment for high school students. **Computers & Education**, v. 58, n. 1, p. 303–320, jan. 2012.

CHEN, C. The implementation and evaluation of a mobile self- and peer-assessment system. **Computers & Education**, v. 55, n. 1, p. 229–236, ago. 2010.

COLLIER-REED, B. (ED. . **Proceedings of the First Biennial Conference of the South African Society for Engineering Education Proceedings of the First Biennial Conference of the South African Society for Engineering Education**. [s.l: s.n.].
HUITT, W. Bloom et al.'s Taxonomy of the Cognitive Domain. **Educational Psychology Interactive**, 2011.

ISAACSON, J. J.; STACY, A. S. Rubrics for clinical evaluation : Objectifying the subjective experience. **Nurse Education in Practice**, v. 9, n. 2, p. 134–140, 2009.

JONSSON, A.; SVINGBY, G. The use of scoring rubrics : Reliability , validity and educational consequences. v. 2, p. 130–144, 2007.

KARAMUSTAFAOĞLU, S. et al. Analysis of Turkish High-School Chemistry-Examination Questions According To Bloom's Taxonomy. **Chemistry Education Research and Practice**, v. 4, n. 1, p. 25, 2003.

MOSKAL, B. M. Scoring Rubrics : What , When and How ? **Practical Assessment, Research & Evaluation**, v. 7, n. 3, p. 1–7, 2000.

NAPOLES, J. Relationships Among Instructor, Peer, and Self-Evaluations of Undergraduate Music Education Majors'Micro-Teaching Experiences. **Journal of Research in Music Education**, v. 56, n. 1, p. 82–91, 2008a.

NAPOLES, J. Relationships Among Instructor, Peer, and Self-Evaluations of Undergraduate Music Education Majors' Micro-Teaching Experiences. **Journal of Research in Music Education**, v. 56, n. 1, p. 82–91, 1 abr. 2008b.

PALMER, E. J.; DEVITT, P. G. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. **BMC medical education**, v. 7, p. 49, 2007.

SADLER, P. M.; GOOD, E. The Impact of Self- and Peer-Grading on Student Learning. **Educational Assessment2**, v. 11, n. 1, p. 1–31, 1 fev. 2006.

SADLER-SMITH, E. in Management Education. v. 6, n. 2, p. 186–205, 2007.

TARAS, M. Assessment – Summative and Formative – Some Theoretical Reflections. **British Journal of Educational Studies**, v. 53, n. 4, p. 466–478, 2005.

THEISING, K.; WU, K.; HECK SHEEHAN, A. Impact of peer assessment on student pharmacists' behaviors and self-confidence. **Currents in Pharmacy Teaching and Learning**, v. 6, n. 1, p. 10–14, jan. 2014.