

ENSINO DE METODOLOGIA DA CIÊNCIA: DEMONSTRAÇÃO DA MAIOR VALIDADE DA AVALIAÇÃO INDEPENDENTE EM RELAÇÃO À FEITA POR CONSENSO

Geraldina Porto WITTER*

RESUMO: Em várias situações o pesquisador usa juízes ou avaliadores distintos para medir algum objeto ou aspecto de um estímulo. No ensino de metodologia da ciência esta questão deve ser considerada. O objetivo do presente estudo foi tentar uma demonstração simples para ensinar este assunto de modo a: (1) verificar comparativamente a validade de dois procedimentos de avaliação com juízes (independente e consensual); (2) - comparar o comportamento de estudantes de psicologia e de biblioteconomia, e (3) correlacionar avaliações de dois objetos. Os resultados mostraram que a avaliação independente é mais precisa, que não houve diferença entre os grupos de sujeitos e que as correlações foram significantes.

PALAVRAS-CHAVE: Procedimento de ensino. Metodologia da Ciência. Avaliação independente. Avaliação por consenso.

Em trabalho anterior (WITTER, 1987, p. 147) foi enfatizada a importância de que o ensino de metodologia da ciência deveria começar sistematicamente já na graduação e ter também um caráter de prática efetiva no fazer ciência, devendo a programação "ênfatisar os vários tipos de modalidades de pesquisa, indo das pesquisas descritivas mais simples de levantamento até os mais complexos delineamentos experimentais". Nestes vários tipos o pesquisador frequentemente depara-se com a questão da avaliação particularmente quando ela implica em aspectos subjetivos, portanto mais discutíveis e sujeitos a controvérsias. Esta problemática é mais usual nas chamadas ciências humanas.

A frequência com que esta questão surge na pesquisa é suficiente para justificar a sua inclusão nos cursos de metodologia científica. De fato, é um aspecto relevante tanto no ensino quanto na transmissão do saber e do fazer ciência nas várias áreas do conhecimento, posto tratar da experiência de mensuração.

Em muitas circunstâncias, o pesquisador recorre a avaliações feitas por juízes para medir algum aspecto do problema que está estudando. Introduz

* Professora Livre Docente do Instituto de Psicologia da Universidade de São Paulo e da Pontifícia Universidade Católica de Campinas.

assim um controle externo que tem a possibilidade de reduzir um eventual ou mesmo sistemático erro de julgamento do próprio pesquisador ou de quando apenas uma pessoa faz a avaliação. A necessidade de recorrer a juízes pode advir da avaliação de aspectos mais simples e objetivos até aspectos os mais complexos e subjetivos. No primeiro caso, podem ser lembradas situações em que são avaliados: preservação de material didático, número de crianças envolvidas nas atividades acadêmicas em sala de aula, estado de preservação das coleções na biblioteca escolar, comportamento exploratório de alunos transferidos, distância entre dois pontos em espaço aberto ou fechado. No segundo polo podem ser citados: a qualidade da produção artística ou científica, as características de personalidade dos auxiliares de bibliotecas, atitudes de professores, ideologia inferida pela análise de discurso político ou pedagógico.

Há duas maneiras freqüentemente empregadas para fazer estas avaliações quando não se tem um instrumento mais preciso ou quando as respostas são abertas a muitas interpretações.

Uma destas maneiras implica em pedir a juízes independentes que avaliem ou categorizem o objeto de estudo e manifestem suas avaliações independentemente, isto é, sem que um conheça a avaliação feita pelo outro. Neste caso, após as avaliações independentes terem sido concluídas, há duas alternativas de ação: aplica-se uma das fórmulas matemáticas para verificar o índice de concordância entre os juízes, ou faz-se a média das avaliações independentes. A primeira opção é mais simples e segura, esperando-se encontrar índices iguais ou superiores a 75% ou correlações significantes para dar prosseguimento ao trabalho. Quando estes índices são alcançados as análises qualitativas e quantitativas podem ser feitas, com segurança, usando-se apenas os dados de um dos juízes. Quando o índice não é atingido deve-se rever os critérios de avaliação, reavaliar os elementos ou, se isto não for possível, fica-se com a alternativa de tirar a média da opinião dos juízes.

A outra maneira de trabalhar quando se está diante do problema de mensuração aqui enfocado requer que os juízes trabalhem conjuntamente, discutam a avaliação e emitam um parecer que seja resultado do consenso.

A primeira maneira (independente) tem maior sustentação estatística e se mostrado cientificamente mais válida que a segunda (consensual), embora, na prática profissional muitas vezes, esta última venha sendo muito empregada. A razão da insistência da indicação da primeira maneira como preferível decorre "do clássico pressuposto de que o valor esperado do erro componente de um conjunto de avaliações é zero, somente se aplicar a escores independentes. A discussão envolvida no procedimento consensual viola este pressuposto" (ROCKLIN, 1985, p.44).

Ao ensinar esta estratégia de controle científico, fundamental para garantir inclusive que o pesquisador mesmo inconscientemente não vize seus dados, muitas vezes o docente esbarra com preconceitos firmemente estabelecidos entre os alunos de que o consenso é melhor, mais democrático, etc. Demonstrar matemática ou estatisticamente que estão apenas sob o controle de seus preconceitos, pode ser particularmente complexo e mesmo difícil, especialmente se a

classe não tiver uma boa base de matemática e de estatística para acompanhar os desdobramentos das equações e análises matemáticas feitas pelo professor. Esta vivência tem sido uma constante ao longo dos quase 30 anos em que a autora vem ensinando metodologia da ciência, em cursos de graduação e de pós-graduação. Exercícios de demonstração em sala de aula têm sido úteis para conscientizar o aluno das variáveis em jogo por trás destes procedimentos e para a discussão de aspectos filosóficos e políticos envolvendo o trabalho do cientista.

Em 1985, Rocklin publicou um artigo em que relata uma demonstração feita em sala de aula que permite ao aluno detectar a maior validade da avaliação independente. Trabalhou pedindo que os alunos de um curso de psicologia, divididos em grupos de 3 a 4 estudantes e outros alunos atuando independentemente avaliassem o comprimento de um tubo (109 cm) e o peso de um livro (213 gramas) tendo verificado com eles, após os cálculos comparativos que o julgamento consensual ficava mais afastado da medida real que a avaliação independente.

Em 1988, estando lecionando metodologia da pesquisa em um curso de pós-graduação de Biblioteconomia e metodologia do ensino da psicologia em um curso de pós-graduação em Psicologia Clínica, a autora deste trabalho decidiu incluir a demonstração elaborada por Rocklin, com pequenas alterações, entre as estratégias de ensino usadas nas duas turmas, tendo ainda como objetivo educacional, para a segunda turma, apresentar uma forma de ensinar os conceitos em jogo. Isto permitiu a formulação dos seguintes objetivos para este estudo: (1) verificar a validade comparada de avaliação independente com a mensuração consensual e (2) comparar a avaliação dos alunos de biblioteconomia com os de psicologia e (3) correlacionar as avaliações de dois objetos da mesma natureza.

Método

Material: Foram utilizados materiais similares aos da pesquisa de Rocklin (1985): um tubo de papelão cinza de 112 cm tendo 5,5 cm de diâmetro, um livro com capa de cartolina (brochura) pesando 280 gramas e outro encadernado em percalina com peso de 780 gramas. Foram usadas fichas (meia página de papel almaço pautado) para os sujeitos lançarem suas respostas.

Sujeitos: Dois grupos de sujeitos adultos participaram do trabalho: (1) Grupo P foi constituído por 17 pós-graduandos de psicologia de uma universidade particular que frequentavam disciplina sobre metodologia do ensino desta matéria no curso superior, sendo três masculinos e 14 femininos; (2) Grupo B foi composto por oito elementos do curso de pós-graduação em biblioteconomia, da mesma instituição, só um do sexo masculino, trabalhando com metodologia da pesquisa.

Procedimento: A tarefa foi apresentada aos dois grupos como um exercício para demonstrar características da avaliação feita por consenso em comparação com avaliações feitas por juízes independentes. Em ambas as classes

foram os sujeitos divididos em dois sub-grupos distintos. No fundo da classe (em oposição à lousa) foram compostos subgrupos de quatro sujeitos para fazer avaliação por consenso. No Grupo P foram compostos três subgrupos para avaliação consensual, tendo três elementos cada, e no Grupo B só um (N=4), dado o número reduzido de sujeitos em classe. Nestas circunstâncias, no GP ficaram oito estudantes como juizes individuais independentes e no GB ficaram quatro elementos atuando neste papel.

Foi enfatizada a importância de, no caso da avaliação em grupo, discutirem o assunto antes de responderem e de, na avaliação individual garantirem o sigilo e a fidedignidade de sua opinião ao emitir a resposta.

Inicialmente foi apresentado o tubo, fazendo-se movimentos de oscilação para a direita e para a esquerda, sendo seguro pela professora-pesquisadora, com a mão esquerda, na altura do meio (55 cm). Em seguida, apoiou-o sobre a mesa, mantendo-o na posição vertical e à sua esquerda. Pediu que os grupos e os juizes individuais comesçassem a trabalhar estimando, sem uso de instrumento ou de contato direto com o objeto, qual era o comprimento do tubo (em centímetros). Quando todos completaram a tarefa pegou o Livro 1 (o mais leve) pelo dorso, movimentou-o para a direita e à esquerda e ficou com o mesmo à altura de sua cabeça. Pediu, então, que estimassem o peso do mesmo em gramas. Concluída a tarefa foi apresentado o Livro 2.

Imediatamente foi feita em classe uma primeira análise dos dados, tendo se verificado resultados semelhantes aos já relatados na literatura, como será apresentado aqui na parte de resultados. Seguiu-se uma breve discussão dos dados obtidos, dos métodos de avaliação comparados e das possibilidades de ensino do procedimento. Abriu-se um espaço para que se manifestassem também quanto à vivência que esta prática lhes propiciou. Em aula posterior, a autora apresentou à classe esse trabalho, o qual foi lido e teve a concordância consensual do grupo.

Resultados e Discussão

Na avaliação do tubo a média dos grupos de GP correspondeu a 117 cm e a média das avaliações individuais a 121 cm; no GB o grupo atribuiu o valor de 130 cm e a média dos juizes independentes foi de 120 cm. No GP os grupos afastaram-se da medida real por + 5 cm e os juizes individuais por + 9 cm, dados estes não consoantes com a literatura. No GB, conforme o esperado, o grupo afastou-se mais da medida real (+ 30 cm) do que os sujeitos que fizeram avaliações independentes (+ 8 cm) confirmando a literatura. Todavia fazendo uma análise do índice de acordo dos grupos entre si em relação à média obteve-se 91, 6% sendo esta média de 95,6% entre os juizes individuais (estes cálculos não puderam ser feitos para GB por só se dispor de um grupo) mostrando maior coerência de avaliação para os juizes independentes.

No que tange à avaliação do peso do Livro 1 a média de avaliação dos sub-grupos de GP foi de 233 gramas, com um afastamento de - 47

gramas do valor real, enquanto que a média dos juízes independentes foi de 324 gramas, com um afastamento de + 44 gramas. O índice de acordo no primeiro caso foi da ordem de 66,6% e no segundo de 100% confirmando-se os dados da literatura. No Grupo B reaplicou-se os resultados constantes da introdução sendo de 350 gramas a resposta do grupo, ou seja, com um afastamento de + 70 gramas, enquanto que os julgamentos individuais resultaram em 312 gramas, isto é, com um desvio de + 32 gramas.

Em relação ao Livro II foi constatada a avaliação média de 900 gramas entre os grupos de consenso do GP, com um índice de acordo de 70% e um desvio de 120 gramas, enquanto que, com um índice de acordo de 94% os juízes independentes avaliaram o livro como tendo exatamente seu peso real, isto é, 780 gramas. Na avaliação conduzida pelo sub-grupo de consenso do Grupo B obteve-se 550 gramas, que representa um erro de 230 gramas, sendo a média dos juízes independentes de 730 gramas, portanto com um distanciamento de apenas 50 gramas do valor real. Estes dados estão de acordo com o esperado e já registrado em pesquisas anteriores quanto à superioridade da avaliação por juízes independentes.

Para atender ao último objetivo deste trabalho foi feita a comparação entre as avaliações dos juízes independentes do GP com o GB. Dado o pouco número de sujeitos, as circunstâncias e características das respostas, recorreu-se ao teste U para amostras independentes, ao nível de significância de 0,05. Definiu-se H_0 : GP = GB e H_a : GP \neq GB (Siegel, 1956); portanto, teste bicaudal, posto não haver qualquer razão para direcionar o teste.

Os cálculos foram feitos em relação ao afastamento do julgamento do juiz e o valor real do objeto, tendo-se tomado o módulo da diferença, isto é, desconsiderado se o erro foi por superestimar ou subestimar o comprimento ou o peso. Desta forma considerou-se como melhor juiz o que menos se afastou, positiva ou negativamente, do valor real da dimensão de objeto que deveria ser avaliado.

No caso da avaliação do tubo ($n_1 = 4$ e $n_2 = 8$) obteve-se $U_0 = 7$ a que está associada a probabilidade de 0,08, portanto superior ao nível estabelecido (0,05) permitindo concluir pela não rejeição de H_0 , ou seja, não houve diferença de precisão de avaliação entre os dois grupos de sujeitos que atuaram como juízes independentes.

No que concerne ao Livro 1 os cálculos resultaram em $U_4 = 14$, associado a $p = 0,80$, permitindo concluir pela não rejeição da hipótese nula, ou seja, não houve diferença significativa entre os juízes independentes dos dois grupos no que diz respeito à avaliação do peso do objeto em questão. Em relação ao Livro 2 obteve-se resultado similar posto que $U_0 = 11$, $p = 0,92$, então H_0 não é rejeitada, ou seja os juízes do GP e GB não apresentaram diferença significativa de julgamento.

Os resultados do Teste U confirmaram que, independentemente da formação anterior dos sujeitos, grupos de juízes independentes tendem a fazer avaliações de peso e comprimento que não diferem significativamente entre si.

Os dados obtidos confirmam os relatados por Rocklin (1985), estão consoantes com a literatura sobre metodologia da ciência e apóiam a utilização de juízes independentes como critério de segurança em pesquisas científicas, especialmente quando há mensuração a nível nominal ou complexidade na categorização de avaliação usada.

As manifestações da grande maioria dos sujeitos (dois do GP não fizeram comentários verbais) foi favorável a esta prática como veículo para a aprendizagem dos conceitos envolvidos, tendo apresentado atitudes verbais positivas e entusiasmo na análise dos dados.

Esta mobilidade de prática permite ainda o estudo de aspectos psicofísicos da avaliação bem como a análise do erro de equação pessoal. Esta última sugestão é particularmente viável quando são usados dois objetos da mesma natureza como no presente trabalho. Registrou-se sujeitos que erram por igual número de gramas e na mesma direção (super ou subestimando o peso) ao avaliar os dois objetos, nos demais a variação foi pequena, tendo se alcançado uma correlação de 0,98, significativa ($r_c = 0,53$) ao nível de 0,05. Este aspecto também pode ser planejado como elemento a ser explorado pelo docente-pesquisador nas suas aulas usando esta prática de ensino.

Pode-se concluir pela validade do procedimento de ensino sugerido por Rocklin (1985), isto é, que os resultados confirmam a superioridade de julgamentos feitos por juízes independentes em relação à avaliação feita em grupo (por consenso), bem como, que isto independe de se estar trabalhando com pessoas com formação em Psicologia ou em Biblioteconomia, possivelmente por se tratar de conceitos e aspectos basilares em Ciência.

Concorda-se com Rocklin (1985) em que possivelmente esta demonstração seja particularmente efetiva por fornecer uma vivência concreta e comparativa dos dois procedimentos; a superioridade da avaliação independente é geralmente contra a intuição de modo que o resultado da demonstração foi surpreendente para os estudantes, o que pode auxiliar na retenção da informação e na conscientização dos muitos aspectos envolvidos na questão.

SUMMARY: In a variety of situations, the researchers use multiple raters or judges to measure some object or aspect of a stimulus. In the teaching of science methodology the question must be considered. The aim of the present study was to test a simple demonstration to teach that issue in order (1) to verify comparatively the validity of two procedures of evaluation with judges (independent and consensual); (2) to compare the behavior of students of Librarian Science and Psychology and (3) to study the correlation of the evaluation of two objects. The results showed that independent procedures is more accurate, there was no difference between the two groups of subjects and the correlation was significant. The demonstration appears to be particularly effective.

KEY-WORDS: Teaching procedures. Science methodology. Independent evaluation. Consensual evaluation.

REFERÊNCIAS BIBLIOGRÁFICAS

- ROCKLIN, Thomas – Independent rating are more valid than group consensus: a classroom demonstration. *Teaching Psychology*, 1985, 12 (1): 44-45.
- SIEGEL, Sidney – *Nonparametric Statistics for the Behavioral Sciences* – McGraw-Hill Book Company, Inc., N.Y., 1956.
- WITTER, Geraldina P. – O ensino de metodologia científica em biblioteconomia: algumas considerações. *Ciência da Informação*, 1987, 10 (2): 145-150.

(Recebido para publicação em 26-01-89
e liberado em 09-05-89)