Comparação de Métodos para Definição de Linhas de Corte para Testes baseados em Critérios

Lilia Carolina C. da Costa
Jonatas Silva do Espirito Santo
Universidade Federal da Bahia

O estabelecimento da linha de corte de um teste se refere ao processo pelo qual uma nota para a aprovação dos examinados é determinada. Esta é uma área de estudo da psicometria conhecida como *Standard Setting*. No caso de testes baseados em norma, i.e., quando se deseja comparar os examinados e selecionar os melhores, a linha de corte é usualmente definida como o número de vagas, como por exemplo, vestibular ou concurso. Por outro lado, quando se está trabalhando com testes baseados em critérios, i.e, quando se deseja, por exemplo, conhecer o desempenho individual dos examinados, como certificação ocupacional ou habilitação para dirigir, há na literatura uma gama de métodos. Se essa nota de aprovação for definida sem que o teste propriamente dito seja a referência, um teste fácil poderá produzir um número substancialmente maior de aprovados, em detrimento dos examinados minimamente qualificados; e um teste difícil poderá ter o efeito oposto, reprovando examinados injustamente que teriam méritos para aprovação. Esse estudo tem o objetivo de descrever e comparar três métodos existentes na literatura para a definição de linhas de corte, a saber: método holístico, método Neldesky e método Angoff.

No método holístico, um grupo de Juízes analisa o conteúdo do teste e com base numa impressão geral do mesmo, estabelece o percentual de itens que deve ser respondido corretamente por um examinado que tenha um mínimo de conhecimento desejado no assunto de interesse. Shepard(1976) sugere que seja utilizado diversos grupos de juízes que represente vários estratos da população que se interessa pelo resultado do teste. Por exemplo, em um teste de conclusão do ensino médio, grupos de alunos, professores, pais e até representantes da comunidade podem ser consultados. No ajuste, deve ser levado em consideração o erro tolerável, isto é, quando se elabora um teste, deseja-se que os examinados respondam a todos os itens corretamente. Mas devem-se levar em conta algumas respostas incorretas devidas á leitura errada, marcação errada, desatenção, contagem errada, e assim por diante. Um dos problemas do uso desse método é o fato de o desenvolvedor do teste nunca saber se outra amostra de juízes fixaria outro ponto de corte. Uma solução é utilizar mais de um grupo de juízes e repetir o estudo do ajuste para cada grupo, no entanto o número de juízes disponíveis (ou acessíveis) geralmente é fixo ou limitado. Se forem, por exemplo, feita duas replicações, em uma mesma amostra de juízes, o número de juízes em cada grupo é reduzido pela metade. Isto poderia ser um problema, visto que padrões estabelecidos por um número reduzido de juízes terão uma maior variação de amostra para outra do que padrões ajustados por um número maior de juízes. Outro problema é que não há nenhuma certeza que juízes diferentes basearão suas impressões nos mesmos aspectos do teste ou terão percepções similares da área estudada. Isto pode também resultar numa variação indesejável do padrão de uma amostra de juízes a outra.

O método Nedelsky (1954) foi especificamente elaborado para testes com questões de múltipla escolha. A linha de corte é determinada como segue:

1. Cada juiz (em geral, um especialista qualificado na área de interesse do teste) é instruído a marcar, em cada item, o número de opções de respostas que um examinado minimamente competente deve ser capaz de eliminar facilmente, isto é, as opções, que sem dúvida, o examinado sabe que são falsas.

2. Para cada item, o juiz registra a probabilidade de um examinado minimamente competente acertar a questão levando em consideração o número de opções de respostas restantes, ou seja, aquelas com algum nível de dificuldade para responder. Por exemplo, em uma questão com cinco opções de respostas, se duas forem marcadas como respostas que serão facilmente eliminadas pelo examinado, o valor registrado pelo juiz para aquele item seria de um terço.

3. A soma das probabilidades sobre todos os itens do teste é denotada como M e pode ser considerada como o provável resultado de um examinado minimamente competente, conforme determinado a partir de avaliações de um único juiz.

4. Os procedimentos descritos nas etapas anteriores são feitos por cada juiz e após obter a distribuição de respostas M, Calcula-se a média (μM) e o desvio padrão (σM) de M sobre todos os juízes . Nedelsky inicialmente sugeriu que o ponto de corte deveria ser fixado em $\mu M + k\sigma M$ onde o valor de k seria arbitrariamente escolhido, provavelmente entre o intervalo de 0,5 e 1,0. A lógica pressupostamente envolvida na seleção de k e o ajustamento do valor de μM foram criticados e, conseqüentemente, alguns usuários desta técnica preferem simplesmente definir o ponto de corte em μM (Meskauskas, 1976).

Outra técnica, também baseada no conteúdo individual de cada item do teste, foi proposta por Angoff (1971). Ele sugere que cada juiz deve estimar, para cada item do teste, a proporção de examinados com conhecimento mínimo desejado que iriam responder corretamente o item (pode-se pensar também como a probabilidade deste indivíduo responder corretamente o item). Estas probabilidades serão resumidas (utiliza-se alguma medida estatística ou função de uma, por exemplo, a média aritmética) sobre todos os itens para obter a pontuação mínima atribuída por um único juiz. Após analisar a pontuação mínima de cada juiz, resumem-se essas pontuações através da medida adotada anteriormente e esse será o ponto de corte final.

Com a finalidade de comparar os três métodos, um grupo de juízes foi formado por 6 formandos do Curso de Estatística, do semestre 2008.2, da Universidade Federal da Bahia. Este grupo analisou um teste de 9 questões de múltipla-escolha, com 5 opções, cujo conteúdo foi estatística descritiva. Primeiramente, os juízes discutiram o que seria um respondente com o conhecimento mínimo em estatística descritiva que deveria ser aprovado neste teste. Em seguida, os procedimentos descritos acima foram utilizados

para a definição da linha de corte mediante os três métodos. Duas rodadas foram realizadas para cada método, com uma discussão entre uma rodada e outra, sobre os itens de maior discrepância entre os juízes.

Diferente do primeiro método, holístico, os dois seguintes métodos fornecem informações para cada item do teste. A tabela 1 apresenta o resultado do método Nedelsky, na segunda rodada, em que os juízes disseram qual o número de opções que os respondentes com conhecimento mínimo em estatística descritiva eliminaria por ser obviamente errada. Com base nesses resultados, a tabela 2 foi construída calculando a chance de o respondente acertar o item como 1/(no. de opções restantes). Por exemplo, o juiz 1 (COD01) afirmou que o respondente eliminaria 3 das 5 opções do item 1, portanto a chance dele acertar seria ½.

Tabela 1 – Método Nedelsky, 2ª. Rodada, Número de opções prováveis de serem eliminadas pelos respondentes com mínimas competências, segundo os juízes.

Juiz	Item								
	1	2	3	4	5	6	7	8	9
COD01	3	3	3	1	4	2	0	3	1
COD02	3	3	3	3	4	2	3	2	1
COD03	4	3	3	3	4	2	3	2	2
COD04	4	3	3	3	4	3	2	3	1
COD05	4	3	3	3	4	2	2	2	1
COD06	4	3	3	3	4	3	2	2	1

Tabela 2 – Método Nedelsky, 2ª. Rodada, Chance dos respondentes acertarem os itens.

Juiz	Item									pto de corte
	1	2	3	4	5	6	7	8	9	
COD01	0,50	0,50	0,50	0,25	1,00	0,33	0,20	0,50	0,25	4,03
COD02	0,50	0,50	0,50	0,50	1,00	0,33	0,50	0,33	0,25	4,42
COD03	1,00	0,50	0,50	0,50	1,00	0,33	0,50	0,33	0,33	5,00
COD04	1,00	0,50	0,50	0,50	1,00	0,50	0,33	0,50	0,25	5,08
COD05	1,00	0,50	0,50	0,50	1,00	0,33	0,33	0,33	0,25	4,75
COD06	1,00	0,50	0,50	0,50	1,00	0,50	0,33	0,33	0,25	4,92

A tabela 3 apresenta o resultado do método Angoff, na segunda rodada, em que os juízes disseram qual o percentual de respondentes, com conhecimento mínimo no assunto do teste, que responderiam corretamente ao item.

Este estudo utilizou duas estatísticas para estimar a concordância entre os juízes. A primeira foi Coeficiente de Concordância W de Kendall, que é uma medida não-paramétrica que varia entre 0 e 1 (Siegel, 1975). A segunda foi o Coeficiente de Correlação Intraclasse (ICC), que é uma medida de confiabilidade das avaliações realizadas por diferentes especialistas (Shrout & Fleiss, 1979). Os resultados estão na tabela 4.

Tabela 3 – Método Angoff, 2ª. Rodada, Proporção dos respondentes com mínimas competências que acertariam os itens.

Juiz	Item									pto de corte
	1	2	3	4	5	6	7	8	9	
COD01	0,90	0,80	0,90	0,60	0,90	0,70	0,90	0,80	0,50	7,00
COD02	0,90	0,80	0,70	0,60	0,98	0,63	0,70	0,60	0,59	6,50
COD03	0,98	0,90	0,95	0,90	0,98	0,80	0,90	0,75	0,70	7,86
COD04	0,90	0,80	0,60	0,60	0,90	0,60	0,60	0,60	0,40	6,00
COD05	0,95	0,80	0,85	0,80	0,87	0,70	0,75	0,60	0,30	6,62
COD06	0,95	0,95	0,80	0,70	0,95	0,65	0,85	0,70	0,40	6,95

Tabela 4 – Estatísticas de concordância entre os juízes por método e rodada.

Método	**Rodada**	**W de Kendall**	**ICC (IC 95%)**
Nedelsky	1a.	0,570	-0,467 (-3,261 ;0,762)
	2a.	0,790	0,379 (-0,804 ;0,899)
Angoff	1a.	0,679	0,907 (0,731 ;0,985)
	2a.	0,857	0,867 (0,612 ;0,978)

Mediante as estatísticas W de Kendall e ICC, verificou-se que a discussão entre a primeira e a segunda rodada foi relevante, visto que os juízes concordaram mais na segunda rodada. Além disso, O método Angoff se mostrou menos subjetivo que o Nedelsky, tendo os maiores valores das estatísticas apresentadas, e o ICC foi maior que 0,75, que representa, segundo Pinto *et al.* (2008), uma excelente concordância.

Neste estudo não é possível saber o verdadeiro valor do parâmetro, ou seja, qual deve ser o ponto de corte ideal, tal que o respondente com uma nota acima desse valor seria classificado corretamente como tendo o conhecimento mínimo em estatística descritiva. Entretanto, o método que fornecer resultados mais diferentes em relação aos outros, pode estar mais distante desse valor ideal (tabela5). Utilizando a prova de Friedman (Siegel, 1975), a hipótese de que os resultados encontrados, em cada método e em cada rodada, foram extraídos de uma mesma população foi rejeitada ($\chi 2 = 18,8$, com 5 graus de liberdade e p_valor=0,02). Analisando por pares os resultados, mediante a estatística de Wilcoxon, observa-se que o Nedelsky foi o método que apresentou resultados mais diferentes (tabela 6).

Tabela 5 – Linhas de corte definidas pelos juízes para cada método em cada rodada.

Juiz	Holístico		Nedelsky		Angoff	
	1a. Rodada	2a. Rodada	1a. Rodada	2a. Rodada	1a. Rodada	2a. Rodada
COD01	6,3	6,3	4,4	4,0	7,0	7,0
COD02	4,5	5,4	4,7	4,4	6,6	6,5
COD03	7,9	7,0	5,4	5,0	8,2	7,9
COD04	6,3	6,3	5,3	5,1	5,8	6,0
COD05	5,4	5,4	5,2	4,8	7,1	6,6
COD06	6,0	6,0	5,7	4,9	4,0	7,0
Média	6,1	6,1	5,1	4,7	6,4	6,8
Mediana	6,2	6,2	5,2	4,8	6,8	6,8
DP	1,1	0,6	0,5	0,4	1,4	0,6
Mínimo	4,5	5,4	4,4	4,0	4,0	6,0
Máximo	7,9	7,0	5,4	5,1	8,2	7,9

Tabela 6 – Estatística de Wilcoxon e o valor p calculados para cada par de métodos e rodada.

Métodos		Holístico		Nedelsky		Angoff	
		1a. Rodada	2a. Rodada	1a. Rodada	2a. Rodada	1a. Rodada	2a. Rodada
Holístico	1a. Rodada	--	0 (1)	-1,89 (0,058)	-2,2 (0,028)	-0,73 (0,463)	-1,75 (0,08)
	2a. Rodada		--	-2,2 (0,028)	-2,2 (0,028)	-0,74 (0,462)	-1,99 (0,046)
Nedelsky	1a. Rodada			--	-2,23 (0,026)	-1,79 (0,074)	-2,2 (0,028)
	2a. Rodada				--	-1,78 (0,075)	-2,21 (0,027)
Angoff	1a. Rodada					--	-0,13 (0,893)
	2a. Rodada						--

Referências Bibliográficas

ANGOFF, W.H. (1971) Norms, scales, and equivalent scores. In R.L. Thorndike(Ed.) Educational measurement(2nd Ed.)Washinton, D.C.: American Council on Education.

MESKAUSKAS, J.A. (1976) Evaluation models for criterion-referenced terting: Views regarding mastery and standard setting. Review of Educational Research; 46:133-158.

NEDELSKY, L.(1954) Absolute grading standards for objective tests. Educational and Psychological Measurement; 14:3-19.

Pinto, J. S., Lopes, J. M., Oliveira, J.V., Amaro, J. P., Costa, L. D. (2008). Métodos para Estimação de Reprodutividade de Medida. Disponível em: http://users.med.up.pt/joakim/intromed/coeficientecorrelacaointraclasse.htm. Acesso em: 24 de abril de 2008.

SHEPARD, L.A. (1979) Setting standards. In M.A. Bunda and J.R. Sanders(Eds.). Practices and problems in competency-based measurement. National Council of Measurement in education.

Shrout, P.E. & Fleiss, J.L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability, *Psychological Bulletin*, Vol. 86.

SIEGEL, S. (1975). Estatística não-paramétrica (para as ciências do comportamento). São Paulo: McGraw-Hill.