Fidedignidade de Notas Atribuídas a Redações: enfoque teórico e empírico

Nícia Maria Bessa*

Desde a primeira década deste século, uma vasta literatura tem focalizado a avaliação da fidedignidade das medidas, em educação e em psicologia. Os fundamentos de uma teoria das medidas educacionais, lançados por Spearman, são sistematizados pela primeira vez por Edward L. Thorndike em 1904 (Stanley, 1971).

A partir de então, vários métodos de avaliação da fidedignidade têm sido propostos no contexto de teorias que conceituam de modo diverso o erro de medida. O enfoque correlacional passa a ser combinado com a aplicação dos processos de análise da variância.

O desenvolvimento recente das teorias da generalização (Cronbach, Gleser, Nanda, Rajaratnam, 1972) e dos testes congenéricos (Joreskog, 1971), ou das aplicações da análise da variância multivariada (Bock, 1966), atestam que persiste o interesse pelos processos de avaliação do erro das medidas educacionais. Dentro da perspectiva das diferentes teorias, são empregados métodos de análise de correlação, de análise da variância, ou de análise fatorial, em estudos cujo objetivo é aferir o erro de medida. A estrutura dos experimentos, que se utilizam na coleta de dados, pode variar ou não, de acordo com o contexto teórico; certamente haverá, entretanto, diferenças de interpretação dos resultados obtidos (Bessa, 1984).

O presente trabalho é parte de um conjunto de estudos, em que os resultados de um experimento são interpretados no contexto da teoria das medidas educacionais. Este artigo, em particular, limita-se à apresentação de um estudo empírico sobre a fidedignidade de notas atribuídas a redações; a interpretação dos resultados, à luz da teoria, será objeto de artigo subseqüente.

1 Trabalho executado com apoio do CNPq.

* Fundação CESGRANRIO; PUC/RJ

A FIDEDIGNIDADE DE NOTAS ATRIBUÍDAS A REDAÇÕES: UM ESTUDO EMPÍRICO

Embora as pesquisas sobre a fidedignidade de notas atribuídas a provas discursivas sejam numerosas e remontem a mais de cinco décadas (Coffman, 1971, p.277-279), são ainda pouco comuns em nosso país (Vianna, 1982). Em recente revisão bibliográfica, Vianna (1982) concluiu que as poucas realizadas no Brasil reproduzem os resultados daquelas realizadas em outros países. De fato, os estudos de Angotti (1983), de Castilhos (1982), de Ciacaglia (1981), de Costa Ribeiro *et alii* (1981), de Vianna (1978, 1976a, 1976b) sugerem que as distribuições de notas atribuídas, através de julgamento de examinadores, a provas dissertativas refletem considerável proporção da variância devida a erros de medida.

Conclui-se dos trabalhos de Angotti (1983) e de Vianna (1982, 1976a) que, em geral, as notas atribuídas a redações pelos mesmos juízes, em duas ocasiões, sofrem considerável flutuação.

Com base na análise da variância de notas atribuídas a redações, Vianna (1976b) encontra diferenças significativas entre médias das notas atribuídas por examinadores diversos e coeficiente de fidedignidade de 0,53, com relação às notas de um só avaliador.

Castilhos (1978) estuda o efeito do treinamento dos examinadores, mas os coeficientes de generalização obtidos levam a conclusões semelhantes às de Vianna (1976b).

O estudo de Costa Ribeiro *et alii* (1981) indica haver diferenças entre equipes de examinadores, numa dimensão de severidade/benevolência, assim como entre as distribuições de notas atribuídas por diversos examinadores dentro da mesma equipe.

Estes resultados se aproximam daqueles dos estudos citados por Coffman (1971), assim como daqueles apresentados por Biggs & Collis (1982), ou por Diederich (1974). De modo geral, confirmam uma longa história de estudos realizados em vários países (Chediak, Bessa *et alii*, 1975; Vianna, 1976a, 1982).

Dada a larga utilização das notas atribuídas a redações em processos seletivos – como é o caso dos exames Vestibulares às escolas superiores – justifica-se a persistência do interesse pelo estudo da respectiva confiabilidade.

O objetivo do presente estudo empírico é verificar a fidedignidade de notas atribuídas a redações através da investigação da consistência dos julgamentos de diversos juízes em relação às mesmas redações. Procura-se estimar a fidedignidade das notas atribuídas por um ou mais examinadores ao julgar redações, todas sobre o mesmo tema.

METODOLOGIA

Em 1979, os candidatos inscritos para os exames Vestibulares da Fundação CESGRANRIO foram submetidos a uma prova de redação, em Português, cujo tema foi: A PAZ.

Estando arquivados os originais, identificados com código individualizado, foi possível utilizá-los na presente investigação.

População e amostra

Em estudos sobre grupos que se candidatam a cursos diversos, ao se inscreverem para os Vestibulares da CESGRANRIO, Costa Ribeiro (1981) e Costa Ribeiro e Klein (1982) constatam haver diferenças entre as respectivas médias obtidas no conjunto das provas.

De outro lado, a seleção para a universidade se opera dentro de cada grupo de candidatos que optam por determinada carreira.

Nestas condições, preferiu-se realizar o estudo da fidedignidade das notas atribuídas a redações em populações de candidatos que, em primeira opção, escolhem a mesma carreira. Definiram-se seis populações de candidatos, compostas pelos inscritos para o Vestibular de 1979 nos seguintes cursos: Ciências Biológicas, Educação, Educação Física Masculina, Engenharia, Letras e Medicina.

De acordo com Costa Ribeiro e Klein (1982), estes grupos se situam em pontos diferentes

de um contínuo que pode representar as variações das respectivas médias nas provas. O estudo por grupo pode revelar se as diferenças de desempenho se refletem na fidedignidade das notas atribuídas a redações.

De cada população P_ℓ ($\ell = 1,..,6$) de candidatos selecionou-se uma amostra aleatória. As respectivas frações de amostragem são apresentadas no Quadro 1. Essas frações foram determinadas com base nas médias e variâncias das notas obtidas na prova objetiva de Português, do mesmo anó, pelos componentes de cada população P_ℓ; em cada caso esperava-se um erro de amostragem de, no máximo, 0,05 em relação à média. Das listas de candidatos, organizadas por ordem segundo o número de inscrição, foram selecionados aleatoriamente aqueles que deveriam compor a amostra de cada população P_ℓ.

QUADRO 1
Número de Candidatos na População P_ℓ e Respectiva Fração de Amostragem

Curso	População	Amostra	
	N	N	Fração
Ciências Biológicas	2.496	249	1/10
Educação	1.888	269	1/7
Educ. Física Masc.	692	230	1/3
Engenharia	27.562	275	1/100
Letras	4.104	820	1/5
Medicina	14.993	249	1/60

Foram excluídos os casos de candidatos qua não fizeram a prova de redação, por qualquer motivo, não sendo substituídos por outros, na amostra.

Além das amostras de candidatos, foi selecionada uma amostra aleatória simples de 16 professores de português, dentre os 155 que haviam realizado o trabalho de atribuição de conceitos (A, B ou C) às redações, no Vestibular de 1979 da CESGRANRIO.

Delineamento e estruturação da pesquisa

É propósito desta pesquisa avaliar a contribuição da variância devida a erros de medida σ^2_{ψ} em relação à variância das notas atribuídas às redações σ^2_x. Para tanto, empregou-se um esquema experimental de um só fator, com medidas repetidas dos mesmos sujeitos. Os professores-juízes constituíram o fator PROFESSOR; a média das notas atribuídas por um professor a todas as redações serviu de base para a definição do efeito do fator PROFESSOR.

Cada um dos 16 professores atribuiu, independentemente, uma nota a cada redação; essas notas foram consideradas como medidas repetidas, independentes, de um mesmo sujeito (ou redação). A diferença entre uma nota atribuída a determinada redação e a média das 16 notas conferidas à mesma redação serviu à aferição da variância devida a erros de medida.

Os sujeitos – autores das redações – constituíam amostra aleatória de uma população P_ℓ; também os, professores formavam amostra aleatória de uma população definida. Este esquema permite a utilização de uma análise dos componentes da variância das notas σ^2_x, com referência a

cada população P_ℓ .

No modelo de análise dos componentes da variância, represente-se a nota atribuída à redação do indivíduo i pelo juiz g por:

$$X_{ig} = \mu + (\mu_i - \mu) + (\mu_g - \mu) + (X_{ig} - \mu_i - \mu_g + \mu)$$

onde

$$\mu = E_i E_g (X_{ig}) \qquad i = 1,...,N$$

$$\mu_i = E_g (X_{ig}) \qquad g = 1,...,G$$

$$\mu_g = E_i (X_{ig})$$

De acordo com o modelo, X_{ig} é uma variável aleatória na população P_ℓ de sujeitos e na população de juízes que atribuem as notas às redações. São também variáveis aleatórias os efeitos dos professores (ou juízes) e os efeitos residuais – estes últimos confundidos com a interação entre provas individuais e juízes. Pressupõe-se que os efeitos dos juízes e os resíduos sejam variáveis aleatórias independentes, tenham distribuição com média zero e variâncias σ^2_γ e σ^2_ψ respectivamente; pressupõe-se, ainda, que σ^2_ψ seja igual para todas as repetições da medida g. Acrescenta-se o pressuposto da distribuição normal dessas variáveis, para efeito de estimativas de variâncias e de testes de hipóteses.

Ao estruturar a pesquisa, a seleção de candidatos e a seleção de professores-juízes foram realizadas por processo aleatório dentro das respectivas populações. Como a pesquisa se iniciou em 1981, havia pouca probabilidade dos antigos examinadores se recordarem das provas examinadas dois anos antes; a probabilidade de uma prova ser julgada pelo mesmo professor também era pequena, pois o total de candidatos era de 123.707.

Controlou-se a contaminação entre critérios de julgamentos dos juízes, fazendo-os ignorar quais eram os demais professores selecionados e qual a fonte a que se recorria para a seleção; além disso, cada um recebeu o lote de redações para trabalhar independentemente dos demais. Os contatos com os professores foram realizados individualmente. Cada professor trabalhou sozinho, sem contato com qualquer outro, julgando uma média de 90 redações por dia.

Procurou-se evitar um efeito do conhecimento da carreira escolhida pelo candidato e um efeito da ordem de apresentação das redações para julgamento. A investigação se estendeu a candidatos de várias carreiras, estando as provas misturadas. Os documentos arquivados na Fundação CESGRANRIO eram identificados por um número diferente daquele recebido pelo candidato ao inscrever-se; a identificação do candidato não era fácil, nem direta, exigindo que se estabelecesse a correspondência entre os dois códigos. Não era possível, portanto, que o professor-juiz percebesse, ao examinar a redação, quaisquer informações sobre o candidato que pudessem influenciar seu julgamento. As notas eram escritas em listas, à parte, contendo os códigos de todas as redações; não foram feitas marcas ou comentários nas cópias das provas. Para maior segurança, antes de serem entregues a um professor-juiz, os documentos eram examinados para verificar se havia marcas que revelassem julgamento dos demais examinadores. O possível efeito da ordem em que as redações se apresentavam ao examinador para julgamento foi controlado fazendo-se uma arrumação do lote de documentos em ordem obtida aleatoriamente antes de enviá-lo a cada professor-juiz.

As instruções para o julgamento foram distribuídas aos professores, por escrito, em cópias datilografadas. Ao passá-las ao professor, possíveis dúvidas eram discutidas com o pesquisador. O método de julgamento com base na impressão geral (Coffman, 1971) foi adotado. As instruções, por escrito, solicitavam que o professor expressasse o julgamento em nota entre zero a dez, sempre em número inteiro. Conforme a impressão global que tivesse de cada redação, o professor deveria colocá-la em uma das 11 pilhas correspondentes aos pontos de escala de notas de zero a dez. Depois de reler as redações de cada pilha, deveria anotar a respectiva nota na lista apropriada.

Confundem-se com outras fontes de variação, não podendo ser aferidos separadamente, os

efeitos da interação entre o examinador e as condições de apresentação da redação – por exemplo, qualidade de caligrafia, tipo de erros gramaticais, limpeza. A influência de condições como estas tem sido constatada em estudos sobre atribuição de notas a provas dissertativas em geral (Marshall, 1967; Marshall & Powers, 1969; Chase, 1979).

A perda de sujeitos selecionados para as amostras deu-se, em maior parte, antes de ser iniciado o julgamento pelos examinadores: alguns candidatos inscritos não fizeram a prova de redação; outros produziram documentos cujas cópias, em xerox, não ficaram suficientemente nítidas. Alguns documentos foram considerados ilegíveis pelos juízes. Os totais de perdas variam entre 28% e 47%, dentre os candidatos amostrados.

Durante a realização dos exames vestibulares, os candidatos submeteram-se às provas em locais diferentes do Estado do Rio de Janeiro e sob a fiscalização de equipes diversas. As instruções, o horário, o material distribuídos e o tempo concedido (60 minutos) foram iguais para todos os candidatos. Podem ter variado, portanto, as condições ambientais, de um local de aplicação de prova para outro, embora os fiscais procurassem manter condições padronizadas, tanto quanto possível.

RESULTADOS

Cada população P ℓ de candidatos foi analisada separadamente. Das respectivas amostras perderam-se elementos, por motivos vários. O Quadro 2 mostra o número de casos utilizados para fins de análise das notas atribuídas pelos professores-juízes.

QUADRO 2
Número e Porcentagem de Redações Utilizadas no Estudo

Curso	Amostra N	Redações N	Utilizadas %
Ciências Biológicas	249	178	71
Educação	269	193	72
Educ. Física Masc.	230	123	53
Engenharia	275	198	72
Letras	820	566	69
Medicina	249	179	72

A perda de casos foi nitidamente maior no grupo de candidatos aos cursos de Educação Física Masculina. Não se pode concluir, porém, que isto se deva exclusivamente a algum fator particular pois cópias foram excluídas por motivos vários e alguns originais não foram identificados no arquivo.

Distribuição de notas

As médias e desvios-padrão das notas atribuídas pelos professores-juízes se apresentam no

QUADRO 3.

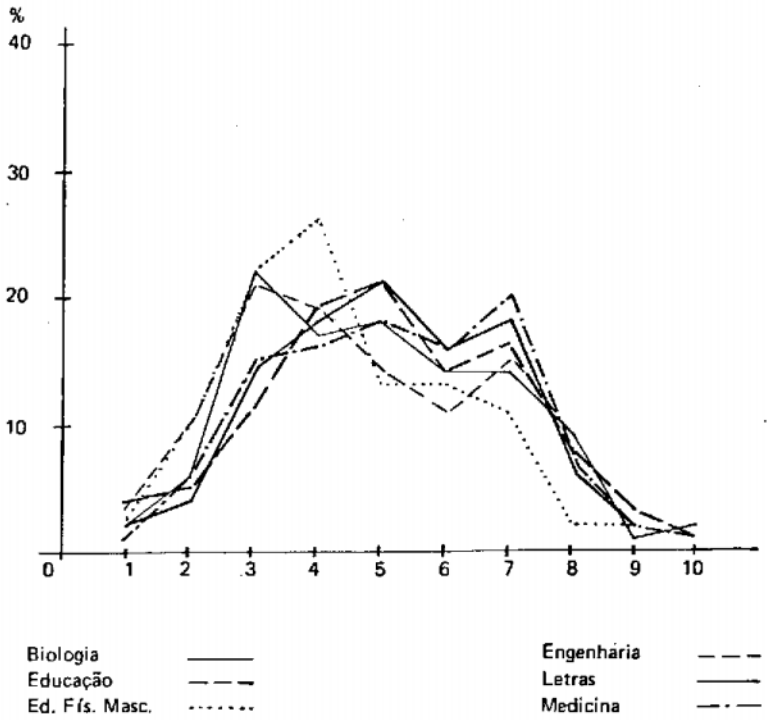
QUADRO 3

Média e desvio-padrão das notas atribuídas às redações, por curso e por professor.

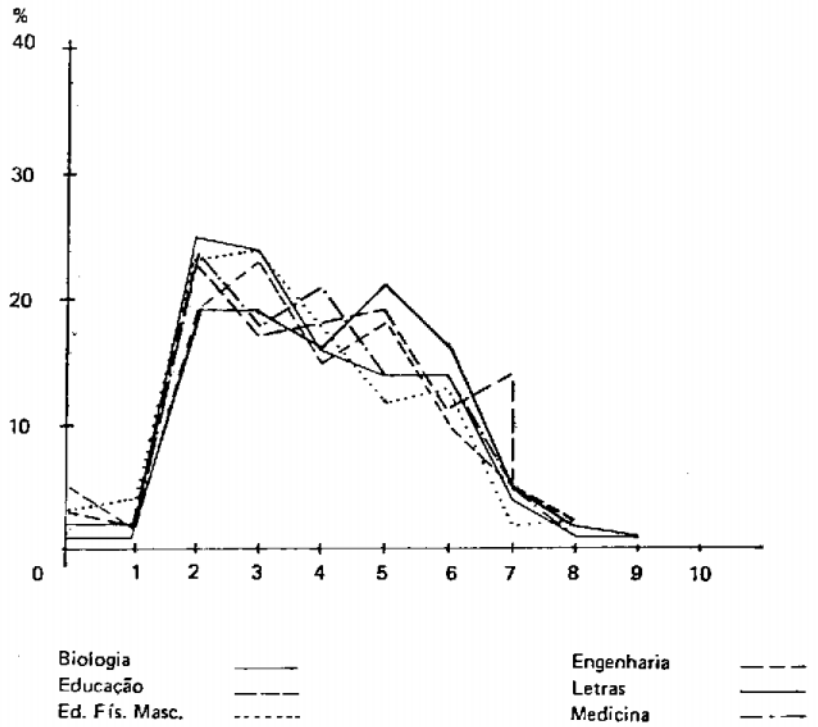
Curso	Professor																												Média do Curso
	1		2		3		4		5		6		7		8		9		10		11		12		13		14		15		16		
	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	M	DP	
Biologia (178)	4,8	1,9	3,8	1,7	4,7	2,5	3,4	2,2	3,7	2,6	4,3	2,0	4,8	1,7	3,7	1,3	4,8	2,4	4,4	2,0	4,7	2,2	4,5	2,2	2,8	2,3	5,0	1,8	4,3	2,3	4,0	1,8	4,23
Educação (193)	5,1	1,9	3,8	1,7	5,0	2,5	3,6	2,1	4,1	2,5	4,6	2,1	5,1	1,8	4,0	1,5	5,2	2,3	4,5	2,2	4,6	2,3	4,7	2,3	2,9	2,3	5,4	1,9	4,6	2,3	4,3	1,8	4,47
Ed. Fís. Masc. (123) ...	4,4	1,7	3,6	1,7	3,8	2,6	2,9	2,2	2,7	2,4	3,5	2,0	4,6	1,6	3,3	1,4	4,0	2,3	3,7	2,1	3,8	2,2	3,8	2,1	2,1	2,1	4,4	1,8	3,4	2,3	3,4	1,7	3,59
Engenharia (198)	4,7	1,9	3,8	1,8	4,4	2,6	3,2	2,2	3,2	2,4	3,9	2,1	4,9	1,6	3,5	1,4	4,4	2,4	4,0	2,1	4,4	2,2	4,2	2,2	2,6	2,2	4,8	1,8	4,0	2,3	3,8	1,7	3,99
Letras (566)	5,1	1,8	4,0	1,7	5,1	2,6	3,8	2,3	4,1	2,6	4,6	2,0	5,3	1,7	3,9	1,5	5,1	2,2	4,6	2,1	4,8	2,3	4,8	2,2	3,2	2,3	5,4	1,8	4,7	2,3	4,3	1,7	4,55
Medicina (179)	5,2	1,8	4,0	1,7	5,1	2,6	3,8	2,1	4,0	2,5	4,7	2,0	5,1	1,6	3,9	1,4	5,3	2,3	4,9	2,0	5,0	2,2	4,9	2,1	3,2	2,3	5,3	1,7	4,7	2,3	4,3	1,6	4,58
Média	4,96		3,88		4,83		3,56		3,79		4,38		5,07		3,78		4,91		4,44		4,64		4,59		2,93		5,17		4,43		4,12		

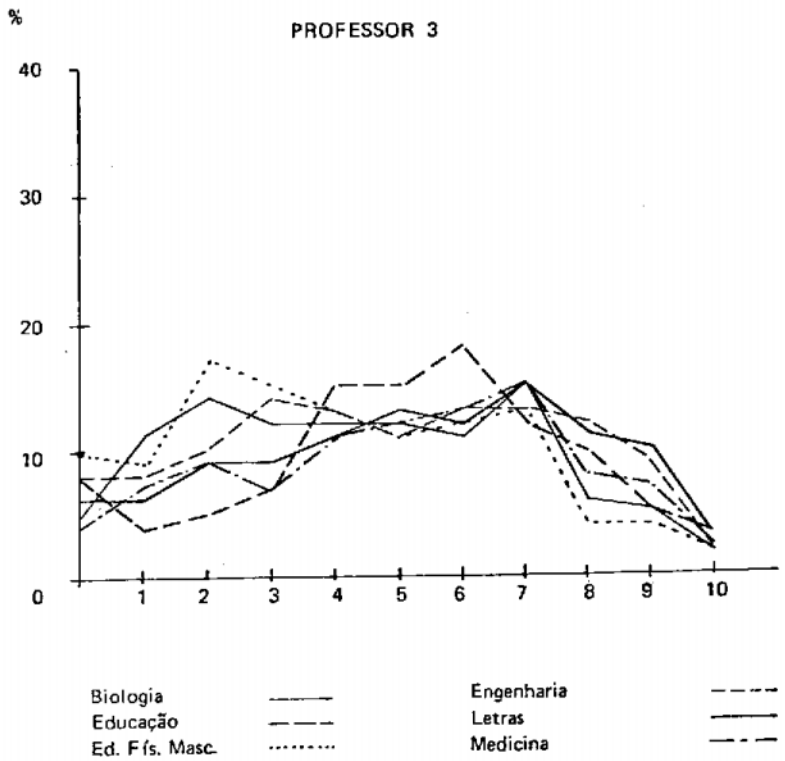
Observa-se que a amplitude da variação entre as médias das notas atribuídas pelos professores-juízes é de cerca de dois pontos, na escala utilizada. Na Figura 1, as distribuições das notas atribuídas, por professor-juiz, mostram que os valores modais variam de zero a seis, e que a proporção de notas zero varia enormemente de um avaliador a outro. Os gráficos indicam claramente a tendência de certos juízes a atribuírem notas com maior severidade, a todos os candidatos, qualquer que seja o curso escolhido. Note-se que Costa Ribeiro *et alii* (1981) identificam uma dimensão de "severidade-benevolência" em que podem ser posicionados os julgamentos de diversos juízes, ao atribuirem conceitos A, B e C a redações, usando critérios e processos de avaliação que divergem daqueles empregados no presente estudo.

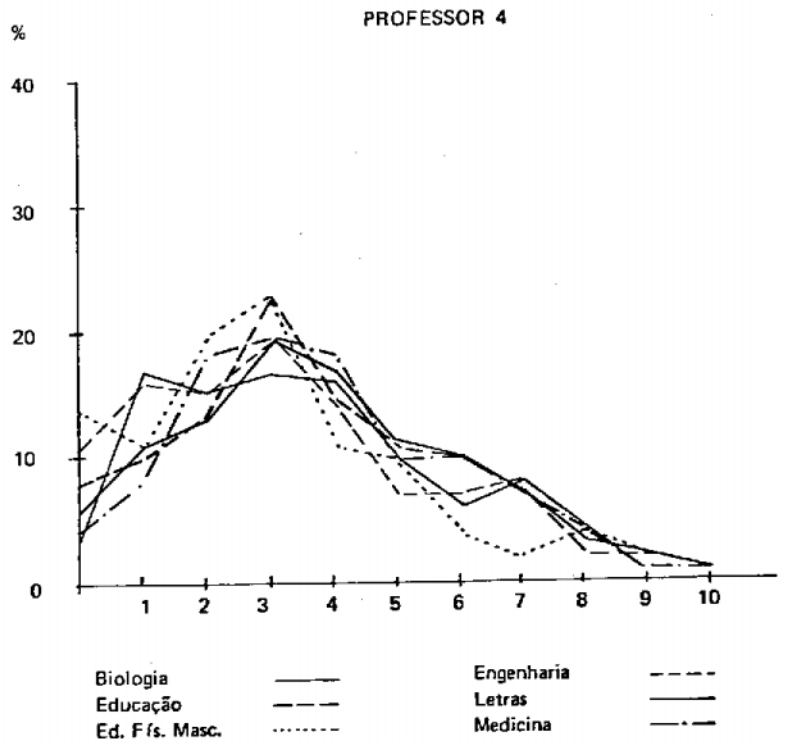
PROFESSOR 1



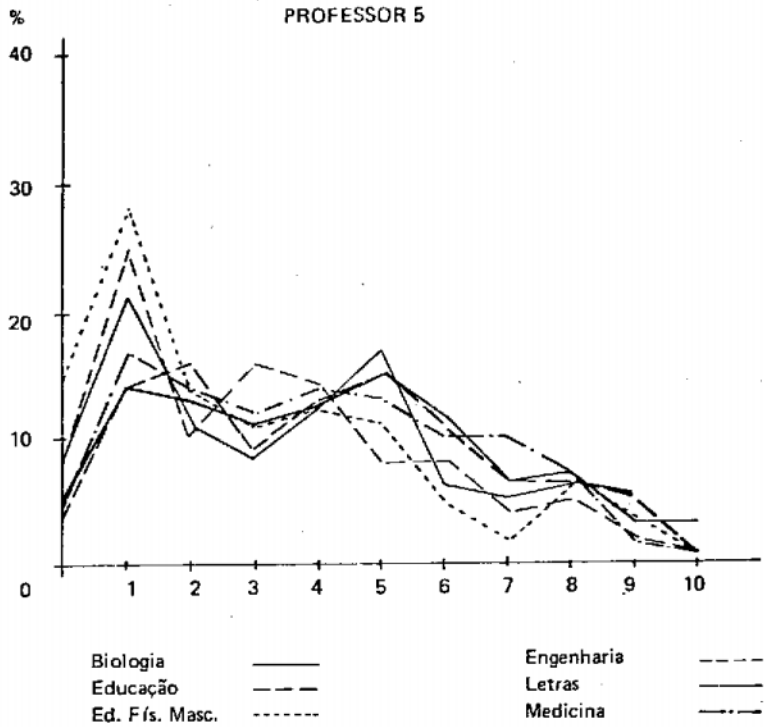
PROFESSOR 2



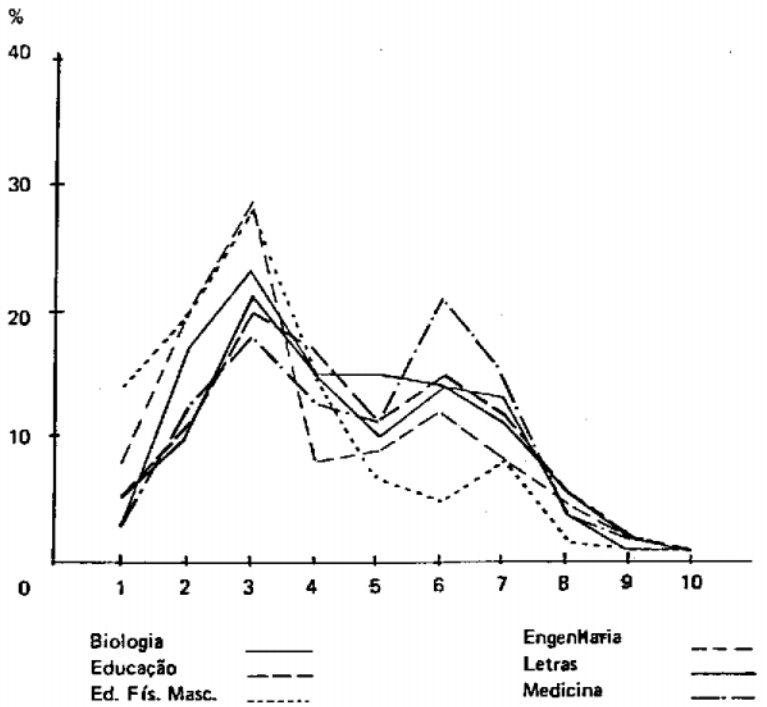

%
PROFESSOR 3
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
Biologia
Educação
Ed. Fís. Masc.
Engenharia
Letras
Medicina


%
PROFESSOR 4
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
Biologia
Educação
Ed. Fís. Masc.
Engenharia
Letras
Medicina

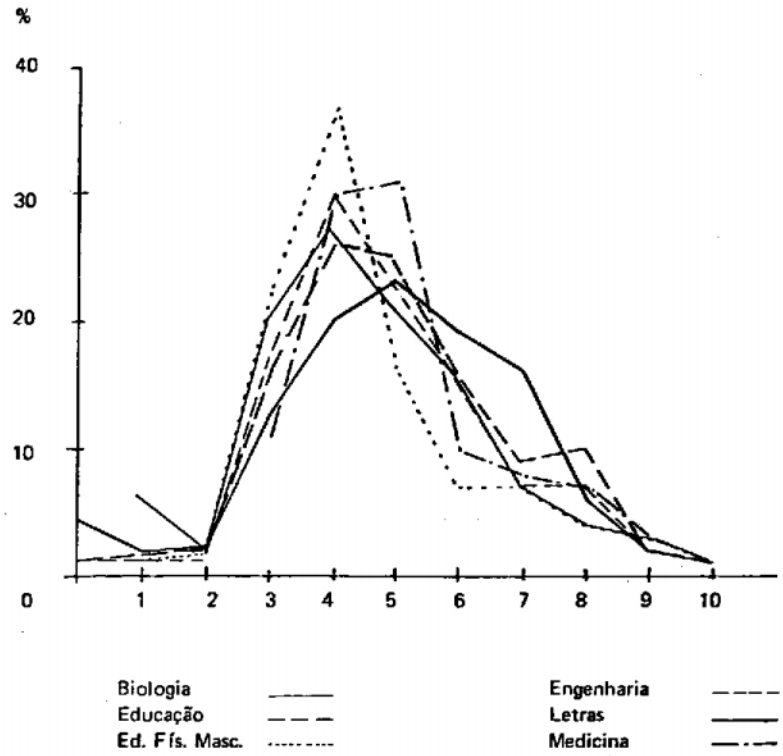
PROFESSOR 5



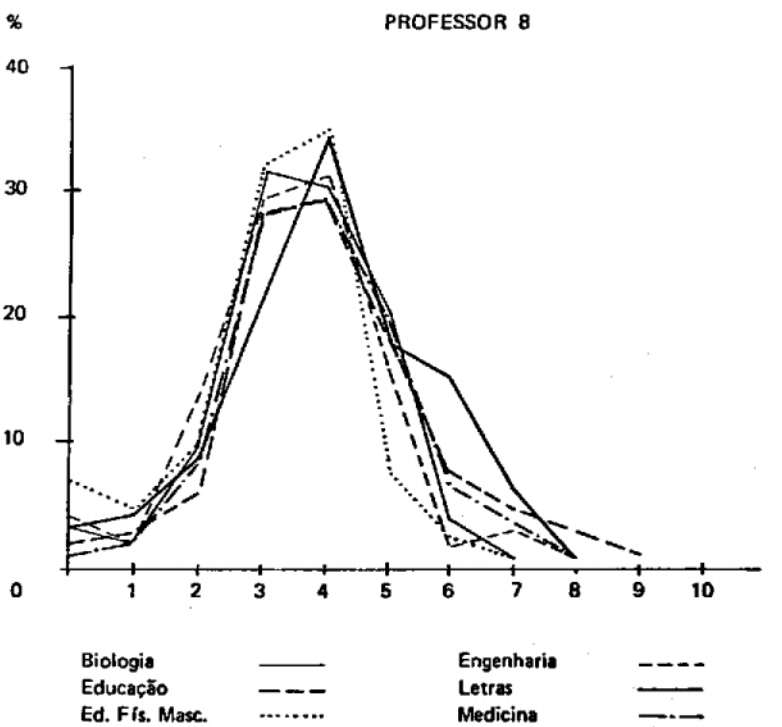
PROFESSOR 6



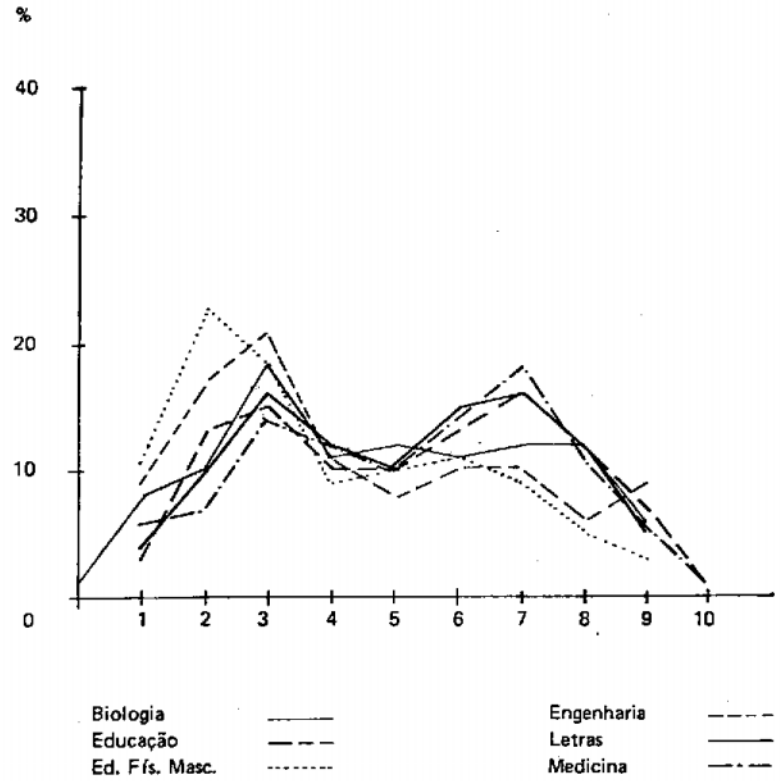
PROFESSOR 7



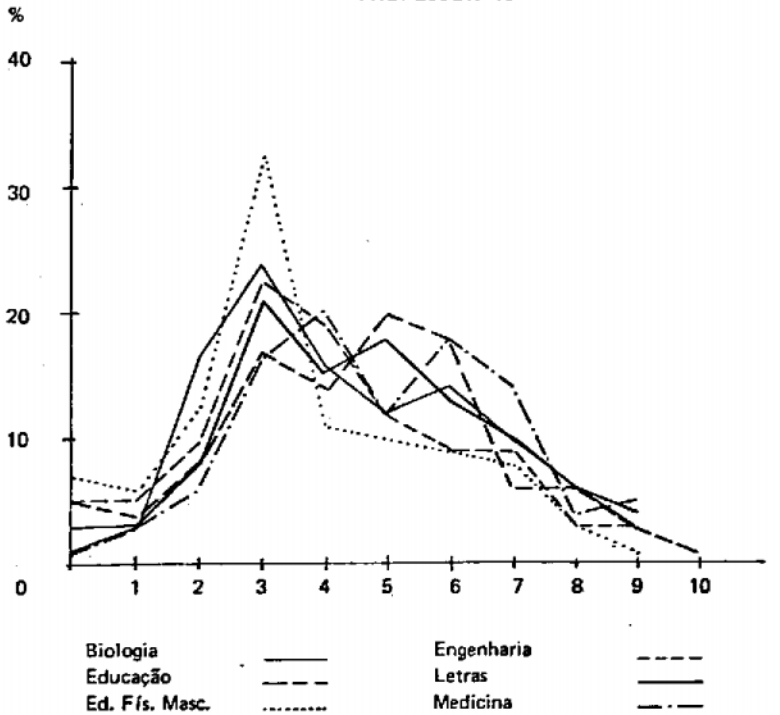
PROFESSOR 8



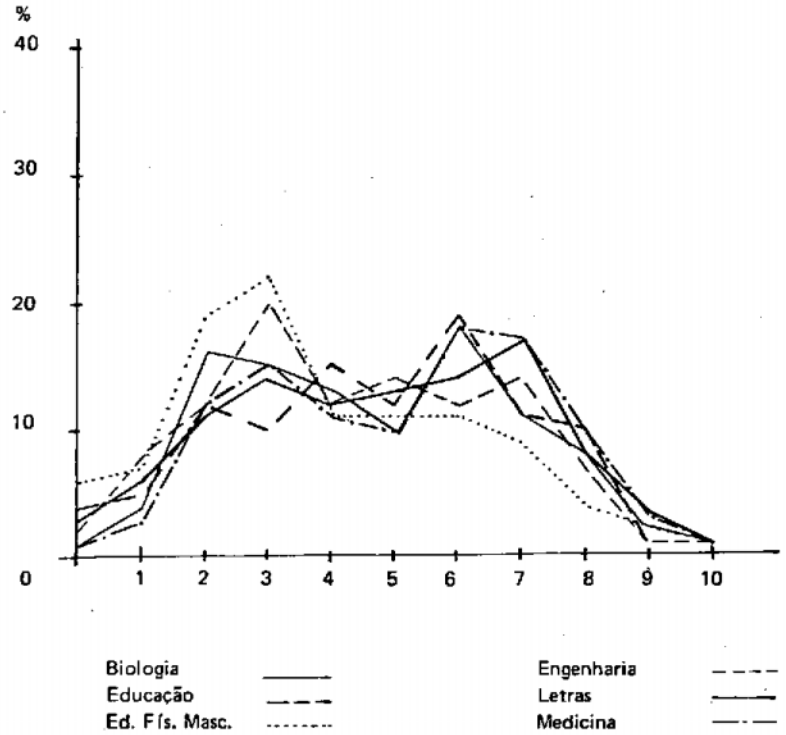
PROFESSOR 9



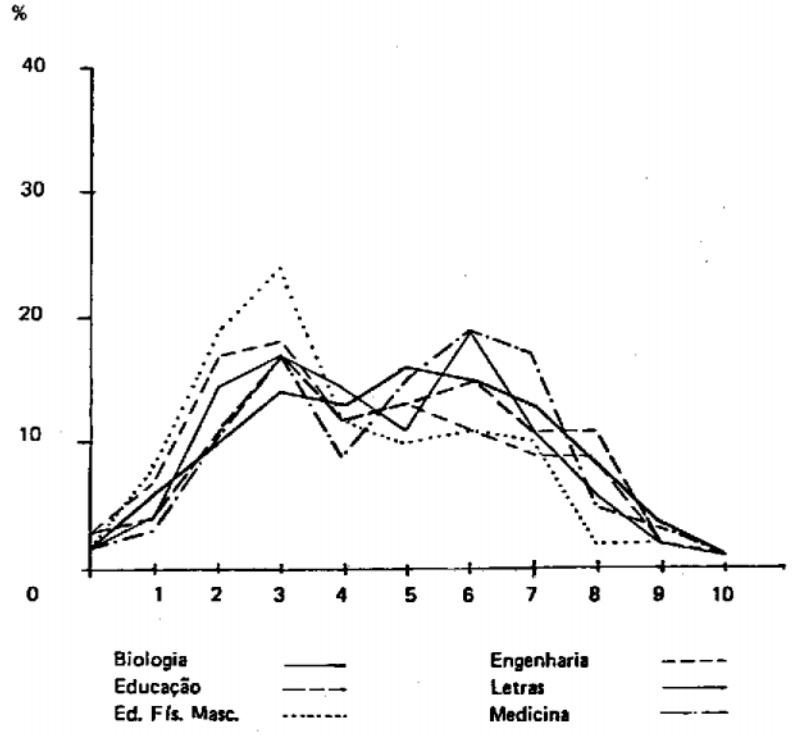
PROFESSOR 10



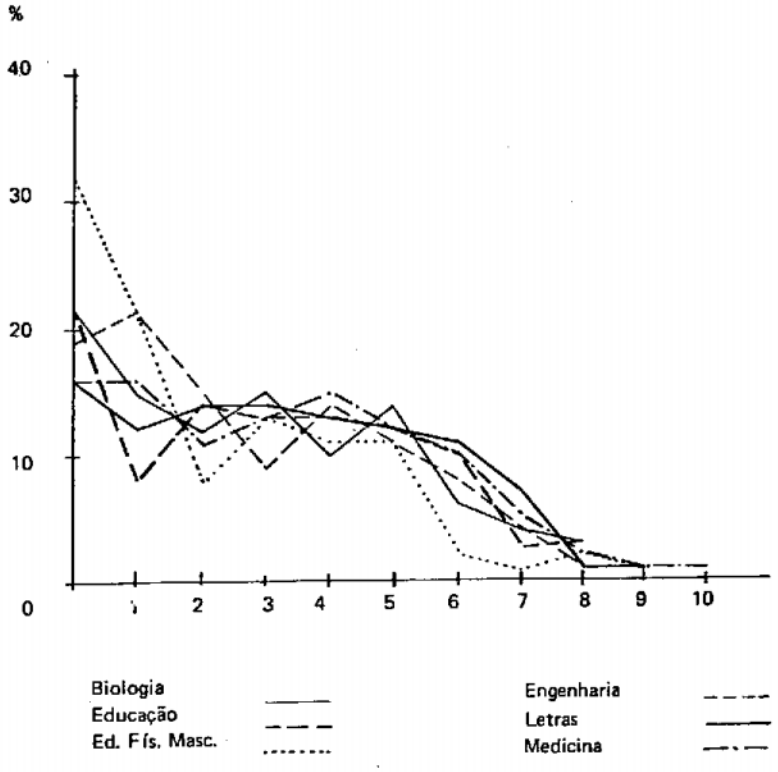
PROFESSOR 11



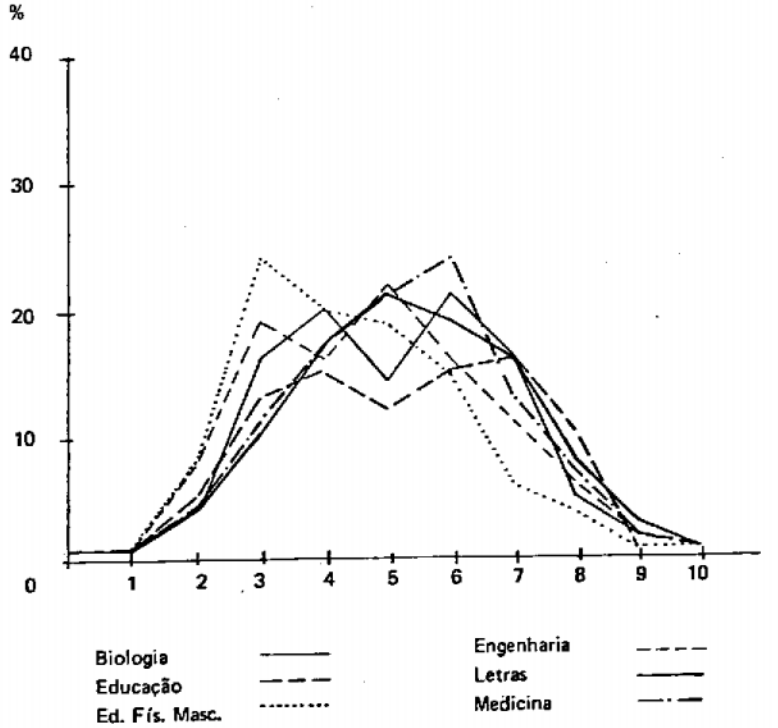
PROFESSOR 12



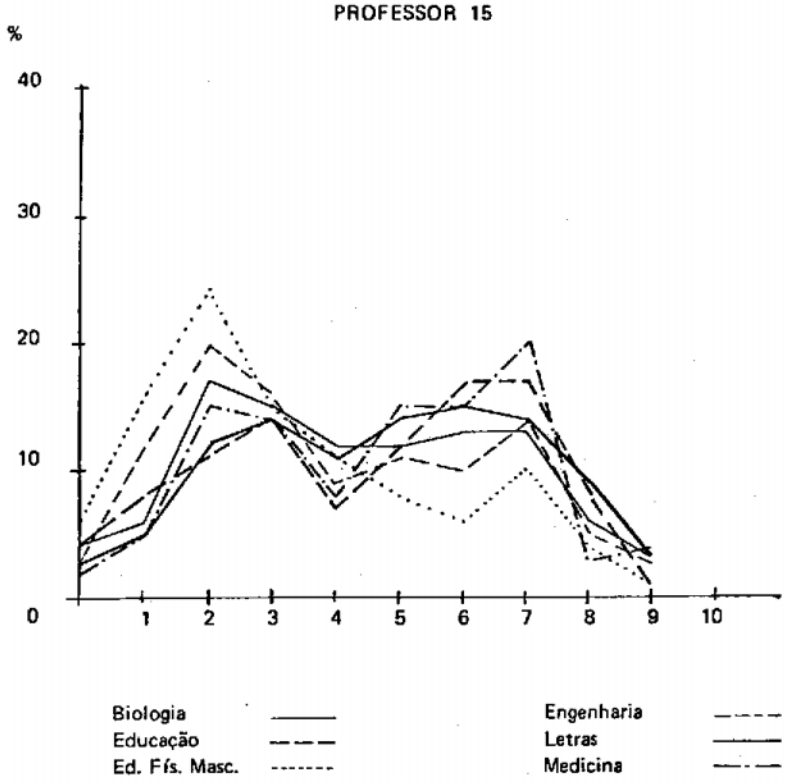
PROFESSOR 13



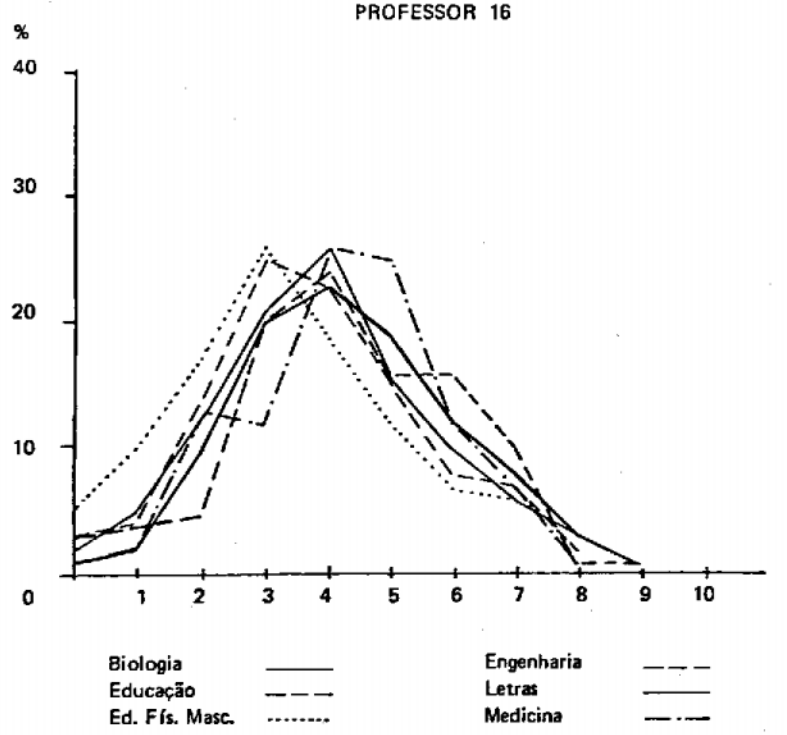
PROFESSOR 14



PROFESSOR 15



PROFESSOR 16



Observa-se, além disso, que em cada gráfico da Figura 1 há notável semelhança quanto à forma das distribuições de notas atribuídas por um professor aos candidatos dos seis cursos. Há muito maior semelhança entre a forma das distribuições de notas atribuídas pelo mesmo professor, aos candidatos dos seis cursos, do que entre a forma das distribuições de notas atribuídas aos candidatos de um só curso, por diferentes professores-juízes. São exemplos marcantes as distribuições de notas dos candidatos a Medicina, que apresentam assimetria negativa ou tendem a uma assimetria positiva conforme o professor avaliador.

Seria difícil interpretar as diferenças entre as médias obtidas por candidatos a diferentes cursos (Quadro 3), com base nos resultados das amostras selecionadas para esta pesquisa, pois não foi possível controlar o efeito da interação das redações com os professores-juízes. Estes últimos são professores de Português; não seria impossível um efeito positivo do estilo de redações daqueles que se destinam a determinados cursos (Marshall, 1967; Coffman, 1971). De qualquer modo, neste trabalho, a análise focaliza cada uma das populações separadamente, sem procurar compará-las.

Diferenças entre professores-juízes.

Os resultados das análises da variância são apresentados por curso, no Quadro 4. Essas análises tomaram por base as notas atribuídas pelos 16 professores. Em todos os grupos, os valores de F obtidos, com os graus de liberdade indicados para "juízes" e "resíduo", são significativos ao nível de 0,01.

QUADRO 4

Resultados da ANOVA das Notas Atribuídas por 16 Professores – Juízes

Grupo: **Ciências Biológicas**

Fonte	S Q	G L	M Q	F
Entre sujeitos	8232,836	177	46,513	
Dentro	5004,563	2670	1,874	
Entre juízes	1027,674	15	68,512	45,739 **
Resíduo	3976,887	2655	1,498	
Total	13237,399	2847	4,650	

(a)

Grupo: **Educação**

Fonte	S Q	G L	M Q	F
Entre sujeitos	9356,759	192	48,733	
Dentro	5742,500	2895	1,984	
Entre juízes	1321,404	15	88,094	57,386 **
Resíduo	4421,096	2880	1,535	
Total	15099,259	3087	4,891	

(b)

Grupo: **Educação Física Masculina**

Fonte	S Q	G L	M Q	F
Entre sujeitos	5380,038	122	44,099	
Dentro	3407,438	1845	1,847	
Entre juízes	792,776	15	52,852	36,991 **
Resíduo	2614,662	1830	1,429	
Total	8787,475	1967	4,467	

(c)

** Significativo ao nível de 0,01

Grupo: **Engenharia**

Fonte	S Q	G L	M Q	F
Entre sujeitos	9609,125	197	48,777	
Dentro	5370,875	2970	1,808	
Entre juízes	1145,768	15	76,385	53,423 **
Resíduo	4225,107	2955	1,430	
Total	14980,000	3167	4,730	

(d)

Grupo: **Letras**

Fonte	S Q	G L	M Q	F
Entre sujeitos	27349,589	565	48,406	
Dentro	24751,188	8490	2,915	
Entre juízes	3314,439	15	220,963	87,357 **
Resíduo	21436,749	8475	2,529	
Total	52100,776	9055	5,754	

(e)

Grupo: **Medicina**

Fonte	S Q	G L	M Q	F
Entre sujeitos	7683,381	178	43,165	
Dentro	5297,500	2685	1,973	
Entre juízes	1112,647	15	74,176	47,326 **
Resíduo	4184,853	2670	1,567	
Total	12980,881	2863	4,534	

(f)

** Significativo ao nível de 1%.

Com base nos valores de F obtidos, pode-se aceitar a hipótese de que há diferenças entre as médias das notas atribuídas pelos professores-juízes, em cada uma das seis populações estudadas; ou seja, que a variância dos efeitos devidos ao fator PROFESSOR é maior do que zero.

Confirmam-se as conclusões baseadas no Quadro 3 e nos gráficos da Figura 1. Estes resultados situam-se na mesma linha daqueles obtidos por Costa Ribeiro *et alii* (1981) e por Vianna (1976; 1978), no Brasil, assim como por outros pesquisadores em outros países (Coffman, 1971).

Conforme o delineamento da pesquisa, a diferença entre a média das 16 notas atribuídas a uma redação e cada uma dessas notas deve-se, em parte, a diferenças entre julgamentos dos professores-juízes e, em parte, a erro residual. Na parte da variância devida a erro confundem-se os efeitos de diversas fontes não controladas, que podem afetar o julgamento de um professor-juiz, incluindo-se aí a interação do avaliador com a redação do candidato.

No caso particular, em que a mesma redação é submetida aos 16 professores, as diferenças entre as respectivas notas podem ser ou não interpretadas como erros de medida. A parte da variância devida a diferenças entre juízes deve ser removida, antes da avaliação do erro de medida, quando se utiliza a nota média atribuída por vários professores-juízes a cada redação, ou quando as notas são equacionadas. Pelo contrário, a parte da variância devida a diferenças entre professores-juízes deve fazer parte do erro de medida (Ebel, 1967, p. 121; Winer, 1962, p. 131) quando é utilizada a nota atribuída por um só juiz.

Fidedignidade das notas

O delineamento da pesquisa permite que se identifiquem, como fonte de variância das notas obtidas, as diferenças entre sujeitos. Da comparação da variância entre sujeitos com a variância devida a erros de medida pode-se obter a proporção desta última na variância das notas. Os valores obtidos na análise da variância de medidas repetidas são utilizados para estimar a fidedignidade das notas (Ebel, 1967; Winer, 1962):

$$\hat{\rho}_T = \frac{MQ_I - MQ_T}{MQ_I} \qquad T = 2, \ldots, 16$$

onde:

MQ_I = média dos quadrados das diferenças entre sujeitos.

MQ_T = média dos quadrados das diferenças entre repetições ou "dentro-sujeitos"

Ao estimar $\hat{\rho}_T$, neste caso, MQ_T permite estimar a variância devida a erros de medida que inclui as diferenças entre T professores-juízes. Excluindo-se da variância devida a erros de medida as diferenças entre os T professores-juízes, obtém-se:

$$\hat{\rho}_{TD} = \frac{MQ_I - MQ_R}{MQ_I}$$

onde:

MQ_I = média dos quadrados das diferenças entre sujeitos.

MQ_R = média dos quadrados das diferenças residuais

O estimador $\hat{\rho}_{TD}$ é útil quando as diferenças entre juízes não são importantes na comparação das notas das redações, ou quando são utilizadas as médias das notas atribuídas pelos T juízes.

De modo semelhante, definem-se os estimadores da fidedignidade das notas atribuídas por um só juiz, com base na análise da variância das notas atribuídas por T professores-juízes (Winer, 1962):

$$\hat{\rho}_{1D} = \frac{MQ_I - MQ_R}{MQ_I + (T-1)\, MQ_R}$$

$$\hat{\rho}_1 = \frac{MQ_I - MQ_T}{MQ_I + (T-1)\, MQ_T}$$

Os estimadores $\hat{\rho}_1$ e $\hat{\rho}_{1D}$ se aplicam na aferição da proporção da variância das notas atribuídas por um só dos T juízes.

Para cada um dos seis cursos, foram obtidos valores $\hat{\rho}_1$, $\hat{\rho}_{1D}$, $\hat{\rho}_T$ e $\hat{\rho}_{TD}$, para $T = 2, \ldots, 16$.

Os professores-juízes foram selecionados aleatoriamente para a formação dos grupos analisados, adicionando-se, sempre, mais um ao grupo constituído anteriormente.

Os coeficientes de fidedignidade encontrados com relação ao julgamento de um só professor se apresentam no Quadro 5. Em geral, verifica-se que, ao subtrair da variância devida a erros a parte devida a diferenças entre juízes, obtém-se um coeficiente maior; entretanto, essa discrepância é observada mais nitidamente nos valores de $\hat{\rho}_1$ e $\hat{\rho}_{1D}$ (Quadro 5) e um pouco menos em $\hat{\rho}_T$ e $\hat{\rho}_{TD}$ (Quadro 6).

QUADRO 5

Coeficiente de Fidedignidade das Notas Atribuídas por um Juiz, com base nas Notas de 2 ou Mais Juízes

Carreira / Número de Professores		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Biologia	R_1	0,68	0,65	0,65	0,61	0,62	0,60	0,60	0,57	0,58	0,60	0,62	0,63	0,60	0,61	0,60
	R_{1D}	0,71	0,71	0,68	0,68	0,68	0,68	0,68	0,65	0,65	0,66	0,68	0,68	0,66	0,66	0,65
Educação	R_1	0,70	0,62	0,65	0,59	0,62	0,58	0,59	0,58	0,58	0,60	0,62	0,63	0,61	0,60	0,60
	R_{1D}	0,71	0,68	0,69	0,67	0,68	0,68	0,69	0,67	0,66	0,68	0,69	0,70	0,67	0,67	0,66
Educ. Fís. Masc.	R_1	0,65	0,60	0,62	0,57	0,60	0,56	0,58	0,56	0,56	0,58	0,61	0,62	0,60	0,59	0,59
	R_{1D}	0,67	0,66	0,65	0,62	0,64	0,65	0,66	0,63	0,63	0,65	0,67	0,67	0,66	0,66	0,65
Engenharia	R_1	0,64	0,63	0,66	0,62	0,65	0,62	0,63	0,61	0,61	0,63	0,64	0,66	0,62	0,62	0,62
	R_{1D}	0,66	0,68	0,69	0,68	0,70	0,70	0,70	0,67	0,68	0,69	0,70	0,71	0,68	0,68	0,67
Letras	R_1	0,69	0,63	0,66	0,61	0,64	0,60	0,61	0,59	0,60	0,62	0,50	0,52	0,50	0,50	0,49
	R_{1D}	0,72	0,69	0,70	0,68	0,69	0,69	0,69	0,67	0,67	0,68	0,53	0,55	0,53	0,54	0,53
Medicina	R_1	0,67	0,62	0,61	0,57	0,60	0,56	0,58	0,56	0,57	0,58	0,61	0,62	0,58	0,58	0,57
	R_{1D}	0,71	0,68	0,65	0,64	0,66	0,66	0,66	0,65	0,64	0,65	0,67	0,68	0,64	0,64	0,62

Nota: R_1 = coeficiente de fidedignidade = $\hat{\rho}_1$
R_{1D} = coeficiente de fidedignidade ajustado, excluindo-se o efeito dos diferentes juízes = $\hat{\rho}_{1D}$

QUADRO 6

Coeficiente de Fidedignidade de Notas Médias Atribuídas por 2 ou Mais Juízes

Carreira / Número de Professores		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Biologia	R_T	81	85	88	89	91	91	92	92	93	94	95	96	95	96	96
	R_{TD}	83	88	90	91	93	94	94	94	95	96	96	97	96	97	97
Educação	R_T	82	83	88	88	91	90	92	93	93	94	95	96	96	96	96
	R_{TD}	83	86	90	91	93	94	95	95	95	96	96	97	97	97	97
Educ. Fís. Masc.	R_T	79	82	87	87	90	90	92	92	93	94	95	95	95	96	96
	R_{TD}	80	85	88	89	91	93	94	94	94	95	96	96	96	97	97
Engenharia	R_T	78	83	89	89	92	92	93	93	94	95	96	96	96	96	96
	R_{TD}	79	86	90	91	93	94	95	95	95	96	97	97	97	97	97
Letras	R_T	82	84	89	89	91	91	93	93	94	95	92	93	93	94	94
	R_{TD}	84	87	90	91	93	94	95	95	95	96	93	94	94	95	95
Medicina	R_T	80	83	86	87	90	90	92	92	93	94	95	95	95	95	95
	R_{TD}	83	87	88	90	92	93	94	94	95	95	96	97	96	96	96

Nota: R_T = coeficiente de fidedignidade da nota média = $\hat{\rho}_T$
R_{TD} = coeficiente de fidedignidade da nota média ajustado, excluindo-se o efeito das diferenças entre juízes = $\hat{\rho}_{TD}$

No Quadro 6 apresentam-se os valores dos coeficientes de fidedignidade encontrados quando varia o número de professores-juízes tomado por base para se obter a média das notas atribuídas a cada redação. Fica evidente que os coeficientes de fidedignidade crescem à medida que as médias das notas se baseiam em um número maior de juízes. A tendência geral é de um crescimento mais acelerado dos coeficientes obtidos com base nas notas dadas por 2 a 6 professores-juízes aproximadamente.

As notas atribuídas por um único juiz apresentam coeficientes de fidedignidade inferiores àqueles das notas médias em geral, como também se observa nos estudos de Vianna (1976) e de Coffman (1971).

Obteve-se, ainda, para cada um dos seis grupos, estimativa da variância residual e da variância devida a diferenças entre sujeitos e entre os professores-juízes (Quadro 7).

QUADRO 7

Estimativas de Variância, em Relação à Variância Total

Curso	Fonte	Estimativa da Variância	Proporção da Variância Total
Biologia	Sujeitos	2,81	0,65
	Juízes	0,03	0,01
	Resíduo	1,49	0,34
	Total	4,33	
Educação	Sujeitos	2,95	0,60
	Juízes	0,45	0,09
	Resíduo	1,53	0,31
	Total	4,93	
Educação Física Masculina	Sujeitos	2,67	0,59
	Juízes	0,42	0,09
	Resíduo	1,43	0,32
	Total	4,52	
Engenharia	Sujeitos	2,96	0,62
	Juízes	0,38	0,08
	Resíduo	1,43	0,30
	Total	4,77	
Letras	Sujeitos	2,87	0,49
	Juízes	0,39	0,07
	Resíduo	2,53	0,44
	Total	5,79	
Medicina	Sujeitos	2,60	0,57
	Juízes	0,41	0,09
	Resíduo	1,56	0,34
	Total	4,57	

DISCUSSÃO

No presente trabalho, estimam-se coeficientes de fidedignidade das notas atribuídas às redações de populações de candidatos a seis cursos universitários, por 16 professores de Português, em uma escala de zero a 10 pontos. As inferências são válidas apenas com relação ao que se costuma denominar de "processo global" de julgamento de redações; além disso, restringem-se às demais condições em que se realizou a pesquisa.

Os resultados mostram que cerca de 63% dos professores-juízes utilizam toda a escala de notas, de zero a 10; na maioria, não se observa tendência a evitar os valores extremos. Não se pode generalizar estes resultados, pois se trata de população de juízes que haviam participado, em anos anteriores, das sessões de treinamento e de julgamentos de redações, no Vestibular da CESGRANRIO. Esse treinamento, segundo Castilhos (1982, p. 105), pode levar a uma certa homogeneidade de julgamento dentro de um grupo que realiza o trabalho sob a coordenação de um mesmo supervisor. Entretanto, conforme notam Castilhos (1982, p. 105) e Costa Ribeiro *et alii* (1981), esse efeito nem sempre ocorre, supondo-se que dependa do supervisor do grupo; ainda mais, são observadas diferenças entre grupos que trabalham com supervisores diferentes. Como a amostra de professores-juízes, no presente estudo, foi selecionada aleatoriamente de uma mesma lista de nomes de professores, ignorando-se a organização dos grupos de trabalho que se formaram em 1979, é razoável supor que estes 16 professores tenham trabalhado anteriormente em grupos diversos, sob supervisão diversificada. De qualquer modo, não se pode descartar a possibilidade de ocorrerem certos efeitos da experiência anterior no julgamento das redações – como é o caso da tendência ao uso de todos os valores da escala de notas.

Ressalta a tendência dos gráficos das freqüências relativas das notas atribuídas por um mesmo juíz a representarem distribuições de forma muito semelhante para os grupos de candidatos aos seis cursos. Nota-se, por exemplo, que a proporção de notas zero e de notas 10 parece mais característica do professor-juiz do que do grupo de candidatos. Além disso, as diferenças entre a forma das distribuições das notas atribuídas pelos diversos professores-juízes a um mesmo grupo de candidatos reforçam a hipótese de que a freqüência relativa das notas depende mais do avaliador do que da distribuição da habilidade de redigir em cada um dos seis grupos. Estes resulados coincidem com os de Vianna (1978), obtidos em condições que não são idênticas às da presente investigação. Reforçam, ainda, as observações de Coffman (1971, p. 277) sobre a tendência dos juizes a distribuírem suas notas de formas diversas.

As diferenças encontradas entre as médias das notas atribuídas pelos 16 professores permitem inferir que os julgamentos podem variar numa dimensão de severidade-benevolência, conforme o professor-juiz. Confirmam-se os resultados das pesquisas de Costa Ribeiro *et alii* (1981) e de Vianna (1976; 1978). As diferenças entre a forma das distribuições de notas atribuídas por diversos professores-juízes a um mesmo grupo sugerem que ocorram, também, outras variações nos critérios adotados, no julgamento da mesma redação. Indicação semelhante emerge do estudo de Vianna (1978), em que uma só redação recebe notas que variam com uma amplitude de 41 pontos quando avaliada como base na impressão geral, em uma escala de zero a 100. Tais observações levam à hipótese de que o julgamento dos juízes pode variar não apenas em uma dimensão de severidade-benevolência, mas também quanto a critérios com que são avaliadas diversas facetas, que possam ser consideradas importantes em uma redação – como, por exemplo, a riqueza do vocabulário ou a correção gramatical. Pesquisas como as de Marshall (1967) e de Fench (1986) apoiam essa conclusão.

Dada a estruturação da presente pesquisa, na análise da variância realizada podem ser, ou não, consideradas como parte do erro de medida as diferenças entre a média geral e as médias, por juiz, das notas atribuídas aos candidatos. Se incluídas no erro de medida, os coeficientes de fidedignidade obtidos com relação às notas de um só professor-juiz vaiam de 0,49 a 0,70; caso excluídas, esses coeficientes se mostram um pouco mais altos, em geral, variando de 0,53 a 0,72. Os coeficientes obtidos com relação às notas médias, por candidato, quando baseados nos julgamentos de dois ou mais professores-juízes, são superiores, variando entre 0,78 a 0,97; ou seja, indicando uma proporção de 0,22 a 0,03 da variância dos erros de medida na variância total das notas atribuídas. As notas médias, mesmo quando baseadas nos julgamentos de dois juízes apenas, mostram-se nitidamente mais confiáveis do que aquelas atribuídas por um só juiz.

Na prática, quando se examina um grande número de redações, a adoção de um processo em que dois juízes, no mínimo, atribuam notas a cada uma delas, significa dobrar o custo (em cruzados), aproximadamente. Castilhos (1982, p. 109-110) sugere maneiras de diminuir o custo do treinamento dos professores-juízes para compensar o aumento do custo ao se duplicar o número de examinadores. Dadas as proporções que os exames Vestibulares assumem, no Brasil, o aumento do custo do julgamento das redações constitui problema a ser equacionado ao da fidedignidade das notas obtidas.

Como resultado da estimativa da variância entre professores-juízes e da estimativa da variância residual, para cada curso, tem-se que a proporção da primeira na variância total das notas representa menos de um terço da proporção da segunda. Aparentemente, a maior ou menor severidade do julgamento dos 16 professores-juízes contribui em menor proporção para o erro de medida do que outras fontes, não controladas, que se refletem na variância residual. Este resultado explica a fidedignidade, geralmente insatisfatória, das notas atribuídas por uma só pessoa, mesmo quando os professores-juízes são submetidos a um treinamento adequado; trata-se da extrema dificuldade de serem controlados, na prática, efeitos de fatores que afetam o julgamento de modo aleatório, além da interação do juiz com o texto a ser avaliado. A proporção da variância total das notas devida a diferenças entre juízes na dimensão de severidade-benevolência pode ser somada à da variância residual; assim composta, a parte atribuída a erros de medida totaliza de 35% a 51%, neste estudo. Mesmo que se consiga controlar as diferenças devidas à maior ou menor severidade dos juízes e que sejam estas excluídas da variância total das notas, a proporção da variância devida a erros de medida ainda é considerável (entre 30% e 44%).

Estes resultados situam-se na mesma linha dos estudos psicométricos relacionados por Vianna (1982), ou por Chediak, Bessa *et alii* (1975) e sugerem que se procure equacionar os problemas do custo do julgamento das redações e da obtenção de notas cujo nível de fidedignidade seja adequado ao uso que delas se pretende fazer.

REFERÊNCIAS BIBLIOGRÁFICAS

ANGOTTI, D. (1983) *Análise dos resultados do teste de seleção ao curso de formação de professores de 1ª a 4ª séries do 1º grau de um colégio estadual do Rio de Janeiro.* Dissertação de Mestrado, não publicada. PUC/RJ, Rio de Janeiro.

BESSA, N. M. (1984) *Erros de medida em notas de provas educacionais: conceituação e avaliação.* Trabalho apresentado no 6º Simpósio Nacional de Probabilidade e Estatística, UFRJ, Rio de Janeiro.

BIGGS, J. B., Collis, K. F. (1982) *Evaluating the quality of learning.* N.Y.: Academic Press.

BOCK, R. D. (1966) Contributions of multivariate experimental designs to educational research. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology.* Chicago: Rand McNally, 820-840.

CASTILHOS, M. T. de J. (1982) *The effect of a training program on the fluctuation of raters' scoring of writing compositions.* Tese de doutorado, não publicada, University of California at Los Angeles.

CHASE, C. I. (1979) The impact of achievement expectations and handwriting quality on scoring essay tests. *Journal of Educational Measurement,* 16(1), 39-42.

CHEDIAK, A. J., Bessa, N. M. (Relatores) (1975) *Parecer sobre a inclusão de prova discursiva em concursos vestibulares,* da Comissão Especial instituída pela Fundação CESGRANRIO. Rio de Janeiro: Fundação CESGRANRIO, *(Mimeo).*

CIACAGLIA, L. R. A. (1981) *Correção de redações – fidedignidade de medidas subjetivas.* São Paulo: Edicon.

COFFMAN, W. E. (1971) Essay examinations. In R. L. Thorndike (Ed.), *Educational Measurement.* Washington, D. C.: American Council on Education, 271-302.

COSTA RIBEIRO, S. (1981) Mecanismos da escolha na carreira e estrutura social da universidade. *Educação e Seleção,* jul. (3), 93-103.

COSTA RIBEIRO, S., Klein, R. (1982) A divisão interna da universidade: posição social das carreiras. *Educação e Seleção,* jan-jul. (5), 29-43.

COSTA RIBEIRO, S., Pessoa, D., Klein, R., Uchoa, C. E. F., Fontanive, N. S. (1981) Flutuação de critérios na avaliação de redações. *Educação e Seleção,* jul-dez (4), 27-42.

CRONBACH, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972) *The dependability of behavioral measurements.* New York: John Wiley.

DIEDERICH, P. B. (1974) *Measuring growth in English.* Urbana, III.: National Council of Teachers of English.

EBEL, R. L. (1967) Estimation of the reliability of ratings. In W. A. Mehrens; R. L. Ebel (Eds.), *Principles of educational and psychological measurement.* Chicago, III.: Rand McNally, 116-131.

FRENCH, J. W. (1966) Schools of thought in judging excellence of English themes. In A. Anastasi (Ed.), *Testing problems in perspective.* Washington, D. C.: American Council on Education, 587-596.

JÖRESKOG, K. G. (1971) Statistical analysis of sets of congeneric test. *Psychometrika,* 36(2), 109-133.

MARSHALL, J. C. (1967) Composition errors and essay examination grades re-examined. *American Education Research Journal,* 4(4), 375-385.

MARSHALL, J. C., POWERS, J. M. (1969) Writing neatness, composition errors and essay grades. *Journal of Educational Measurement,* 6,97-101.

STANLEY, J. C. (1971) Reliability. In R. L. Thorndike (ed.), *Educational Measurement.* Washington D. C.: American Council on Education, 356-442.

VIANNA. H. M. (1976) Redação e medida da expressão escrita: algumas contribuições da pesquisa educacional. *Cadernos de Pesquisa,* (16), 41-47. (a)

VIANNA, H. M. (1976) Flutuações de julgamentos em provas de redação. *Cadernos de Pesquisa,* (19), 5-9. (b).

VIANNA, H. M. (1978) Aplicação de critérios de correção em provas de redação. *Cadernos de Pesquisa,* (26), 29-34.

VIANNA, H. M.(1982)Redação e medida de expressão escrita: algumas contribuições da pesquisa educacional (II). *Educação e Seleção,* (6), 15-25.

WINER, B. J. (1962) *Statistical principles in experimental design.* New York: MacGraw-Hill.