FLUTUAÇÕES DE JULGAMENTOS EM PROVAS DE REDAÇÃO *

HERALDO MARELIM VIANNA **

RESUMO

Um conjunto de 161 provas de dissertação, elaboradas por candidatos ao Concurso Vestibular (1975), foi corrigido independentemente por quatro professores, segundo procedimentos padronizados. A análise da variância das notas mostrou diferença significativa entre os professores. A fidedignidade das notas não foi satisfatória para um examinador, mas a combinação de quatro notas mostrou alta fidedignidade.

SUMMARY

A set of 161 essay-tests written by Entrance Examination candidates (1975) was marked independently by four readers, according to standardized procedures. An analysis of variance of the marks showed a significant difference among readers. Reliability of the scores for one reader was not satisfactory, but the composite of four marks presented high reliability.

INTRODUÇÃO: aspectos gerais e hipóteses

As provas de redação, instrumentos geralmente usados para fins de verificar a capacidade de expressão escrita, possuem méritos indiscutíveis; contudo, apresentam algumas dificuldades que, apesar de numerosas tentativas de solução, constituem um desafio para a maioria dos especialistas. A literatura brasileira sobre instrumentos de medidas educacionais ainda é bem reduzida e inexistem, praticamente, estudos empíricos sobre esse tipo de prova, apesar de bastante empregado em nosso contexto educacional, num passado recente. Um levantamento das contribuições da literatura estrangeira foi realizado (Vianna, 1976) e constatou-se que um dos problemas centrais das provas de redação consiste na flutuação dos julgamentos, mesmo quando realizados por professores capazes e experientes. Assim sendo, procurou-se, na presente pesquisa, verificar, basicamente, aspectos relativos à fidedignidade das notas em provas de redação, formulando-se, para esse fim, as seguintes hipóteses: a) os professores, ao corrigirem provas de redação, usam os mesmos padrões; e 2) os coeficientes de fidedignidade das notas de provas de redação podem ser considerados satisfatórios.

* Expressamos os nossos agradecimentos às Professoras Dulce de Godoy Alves, Flávia de Barros Carone, Ilka Brunilda de Gallo Laurito, Lygia Corrêa Dias de Moraes, Vilma Fagundes Sanchez, e ao Professor Amauri M. T. Sanchez, que colaboraram ativamente no desenvolvimento da presente pesquisa.

** Do Departamento de Seleção de Recursos Humanos da Fundação Carlos Chagas.

METODOLOGIA: amostra, sistema de correção, aplicação e características do instrumento.

A amostra foi organizada com 161 sujeitos, selecionados entre candidatos inscritos no Concurso Vestibular para carreiras da Área Biomédica, realizado em São Paulo pela Fundação Carlos Chagas, em janeiro de 1975, e que fizeram todas as provas, independentemente das suas classificações finais. A amostra do tipo aleatório foi estratificada pela primeira opção dos candidatos para os diversos cursos. Assim sendo, de acordo com as opções de curso, posteriormente agrupadas em carreiras, a amostra ficou constituída da seguinte forma: — Medicina (43%), Medicina Veterinária (2%), Farmácia e Bioquímica (8%), Odontologia (12%), Biologia e História Natural (8%), Enfermagem e Obstetrícia (7%), Agronomia e Engenharia Florestal (3%), Nutrição (4%), Psicologia (7%), Especialidades Para-médicas (4%) e Educação Física (2%).

Alguns princípios foram definidos para que os julgadores das redações pudessem segui-los uniformemente. Tal procedimento objetivou controlar a fidedignidade dos julgadores, isto é, a coerência dos julgadores entre si. A especificação detalhada dos elementos a verificar encontra sua justificativa no fato de que, freqüentemente, na correção de uma dissertação, fatores irrelevantes ou estranhos ao processo de avaliação costumam influenciar na decisão dos julgadores. Sims (1933) mostrou que os escores resultantes de julgamentos independentes de dois julgadores são quase que semelhantes quando um método específico é estabelecido para orien-

tar a correção de um trabalho dissertativo. A influência de fatores irrelevantes na correção de dissertações foi estudada por Chase (1968), que demonstrou ser ela bem pequena quando uma chave de correção dos trabalhos é inicialmente estabelecida. Assim sendo, para fins do presente trabalho, cada julgador examinou e quantificou 20 itens relativos a 4 aspectos da dissertação, que foram definidos da seguinte forma:

A — *Estrutura geral*

Os julgadores verificaram se a dissertação constituía um todo orgânico e formava um conjunto articulado e completo de idéias *(organicidade do texto)*, em torno de um único tema *(unidade)*.

B — *Estrutura interna*

A estrutura interna da dissertação foi avaliada tendo em vista o relacionamento lógico das várias proposições, considerando-se a organização e o encadeamento dos parágrafos que as contêm *(metodologia do texto)*.

C — *Conteúdo*

A complexidade da avaliação do conteúdo da dissertação exigiu o desdobramento do presente tópico em diferentes aspectos. Assim, procurou-se verificar se o texto dissertativo:

1 — refletia ponderação a respeito do tema ou se, ao contrário, traduzia apenas uma opinião (*necessidade do texto*);
2 — apresentava idéias fundamentadas e coerentes (*coerência interna*);
3 — permitia estabelecer uma perfeita relação de entendimento entre o expositor e o leitor (*clareza*);
4 — concentrava a exposição em aspectos relevantes do tema, evitando digressões alheias ao assunto principal (*concentração*);
5 — representava uma contribuição nova, acrescentando alguma coisa ao que já é de domínio comum (*pensamento divergente*);
6 — revelava uma nova forma de apresentação para idéias já conhecidas (*individualidade*).

D — *Expressão*

O julgamento das dissertações, no seu aspecto formal, revelou-se, também, tarefa complexa. A importância desse aspecto, na caracterização da capacidade de expressão escrita, determinou o seu parcelamento em 2 subtópicos:

1 — propriedade do vocabulário (*léxico*); e
2 — correção gramatical (*ortografia, pontuação, flexão do substantivo e do adjetivo, emprego do pronome, emprego de tempos e modos, concordância verbal e nominal, regência verbal e nominal, e estrutura da frase*).

A prova de dissertação foi aplicada no dia 27 de abril de 1975 e teve a duração de duas horas e trinta minutos. Ao grupo pesquisado exigiu-se a elaboração de apenas uma única dissertação; entretanto, tendo em vista que a capacidade de expressão escrita pode variar, em função do tema apresentado e das capacidades exigidas, procurou-se atenuar a possível influência dessa problemática por meio da seleção de um tema de ordem geral, que independesse de conhecimentos prévios. Por outro lado, tratando-se de uma amostra de indivíduos com experiência recente de Concurso Vestibular, o que implica, necessariamente, numa opção profissional, procurou-se apresentar um assunto que, possivelmente, já tivesse sido objeto de análise e discussão. Assim sendo, foi solicitado aos participantes que analisassem e discutissem o tema: PAPEL DO INDIVÍDUO DE FORMAÇÃO UNIVERSITÁRIA NA SOCIEDADE A QUE PERTENCE, e expusessem, sob a forma de dissertação, as suas conclusões.

A extensão média das dissertações foi de 30 linhas; entretanto, 50% dos indivíduos situaram-se na faixa sugerida de 30 a 35 linhas. Alguns poucos (7%), alongaram-se até 51 linhas, mas um grupo maior, constituído por 43% dos pesquisados, elaborou trabalhos com extensão variável entre 15 e 29 linhas. Apesar de orientados inicialmente, apenas 17% esboçaram um esquema da possível estrutura de dissertação; contudo, 96% redigiram um rascunho para o trabalho. O tempo gasto na elaboração das dissertações não foi rigorosamente controlado, mas tudo indica ter sido adequado, pois ao término de 2:00 horas estavam concluídos os trabalhos de aplicação.

As dissertações foram corrigidas, no período de 29-04 a 30-05 de 1975, por uma equipe de quatro professores, dos quais três de nível universitário e um de ensino de 2º Ciclo. Os trabalhos de avaliação foram independentes, isto é, sem que um professor conhecesse as notas do outro julgador. Ainda que uma correção holística talvez proporcionasse resultados equivalentes ao de uma correção analítica, preferiu-se esta última, porque, sendo uma experiência nova, se desejava que todos os julgadores seguissem a mesma orientação e observassem os aspectos anteriormente definidos e discutidos pela equipe. Procurou-se evitar, na fase de correção, qualquer comentário sobre as notas atribuídas inclusive médias parciais e medidas de variabilidade, a fim de evitar que, a partir das informações, se modificassem os comportamentos dos julgadores. As notas foram atribuídas numa escala de 100 pontos e cada sujeito recebeu quatro notas, que, posteriormente, foram sintetizadas numa nota média final, cujas freqüências estão representadas na Tabela 1.

TABELA 1 — *DISTRIBUIÇAO DAS NOTAS FINAIS DA PROVA DE DISSERTAÇAO (MÉDIA DAS NOTAS DE QUATRO PROFESSORES).*
São Paulo. 1975

Notas	*F*
90 — 99	2
80 — 89	11
70 — 79	19
60 — 69	38
50 — 59	38
40 — 49	31
30 — 39	15
20 — 29	3
10 — 19	3
0 — 9	1
N	161

As notas médias finais variaram de 1,25 a 92,25, com uma amplitude, portanto, de 92 pontos. A média do grupo foi 56,95 e o desvio padrão 16,15. A partir desses dados, analisaram-se duas características quantitativas da dissertação: o grau de dificuldade e o poder discriminativo do instrumento. Utilizaram-se os índices P (dificuldade) e D (discriminação), sugeridos por Whitney e Sabers (1970).

O grau de dificuldade foi obtido comparando-se o desempenho dos sujeitos da amostra com o nível de desempenho mais alto possível. Assim, em termos operacionais, a dificuldade foi definida como a razão entre a diferença da média ($\overline{X}$) do grupo e o escore mais baixo possível (Xmin) — *nível de desempenho* — e a diferença entre o escore mais alto possível (Xmax) e o escore mais baixo possível — *nível de desempenho mais alto possível*. Isto é,

$$P = \frac{\overline{X} - Xmin}{Xmax - Xmin} \qquad (1)$$

O índice de dificuldade (P) foi multiplicado por 100 para expressá-lo em unidades de porcentagem. A dificuldade da dissertação foi de 57%, o que pode ser interpretado como traduzindo uma prova de dificuldade mediana. Suplementando-se essa informação com a distribuição da Tabela 1, poder-se-ia dizer que a sua tendência foi de média para fácil.

O índice D — poder discriminativo — foi estabelecido com base no pressuposto de que, em geral, indivíduos com diversos níveis de rendimento apresentam diferentes desempenhos num exame. Assim, procurou-se comparar o desempenho entre grupos extremos. Para a formação desses grupos, adotou-se o método dos 25% superiores e inferiores. Operacionalmente, o índice de discriminação foi definido como a diferença entre a dificuldade da questão para o grupo superior (P_S) e a dificuldade da mesma questão para o grupo inferior (P_I). Ou seja,

$$D = P_S - P_I \qquad (2)$$

Os valores de P_S e P_I, calculados segundo a fórmula (1), foram 0,765 e 0,363, respectivamente. Tais valores determinaram um coeficiente *D* igual a 0,402; assim, as notas das dissertações, que resultaram da média de quatro julgamentos independentes, foram discriminativas, isto é, separaram os bons candidatos dos de desempenho deficiente, pois se admitiu como padrão mínimo desejável, tendo em vista o tamanho da amostra, o índice 0,40, ainda segundo Whitney e Sabers (1970).

ANÁLISE DA CORREÇÃO DA PROVA DE REDAÇÃO

Após a correção independente das dissertações, segundo o critério definido anteriormente, as notas atribuídas foram tratadas estatisticamente, calculando-se, inicialmente, as médias e os desvios-padrão, assim como os coeficientes de intercorrelação momento-produto, conforme a Tabela 2.

TABELA 2 — *MÉDIAS, DESVIOS-PADRÃO E INTERCORRELAÇÕES DAS NOTAS ATRIBUÍDAS POR QUATRO PROFESSORES A 161 DISSERTAÇÕES.*
São Paulo. 1975

P	$\overline{X}$	s	Intercorrelações(x)			
			P_1	P_2	P_3	P_4
P_1	45,96	16,92	—	0,69	0,64	0,62
P_2	53,38	21,68		—	0,75	0,72
P_3	65,10	18,74			—	0,68
P_4	63,37	16,38				—

(x) Significantes ao nível 0,001

As diferenças apresentadas pelas médias e variâncias indicam que, possivelmente, os julgadores não teriam seguido de modo uniforme os critérios por eles estabelecidos. As maiores diferenças de médias estão entre P_1 e P_3. O primeiro foi rigoroso na atribuição das notas, ao passo que P_3 reagiu com certa liberalidade às dissertações. Os professores P_3 e P_4, em geral, atribuíram graus altos; contudo, P_3 distribuiu suas notas com maior variabilidade do que P_4, que, no grupo de professores, apresentou menor variação entre as notas. A média de P_2 aproximou-se da média global dos quatro professores (56,95) e, comparativamente, a variabilidade das notas de P_2 foi a mais bem distribuída. Assim, em relação às médias e variâncias, os dados, em princípio,

confirmam os resultados de pesquisas anteriores (Vianna, 1976), quanto à variabilidade dos avaliadores no julgamento de uma prova de dissertação. A obtenção de um escore alto ou baixo pelo examinando, no caso presente, estaria na dependência do professor que julgou a dissertação.

As intercorrelações para as notas dos quatro professores variaram de 0,62 a 0,75, e sugerem um certo grau de concordância entre os professores, sendo tal ocorrência possível, porquanto as correlações momento-produto podem ser altas mesmo havendo diferenças entre as médias e as variâncias. A média das seis intercorrelações (0,69), usada a transformação *z* de Fisher, poderia ser empregada como uma estimativa da fidedignidade dos julgamentos, se as variâncias fossem aproximadamente iguais. O teste de homogeneidade das variâncias, por intermédio do Fmax de Hartley, demonstrou diferença significante ao nível 0,01, rejeitando-se, portanto, a hipótese de igualdade das variâncias. Assim sendo, a fidedignidade das medidas foi estimada a partir da análise da variância das notas dos quatro professores, usando-se os referidos professores como tratamento.

TABELA 3 — *ANÁLISE DA VARIÂNCIA DAS NOTAS ATRIBUÍDAS POR QUATRO PROFESSORES A 161 DISSERTAÇÕES. São Paulo.* 1975

Fonte de variação	*Soma dos quadrados*	*Grau de Liberdade*	*Quadrados médios*	*F*
Entre alunos	166946,02	160	1043,41	
Intra alunos	92048,10	483	190,58	
Entre professores	38801,91	3	12933,97	116,596xxx
Resíduo	53246,28	480	110,93	
TOTAL	258994,21	643		

xxx Altamente significante ($\alpha \leq 0{,}001$)

Os dados da ANOVA mostraram que o "fator" professor teve influência altamente significante nas notas da prova de dissertação, confirmando-se, assim, a impressão inicial resultante da inspeção das médias e variâncias. A nota dos examinandos, na prova de dissertação, dependeu do "fator" professor, rejeitando-se, *ipso facto,* a hipótese de que o fator professor *não* teria influência na nota da dissertação.

Procedeu-se, também, à comparação (contraste) entre as notas emparelhadas, atribuídas pelos quatro avaliadores das 161 dissertações. Os resultados (Tabela 4) mostraram existir diferenças significantes entre as notas dos pares de professores P_1-P_2, P_1-P_3, P_1-P_4, P_2-P_3 e P_2-P_4. As notas do par de professores P_3-P_4 não apresentaram diferenças significativas.

TABELA 4 — *TESTE DE COMPARAÇÃO PARA AS NOTAS EMPARELHADAS ATRIBUÍDAS POR QUATRO AVALIADORES A 161 DISSERTAÇÕES. VALORES* "t" ($\alpha = 0{,}001$; G.L. = 160).

Professores	P_1	P_2	P_3	P_4
P_1		—5,93xxx	—15,94xxx	—15,29xxx
P_2			—10,15xxx	—8,45xxx
P_3				1,55(ñs)

xxx — diferenças significantes
ñs — diferença não significante

A fidedignidade das notas foi estimada usando-se a ANOVA e as fórmulas sugeridas por Winer (1970). Inicialmente, calculou-se a fidedignidade para as quatro medidas:

$$r_4 = 1 - \frac{\text{Q.M. intra-alunos}}{\text{Q.M. entre-alunos}}$$

$$r_4 = 0{,}82$$

A fidedignidade para UMA medida foi estimada da forma seguinte:

$$r_1 = \frac{\text{Q.M. entre-alunos} - \text{Q.M. intra-alunos}}{\text{Q.M. entre-alunos} + (k-1)\ \text{Q.M. intra-alunos}}$$

k = número de tratamentos (professores, no caso)
$r_1 = 0{,}53$

Os coeficientes calculados demonstram que apenas UMA medida, na correção da presente dissertação, não resultou em notas fidedignas; entretanto, o conjunto das QUATRO medidas apresentou coeficiente superior ao exigido para provas de dissertação, que deve ser de 0,80, segundo Gulliksen (1950).

Estimaram-se, também, as fidedignidades para dois e três julgamentos, usando-se a conhecida fórmula de Spearman-Brown,

$$r_k = \frac{kr_1}{1 + (k-1)\ r_1},$$

que proporcionou para essas medidas os coeficientes 0,69 e 0,77, respectivamente. Desse modo, os coeficientes evidenciaram que, na correção de dissertações, o aumento do número de julgadores concorreu para maior fidedignidade dos resultados.

CONCLUSÕES

Os dados coletados e analisados na pesquisa permitiram concluir que:

1º — os sujeitos pesquisados não tiveram grande dificuldade no desenvolvimento do tema proposto para dissertação, podendo-se considerar a prova discursiva como sendo de dificuldade mediana para o grupo;

2º — as notas finais das dissertações, representadas pela média da correção independente de quatro avaliadores, revelaram-se discriminativas, isto é, separaram diferentes níveis de desempenho;

3º — as notas atribuídas a cada uma das dissertações por quatro julgadores, apresentaram, em geral, grande amplitude de variação, possivelmente por falta de uniformidade na aplicação do critério de correção, não se confirmando, assim, a primeira hipótese estabelecida;

4º — a fidedignidade das notas para um avaliador foi baixa e ficou demonstrado que o *fator* professor tem influência na distribuição das notas em provas de dissertação, não se concretizando, desse modo, a segunda hipótese levantada, quando se trata de um único examinador;

5º — a fidedignidade das médias de dois e três julgadores não foram plenamente satisfatórias, sobretudo no caso de dois avaliadores;

6º — a fidedignidade da média dos quatro avaliadores das dissertações foi superior ao mínimo exigido, evidenciando que quanto maior o número de avaliadores, maior, igualmente, a fidedignidade dos resultados.

Assim sendo, conclui-se, finalmente, que, tendo em vista os resultados da presente pesquisa, seria recomendável que as notas de provas de dissertação resultassem da média de, no mínimo, três examinadores e, se possível, de quatro julgadores, a fim de que os resultados possam merecer confiança e não fiquem sujeitos à influência da equação pessoal de cada examinador, fator principal da flutuação dos resultados em correções de provas de dissertação.

REFERÊNCIAS BIBLIOGRÁFICAS

CHASE, C.I. (1968) — The impact of some obvious variable on essay-test scores. *Journal of Educational Measurement,* 5.

GULLIKSEN, H. (1950) — *Theory of Mental Test.* Nova Iorque. John Willey and Company.

SIMS, M. (1933) — Reducing the variability of essay examination marks through elementary variation in standards of grading. *Journal of Education Research,* 26.

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

WHITNEY, D.R. e SABERS, D.L. (1970) — Improving essay examinations. III — *Use of item analysis.* Technical Bulletin, nº 11. University Evaluation and Examination Service. The University of Iowa. Iowa City, Iowa.

WINER, B.J. (1970) — *Statistical Principles in Experimental Design.* Nova Iorque. McGraw-Hill Book Co.